Received 6 December 2023, accepted 17 December 2023, date of publication 18 December 2023,
date of current version 27 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344653

SURVEY

A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats

RAMI MUBARAK[1], TARIQ ALSBOUI[2], OMAR ALSHAIKH[2], ISA INUWA-DUTSE[2], SAAD KHAN[2], AND SIMON PARKINSON[2]
[1]Royal Academy of Police, Ministry of Interior, Manama 33305, Bahrain
[2]Department of Computer Science, University of Huddersfield, HD1 3DH Huddersfield, U.K.

Corresponding author: Simon Parkinson (s.parkinson@hud.ac.uk)

ABSTRACT In the rapidly evolving digital landscape, the generation of fake visual, audio, and textual content poses a significant threat to the trust of society, political stability, and integrity of information. The generation process has been enhanced and simplified using Artificial Intelligence techniques, which have been termed *deepfake*. Although significant attention has been paid to visual and audio deepfakes, there is also a burgeoning need to consider text-based deepfakes. Due to advancements in natural language processing and large language models, the potential of manipulating textual content to reshape online discourse and misinformation has increased. This study comprehensively examines the multifaceted nature and impacts of deep-fake-generated media. This work explains the broad implications of deepfakes in social, political, economic, and technological domains. State-of-the-art detection methodologies for all types of deepfake are critically reviewed, highlighting the need for unified, real-time, adaptable, and generalised solutions. As the challenges posed by deepfakes intensify, this study underscores the importance of a holistic approach that integrates technical solutions with public awareness and legislative action. By providing a comprehensive overview and establishing a framework for future exploration, this study seeks to assist researchers, policymakers, and practitioners navigate the complexities of deepfake phenomena.

INDEX TERMS Deepfakes, visual, audio, text.

I. INTRODUCTION

In recent years, the rise of social media platforms and the widespread adoption of smart devices have revolutionised how we communicate, share information, and interact with the digital world. Social media platforms such as Facebook, Instagram, Twitter, and Snapchat have become integral parts of our daily lives, connecting us with friends, family, and even strangers from around the world. These platforms share our thoughts, opinions, and personal experiences. The accessibility of smart devices has also played an important role in facilitating the seamless sharing of information. With the proliferation of smartphones, tablets, and other devices connected to the Internet, people can effectively capture and upload photos, record videos, and document their lives in real-time. The convenience of having these devices at our fingertips has empowered individuals to become content creators, broadcasters, and influencers, fostering a culture of digital self-expression.

Although technology has advanced significantly, the emergence of deepfakes has introduced new complexities. Deepfakes are highly realistic synthetic media created using artificial intelligence techniques, which are widely known as deep learning. They involve manipulating media, such as images, videos, audio, and text generated or altered using complex deep neural networks, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), variable autoencoders (VAEs), and diffusion models (DMs). The availability and rapid development of deepfake techniques, such as face-swapping, lip-syncing, puppeteering, voice conversion, and natural language processing (NLP), have both positive and negative consequences. On the positive side, fake core technologies offer novel opportunities in various fields, such

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

TABLE 1. Summary of key related surveys.

Source	Type	Focus
Verdoliva, 2020 [3]	Visual	Focussing on deepfakes from a forensic viewpoint. Highlighting the limitations of current forensic detection methods.
Tolosana et al., 2020 [14]	Visual	Reviewed facial manipulations and detection techniques.
Mirsky et al., 2021 [13]	Visual	Focused on field-specific generation technicalities in depth. Detection is briefly discussed.
Rana et al., 2022 [24]	Visual	A systematic review and comprehensive classification of deepfakes detection methods from 2018-2020.
Zhang, 2022 [25]	Visual	Face swapping and face re-enactment only. Focusses on methods to generate and detect these types of deepfakes.
Nguyen et al., 2022 [26]	Visual	Discusses techniques and challenges in deepfakes detection, with methods classified by data type. For images, the focus is on hand-made or deep features. For videos, methods use either temporal features across frames or visual artefacts within a single frame.
Masood et al., 2024 [11]	Visual and Audio	Generating and detection are discussed in depth, including reviewing datasets.

as entertainment, education, and advertising [1]. They enable realistic dubbing, animation, virtual experiences [2], content generation, and innovative branding techniques [3], [4]. The emergence of user-friendly deepfake tools (e.g., Deepfakes Web [5], DeepFaceLab [6], FaceApp [7], ChatGPT [8], DALL-E2 [9], and Midjourney [10]) powered by deep learning methods has made it easier for non-expert users to create synthetic content that is indistinguishable and innovative [11]. However, fake content also poses significant threats, including identity theft, revenge pornography, fraud, and threats to national security [1], [12], [13]. The increasing prevalence and consequences of deepfakes have captured the attention of researchers, policymakers, and society as a whole, leading to efforts to understand their nature, impacts, and methods to detect and mitigate their effects [14], [15], [16], [17]. Large multinational organisations, such as Facebook, Microsoft, and Amazon, have organised deepfake detection challenges to encourage the development of effective detection methods [1]. Detecting deepfake content remains complex due to their high-quality output and the continuous advancement of generation techniques [18], [19]. Therefore, it is necessary to continually study fakes to understand the evolving landscape of media manipulation, as their use raises critical questions about the truth, trust and reliability of digital content [11], [20]. Failure to address the challenges posed by deepfakes can result in erosion of public trust, spread of misinformation, and potential damage to individuals and organisations [21], [22], [23].

The field of deepfakes has seen a surge in research, as shown in Figure 1, which shows the annual increase in the number of publications related to deepfakes. Several studies have extensively explored the detection of deepfakes and have provided valuable information on this rapidly growing field. Table 1 provides an overview of recent significant surveys and their focus. In this table, we provide an overview of the main focus of key and related surveys. The motivation for focussing on these few is that through this survey we have identified them as having close alignment, yet are unable to provide a comprehensive survey on the impacts of deepfakes and their detection mechanisms across the three types (visual, audio, and text), as presented in this paper. These surveys have significantly improved our knowledge of deepfake creation, detection techniques, and general aspects of this technology. Although surveys have explored various aspects of deepfakes, primarily focussing on the generation and detection of visual and audio deepfakes, they have often overlooked text-based deepfakes; therefore, there is an absence of a holistic survey focused on the potential impact of deepfakes. It should be noted that the term ‘‘text-based deepfakes’’ is not as widely used as ‘‘audio’’ or ‘‘video’’ deepfakes, since the emphasis has been on manipulating audio or visual content. However, the concepts of text-based manipulation and generation have become increasingly relevant to recent advances in natural language processing and large language model technologies. This article presents an all-encompassing analysis of audio, visual, and text-based deepfakes, focused on both detection methodologies and subsequent impacts. This was achieved by thoroughly examining their various types and definitions, along with exploring recent detection techniques. In addition, it investigates the social, political, and economic impact of deep-fakes in various sectors. By analysing the motivations behind the creation and proliferation of deepfakes and delving into their consequences, this study highlights the urgent need for further research to address the identified challenges.

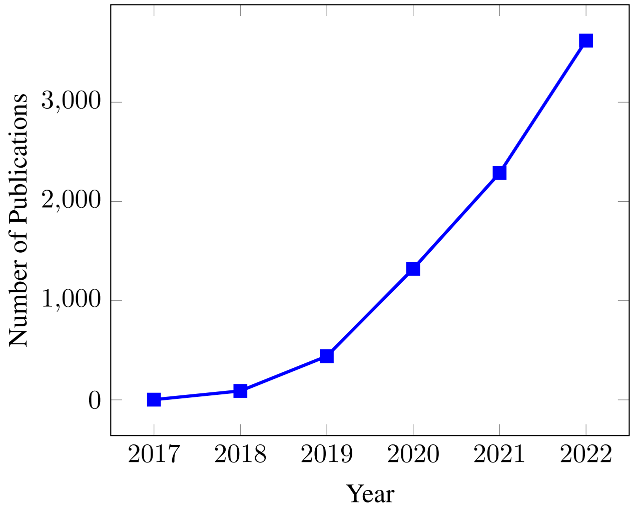


FIGURE 1. Number of publications related to deepfakes by year.

The survey presented in this work is of great significance to governments, industry, and society. The ability to acquire a comprehensive understanding of how deepfake generation can be used enables people to understand how they might become prominent in their individual lives. Furthermore, by understanding the potential impact that techniques can generate, individuals can make informed decisions on how they might implement defensive measures and their limitations. For citizens, this can help prevent them from following fowl to an adversary using fake content to instil trust. For governments, it can ensure they are aware of the possibilities of deepfake content generation as well as its impacts, especially in terms of adversarial influence. The knowledge generated in this article is also of great significance to the industry. It is important that industry knows the capabilities of deepfakes so that they can benefit from their legitimate and efficiency-improving functionality, but also so that they are aware of how they could be used against them by an adversarial. For example, its use in phishing attempts will be significant as digital media (text, audio, video) becomes increasingly realistic and believable to employees, making it very difficult for them to easily identify that an adversary is trying to convince them to do something damaging to their organisation, such as transfer of money.

This study presents contributions to the existing body of research on deepfakes and their detection. These contributions are intended to help researchers, policymakers, and practitioners understand the complexities of deep-fake landscapes and navigate potential solutions. Specifically, includes the following contributions:

- A first-of-a-kind survey focussing on all three types of deepfake generation. More specifically, audio, visual, and textual formats. The survey provides a unique focus on analysing a broad range of social, political, economic, and technological impacts of deepfakes.
- The survey provides an up-to-date overview of state-of-the-art detection methodologies before identifying potential future research directions in the deepfake detection field, particularly those areas that can benefit from further development and refinement.

II. METHODOLOGY

To ensure the rigour and completeness of this literature survey, a systematic approach was adopted to collect, analyse, and synthesise publications related to deepfakes. This section details the methodology used during the research process to ensure transparency and repeatability. Table 2 provides a breakdown of the research methodologies used in this study. This includes details on the identification and analysis of publications of interest. As evident in the table, the PRISMA approach is adopted in this work [27]. More specifically, 590 papers were initially identified using a pre-defined search query using well-established repositories. Inclusion and exclusion criteria resulted in 255 articles being included in this review. The remainder of this paper is as follows: Section III investigated the multifaceted nature of deepfakes, providing foundational insights and emphasising their definitions and the underpinning technology. Section IV classifies and discusses various types of deepfakes and provides a comprehensive summary of their range. The manifold impacts of deepfakes on critical domains are critically examined in Section V. Section VI provides an overview of the state-of-the-art methods used for the detection of deepfakes. The preceding sections and key insights are discussed in Section VII as part of a comprehensive analysis. Finally, we conclude by highlighting the primary insights and identifying potential avenues for future research in the realm of deep-fakes.

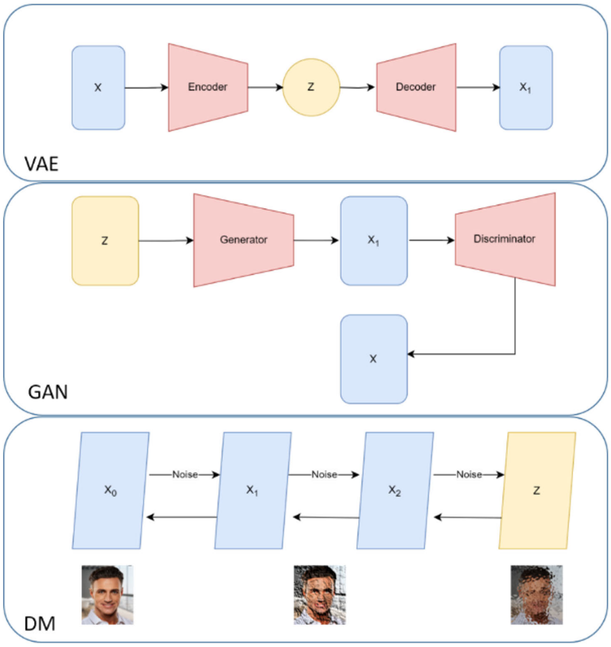


FIGURE 2. Illustration of Generative Adversarial Networks (GANs), influenced by GAN [36], VAE [37] and DM [38].

III. DEEPFAKES

A narrow definition of deepfake media includes the use of artificial intelligence and deep learning techniques to manipulate or synthesise multimedia content, specifically visual (images and videos), audio, and text. These techniques involve creating highly realistic and often deceptive content that is difficult to distinguish from authentic or authentic media [28]. Deepfakes first gained widespread attention in 2017, when a Reddit user named "deepfakes" used this technology to create pornographic content by swapping the faces of celebrities using deep learning models (DL) [29], [30]. The deep learning models are trained to generate deceptively realistic counterfeits by combining and overlaying objects in the media [31]. Deepfakes have gained significant attention because of their ability to create compelling and often indistinguishable fake content [32], [33], [34], [35].

The technology underlying deepfakes primarily revolves around deep learning and generative models. The deep learning models used for visual and audio deepfakes are shown in Figure 2, namely, Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE). GANs, introduced by Goodfellow et al. [36] in 2014, consist of two neural

TABLE 2. Focus of other surveys.

Step	Description
Data Collection	• Peer-reviewed articles, research papers, conference proceedings, and official reports focused on deepfakes. • Google Scholar, IEEE Xplore, ACM digital library and ScienceDirect were with the following search query: deepfakes/AI-generated content/ deepfakes impacts OR consequences /text-based deepfakes/synthetic text/Bot generated text AND social bots/ face swap/lip-syncing/face generation/reenactment/text to speech/speech to speech AND voice conversion/deepfakes detection. • From the initial search, a pool of 590 research articles related to deepfakes was assembled.
inclusion and Exclusion Criteria	• Inclusion: Papers in five years between January 2017 to June 2023. • Exclusion: Non-English articles and opinion pieces without empirical data.
Analysis	• Detailed reading of selected articles and extracted data related to deepfakes generation, impacts, and detection, then organised data into themes.
Synthesis	• Synthesized data from articles to identify patterns, trends, limitations, and gaps. • Highlight prevalent deepfake detection methodologies, implications and future research areas.

networks: a generator and a discriminator. The generator network produces fake content, whereas the discriminator network attempts to distinguish between real and fake content. Through an iterative training process, both networks improve their performance, resulting in the generation of more realistic fake content [36], [39]. VAEs, also introduced in 2014 by Kingma and Welling [37], are based on neural network autoencoders. These auto-encoders consist of an encoder and decoder network. The encoder network learns to represent the input data in a lower-dimensional latent space, whereas the decoder network reconstructs the original data from the latent space representation. VAEs extend this concept by introducing probabilistic modelling, which allows the generation of new data points in addition to reconstruction. VAEs are often used for signal analysis tasks in deepfake generation [40], [41].

In 2015, Sohl-Dickstein et al. introduced probabilistic diffusion models, known as diffusion models (DM), a class of generative models that excel in matching the distribution of data by progressively reversing the multistep process of introducing noise [38]. These models were designed to estimate the underlying probability distribution of a dataset by iteratively transforming a noise distribution to resemble the target distribution. This gradual noise-reversal process allows the model to capture the intricate patterns and structure of the data. DMs have recently been shown to generate images of high quality while providing attractive characteristics, such as distribution coverage, a stationary training objective, and simple scalability [42]. DMs can outperform or are comparable to (GANs), and they enable sophisticated text-to-image synthesis models, such as DALL-E 2 [9] and Midjourney [10].

On the contrary, text-based deep-learning models are generated using natural language processing (NLP)-based deep learning models [44], which are designed to capture the sequential nature of text data. These models include Conventional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [45]. However, the recent use of transformer models introduced in 2017 by Vaswani et al. has revolutionised (NLP) tasks [43]. This is because, as shown in Figure 3, the transformer model uses an encoder-decoder architecture with self-attention mechanisms to capture contextual relationships between words in a text. The model understands word dependencies by attending to different parts of the input sequence and generates coherent and contextually relevant texts. Text deepfakes are created by fine-tuning a pre-trained transformer model on specific text

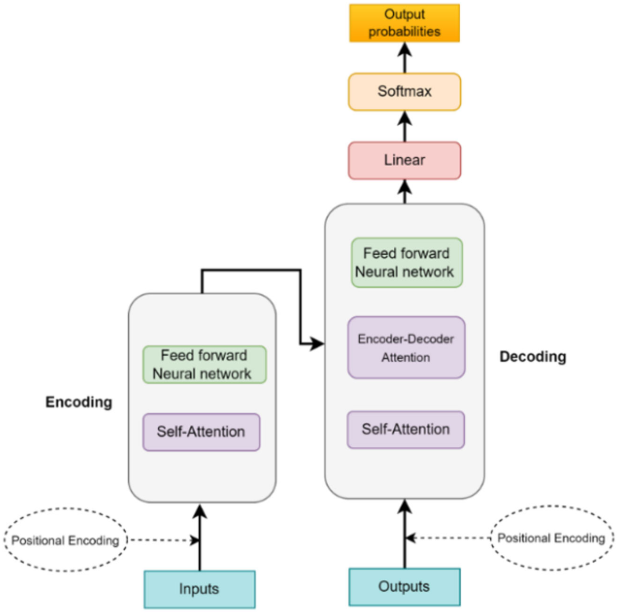


FIGURE 3. Simplified transformer model architecture, influenced by Vaswani et al., 2017 [43].

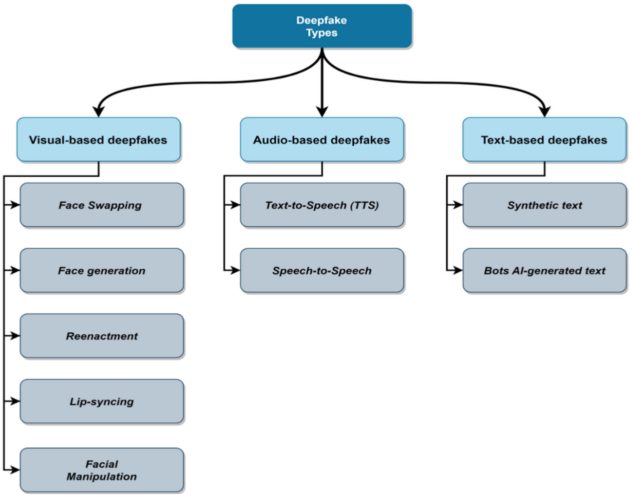


FIGURE 4. Deepfake types.

data, allowing it to mimic the style and characteristics of a particular author or writing style. Conditional text generation techniques can also guide the output of fake text. More advanced models, such as transformer models [46], [47] and generative pre-training transformers (GPTs) [48], can be applied in text deepfake generation.

IV. TYPES OF DEEPFAKES

Deepfakes can be divided into three main categories based on the medium they mainly manipulate. These are the categories of visual, audio, and text-based deepfakes. Each category includes different manipulation techniques, as shown in Figure 4, which illustrates each type and its applications.

A. VISUAL-BASED

Visual deepfakes involve manipulating or synthesising images and videos using DL-based techniques to alter facial expressions, gestures, lip movements, and even entire body movements of individuals. They appear to say or do things that they never actually do. These deepfakes can also be used to impersonate individuals by superimposing their faces on someone else's body in videos and images. A prominent example of a video deepfake that went viral in 2018 depicted Mark Zuckerberg, the founder of Facebook, in a manipulated video supposedly announcing plans to shut down the social media platform and pay for all its services. However, the video was entirely fabricated using deepfake technology, and Mark Zuckerberg did not make any such announcement [49]. Visual deepfake manipulation techniques were discussed in depth in [1], [11], and [13]. This section explores the types and applications of deep-fake visual manipulation.

1) FACE SWAPPING

Face swapping is one of the most widely discussed deepfake types. It involves replacing a person's face in an image or video with someone else's face, creating a seamless and realistic blend [50]. Various face-swapping techniques employ deep neural networks to swap facial attributes while preserving the original visual context [51]. An example is shown in Figure 5 for demonstration purposes. The main goal of this manipulation is to achieve a convincing identity swap in which the naked eye cannot differentiate between fake and real. A good application of this type can be seen in the entertainment industry, where it can be used to create impressive visual effects and realistic character transformations [3], [13]. For example, in movies, face-swapping can be used to seamlessly integrate deceased individuals [22] or to seamlessly replace actors with digital doubles for specific scenes or stunts. However, since the introduction of publicly available face-swapping tools such as Faceswap [52], RSGAN [53], FakeApp [30], DeepFaceLab [6], and Deepfakes Web [5], it has become very easy for average users to try and use these tools to swap faces for the good or bad. A bad application is when face-swapping is used for malicious purposes, such as creating deepfake videos to spread misinformation and defame individuals. One prominent example is the use of facial swapping to create non-consensual pornographic videos [54].

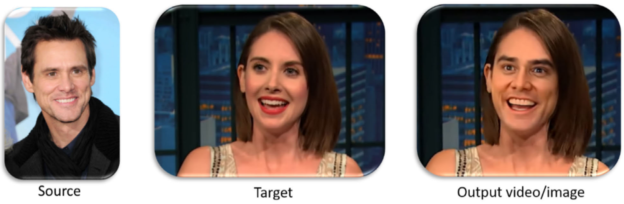


FIGURE 5. Face swap example where Jim Carry's face was swapped into Alison Brie's video using DeepFaceLab [6].

2) FACE GENERATION

The other name for this type is face synthesis, which uses deep learning and involves generating new photorealistic facial images that do not exist in reality and do not represent existing identities [14]. To create realistic images, deep learning models are trained on vast datasets of human faces to learn patterns and characteristics [55]. The generator network then produces new faces by sampling the learnt latent space. Identity preservation, symmetry, and texture are a few challenges posed by the face generation process. For a realistic generation, numerous semantics, including age, position, expression, and style, must be considered [56]. A positive application of face synthesis is in digital art and character creation for video games, in which artists can generate unique and diverse character designs without relying on real-world references. However, a bad application arises when face synthesis is used to create fake profile pictures for fraudulent social media accounts, leading to identity theft or the spread of disinformation [57]. Notable publicly available face generation tools include "This Person Does Not Exist" (TPDNE) [58], using GANs to produce realistic human faces that do not correspond to real individuals. Additionally, advanced models, including Midjourney [10] and

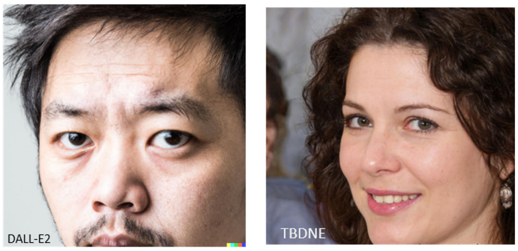


FIGURE 6. Face generation with DALL-E2 [9] and TBDE [58].

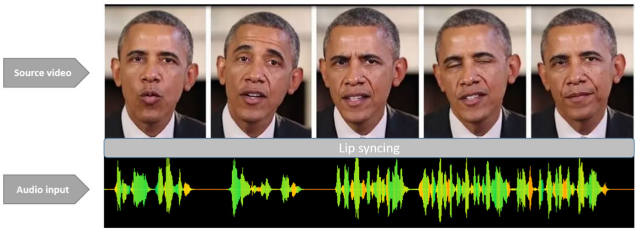


FIGURE 7. Lip-syncing illustration.

DALL-E2 [9] are readily accessible online, generating faces instantly from simple text prompts. For example, In DALL-E2 [9], writing "generates a face for a 30-year-old Asian male with a moustache" in seconds and produces a very realistic face, as shown in Figure 6.

3) LIP-SYNCING

Manipulating a person's lip movements in a video to synchronise them with different audio clips creates an illusion of accurate speech alignment. For example, Figure 7 shows the creation of a smooth lip-syncing video that matches a different audio source. This is a challenging process because the appearance and movement of the lower face, specifically the lip region, are essential for the formation of realistic lip synthesis [11]. Lip-sync deepfakes have legitimate uses in the entertainment industry, dubbing, and video games [59]. They can also be misused to spread misinformation. Manipulated videos of public figures or celebrities can be used to distort their speech and propagate false information, undermining trust in the media and public figures [60]. An infamous example of a lip-sync deepfake is Jordan Peele's deepfake video [49], in which former President Barack Obama's speech is manipulated. Created using lip-syncing technology, the video showcased Peele, mimicking Obama's speech and spreading false statements attributed to the former president. This video was generated using a technique pioneered by Suwajanakorn et al. [61], based on a 3D lip-sync technique that used recurrent neural networks to map audio features to different mouth textures. Although this example was intended to warn against fake news, it also highlights the potential for the deceptive use of lip-syncing deepfakes and their effects and sequences.

4) REENACTMENT

The other name is puppeteering which involves animating a target face or body in a video by replicating the facial expressions or body movements of a source driver. This generates realistic and coordinated movements throughout the video sequence, ensuring temporal coherence and natural actions. The goal is to achieve a realistic and seamless reenactment, in which the target mimics the movements and expressions of the source driver. However, achieving this is challenging because of identity-preserving issues such as identity mismatch or leaking [62]. Examples of publicly available reenactment tools include Face2Face [63], FSGANv2 [64], MarioNETte [65] and DeepFaceLab [6]. These tools use facial tracking and reenactment techniques to demonstrate facial expression and movement transfer. "Everybody Dance Now" [66] is a body reenactment deepfake method developed by researchers from UC Berkeley and Adobe Research. It is designed to reenact a person's dance moves onto a target video, essentially allowing the target person to dance like the source person. This technique utilises generative adversarial networks (GANs) and estimation techniques to achieve this. A beneficial application of reenactment can be seen in animations and movies, where it can streamline the creation of lifelike and expressive character animations. For example, they can be used in films or television shows to facilitate challenging or risky scenes, allowing actors to perform actions that would otherwise be physically difficult [29]. However, a harmful application is when this deepfake is used to create videos of individuals saying or doing things they never actually did. They can even be used to create fabricated evidence or videos that falsely implicate individuals in crimes or unethical behaviour [67].

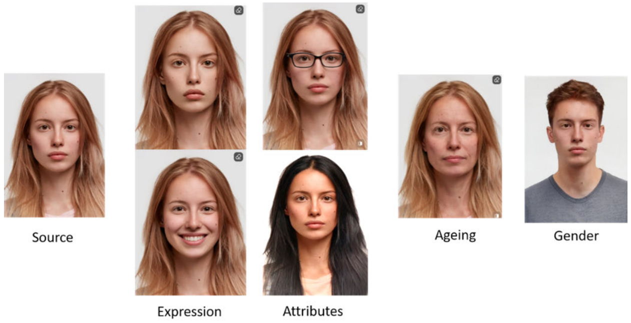


FIGURE 8. Facial manipulation examples using FaceApp [7].

5) FACIAL MANIPULATION

This manipulation type involves altering individuals' facial expressions or attributes in images and videos, allowing for modifications, such as changing emotions or mimicking specific facial gestures. Deep learning techniques are utilised to analyse and modify facial expressions while preserving the overall appearance of a person [68]. A real-world example is FaceApp [7], a popular mobile app primarily used for entertainment. It uses artificial intelligence (AI) and deep learning to apply filters and effects to user faces (e.g., Figure 8) to generate realistic transformations, such as making users look younger or changing their gender. It is primarily used for entertainment. Although facial expression manipulation deepfakes have legitimate applications in fields such as entertainment and advertising, it is crucial to acknowledge their potential for misuse.

B. AUDIO-BASED

Audio deepfakes are created using artificial intelligence and machine learning techniques. The process involved training a deep learning model on a dataset of audio recordings of a specific person's voice. This data set helps the model learn the unique characteristics, tone and speech patterns of the target person [69]. Once the model is trained, it can generate new audio content that closely resembles the voice of the person on which it was trained. The generated audio can be used to create synthetic speech or manipulate existing audio recordings, making it appear that the person said or recorded something that they did not. An example of a deepfake incident based on audio occurred when the CEO of a UK-based energy firm was scammed to transfer €220,000 to a Hungary supplier. The scammer used voice conversion techniques to mimic the voice of the firm's German boss [70], [71]. Audio deepfakes can be used alone as audio clips or can be integrated with video deepfakes to generate audio-visual content. Within this category, two types of manipulation were identified: text-to-speech and speech-to-speech or voice conversion.

1) TEXT-TO-SPEECH

This manipulation technique involves synthesising typed-in text into speech that mimics a specific person. Text-to-Speech (TTS) is an old technology used in various applications such as voice assistants, navigation systems, and speech-to-speech translation [11], [72]. However, traditional TTS approaches lack naturalness and fail to resemble human-like speech [73]; speech sounds are artificial and can easily be noticed. Recent advancements in deep learning-based approaches, such as WaveNet [73], Deepvoice [74], Tacotron [75], and NaturalSpeech [76], have significantly enhanced the TTS. These models analyse extensive audio data to accurately mimic a person's and capture subtle nuances, intonations, and speech patterns for highly natural and expressive voice imitations. This technology has revolutionised voice synthesis in applications such as virtual assistants [77] and personalised voice avatars [78]. However, the ease of access to DL-based TTS synthesis models raises concerns regarding potential misuse. They can be used to create fake audio messages and impersonate voices for threatening or defamatory purposes. Such abuse could facilitate social engineering attacks [13], scam calls [79], and dissemination of false information through manipulated voice recordings.

2) SPEECH-TO-SPEECH / VOICE CONVERSION

Unlike Text-to-speech, this technique modifies a speaker's voice characteristics to match another person's or a pre-defined target voice. It clones the target voice and speech characteristics by transforming the speech signal while preserving linguistic content and other aspects of the target speaker's identity [80], [81]. The speech-to-speech conversion model pipeline, as illustrated in Figure 9, enabled by AI-driven algorithms, finds applications in various domains. For instance, it allows individuals to create personalised voice assistants for a more engaging user experience [82]. In the entertainment industry, voice conversion is used to add subtitles to foreign films [13] or TV shows with the original actors' voices. Additionally, voice conversion can aid individuals with speech disabilities [83] in effectively communicating by generating speech that matches their authentic voice. However, user-friendly online voice synthesis tools such as Resemble.AI [84] and MURF.AI [85] can be hazardous because they can be used to create audio content to deceive individuals by impersonating the target's voice for fraudulent purposes. These deepfakes can be employed for malicious activities, including identity fraud and the spreading of misinformation.

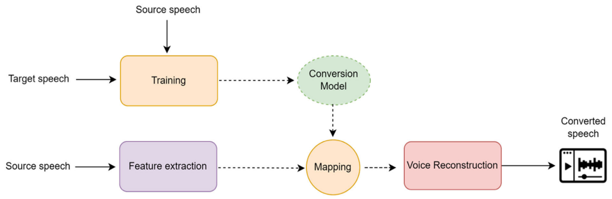


FIGURE 9. Typical Speech-to-Speech conversion pipeline influenced by [77].

C. TEXT-BASED

Text deepfakes pertain to generating written content, such as articles, reviews, forums, or social media posts, using DL language models [86]. These models can replicate human beings' writing style, tone, and vocabulary, creating text that appears to be authored by a specific person when it is generated by AI. Some text-based deepfakes are created and distributed by bots, presenting significant challenges to online platforms [87]. This section identifies two text-based deepfakes: synthetic text and bot-AI-generated text.

1) SYNTHETIC TEXT

These deepfakes are artificial texts generated by deep learning models, specifically large language models (LLM) trained on large corpora of human-written content. A prominent example is the Generative pretrained transformer (GPT) model developed by OpenAI [48], [88]. These models can generate human-like text, complete with contextually relevant responses, making it challenging to distinguish between human and AI-generated text [88]. The primary concern with synthetic text lies in its misuse in creating persuasive and seemingly genuine fake content. Given the proficiency of LLM in generating human-like text, these models can be used to generate misinformation, produce fake news [89], or create seemingly genuine reproduced content such as articles or blogs. For instance, an adversarial actor can use these models to create a credible fake news article, potentially spreading misinformation or sowing discord. The same technology can also be used to generate convincing phishing attacks, which pose a cybersecurity threat [90]. However, it is important to remember that synthetic text generation has positive applications, including content creation, translation, summarisation, personalised learning, and more [48].

2) BOTS AI-GENERATED TEXT

It primarily involves the use of automated programs, known as bots, to create and disseminate text on a large scale. Social bots and chatbots are often used on social media platforms and websites and can post, comment, and interact with other users or content. The advancement of DL and LLMs has enhanced the capabilities of bots as they can now generate automated text that closely mimics human written text [8]. On the one hand, bots have positive uses, such as automating customer service responses, providing personal assistance, or automating routine posts for businesses on social media. However, the primary misuse of bots in text-based deepfakes involves the propagation of misinformation, spamming, or manipulation of public sentiment on social platforms [91], [92]. For instance, malicious social bots can be utilised during elections to amplify specific narratives or disseminate false information about candidates, thereby influencing public opinion [93]. In another example, during periods of public comment on official governmental websites, bots can exploit system vulnerabilities and inundate the platform with deepfake comments. These comments can be challenging to differentiate from genuine human submissions, further exacerbating the issue [94].

V. IMPACTS OF DEEPFAKES

In our digitally connected world, rapid advancements in artificial intelligence (AI) and the widespread use of social media platforms have reshaped the creation, sharing, and consumption of information. Deepfakes have emerged as potent tools with far-reaching implications. Owing to the sophistication and accessibility of deepfake technology, it is now possible to generate highly convincing yet entirely falsified media content [29]. According to Reuters [95], "DeepMedia", a company specialising in synthetic media detection tools, has reported a significant increase in the number of deepfakes online in 2023. Compared to 2022, there has been a tripling in video deepfakes and an eight-fold increase in voice deepfakes this year. The company estimates that approximately half a million video and voice deepfakes will be disseminated on global social media sites by 2023. This capacity to distort reality has the potential to significantly impact various aspects of society.

The effects of deepfakes are felt in various areas including politics, society, economics, and technology. Deepfakes pose a significant threat [96] with far-reaching consequences, such as the spread of false information, manipulation of public opinion, the potential for financial fraud, and the loss of trust in digital media [97]. These risks are substantial, and addressing them is complex. Considering the various aspects involved, this section offers a thorough analysis of the effects of deepfakes. It explores the dangers and possible risks in these significant areas. Table 3 illustrates a summary of the impacts and scale, which is used throughout the following subsections.

A. POLITICAL IMPACT

Politics represents one of the most critical areas affected by deepfakes because of its significant influence on public opinion [20], policy making, and international relations [101]. One of the primary concerns is the use of deepfakes to spread misinformation and disinformation by creating false speeches, actions, or scenarios involving political figures [122]. An alarming example of this scenario is the circulated video depicting a deepfake version of the Ukrainian president [123]. In the video, which lasted approximately one minute, the fabricated representation of the president appeared to instruct Ukrainian troops to cease fighting against Russia and surrender their weapons. Deepfakes of this type can distort reality, mislead the public, and manipulate political views. In the context of elections, the misuse of deepfakes is particularly consequential. For example, a fake video could falsely portray a politician making controversial remarks, potentially damaging his reputation [26] and influencing public opinion or voting behaviour [3].

According to Dobber et al., the combination of political micro-targeting techniques and deepfakes can pose a significant threat to the sanctity of elections [102]. With the ability to create and disseminate false narratives about candidates, deepfakes can undermine political campaigns and sway public opinion [102]. Another aspect of the ability of deepfakes to potentially influence the results of an election is that when the distribution of falsified content is timed correctly, it could circulate widely before there is an adequate opportunity for the victim to discredit it effectively [124] because election periods are usually limited in time and the impact of deepfakes in such scenarios can be irreversible. Moreover, the growing sophistication and use of social bots, especially with the rise of deepfakes, raises significant concerns due to their potential to create and disseminate deceptive information [99]. As evidenced during the 2016 US Presidential Election, social bots comprised 14% active users and generated approximately 20% of all tweets [125]. These bots can mimic human behaviour, leverage sentiment analysis techniques to align their content with public opinion [126], and strategically target influential users to broaden their impact [99].

Integrating deepfakes can further enhance their content's persuasiveness and perceived authenticity, making it even more challenging to discern the truth from fabrication. The significant role of social bots in shaping public

TABLE 3. Deepfakes impacts in terms of scale.

Impact	Low	Medium	High
Political	• International diplomacy [82], [98]	• Reputation damage [26] • Voting behaviour [3]	• Public opinion [20], [99], [87], [100] • Policy-making [101] • Election results [102] • Inciting violence or unrest [103], [104], [105]
Social	• Fraudulent schemes [106] • Revenge porn [107] • Mental well-being [108] • Anxiety [105] • Feelings of violation [109]	• Media ecosystem [110] • Influence perceptions about Journalism and figures [20], [111] • Cyberbullying, harassment [112] • Evidence credibility in courts [26], [113]	• Trust in online information and digital, media [20], [114] • Criminal activities or acts of terrorism [67] • Defamation and privacy rights [93] • Regulatory jurisdiction [107], [115]
Economic	• Corporate espionage and damage to brand reputations [54], [116] • Unauthorised access to secure systems [117]	• Financial frauds [118], [70], [104]. • Misleading consumers [116] • Cybersecurity threats [118], [119]	• Consumer perceptions and disrupt market, Dynamics [29] • Dynamics [29] • Economic instability [117] • Sophisticated phishing attacks [120] • Content verification [121]

discourse, particularly when amplified by deepfakes [87], [99], [100], underscores the urgent need for robust detection and mitigation strategies to safeguard the integrity of online discourse and democratic processes. Deepfakes can significantly impact uprisings and social movements. In situations where public sentiment is already charged, the dissemination of realistic deepfakes could further inflame tensions and potentially incite violence or unrest [103], [104], [105].

Furthermore, the impact of deepfakes extends to international diplomacy [82]. Deepfakes have the potential to create diplomatic incidents or falsely portray a leader's statements or actions, inciting international tension or even conflict. For example, a strategically released deepfake designed to stir public sentiment can be disseminated during a crucial international summit. The controversial content of deepfake might create such a political stir that it becomes practically impossible for one party to pursue its planned agenda [98].

B. SOCIAL IMPACT

In today's digital era, the speed and scale of information dissemination have been greatly amplified by social media platforms [89]. However, these advantages also come with notable challenges, especially with the advent of artificial intelligence, which has profound implications for social landscapes. Concerns about maintaining trust in online information have grown among social media platforms and government bodies [114]. The European Commission took proactive measures by creating an expert panel to develop 'ethics guidelines for trustworthy AI' [127]. These guidelines serve as a preliminary structure to oversee AI development within the European Union. The problem of untrustworthy AI reaches a critical point with the emergence of deepfakes, which are frequently depicted as a significant threat to online trust [114].

Recent research has provided useful information on the possible effects of deepfakes on journalism within controlled environments, illustrating that exposure to such manipulated content can influence the perceptions of news outlets and political leaders, for example, the following studies present work on this topic: [20], [102], [111]. Deepfakes, with their potential to create realistic but falsified audio-visual and text content, have introduced a new level of uncertainty into the media ecosystem [110]. Trust in digital media, which has been the foundation of online interactions, is eroded as deepfakes increase [20]. People are increasingly questioning the authenticity of digital content. As deep fakes become more common, they plant seeds of doubt and confusion, affecting the open exchange of information and compromising the integrity of online content [15].

The social dynamics of online interactions were also significantly affected. Consider a situation in which deepfakes are used to fabricate video calls or to put words into people's digital mouths. Such manipulations can lead to misunderstandings, mistrust, and strained interpersonal relationships, thereby altering the landscape of online communication and interaction. Recently, there has been an instance of deepfake misuse involving Martin Lewis, a well-known consumer finance expert. A fake advertisement featuring an incredibly realistic deepfake version by Lewis was circulated on Facebook. The fabricated Lewis can be seen backing a supposed investment scheme, claiming that it is supported by Elon Musk and labelling it as a ''great investment''. This video is dangerous due to its convincing nature and its potential to deceive unsuspecting individuals into fraudulent schemes [106]. Furthermore, deepfakes have emerged as new tools for cyberbullying, harassment [112], and the perpetration of revenge porn [107]. Deepfakes can be used to publicly shame or bully targets by superimposing an individual's face onto inappropriate or explicit content. A notable example is the disturbing trend of ''deepfake pornography'' in which an individual's face, often a female, is grafted onto explicit content without consent. Such uses can inflict significant emotional trauma, invade privacy, and cause reputational damage [128].

According to a recent report from University College London (UCL), deepfakes have been identified by experts as the most concerning application of artificial intelligence, mainly because of their potential use in criminal activities or acts of terrorism [67]. Furthermore, deepfakes pose challenges to mental well-being [108]. Fear of becoming a fake target can induce anxiety [105], and being a victim of a fake attack can cause feelings of violation [109]. In a broader sense, the general mistrust and uncertainty bred by deepfakes can foster a sense of cynicism and disconnection in society.

In the documentary ''My Blonde GF'' [129], directed by Rosie Morris, the adverse effects of deepfake technologies are brought to light. The film follows the ordeal of writer Helen Mort, who discovered that her face had been used without her consent to deepfake pornographic images. The images, believed to have been sourced from Mort's old Facebook account and other public photos, were manipulated into explicit scenes, causing significant distress. The documentary also emphasises Mort's feelings of powerlessness and violation, as she did not know who was responsible for creating and distributing the deepfake images. Despite her ordeal, the police could not prosecute the perpetrator, as creating deepfake images is not currently classified as a crime. This incident underscores the legal loopholes and the urgent need for legislation that addresses the issues raised by deepfakes [130].

The proliferation of deepfakes within the legal domain raises serious concerns, particularly regarding the credibility of evidence in courts, defamation, privacy rights, and regulatory measures [113], [115]. The ability of deepfakes to create convincingly authentic media of events or statements that never occurred calls into question the trustworthiness of traditionally regarded photographic and video evidence in court proceedings [26]. Furthermore, this distortion of reality poses critical issues related to defamation and invasion of privacy [93]. Moreover, with the rapid advancement of artificial intelligence and deepfake technology, regulations often struggle to keep up, leaving many jurisdictions without clear rules on the creation and dissemination of deepfakes [107]. This regulatory gap can leave victims of deepfake-induced harm with limited legal options. There is an emerging debate regarding the legal responsibilities of AI developers and social media platforms. These parties might be held accountable to prevent the misuse of deepfake technology and mitigate its damaging effects [131]. However, the legal landscape is beginning to adapt to the threats posed by deepfakes. For example, amendments to the Online Safety Bill have criminalised the sharing of explicit deep-fake images or videos in England and Wales without the depicted person's consent [132]. The increasing prevalence of deep fakes underscores the urgency of developing robust legal frameworks that can adapt to the evolving digital media landscape and effectively address the challenges that this technology presents.

C. ECONOMIC IMPACT

According to the World Economic Forum, there has been a substantial increase in deepfake content, with an increase 900% between 2019 and 2020. Forecast trends indicate that this alarming surge is likely to continue, and some researchers predict that almost ''90% of online content could be synthetically generated by 2026''. This rapid proliferation of deepfakes, frequently employed for deceptive and social engineering purposes, has become a significant concern for businesses [118].

In the economic domain, deepfakes present complex challenges that span the corporate, financial, and marketplace spheres. Corporate espionage or sabotage facilitated by deepfakes can lead to significant financial losses and damage the brand's reputation. For example, a fake video that falsely depicts a CEO making controversial statements or revealing sensitive information can severely affect company stock prices and stakeholder trust [54]. Furthermore, deepfakes can be manipulated for financial fraud, creating opportunities for identity theft and unauthorised transactions [104]. For example, by synthesising voices or creating counterfeit identities, deepfakes can deceive individuals and systems alike. A prominent case occurred when a fake audio clip that mimicked the voice of a CEO tricked an executive into transferring €220,000 to a fraudulent account [70]. The banking sector is particularly susceptible, with 92% of practitioners expressing concerns about the potential misuse of deepfakes. According to a report by the World Economic Forum, companies across various sectors have experienced significant financial losses due to deepfake fraud, with some large companies losing up to $480,000 in the past year [118]. The impact of deepfakes also extends to the marketplace, where they manipulate consumer perceptions and disrupt market dynamics. Deepfake videos can falsely depict a product malfunction or unethical business practice, leading to a loss of consumer trust and reputation damage [29]. Conversely, these technologies could be used unethically to falsely promote a product or service, misleading consumers [116]. In addition, deepfakes could be deployed in disinformation campaigns to spread damaging false information about competitors, distorting consumer choices [54]. On a broader economic level, large-scale misuse of deepfakes can undermine public confidence in critical financial institutions. For example, spreading false information about economic indicators or policy decisions could trigger an undue panic or overconfidence among investors, leading to economic instability [117].

D. TECHNOLOGICAL IMPACT

Deepfakes pose significant challenges within the technological domain, particularly concerning cybersecurity. They represent a potent evolution in the complexity of cyber threats and can fuel various cybercrimes, such as phishing, identity theft, and digital espionage. As the World Economic Forum highlighted, deepfake attacks have become a significant concern within organisations, with 66% of cybersecurity professionals reporting such encounters in 2022 alone [118]. These attacks frequently involve fraudulent audio messages crafted using voice-altering software to impersonate high-ranking executives, coercing unauthorised money transfers, or sensitive information [118]. Phishing attempts, traditionally characterised by deceptive emails or messages aimed at extracting sensitive information [133], could be exponentially enhanced using deepfakes [120]. Attackers can exploit deepfake technology to create convincing audio or video impersonations of trusted individuals or authority figures, significantly increasing the persuasiveness of these cyber traps. The implications of such sophisticated phishing tactics can range from substantial security breaches to data theft or financial losses.

According to the Financial Times, in June 2023, Progress Corp., a software company, fell victim to a security breach by a hacking group called Cl0p. They exploited system vulnerability and stole sensitive data from multiple organisations, including British Airways, Shell, and PwC. Experts believe that stolen data, such as millions of American driving licenses and health records, could be used in identity theft scams combined with deepfake software, yielding greater profits than standard corporate ransom demands [119]. Deepfakes also pose considerable threats to identity theft and digital espionage. Malicious actors can manipulate deepfakes to create convincing fake identities or impersonate existing ones, thus facilitating unauthorised access to secure systems [117]. In addition, deepfakes significantly undermine content verification. As they become increasingly realistic and challenging to detect, they can erode the authenticity and credibility of digital content [121]. This intensifies the difficulty of verifying online information, exacerbates misinformation issues, and undermines trust in digital spaces.

VI. DEEPFAKE DETECTION

The emergence of deep learning technology has revolutionised the field of multimedia forensics [134], [135], prompting the need for innovative and timely solutions. Although a multitude of research efforts and forensic tools have been dedicated to detecting anomalies, such as lighting variations [136], shadow inconsistencies [137], and colour illuminations [138], they now face the challenges brought on by the sophistication of deep learning [134]. This dynamic landscape has stimulated a surge in research in multimedia forensics, with a marked focus on exploiting deep learning techniques. Comprehensive explorations of these advancements have already been provided in numerous studies [139], [140], [141]. In tandem with these developments, protecting the integrity of digital media assets has become of critical importance, sparking growing interest in various safeguarding methods, such as image hashing [142], [143]. Innovative technologies such as blockchain [144], smart contracts [145], and cryptography [146] are also utilised for authentication purposes. Furthermore, pioneering active techniques have emerged to fortify the integrity of digital media [147], which signifies promising avenues for future exploration. However, this study primarily provides an overview of specific deepfake detection approaches that employ machine learning and deep learning techniques.

A. VISUAL DETECTION

Venturing the domain of deepfake visual detection reveals a sophisticated intersection between digital security and media integrity. The methodologies employed primarily gravitate towards two pivotal categories: Machine Learning (ML) and Deep Learning (DL) methods. Each tactic demonstrates unique capabilities to address the intricate challenges of deepfakes. For instance, traditional algorithms such as Support Vector Machines or Decision Trees often underpin ML-based methods. On the other hand, DL-based techniques harness advanced architectures, such as Convolutional Neural Networks or Autoencoders, offering a unique perspective to tackle this problem. Subsequent exploration in this section specifically emphasises an overview of state-of-the-art visual deepfake detection techniques. This primarily includes an analysis of two distinctive categories of detection methods. First, methods based on handcrafted features, an approach founded on manual design and extracting specific, recognisable features from visual content, provide tangible manipulation indicators. Second, methods based on deep features are a contrasting approach in which deep learning models automatically learn intricate, high-level patterns from extensive data, thereby distinguishing authentic visuals from manipulated ones. As we dive deeper into these domains, we will explore their principles, applications, and collective significance in the ongoing battle against deepfakes.

1) METHODS USING HANDCRAFTED FEATURES

The main idea behind face swap detection is to identify and flag instances in which the original face in an image or video has been replaced or overlaid with another face, typically using machine learning or deep learning techniques. At one end of the spectrum, Zhang et al. [148] and Yang et al. [35] leveraged the classification process of the Support Vector Machine (SVM). The former developed a novel approach based on Speeded Up Robust Features (SURF) to detect swapped faces in images, albeit with limited success when dealing with manipulated videos. The latter takes a distinctive path by extracting features from the 3D head positions calculated from 2D facial landmarks. While demonstrating promising results, this technique faces challenges in estimating the landmark orientation in blurred images. Matern et al. [149] used a logistic regression classifier to detect simple visual discrepancies in facial video frames, achieving commendable accuracy in simple face swaps or reenactment scenarios. However, its performance may diminish with more sophisticated deepfakes.

Güera and Delp [150] and Ciftci et al. [151] proposed unique solutions in the field of video manipulation. The technique proposed by Guera et al. is built on multimedia stream descriptors, extracting features to differentiate real and manipulated faces within video samples. However, this method underperforms when confronted with video re-encoding attacks. The approach of Ciftci et al. revolves around computing biological signals from facial portions of videos. Despite its potential, its effectiveness diminishes when dimensionality reduction techniques are applied to an extensive feature vector space. By exploring anomaly-based techniques, Jung et al. [152] introduced an innovative method that identifies deepfakes by spotting abnormal eye-blinking durations within videos. Although it shows potential, the effectiveness of the method diminishes for subjects suffering from mental illnesses, which are known to exhibit atypical patterns of eye blinking.

Amerini et al. [153] proposed a deepfake detection technique that employs the difference in optical flow fields to distinguish between genuine and manipulated videos. This method is particularly sensitive to anomalies in the temporal dimension of video sequences. By estimating the optical flow fields of frames, they used these representations to train convolutional networks, specifically "VGG16 and ResNet50", to differentiate between real and fake content. This innovative approach uses both spatial and temporal cues inherent in videos to detect deepfakes. Agarwal et al. [32] proposed a deepfake detection method that leverages OpenFace2 [154], a toolkit for facial behaviour analysis. These techniques extract and analyse spatial and temporal features related to facial movements and head positions to identify deepfakes, particularly those involving world leaders. Its primary limitation is that it is optimised to detect manipulations in videos with faces in direct frontal positions. In another study, Agarwal et al. [155] addressed deepfake detection by targeting lip-sync discrepancies. Specifically, they identified mismatches between phonemes (auditory speech units) and visemes (visual speech units). Despite demonstrating robust detection accuracy, their method underperforms when applied to previously unseen videos.

Korshunov et al. [156] developed a method to detect tampered speakers using phonetically aware audiovisual features. This approach identifies inconsistencies between audio and visual speech, proving to be effective for detecting manipulations, including lip-sync and reenactment. However, they may struggle with low-quality videos, complex languages, or sophisticated deepfake generation techniques. Shahzad et al. [157] presented a multimodal deepfake detection technique that identifies mismatches between video-extracted lip sequences and synthetic lip sequences generated from audio using the Wav2lip model [158]. Despite outperforming many existing methods on the FakeAVCeleb dataset, its performance can be impacted by lip occlusion, non-frontal faces, and adversarial attacks.

Face synthesis detection involves identifying synthetic facial images or alterations made using deep learning algorithms such as Generative Adversarial Networks (GANs). This is done by analysing specific technical features, such as inconsistencies in lighting, texture, or positioning of facial landmarks. Guarnera et al. [159] and McCloskey and Albright [160] contributed to this field using different methodologies. Guarnera et al. utilised an Expectation Maximization (EM) algorithm along with KNN and SVM classifiers to uncover a unique "fingerprint" in deepfake-generated images. Despite its ability to differentiate among various GAN architectures, it struggles when faced with image compression. McCloskey and Albright introduced a new method that uses saturation cues to detect GAN-manufactured images. They exploit distinct patterns of colour saturation as manipulation markers, demonstrating superior capabilities to other techniques mentioned in [148] and [161]. Zhang et al. [162] also focused on identifying and simulating artefacts unique to GAN-generated images using frequency spectrum features; however, their model was limited to image detection only. Refer to Table 4 for a comparison summary of the visual deepfake detection methods based on handcrafted features.

2) METHODS USING DEEP LEARNING

In contrast to handcrafted feature-based methods, deep learning feature-based methods utilise algorithms to independently learn intricate patterns from extensive datasets. Instead of employing predetermined attributes, these methods dynamically learn features from data and capture complex patterns that differentiate between authentic and manipulated content. This flexibility makes deep-feature-based methods a powerful asset for combating deepfakes. Afchar et al. [68] proposed a compact neural network, MesoNet, designed to detect video deepfakes. MesoNet identifies subtle changes characteristic of deepfakes by analysing mid-level features from the video data. Its compact structure offers efficient deepfake detection even in scenarios with limited computational resources. However, it might face challenges when dealing with low-quality videos, where subtle alterations characteristic of deepfakes could be harder to discern.

Cozzolino et al. [163] contributed to deep-feature-based detection methods with their approach, which uses convolutional neural networks (CNNs) for weakly supervised domain adaptation. The DL model learns to recognise intricate features in the source domain and adapts these insights to the target domain. By discerning the deep features associated with data alterations, the model enhances its ability to detect manipulations or forgeries and requires only a few training samples. Other notable CNN-based methods include multitask learning and segmenting [179], co-occurrence matrices [180], GAN stable fingerprints [181], incremental learning [19], attention mechanism [182], [183], 3D attention [184], and CNN in combination with SVM [185], [186]. Nguyen et al. [164] proposed an approach to detect

TABLE 4. Visual deepfakes detection methods based on handcrafted features.

Source	Method	Features	Dataset	Performance	Limitation
Zhang et al. [148]	Detecting face swap in images using SVM classifier.	SURF local descriptor, 64 key points	Self-generated based on LFW	Accuracy = 92%	Works on images only, not tested on more advanced face swaps.
Yang et al. [35]	Face swap detection in image and video - SVM Classifier	3D head positions from facial landmarks	UADFV, MediFor	Area Under the Curve (AUC) = 0.89, 0.84	Degrade performance with less quality or blurred faces.
Matern et al. [149]	Log. Regression and MLP classifiers to detect face swap or reenactment in videos	Texture energy facial features	FaceForensics, CelebA	AUC = 0.866	Diminishing performance with more sophisticated deepfakes.
Guera et al. [150]	Random forest /SVM on videos	stream descriptors extracted vector features	Custom MFC	AUC = 0.93-0.96	Degraded performance with re-encoding attacks.
Ciftci et al. [151]	CNN to classify videos	Biological features from faces	FaceForensics	Accuracy = 96%	Performance is affected when dimensionality reduction is applied.
Jung et al. [152]	CNN, SVM	Eye blinking patterns, eye aspect ratio	Eye Blinking Prediction	Accuracy = 87.5%	Degraded with atypical eye blinking patterns
Amerini et al. [153]	Uses CNN-VGG16 and Resnet50 to detect manipulated	Optical flow clues	FaceForenics++	Accuracy = 81.6%	struggles with rapidly moving or hidden objects and uneven lighting.
Agarwal et al. [32]	SVM to detect manipulated faces in videos	spatial and temporal features related to facial movements	Self-generated	Accuracy = 93%	It is identity-specific and only detects faces in direct frontal positions.
Agarwal et al. [155]	CNN Detects lip-synced videos	Phonemes and Visemes mismatches	Self-generated	Accuracy = 99.6%	Underperforms with unseen data.
Korshunov et al. [156]	RNN- LSTM to detect Lip-synced videos	Phonetic audio-visual features	VidTIMIT, AMI, GRID	Equal Error Rate (EER) = 14.1	Struggles with low-quality videos and complex languages.
Shahzad et al. [157]	3D-CNN to detect lip-syncing	Semantic features from lip sequence	FakeAVCeleb	Accuracy = 98.0%	Performance is impacted by lip occlusion and non-frontal face samples.
Guarnera et al. [159]	EM, KNN, SVM to detect face synthesis	GAN- local features	CelebA	Accuracy = 99.8%	Low performance when the image is compressed.

fake images and videos using a Capsule Network. Unlike traditional Convolutional Neural Networks (CNNs), Capsule Networks consider hierarchical relationships between features, preserving these relationships and making them an effective tool for tasks such as detecting sophisticated manipulations in images and videos. Fernandes et al. [165] presented a unique approach to deepfake detection, using RNN and neural ordinary differential equations (N-ODE) to predict heart rate variations in videos, using discrepancies between these predictions and actual heart rates to discern manipulated faces in videos. Despite its innovative approach, this method could face limitations due to its computational complexity, potentially making it less suitable for real-time or large-scale applications. In their study, de Lima et al. [166] introduced a novel approach for deepfake detection. The authors employed Convolutional Networks, which extract spatial (within a frame) and temporal (across frames) information from videos. Specifically, they used a combination of VGG11, a variant of the VGG model, for spatial feature extraction and Long Short-Term Memory (LSTM) networks to capture temporal dependencies. In a similar study, Chintha et al. [167] proposed an approach for detecting video/audio deepfakes, that combines convolutional latent representations with recurrent structures and entropy-based cost functions. This method detects spatial and temporal deepfake signatures using audio and videos. Tested on the FaceForensics++, Celeb-DF video and ASVSpoof 2019 audio datasets, it sets new benchmarks in all categories, drawing inspiration from the XceptionNet [187] architecture. This blend of methods enhances the model's ability to discern between genuine and deepfake videos and fight against deepfake manipulation.

Agarwal et al. [168] proposed a face swap detection method that combines deep learning features and behavioural biometrics to identify manipulated videos. Behaviour biometrics have received great interest lately in the areas of health [188] and key-timing [189], [190], [191], but the behavioural features extracted from videos are also demonstrating to be of great potential. Their approach involves a VGG encoder-decoder network that helps to analyse the

TABLE 5. Visual deepfakes detection methods based on deep learning features.

Source	Method	Dataset	Performance	Limitation
Afchar et al. [68]	A compact neural network Mesonet based on an inception module	FaceForensic++	Accuracy = 98.4%	subtle alterations characteristic of deepfakes could be harder to discern.
Cozzolino et al. [163]	CNN – forensic transfer NN	Self-generated	Accuracy = 70.6-100%	Degrading performance with compressed images.
Nguyen et al. [164]	Capsule Networks - Hierarchical relations between deep features	CGI, RAISE	Accuracy = 98.5-100%	Complex architecture.
Fernandes et al. [165]	RNN-Neural-ODE - heart rate discrepancies	COHFACE,VidTIMIT	Loss value = 0.015-0.056	Computationally intensive.
de Lima et al. [166]	RCNN-LSTM-VGG-Spatiotemporal	Celeb-DF	Accuracy = 98%	Computationally intensive
Chintha et al. [167]	Based on XceptionNet CNN-LSTM - Spatiotemporal	Celeb-DF, Face Forensics	Accuracy = 97.8%	Computationally intensive.
Agarwal et al. [168]	VGG-Encoder-decoder network, Behavioural biometrics	WLDR, Faceforensic, Celeb-DF	Accuracy = 93-99%	Difficulty generalising to unseen data.
Yang et al. [169]	MTD-Net CNN and CDC	Faceforensics++, DeeperForensics Celeb-DF, DFDC	AUC = 0.60-0.99	Degrading performance with cross–database training/testing.
Yang et al. [170]	MSTA-Net encoder-decoder trace generator and BCE loss classifier	Faceforensics++ , DeeperForensics, Celeb-DF, DFDC	AUC = 0.92-0.99	Low performance with fully synthetic faces.
Mittal et al. [171]	Detects lip-syncing - Siamese neural network with triple loss function	DFDC, DF-TIMIT	AUC = 0.84-0.96	Human emotions can be complex and cause detection failures.
Sabir et al. [172]	CNN with RNN, Detects face manipulations in image frames	FaceForensics++	Accuracy = 94.3%	Computationally intensive, degraded performance with disjointed transitions of the frames.
Guera et al. [173]	RNN-CNN-LSTM at the frame level	Self-generated, HOHA	Accuracy = 97%	Only works for very short videos of less than 2 seconds.
Montserrat et al. [174]	RNN-CNN-LSTM - weighting mechanism	Celeb-DF	Accuracy = 98%	Computationally intensive.
Wang et al. [175]	DNN-MNC - Neuron behaviours	Self-collected	Accuracy = 68-98%	Detection can be evaded with adversarial noise.
Haliassos et al. [176]	Multi-input CNN - Irregular semantics	FaceForensics++	AUC = 0.97	Memory storage limitations
Trinh et al. [177]	DPNet - Temporal inconsistencies	Google's DeepFakeDetection DeeperForensics, Celeb-DF	AUC = 90.9-99.2	Computationally intensive.
Ma et al. [178]	3D-CNN - Spatiotemporal	FaceForensics++, Celeb-DF	Accuracy = 99.3%	Limitation in real time and scalability.

appearance and behaviour of individuals in videos. The effectiveness of their technique is evidenced by an Area Under the Curve (AUC) of 99% on the WLDR, FF, and Celeb-DF datasets, and 93% on the DFD dataset. However, despite the high detection accuracy, one limitation of their approach is the difficulty in generalising well to unseen deepfakes. Yang et al. [169] explored two unique methods centred on texture. First, the MTD-Net model [169] emphasises the value of multiscale texture differences in identifying deepfake images. By harnessing the fact that synthetic processes such as deepfake generation often introduce noticeable texture alterations, this model learns to identify these discrepancies, which are often subtle and vary across different regions and scales of an image. Their second contribution, the MSTA-Net [170] model, complements the detection capabilities of the MTD-Net model by additionally generating manipulation traces through multiscale self-texture attention. This innovative strategy aims to unearth the inconsistencies that manipulated content often presents. This model has better generalizability than MTD-Net. These two studies underscore the potential of multiscale texture analysis as a pivotal tool in the detection of deepfake images, highlighting the critical role of texture-based features in distinguishing between genuine and manipulated content. However, both models exhibited significantly lower performance with fully synthetic faces.

Introducing a novel approach to deep-fake detection for lip sync, Mittal et al. [171] employed a deep-learning network inspired by a Siamese neural network and a triplet loss function. Their methodology was unique due to the simultaneous use of audio and video modalities and their perceived emotions. Performance evaluations on two significant datasets, DFDC and DF-TIMIT, revealed high AUC scores, demonstrating the effectiveness of their approach. However, they reported some detection failure cases due to the complexity of human-perceived emotions.

Several compelling approaches that combine Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been presented. Sabir et al. [172] used recurrent convolutional strategies to discern inconsistencies in facial movements in video frames and highlighted the importance of temporal dynamics in video content. However, this approach has high computational requirements and is dependent on the quality of the training data and temporal continuity, potentially posing challenges in real-time detection scenarios or when analysing videos with disjointed transitions. Similarly, Guera et al. [173] proposed a method in which a CNN was used to extract frame-level features, which were then processed by an LSTM network to identify potential inconsistencies between frames. Although this effective use of both spatial and temporal information aids in deepfake detection, it is mainly applicable to short videos due to computational limitations.

Building on these techniques, Montserrat et al. [174] introduced a method that automatically assigns weights to different regions of the face, further refining the focus of the CNN-RNN model. This strategy amplifies sensitivity to subtle signs of manipulation by concentrating on areas that deep-fake algorithms commonly struggle to accurately reproduce. These studies underscore the potential of integrating CNNs and RNNs for effective deepfake detection. They highlighted the crucial role of spatial and temporal feature extraction, the ability to discern subtle manipulations, and the need to address computational demands and adaptability to different video types and lengths. Wang et al. [175] demonstrated a unique approach to detecting AI-synthesized fake faces by monitoring neuron behaviours. The method uses a Mean Neuron Coverage (MNC) metric to capture neuron activation behaviours in a layer-by-layer manner, helping to distinguish between real and fake faces. However, the performance decreases when dealing with the DFDC dataset, which includes voice-swapped videos, an area not covered by their image-focused method.

In an attempt to overcome the generalisation limitations of most models, Haliassos et al. [176] utilised a multi-input CNN. A two-step approach was adopted to focus the network on mouth movements rather than other manipulation-specific cues. Initially, a spatiotemporal CNN, trained in lipreading, is deployed, creating high-level semantic internal representations sensitive to anomalies in mouth movements. This model demonstrated superior accuracy and better generalisation than other CNN approaches. However, it grapples with challenges associated with memory storage limitations, potentially impacting scalability and efficiency when handling large datasets. Trinh et al. [177] contributed a novel technique to deepfake detection with the Dynamic Prototype Network (DPNet), prioritising interpretable visual explanations to support human understanding. In contrast to many current models that rely on a black-box approach, DPNet highlights the temporal inconsistencies that deepfakes often exhibit. The model identifies prototypical real and deepfake videos based on training video distributions by mapping videos to a latent space using a pretrained neural network. Comparing the distance of the test videos to these prototypes allows for a classifiable understanding of deepfakes and provides a defence against adversarial attacks. The effectiveness of this approach is underscored by its robust performance on various unseen datasets. However, a key challenge associated with this method is the complexity and computational complexity of the model architecture.

In their 2023 study, Ma et al. [178] demonstrated that their 3D Attention Network model performs robustly on videos of varying quality, illustrating its strong generalisability. The effectiveness of their approach was validated through cross-dataset testing, where it outperformed many existing methods in the field. Although this intricate structure contributes to its superior detection capabilities, it may constrain the scalability of the model and the real-time application potential. This factor emphasises the need for continued research to balance model performance with computational efficiency in deepfake detection methods. Refer to Table 5 for a comparison of visual deepfake detection methods based on deep learning features.

B. AUDIO DETECTION

As the sophistication of Text-to-Speech [74] and Voice Conversion [73] technologies has escalated, the potential threats posed by audio deepfakes have become more potent, posing significant risks to voice biometric systems and broader social contexts [192]. Various methodologies for audio forensics have been proposed to identify and counter the spoofed audio content. However, the complexity and realism of synthetic speech continue to present substantial challenges that many existing techniques struggle to address adequately [11]. In this section, we delve into the state-of-the-art strategies employed for audio deepfake detection, categorising them into two primary camps: methods based on handcrafted features and methods leveraging deep learning features.

1) METHODS USING HANDCRAFTED FEATURES

Alzantot et al. [193] developed a method for detecting deep sound effects using deep residual neural networks (ResNet) [194]. The method begins by extracting log-Mel-scale spectrogram features from the input audio, which is then processed by ResNet. This model exploits the ability of the network to learn intricate nonlinear relationships, thereby discerning subtle audio cues indicative of a deepfake. Despite its effectiveness and subsequent adoption as a basis for more recent models, [195], [196], [197] This method grapples with generalisation limits, thus potentially impacting its performance on diverse or unseen data sets.

Lai et al. [198] introduced a method using squeeze excitation and residual networks for audio deepfake detection. The process begins by extracting low-level acoustic features, which are then used to create a unified feature map. This map was segmented for a more detailed analysis before input into the DNN. Although effective, it struggles with overfitting, limiting its generalisation across diverse data. AlBadawy et al. [199] proposed a method to detect AI-synthesised speech using bispectral analysis. This technique is based on the premise that synthetic and natural speech exhibit distinct bispectral properties. Although the bispectrum of a signal is a higher-order statistic that provides additional information beyond power spectral analysis, its application in detecting AI-synthesised speech is relatively novel. By identifying peculiar patterns in the bispectra of synthetic speech, this method helps distinguish it from natural human speech. However, the complexity of bispectral analysis may limit its real-time applicability.

Wu et al. [200] developed a technique to identify synthetic speech using a Light Convolutional Neural Network (LCNN) and a special feature called ''Genuinization''. This method aims to increase the contrast between authentic and fake speeches. The LCNN is designed to handle complex audio data. At the same time, the 'genuinization process is a unique transformer that maintains genuine speech traits but modifies counterfeit speech, resulting in a more significant differentiation between the two. Monteiro et al. [201] developed a method to detect deep audio distortion in speaker recognition systems using LCNN and spectral feature representations. However, the effectiveness of this approach in various real-world scenarios warrants further investigation. The authors of [202] proposed a detection method that leverages signal compounding for data augmentation, specifically for detecting logical access attacks such as automatic speaker verification (ASV) systems. Their study presented the novel idea of exploiting the non-linear characteristics of -law and mu-law-based signals companding to generate diverse training examples, thus improving the robustness of the detection system.

Borrelli et al. [203] proposed an innovative method for synthetic speech detection that uses short- and long-term prediction traces. This method conducts a predictive analysis of both the immediate and extended characteristics of audio sequences, demonstrating the potential for synthetic speech detection. However, its effectiveness is limited when dealing with compressed audio samples, indicating the need for further enhancements to maintain performance in diverse and complex scenarios. Aljasem et al. [204] presented a secure automatic speaker verification (SASV) system for audio deepfake detection. Leveraging sm-ALTP features and asymmetric bagging, this approach enables the precise detection of voice manipulations, such as voice cloning and replay attacks. However, its effectiveness in various data sets and advanced deepfake techniques remains undetermined. Gao et al. [205] presented a novel method for the detection of audio deep-fake that used long-range spectrotemporal modulation features. Using a 2D discrete cosine transform (DCT) on a log-mel spectrogram, the system outperforms traditional feature methods such as CQCC [206]. The model leverages spectrum augmentation and feature normalisation to reduce overfitting, resulting in a state-of-the-art system for spoof detection and demonstrating its effectiveness on two external datasets.

In a recent study by Hamza et al. [207], they implemented a Support Vector Machine (SVM) as part of their deep learning and machine learning methodologies to identify fake audio. They extracted critical information from audio samples using Mel-frequency cepstral coefficients (MFCCs). The model was trained and tested on the Fake-or-Real dataset, a recent standard collection produced via a text-to-speech model and partitioned into four subsets. In particular, integrating transfer learning into their model bolsters its capability to effectively detect deepfakes. Recently, Pianese et al. [208] introduced a novel deepfake audio detection approach that focusses on the biometric characteristics of a speaker without referencing specific attacks. This method trains only on real data, ensuring a level of generalisation. The method has shown promising performance and robustness against audio impairment using standard speaker verification tools. This approach mirrors video deepfake detection methods, which rely on high-level biometric features, and employs a ResNet-34 architecture. However, a significant limitation of this method is its inability to work effectively on unseen and fake audio samples. Similarly, Blue et al. [209] presented a unique audio deepfake detection approach that took advantage of articulatory phonetics and fluid dynamics.

Their method estimates the configuration of the human vocal tract during speech, revealing that deepfakes often simulate improbable or impossible anatomical arrangements. With 99.5% recall and 99.9% precision, the approach can identify nearly all deepfake samples, using biologically constrained aspects of human speech that current models cannot replicate. It is a generator-independent, explainable, and generalised detection mechanism, showcasing a distinctive lens for deepfake detection. However, this underscores the ongoing challenge of fully capturing the subtleties of human speech using artificial models. Lim et al. [210] applied explainable AI (XAI) methods for deepfake voice detection, focussing on interpretations accessible to human perception. Their approach used a simple model that combined a convolutional neural network and LSTM with spectrograms used for feature extraction from raw audio data. This simplification makes the model more accessible and interpretable for human understanding, aiding deepfake detection. Doan et al. [211] introduced a novel framework called BTS-E for audio deepfake detection in their work.

It capitalises on the natural correlation between breathing, talking, and silence sounds within an audio clip, assuming that human sounds like breathing are complex for text-to-speech (TTS) systems to replicate. The framework was extensively tested on ASVspoof datasets, revealing that the breathing sound feature significantly enhanced the detection performance. However, the efficiency of this method in increasingly advanced TTS systems remains to be explored. Table 6 includes a summary of audio deepfake detection methods based on handcrafted features discussed in this section.

2) METHODS USING DEEP LEARNING

Lai et al. [212] proposed a novel strategy, the attentive filtering networks (AFN), for audio replay attack detection. The method is built around the idea of emphasising relevant information in an audio signal while suppressing irrelevant noise or manipulation. By analysing different aspects of audio, such as frequency, pitch, and tone, this technique provides a high detection rate for audio deepfakes, namely replay attacks. Although the system shows high detection rates for audio deepfakes, its performance depends on the type of non-linear activation function used in the AFN.

In 2019, Gomez-Alanis et al. introduced two distinctive, yet interlinked methods for detecting audio deepfakes. The first approach [213] presented a novel deep feature extractor for automatic speaker verification (ASV) spoofing detection, which merges a lightweight convolutional network with a Gated Recurrent Unit (GRU) - Recurrent Neural Network (RNN). This innovative system capitalises on the strengths of both convolutional networks for extracting local features and recurrent networks to capture temporal dependencies in audio signals. On the other hand, the second method [214] proposes a robust detection mechanism that utilises a Gated Recurrent Convolutional Neural Network (GRCNN). This model uses the power of convolutional neural networks to extract local features and incorporates a gating mechanism to regulate the flow of information across the network, resulting in better performance in the detection of spoofing. Although both models have exhibited impressive results in detecting spoofed speech on the tested datasets, further research is needed to evaluate their performance under varying conditions, such as different audio qualities and against increasingly sophisticated deepfake generation techniques.

In their research, Wang et al. [192] introduced a novel approach, DeepSonar, which takes advantage of layerwise neural behaviour of a speaker recognition system, a deep neural network (DNN), to distinguish fake voices synthesised. This system capitalises on layer-wise neuron activation patterns, hypothesising that they can discern subtle differences between real and fake voices, thereby providing a cleaner signal to classifiers compared to raw inputs. DeepSonar was tested on three datasets including commercial products in English and Chinese. The results highlight high detection rates and low false alarm rates, suggesting the robustness of DeepSonar against various manipulation attacks such as voice conversion and additive real-world noises. However, the performance of the system degrades in the case of adversarial noise attacks.

Subramani and Rao [214] proposed a unique solution to detect synthetic speech by implementing two compact models, EfficientCNN and RES-EfficientCNN. These models stand out for their efficiency, demonstrating high accuracy with minimal resource demand. The research also explores a novel multitask approach that boosts detection performance without requiring extra labelling information. Additionally, this study marks the first application of transfer learning in adversarial speech contexts, highlighting its potential for resource-limited situations, such as mobile devices. However, the robustness of these models under diverse conditions remains to be extensively explored.

The RW-Resnet is a method presented by Ma et al. [215] for speech anti-spoofing that works directly on raw waveforms. This approach contrasts with traditional techniques, which often rely on spectrograms or other handcrafted features. Working directly with raw waveforms allows the model to exploit more granular and potentially unique audio signal characteristics that may be lost in other representations. The model is based on ResNet [194], a famous deep learning architecture known for its ability to model complex patterns. Despite the potential of this method, it is critical to test its robustness under varying conditions considering its complexity.

In a similar approach based on the ResNet technique, Zhang et al. [216] proposed a novel approach for the detection of fake speech, incorporating a residual network with a transformer encoder. This combination leverages the strengths of the Residual Network in handling complex patterns using the Transformer Encoder's ability to process temporal relationships within audio data. The proposed architecture provides a more in-depth understanding of the temporal dynamics inherent to speech, thereby enhancing the accuracy of synthetic voice detection.

Building on the success of the RawNet2 [217] architecture, Tak et al. [218] developed a method for the detection of synthetic voice attacks. This end-to-end approach utilises CNN and feature map scaling, which act as an attention mechanism; it operates directly on raw audio signals. The model seeks to leverage its proven capability to capture intricate nuanced patterns in raw audio data. While promising in the AVS spoof dataset, examining the model's performance across diverse synthetic voice attacks and various datasets remains essential for fully assessing its robustness and generalisability.

Zhang et al. [219] introduced a different synthetic voice-spoofing detection approach using a one-class learning strategy. This method differs from traditional two-class classification tasks, since it trains the model solely on bonafide samples, essentially learning the distribution of

TABLE 6. Audio deepfake detection methods based on handcrafted features.

Source	Method	Features	Dataset	Performance	Limitation
Alzantot et al. [193]	Deep residual neural networks (ResNet)189.	log Mel-scale spectrogram	ASVSpoof 2019	EER = 0.06, t-DCF = 0.157	Difficulty generalising to unseen data
Chen et al. [195]	Deep residual neural networks (ResNet)189.	linear filter banks, Low-level	ASVspoof2019	EER = 0.012	Difficulty generalising to unseen data
P et al. [196]	ResNet-34 architecture	log Mel-scale spectrogram	ASVspoof 2019	EER = 5.32%	Difficulty generalising to unseen data
Lai et al. [198]	Squeeze-Excitation and Residual Networks	Low-level acoustic and whole utterance features	ASVSpoof 2019	EER = 0.59, t-DCF = 0.016	Struggles with overfitting
AlBadawy et al. [199]	Bispectral analysis and logistic regression classifier	Bicoherence features	Self-generated	AUC = 0.99	Complexity of bispectral analysis
Wu et al. [200]	Light CNN and transformer	Genuinization features	ASVspoof 2019	EER = 0.04, t-DCF = 0.102	Requires training on genuine speech and overfitting issues.
Monteiro et al. [201]	LCNN, Cepstral coefficient, Temporal pooling	spectral features	ASVspoof 2019	EER = 0.018, t-DCF = 0.06	Generalizability to unseen and more sophisticated data needs to be tested
Das et al. [202]	LCNN with data augmentation	log power spectrum	ASVspoof 2019	EER = 0.031, t-DCF= 0.094	Generalizability to unseen and more sophisticated data needs to be tested
Borrelli et al. [203]	RF, SVM, RBF-SVM classifiers	STLT features	ASVspoof2019	Accuracy = 0.74-0.95	Generalizability to unseen and more sophisticated data needs to be tested
Aljasem et al. [204]	SVM, asymmetric bagging	sm-ALTP features	ASVspoof 2019, VSDC	Accuracy = 0.88, EER = 0.052, t-DCF = 0.132	Requires training on genuine speaker's speech
Gao et al. [205]	CNN, MLP	long-range spectro-temporal modulation	ASVspoof2019, FoR, RTVCspoof	Accuracy = 90-98%, t-DCF = 0.074	Generalizability to unseen and more sophisticated data needs to be tested
Hamza et al. [207]	SVM	Mel-frequency cepstral coefficients	FoR	Accuracy = 98.8%	Not evaluated on other datasets
Pianese et al. [208]	Deep residual neural networks ResNet-24	high-level biometric features	ASVSpoof2019, FakeAVCelebV2 , IWA	Accuracy = 91.3-99.9%	Generalizability to unseen and more sophisticated data needs to be tested
Blue et al. [209]	RNN - Bigram discriminator	articulatory phonetics and fluid dynamics	TIMIT	Recall = 99.5%, Precision = 99.9%	The accuracy decreases when processing phonemes that are not vowels.
Lim et al. [210]	XAI - CNN and LSTM	log Mel-scale spectrogram	ASVspoof 2021, LJSpeech	Accuracy = 99.9%	The study is focused on providing an interpretation rather than a detection enhancement.
Doan et al. [211]	RNN-CNN-Transformer encoder, GMM	frame-level LFCC feature	ASVspoof 2019, ASVspoof 2021	EER = 0.087, t-DCF = 37.67	Only detects TTS audio deepfakes

authentic voices. During the detection phase, the model identifies any sample that deviates from the learnt distribution as a potential spoof.

This method was proven to be effective, as shown by the results, and offered a unique advantage in adaptability to unseen spoofing attacks. However, it can face challenges when dealing with complex and diverse real-world data, as the assumption of a consistent distribution of authentic voices may not always hold. Furthermore, the sensitivity of the model to parameter tuning could affect its practical application.

Conti et al. [220] recently proposed a new method to detect fake speech using emotion recognition from a semantic perspective. The authors suggested that synthetic speech often lacks the natural emotional variations present in human speech, which can be used as a possible way to detect deepfakes. The study applied a 3D-CRNN network to analyse audio samples in the first step to detect emotional content. In the second step, a random forest classifier is used to identify deepfakes by leveraging the differences between human and synthetic emotional patterns. Although this innovative approach may be effective in some cases, this study does not fully explore its effectiveness in different emotional ranges or languages. It is also unknown whether this method is robust under various conditions and requires further examination.

Mo et al. [221] presented a new perspective for detecting synthetic speech in their study. They framed it as a challenge to generate out-of-distribution (OOD) data. Their approach involves multi-task learning, which tackles three subtasks simultaneously: reconstructing natural speech, converting fake voices, and classifying speakers. This flexible approach can be integrated into different network structures and input features. The experimental results indicate that their method significantly improves synthetic speech detection and effectively handles both familiar and unfamiliar attacks. This study highlighted the potential of multi-task learning in the field of synthetic speech detection. However, further investigation is necessary to understand its scalability and adaptability to different situations, including advanced deepfake generation techniques and application scenarios.

Recently, Papastergiopoulos et al. [222] explored the ability of two-dimensional convolutional neural networks (2D-CNN) to detect synthetic speech. The study investigated the generalizability of 2D-CNNs across different datasets and used several audio feature representations, such as STFTs and Mel spectrograms. Despite achieving robust results with Mel spectrograms and Mel energies, the authors noted a considerable drop in performance during testing. As revealed by the data distribution analysis, this discrepancy was mainly attributed to the differences between the training and testing datasets. Thus, the study underscores the importance of addressing the issue of dataset diversity and similarity when designing deepfake detection models.

In their 2023 study, Salvi et al. [221] introduced a novel technique called "Audio Folding" for detecting synthetic speech. This method manipulates audio signals and reveals unique characteristics that distinguish genuine speech from synthetic variants. They employed RawNet2 [218] as a detector that operates directly on raw audio, enabling the model to capture more intricate and unique features. By integrating this audio folding technique with RawNet2, the authors successfully demonstrate an improved synthetic speech detection. Despite these promising results, further studies are needed to ascertain its robustness under different audio conditions and to develop more sophisticated synthetic speech generation techniques. Table 7 summarises the comparison summary of audio deepfake detection methods based on deep learning features based on the discussion in this section.

C. TEXT DETECTION

The field of Natural Language Processing (NLP) has witnessed a significant breakthrough with the introduction of Large Language Models (LLMs), such as GPT-3, GPT-3.5, GPT-4, and PaLM [8], [225]. These models are extensively trained on text data and can generate contextually relevant and highly accurate text. They have exceptional zero-shot generalisation capabilities, meaning that they can perform tasks without explicit training [226]. However, the ability of these models to generate text that mimics human-written content presents unique challenges, particularly in detecting AI-generated text.

Although this high-level exploration primarily focusses on methodologies for detecting AI-generated text, a specific interest lies in detecting text produced by AI-operated bots, such as chatbots and social bots. These bots, leveraging advanced large language models (LLMs), can accurately mimic human conversations, potentially facilitating manipulative and harmful activities on online platforms, as discussed in Section III. Bot detection and AI-generated text detection share some overlap. However, they are distinct research domains with unique techniques and applications. For example, bot detection strategies often emphasise behavioural patterns and meta-data [227], while AI text detection primarily analyses textual content. Consequently, understanding bot detection methodologies offers valuable insight into the broader landscape of AI-generated text detection. For a more in-depth exploration of bot detection, including social bot detection, we recommend referring to comprehensive reviews and literature in the field, such as the systematic review by Orabi et al. [228]. These resources provide a detailed examination of bot detection methods, their strengths, weaknesses, and areas of future development. This understanding forms a crucial aspect of our exploration, contributing significantly to the comprehensive view of the field.

1) DISTINCTION LANGUAGE CHARACTERISTICS

Large Language Models (LLMs) generate text that detection methods often seek to classify in a binary fashion: authored by humans or generated by machines [229]. The distinction lies in the identification of unique language characteristics within these categories. Statistical characteristics are fundamental to this classification, using metrics such as the Zipfian coefficient (a measure of compliance with Zipf's law) and perplexity (a prediction uncertainty metric) [230], [231]. Furthermore, models such as GLTR assist in identifying

TABLE 7. Audio deepfake detection methods based on deep learning features.

Source	Method	Dataset	Performance	Limitation
Lai et al. [212]	Dilated Residual Network and Attentive Filtering Networks AFN	ASVspoof 2017	EER = 0.089	For replay attack detection only- Performance depends on the type of nonlinear activation function.
Gomez-Alanis et al. [213]	Light Convolutional Gated Recurrent Neural Network (LC-GRNN)	ASVspoof 2015-2017-2019	EER = 0.0, t-DCF = 0.061	It needs to be tested against varying conditions, such as different audio qualities and increasingly sophisticated generation techniques.
Gomez-Alanis et al. [223]	Gated Recurrent Convolutional Neural Networks (GRCNNs)	ASVspoof 2015-2017-2019	EER = 0.01, t-DCF = 0.02	It needs to be tested against varying conditions, such as different audio qualities and increasingly sophisticated generation techniques.
Wang et al. [192]	DNN - neuron behaviours	FoR, Sprocket-VC, MC-TTS	Accuracy = 99%	Degraded performance against adversarial noise attacks
Subramani & Rao. [214]	EfficientCNN and RES-EfficientCNN	ASVSpoof2019, RTVCSpoof	macro-F1 = 97	Need to be tested against varying conditions
Ma et al. [215]	CNN -Resnet34 - Deep Residual Learning based on ResNet 189	ASVspoof2019	EER = 0.03, t-DCF = 0.08	Complex architecture
Zhang et al. [216]	ResNet 189 with Transformer Encoder	FoR, ASVspoof2019	EER = 3.99%	Complex architecture
Tak et al. [218]	CNN based on RawNet2 [217]	ASVspoof 2019	EER = 3.5%, t-DCF = 0.904	Need to be tested against varying conditions and datasets
Zhang et al. [219]	ResNet [194] with SGD	ASVspoof 2019	EER = 0.02	The model is sensitive to parameter tuning and faces challenges when dealing with complex real-world data.
Conti et al. [220]	3D-CRNN with random forest, Speech Emotion Recognition	ASVspoof 2019, LibriSpeech, LJSpeech, IEMOCAP	Accuracy = up to 100%	The study does not fully explore its effectiveness across different emotional ranges or languages.
Mo et al. [224]	CNN - Resnet34 - Autoencoder, OOD data	ASVspoof 2019	EER = 0.01, t-DCF = 0.036	Need to be tested against varying conditions and datasets
Papastergiopoulos et al. [222]	2D-CNN - VGG16	FoR, LJSpeech, VoxForge, TIMIT	Accuracy = 93%	Drop in performance with dissimilarities between datasets
Salvi et al. [221]	CNN Based on RawNet2 [217], Audio folding	ASVspoof 2019	Accuracy = 0.95	It needs to be tested against varying conditions, such as different audio qualities and increasingly sophisticated generation techniques.

generation artefacts using LLM word ranking data [232]. However, these techniques are targeted at document-level detection, which can compromise their effectiveness in granular detection scenarios [233].

To identify language patterns in writing created by humans and AI, it is necessary to examine various contextual characteristics. Vocabulary features, for example, show that human-authored texts have greater linguistic diversity but are shorter in length [234]. An analysis of parts of speech and dependency parsing reveals that ChatGPT texts tend to use more nouns, determiners, conjunctions, and auxiliary relations, suggesting a focus on making arguments and being objective [234]. An analysis of sentiment showed a noticeable contrast in emotional expressions. Language models, including ChatGPT, generally have a neutral tone and less negative emotional language than texts written by humans [235]. Detecting texts created by language models can involve considering other factors such as repetitiveness, readability, and conversational patterns [233], [235].

2) METHODS BASED ON SIMPLE CLASSIFIERS

These are straightforward machine learning models trained to distinguish between two classes: human-generated and AI-generated text. Examples include logistic regression, decision trees, and Support Vector Machines (SVMs). These models often use various features of the text, such as word frequencies, sentence lengths, or more complex linguistic features, for classifications. Nguyen-Son et al. [236] devised an approach that takes advantage of the distinct statistical features inherent in computer-generated text to distinguish it from human-authored content. By focusing on specific language patterns, structure distribution, and frequency, the model reveals notable disparities in the statistical properties of human and machine-generated texts. An SVM classifier is utilised to categorise the text based on these unique features. The model was tested using a corpus of 100 English and Finnish books, with the former serving as human-generated examples and the latter translated to English via Google Translate as instances of machine-generated text. The model achieved an accuracy of 89.0% in detecting machine-generated text. However, this approach has not been evaluated using the LLM.

Solaiman et al. [237] presented a baseline approach using logistic regression trained on TF-IDF unigram and bigram features to detect AI-generated text. Their method achieved an accuracy between 74% and 88% in detecting outputs from models ranging from 124 million to 1.5 billion parameters, respectively. It was evaluated using GPT-2 [48] with an accuracy ranging from 93% to 97%. However, researchers found that shorter outputs of text were more challenging to detect than longer ones, and they anticipated that advanced generation strategies, such as nucleus sampling, might pose further difficulties for detection. Gallé et al. [238] proposed an unsupervised and distributional method to identify machine-generated text within documents, discovering a subtle signal in higher-order n-grams, which tends to surface more in machine-generated text than in human-written content. This signal underpins a self-training setup in which documents with pseudo-labels are used to train an ensemble of classifiers. The method was shown to be effective in accurately ranking suspect documents in their experiments, with precision rates exceeding 90% for the largest GPT-2 model examined. However, this study has its constraints, such as the assumption of a balanced training data set and the strictness of exact repetitions. Fröhling and Zubiaga [235] tested different classifiers, including logistic regression, SVM, random forest, and neural networks. They found their feature-based detection method highly effective. Their model showed a strong ability to distinguish between human-written text and text generated by advanced language models such as GPT-2, GPT-3, and Grover. This approach successfully capitalised on the unique linguistic and stylometric features of machine-generated text, underlining the potential of such feature-based strategies for the detection of machine-generated content. However, they also found that the sampling methods affect the transferability of the detectors. Experiments suggest that compensating for these differences might require separate sub-classifiers for each dataset and an ensemble approach to merge their outputs.

3) METHODS BASED ON ZERO SHOT MODELS

One way to detect whether a text is generated by a machine is to use the generative models themselves, such as GPT-2 or Grover, without any additional fine-tuning. These models can recognise the output of similar generative models [239]. Autoregressive models, such as GPT-2, GPT-3, and Grover, predict the next word in a sequence based on prior words, creating a unique statistical pattern in their generated text.

These patterns, which may include specific phrases, repetitions, or syntax, can be used to differentiate machine-generated text from human-written text. Thus, even without specific fine-tuning, detection algorithms can take advantage of these patterns to identify the content that is likely to be generated by these models [48], [237], [240]. In their 2019 study, Solaiman et al. [237] proposed a baseline zero-shot approach that leveraged the total log probability produced by a transformer-based language generative model to discern machine-generated text from human-authored content. Their system operates by determining a threshold value: if a text's total log probability, as determined by the GPT-2 model, is nearer to the average probability of machine-generated texts than human-authored texts, it classifies the text as machine-generated. However, their previously mentioned simple classifier approach proved to be more accurate [237]. Zellers et al. [240] examined the challenges posed by artificially generated text, particularly in the context of fake news. They introduced Grover, a language model constructed similarly to GPT-2 but specifically designed for both generating and detecting fabricated news articles. Grover's approach is based on the idea that knowing how a text is generated is the key to detecting it. To achieve this, Grover leveraged its ability to produce synthetic text, which helps it spot patterns that are common in machine-generated content. This allows Grover to identify materials that are likely to be produced using similar models. Grover was specifically trained to identify fake news and distinguish it from other zero-shot models. It uses statistical patterns learnt from its training data to detect anomalies that may indicate machine-generated text. According to researchers, Grover has consistently performed well across different model sizes and against various models, proving its ability to identify artificially generated content. However, its effectiveness can be weakened by adversarial actions or generation strategies that differ significantly from those it has been trained.

Gehrmann et al. [232] developed GLTR, a statistical tool designed for the detection and visualisation of machine-generated texts. This tool exploits the unique next-word prediction distributions found in machine-generated text, which are different from those found in human-written content. Using the capabilities of language models such as BERT [47] and GPT-2 [48], the GLTR assesses the likelihood of each subsequent word in a text based on the predictions of these models and visually presents this information. This technique helps to detect and understand the specific characteristics of the pattern of machine-generated text. Despite GLTR's utility, it has some limitations. Its effectiveness can be compromised when applied to texts generated by different or more advanced models. It may struggle with very short or overly complex texts, and its performance may vary according to the sampling methods used by the generative models [229]. Furthermore, GLTR may be vulnerable to adversarial attacks designed to mislead the detection process.

A study conducted by Mitchell et al. [241] in 2023 introduced DetectGPT as an innovative technique to distinguish between machine-generated text and human-written. This approach is based on the concept of 'probability curvature', which helps to identify the differences between the two types of text. DetectGPT has successfully detected outputs from various generative models, including GPT versions and transformer-based models, by examining the shape of the predictive probability distribution for the next word in a sequence. However, this method has two limitations. First, DetectGPT's computational load may pose challenges, particularly with larger datasets or real-time detection scenarios. Second, the effectiveness of the method relies heavily on the availability of raw log probabilities from the large language models it is trying to detect, which may limit its practicality when such probabilities are inaccessible or when dealing with new or proprietary models.

4) METHODS BASED ON FINE-TUNED LANGUAGE MODELS

Fine-tuning refers to the process of providing additional training to a pretrained LLM for a particular task or dataset. In the context of detecting AI-generated text, this could involve training a language model such as GPT [48], BERT [47], or Roberta [242] on a data set containing human and AI-generated text to enhance its ability to differentiate between them [237]. Fine-tuning enables LLM to adapt its overall language comprehension skills to the specific task of text detection.

A study by Solaiman et al. [237] in 2019 investigated detection methods using fine-tuning strategies. They created a sequence classifier based on two variations of the RoBERTa model, which is a base model with 125 million parameters and a model with 356 million parameters. Unlike their tests on GPT-2, Roberta is a masked and non-generative language model with a different architecture and tokeniser. The results of their experiments showed that fine-tuning Roberta consistently yielded better detection accuracy than fine-tuning a GPT-2 model with similar capacity. Researchers have noted that discriminative models, such as those used in their study, have more architectural flexibility than generative models. This flexibility makes them more effective in detecting machine-generated text, even though they are not well-suited for text generation. Interestingly, their findings partially contradict those of the GROVER [240] study, which suggested that the same generative language model used to produce text is the best tool for its detection.

Bakhtin et al. [243] used a variety of techniques and energy-based models [244] coupled with a distinct classifier designed to detect machine-generated texts. They experimented with various strategies, such as a basic linear classifier, Bidirectional Long-Short-Term Memory (BiLSTM), and a pair of transformer models, including one unidirectional GPT-2 [48] and the other bidirectional Roberta [242]. These transformer models were initialised using pre-existing checkpoints and then fine-tuned on data specifically gathered for machine-human text classification.

Consistent with the findings of other studies, they observed that the bi-directional transformer model, Roberta, provided the most reliable results. This suggests that pre-trained models that are further fine-tuned can be highly effective for identifying machine-generated text. Moreover, it underscores the significance of bi-directional understanding in classifying the nuanced discrepancies between human and machine-generated content. In the research conducted by Ippolito et al. [245], They explore three prevalent random decoding strategies: top-k, nucleus, and temperature sampling, applied specifically to detect text generated by GPT-2. They compiled an extensive collection of excerpts generated using each strategy. Subsequently, they trained a suite of binary classifiers based on the BERT model [47] to label these excerpts as human-written or machine-generated. Their study uncovered significant discrepancies in the accuracy of both human writers and trained classifiers, depending on the decoding strategy employed and the length of the generated sequences. This study underscores the influence of decoding strategies and text length on the effectiveness of machine-generated text detection.

A study conducted by Fagni et al. [246] in 2021 addressed the issue of identifying deepfake tweets, which they dubbed the "TweepFake" problem. Their approach focused on detecting artificially created tweets through advanced machine learning models that mimicked human-written content. To accomplish this task, they tested 13 different detectors that utilised various methods, such as machine learning with text representations, deep learning networks, and transformer-based classifiers. The study confirmed that generative methods, particularly those using transformer architecture, such as GPT-2, could produce high-quality short texts difficult for even expert human evaluators to identify as machine generated. The study also emphasised the value of transformer-based language models, as they provided helpful word representations for detection techniques based on text representations or fine-tuning. The fine-tuned RoBERTa-based detector outperformed those based on text representations, with nearly 90% accuracy. However, the authors suggested that there is still much room for additional research in this area, particularly with the advancement of LLM in the generation of indistinguishable short texts.

In a similar approach, Tesfagergish et al. [247] focused on detecting deepfake text within social media tweets using a combination of text augmentation, word embeddings, and deep learning techniques. They used GloVe [248] for word representation and fine-tuned the RoBERTa [242] model for the specific task of deep-fake recognition in tweets. Text augmentation was used to enrich their dataset and enhance the learning process. The proposed approach demonstrates the strength of combining pre-trained models with task-specific fine-tuning and dataset augmentation in deepfake text detection within a social media context, such as Twitter. The results underline the potential of machine learning and natural language processing techniques to tackle the growing issue of deepfakes in digital communication.

In their 2022 study, Kowalczyk et al. [249] proposed an innovative approach for detecting and understanding deepfake reviews. They used machine learning models, such as Random Forest and XGBoost, accompanied by an explainability technique called Shapley Additive Explanations (SHAP). This technique provides a deeper understanding of the model's decision-making process, revealing which features are most significant in classifying a review as fake or real. The integration of SHAP with detection models has several benefits. Not only does it enhance the interpretability of the decisions made by the model, identifying why a specific review is classified as fake, but it also adds a layer of transparency to the algorithmic decision-making process. This is especially important in real-world applications such as moderating social media content and product reviews. This study shows a promising research direction that combines accurate detection with interpretable results.

In their 2023 study, Guo et al. [234] deeply analysed the responses generated by ChatGPT, contrasting them with responses from human experts. They developed a comparison corpus, which encompasses responses to identical prompts from both the ChatGPT and human experts. Evaluating this corpus provided insight into how ChatGPT compares to human experts on content relevance, coherence, and empathy. To distinguish between machine-generated and human-authored texts, they employed a model based on fine-tuned RoBERTa. Their findings provide a valuable understanding of the capabilities and limitations of ChatGPT and suggest areas for improvement and more efficient detection strategies. The application of a fine-tuned Roberta model exemplifies the potential of transformer-based models to detect machine-generated texts.

However, this has some limitations. Despite comprehensive data collection efforts, the resulting data set lacks size, diversity, and balance between sources. To improve the accuracy of analysis and content detection, more diverse and extensive data from different sources and languages would be beneficial. Additionally, because all ChatGPT responses in the study were generated without specific prompts, their conclusions were based on ChatGPT's typical generative behaviour. They also noted that the use of unique prompts could lead to the generation of content that might not be recognised by their detection methods, indicating a need for further research in this area.

5) METHODS BASED ON WATERMARKING

Watermark-based identification has emerged as an intriguing paradigm in the field of text detection [250], [251]. Initially used in image copyright protection [252], it was applied in language with the advent of syntax tree manipulations, a concept introduced by Atallah et al. in 2001 [253]. Meral et al. [254] further explored this avenue in 2009, contributing to the development of watermarking techniques in language.

This approach was recently revolutionised by Kirchenbauer et al. [255] In 2023, they introduced a novel approach by developing a watermarking method tailored for Large Language Models (LLMs). This technique involves manipulating the LLM's logits at each step to embed watermarks, with tokens categorised into 'green' and 'red' lists. A watermarked LLM favours tokens from the 'green' list during text generation, forming distinguishable watermark patterns. The authenticity of these watermarks is verified using a specific hash function. This innovative approach has the potential to improve copyright protection and content authentication, enabling secure communication, and offering new research avenues in language privacy and digital rights management. However, a study by Sadasivan et al. in 2023 [256] underscored the complex challenge involved in reliably detecting AI-generated texts. The researchers designed a paraphraser based on a neural network to modify the outputs of the AI generative models. This paraphraser is intended to bypass a variety of detectors, such as watermarking systems and zero-shot classifiers. Furthermore, this research highlighted the susceptibility of watermarked Large Language Models (LLMs) to spoofing attacks. Additionally, in another 2023 study, Krishna et al. [257] explored the vulnerability of AI-generated text detectors to paraphrase. They found that although detection methods such as watermarking techniques, zero-shot, classifiers, and fine-tuned LLMs had reduced accuracy when tested against paraphrased texts, retrieval-based methods were an effective countermeasure. This underscores the need for the continuous development of detection strategies to keep up with evolving evasion techniques and the potential value of retrieval-based defence. These findings underscore the evolving challenges of detecting AI-generated texts and the need for continuous research and development in this field. Table 8 compares text-based deepfake detection methods.

VII. DISCUSSION

A. THE FAR-REACHING IMPLICATIONS OF DEEPFAKES

The adverse implications of deepfakes in the political realm are multifaceted and far-reaching, particularly concerning the spread of misinformation and manipulation of public opinion. There is an undeniable capacity for deepfakes to distort reality, an aspect that can be harnessed with sinister intent. Politicians can be falsely portrayed, leading to skewed public opinions and potential electoral consequences. The time-sensitivity of political campaigns amplifies this concern, as the timely release of deepfakes can circulate and cause irreversible damage before there is an opportunity to discredit them.

This discussion emphasises the potential role of social bots and deepfakes in tandem – a combination that can significantly amplify the spread of misinformation and create a facade of authenticity. This is an area where further research could be conducted, specifically investigating effective countermeasures to this dual threat. The potential use of deepfakes to incite violence or unrest during uprisings and social movements underscores the importance of this issue. The interplay of deepfakes in international relations has yet to be fully explored. Research could dive into the potential scenarios and repercussions of deep-fake-induced diplomatic incidents or conflicts. From a different perspective, deepfakes can seriously harm trust in online information. They can manipulate perceptions and create a sense of doubt within the media ecosystem. This can hurt relationships between people and increase the likelihood of cyberbullying and harassment. It is important to address these issues to maintain online safety and promote civility.

Deepfakes can introduce a complex matrix of legal concerns, including issues related to evidence credibility, defamation, privacy rights, and regulatory measures. As the legal domain adapts to these challenges, there is a need for more comprehensive research on the regulatory frameworks required to effectively manage the increasing prevalence of deep fakes. For example, the role of AI developers and social media platforms is a contentious issue that merits rigorous debate and legislative attention [131]. This discussion also highlights the extensive economic implications of deepfakes, highlighting the potential for significant financial losses, brand reputation damage, and market dynamics disruptions. Deepfakes can potentially undermine public confidence in critical economic institutions, leading to economic instability. Further research could investigate preventive measures and mitigation strategies to protect the financial sector from threats induced by deepfakes.

Deepfakes present substantial cybersecurity challenges in the technological domain. Their increase escalates the complexity of cyber threats, which requires adaptive and robust cybersecurity measures. With the advent of deepfakes, even traditional cybercrimes such as phishing can transform into sophisticated attacks, potentially resulting in serious security breaches and significant financial loss. This increased threat landscape underscores the urgent need for focused research to improve cybersecurity technologies capable of countering advanced deep-fake-induced attacks. Strengthening detection capabilities, devising innovative defensive mechanisms, and developing proactive measures against potential deepfake threats should be the primary objectives of future research. In this dynamic and fast-paced environment, continuous advancements in cybersecurity are critical to protecting digital integrity in the face of evolving fake threats.

B. VISUAL DEEPFAKE DETECTION

As we investigate the different methods and techniques used to detect visual deep-fakes, it becomes clear that this field is rapidly advancing with its high complexity and sophistication. Researchers are approaching this multidimensional challenge by utilising machine learning with various techniques, from handcrafted features to deep learning feature-based methods. This highlights the dynamic manner in which this issue is addressed.

Methods that rely on handcrafted features use visible elements of images, such as differences in facial markers, disparities in lip movements, and irregularities in texture and lighting. These techniques provide useful perspectives for identifying [258]. However, these methods have limitations and may not perform well with blurry images, rapidly moving objects, sophisticated deepfakes, or unseen data. The effectiveness of these methods can also be affected by the complexity of human behaviours, such as eye-blinking patterns or in the case of mental illnesses where head movement and facial expressions are not normal. Despite these challenges, the contributions of these methods are significant in their ability to identify manipulations in visual content and provide concrete indicators.

Instead of relying on handcrafted features, deep learning feature-based techniques use algorithms to identify subtle patterns in large data sets. This enables them to accurately distinguish between real and manipulated content, making them effective weapons against deepfakes. Although they offer impressive capabilities, such as the ability to recognise complex patterns and adapt to different situations, they also have some limitations. For example, they may struggle with low-quality video footage, require significant computational power, and may not perform well when faced with unfamiliar deepfakes. Additionally, factors such as obscured lips, nonfrontal faces, and adversarial attacks can impact their effectiveness. Furthermore, an interesting approach is to combine Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for effective

TABLE 8. Text-based deepfake detection methods.

Source	Method	Features	Dataset/Model	Performance	Limitation
Simple classifiers					
Nguyen-Son et al. [236]	SVM	Statistical	100 English and Finnish books, Google Translate text	Accuracy = 89%	Not evaluated on text generated by LLMs
Solaiman et al. [237]	logistic regression	Unigram and bigram	TF-IDF, GPT-2	Accuracy = 93-97%	Struggles in detecting shorter outputs of text
Gallé et al. [238]	unsupervised classifiers	N-grams	100 books, GPT-2	Precision = 90%	Requires a balanced dataset for training & the strictness of exact repetitions
Fröhling and Zubiaga [235]	LR, RF, SVM, NN	Linguistic and stylometric	Samples from the internet and GPT-2	Accuracy = 89.7%	Sampling methods impact detectors' transferability
Methods based on Zero-shot Models					
Solaiman et al. [237]	GPT-2	Log probability	GPT-2	Accuracy = 85%	Vulnerable to adversary attacks
Zellers et al. [240]	Grover	statistical	RealNews	Accuracy = 73-92%	Detect fake news articles, Vulnerable to adversary attacks or other text-generation techniques.
Gehrmann et al. [232]	GLTR	Statistical	GPT-2, BERT	Accuracy = 72%, AUC = 0.87	Vulnerable to adversary attacks or other text-generation techniques
Mitchell et al. [241]	GPT-2 - DetectGPT	Probability curvature	GPT	AUC = 0.95	Computationally intensive
Methods based on Language Models					
Solaiman et al. [237]	RoBERTa & sequence classifier	Sequences of texts	GPT-2	Accuracy = 95%	Larger language models' outputs are harder to detect
Bakhtin et al. [243]	Energy-based model & linear classifier	BiLSTM	GPT-2, RoBERTa	Accuracy = 97%	Scalability issues
Ippolito et al. [245]	BERT model &binary classifier	Top-k, nucleus, and temperature sampling	GPT-2	Accuracy = 88%	Larger language models' outputs are harder to detect
Fagni et al. [246]	RoBERTa & classifiers	Text representations	TweepFake GPT-2	Accuracy = 90%	Detects short tweets, not tested against advanced LLM texts
Tesfagergish et al. [247]	RoBERTa & GloVe [252]	Word embeddings, text augmentation	TweepFake	Accuracy = 89.7%	Detects short tweets, not tested against advanced LLM texts
Kowalczyk et al. [249]	GPT-2 & Random Forest - XGBoost	XAI - SHAP	Online reviews of Walmart	Accuracy = 88.6%	Case-specific and not tested against advanced LLM texts
Guo et al. [234]	RoBERTa / GLTR & Logistic regression	Statistical analysis	ChatGPT, HC3	F1 = 98.78%	dataset lacked in size, diversity, and balance across sources
Watermarking method					
Kirchenbauer et al. [255]	LLM's logits to embed watermarks	Tokens /hash function	C4, ChatGPT	AUC 0.998	Failed with paraphrasing attacks

deepfake detection. These approaches make good use of both spatial and temporal feature extraction, and can discern subtle manipulations. However, their high computational requirements and dependency on the quality of the training data and temporal continuity can pose challenges in real-time detection scenarios or when analysing videos with disjointed transitions.

Although current deepfake detection methods represent the latest technology, they still face difficulties, particularly when dealing with more advanced deepfakes, varying video qualities, and large-scale applications. Finding a balance between detection accuracy, computational efficiency, and generalisability to previously unseen deepfakes is vital. Looking ahead, the rapidly advancing field of deepfake detection emphasises the necessity for persistent innovation and development. As deepfake technology progresses, methods for detecting it must evolve correspondingly to preserve the integrity and security of digital media. This requires models that not only accurately identify deepfakes, but also elucidate their findings in an accessible manner, providing insights that people can understand. This is exemplified by the Dynamic Prototype Network (DPNet) [177], which combines detection accuracy with interpretability. Another essential aspect for future research on deepfake detection is addressing the time-sensitive nature of deepfake discretisation in media, a point emphasised in this study's impact section. Given the swift propagation and significant damage potential of deepfakes before their discretisation, future research must prioritise the development of detection methodologies that are both efficient and capable of real-time detection. Moreover, the complexity of balancing detection accuracy with computational efficiency continues to pose a challenge, suggesting that there is still considerable work ahead in this domain. Complex models often have high computational demands, potentially limiting their scalability and real-time application potential, thereby underscoring the need for solutions that balance model performance and computational efficiency.

C. AUDIO DEEPFAKE DETECTION

The complexity and increasingly sophisticated nature of audio deepfakes pose significant challenges to digital security, prompting the development of several detection methodologies. Both hand-crafted and feature-based deep learning methods have unique strengths in combating the pervasive issue of deep-learning audio.

Handcrafted feature-based methods, as shown in the work of Alzantot et al. [193], Lai et al. [198], and AlBadawy et al. [199], often use traditional machine learning algorithms and manual feature extraction techniques to identify manipulative cues. However, despite their promising performance, these techniques commonly encounter limitations regarding overfitting, computational complexity, and generalisation across diverse or unseen datasets. Novel approaches such as the 'Genuinization' process developed by Wu et al. [200], Bispectral analysis used by Aljasem et al. [204], and the 'Breathing, Talking, and Silence sounds' (BTS-E) framework presented by Doan et al. [259] demonstrate the potential of leveraging unique audio characteristics for deepfake detection. However, their robustness under varied real-world conditions and against increasingly sophisticated deepfake techniques requires further investigation.

On the other hand, deep learning feature-based methods, such as Attentive Filtering Networks (AFN) by Lai et al. [212], Deep-Feature Extractor for Automatic Speaker Verification (ASV) spoofing detection by Gomez-Alanis et al. [213], and raw-waveform processing strategy by Ma et al. [215], have demonstrated high detection rates for audio deepfakes. These methods effectively leverage intricate non-linear relationships and temporal dependencies in audio signals, therby maximising the capabilities of deep learning algorithms. However, challenges such as overfitting, dependability on the type of non-linear activation function, and performance degradation in the face of adversarial noise attacks can limit their practical applicability.

Recent developments, such as the one-class learning strategy by Zhang et al. [219] and the emotion recognition method by Conti et al. [220], offer innovative angles to deepfake detection. While they show promising potential, further examinations of their robustness across different emotional ranges, languages, and real-world scenarios are warranted. The growing attention paid to the problem of dataset diversity and similarity, as highlighted by Papastergiopoulos et al. [222], underscores the need for robust models capable of handling diverse and complex datasets.

The advances in synthetic speech detection presented in these studies collectively indicate the continued evolution and improvement of deepfake detection methods. However, they also emphasised the persistent challenges associated with overfitting, computational demands, generalisation to diverse datasets, and robustness under varied conditions. Future research must continue to innovate and enhance these detection methodologies to effectively address the increasingly sophisticated world of audio deepfakes, balancing model performance with computational efficiency, adaptability, and robustness across different data types and deepfake generation techniques. This ongoing battle against deepfake manipulation requires a multidimensional and dynamic approach that adapts continuously to the evolving deepfake landscape.

D. DETECTING SYNTHETIC TEXT AND THEIR OVERLAP WITH BOT DETECTION

In the context of the rapid advancement of Large Language Models (LLMs), the ability to produce highly accurate and contextually relevant text is becoming an increasing concern in online security. The sophisticated text generation capabilities of these models can potentially facilitate harmful activities on online platforms, emphasising the importance of detecting AI-generated texts as a mitigation strategy. Several detection methodologies have been developed, each with varying degrees of effectiveness influenced by factors such as the architecture of the language model, the decoding strategy, the length of the text and the availability of raw log probabilities from the LLMs.

Simple classifiers offer a basic, yet sometimes limited, approach, particularly with shorter texts and complex generation strategies. Zero-shot models provide a robust detection strategy by exploiting unique statistical patterns in machine-generated text; however, they are susceptible to adversarial actions or different generation strategies. Fine-tuned language models represent another approach that provides a more promising route for enhancing the detection accuracy. However, their effectiveness can be influenced by the sampling method and the length of the generated sequences.

Watermarking, a more recent development, has emerged as a unique method for detection. This involves embedding watermarks in the logits of an LLM during text generation, forming a pattern that can potentially identify synthetic text. Despite the promise of watermarking, its vulnerability to spoofing/paraphrasing attacks and evasion techniques presents significant challenges that warrant further research. Interestingly, despite being distinct research domains, bot detection and AI-generated text detection overlap. Both fields aim to distinguish between human- and machine-driven activities in their respective domains. Bot detection primarily identifies automated accounts or actions on online platforms, whereas AI-generated text detection seeks to discern between machine-generated and human-generated content.

The detection methodologies developed for bot detection, such as identifying unique behavioural patterns, could provide valuable insights and potential strategies for AI-generated text detection. Furthermore, techniques such as frequency analysis, sentiment analysis, and behavioural pattern detection from bot detection can be adapted for AI-based text detection. However, recognising that synthetic text and bot activities represent different aspects of online manipulation is also crucial. Although the detection strategies share similarities, the specifics of each domain require customised methodologies designed to address distinct challenges and characteristics.

VIII. CONCLUSION

In a digital era marked by a proliferation of deepfakes, this survey recognises the importance of including text-based deepfakes alongside visual and audio variants. Often overlooked, text deepfakes have the potential to significantly reshape online discourse and misinformation dynamics, making their inclusion in deepfake research both timely and crucial. This comprehensive survey sheds light on the multifaceted nature of deepfakes and the profound implications they pose across the political, social, economic, and technological spheres. Specifically, this study highlighted the Threat Landscape. The work has detailed how deepfakes can distort political discourse, erode trust in media, incite violence, destabilise economic markets, and increase cybersecurity risks. This underscores the urgent need for counteractive solutions. To achieve this, the study involved cataloguing detection methods by performing a comparative analysis of various detection methods by examining visual, audio, and text deepfake detection techniques. The strengths and weaknesses of each method have been outlined, revealing the complexity and challenges inherent in deepfake detection. The work has emphasised overlaps with bot detection, and the unique overlap between synthetic text detection and bot activities offers a critical perspective. By understanding these parallels, there is potential for insights and strategies across the two domains.

This work presents a comprehensive investigation into deepfakes from the perspective of understanding their potential impact and limitations. Although the work has extensively studied recent and closely related literature by following a systematic process, due to the speed and volume of new content, it is possible that key articles may have been missed. However, in an attempt to mitigate this, we followed a systematic review using inclusion and exclusion criteria to ensure that the most relevant and recent articles were studied. It is also worth highlighting that when it comes to understanding the potential impact, literature involving theoretical impacts based on case studies is presented. This means that someone working in defence has the potential to discover a use case and impact that was not previously considered. Considering the nature of cyber attacks and that the adversary often has the upper hand, the findings of this article are important and timely. It is why it is very important to educate all on the possibilities of deepfake techniques so that they are adequately prepared to question content and minimise any potential impact.

In terms of future work, several important areas need to be refocused. This included the development of a unified Detection Framework. A holistic framework that integrates visual, audio, and text detection methodologies is needed. Such a framework could be integrated into social media and online platforms to ensure that the content can be identified and correctly labelled. It has also been established that owing to the ease of generating deepfake content and its potential to spread quickly, there is a need for detection and mitigation solutions that can operate in real-time. This will require significant research effort to develop techniques that can operate at scale. This will require the investigation and development of real-time embedded electronics that are capable of processing large volumes of data during transmission. The adaptability and generalizability of the detection tools were identified as key areas. To effectively combat the evolving threats of deepfakes, detection methods must exhibit two key traits: adaptability to rapidly changing and sophisticated techniques and generalizability, ensuring consistent performance across diverse deepfake methods and scenarios. Finally, and certainly not the least important, public awareness and legislation should receive increased focus. Beyond technical solutions, raising public awareness and enacting appropriate legislation are crucial to effectively combating deepfakes. This work has laid a foundational understanding in the face of rapidly advancing deepfake technology. It is now imperative that the research community builds upon this foundation, striving for innovations that ensure that the digital realm remains trustworthy and secure.

REFERENCES

[1] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: Generation, detection, and applications," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 3, pp. 219–289, Sep. 2022.

[2] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4174–4184.

[3] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[4] L. Bode, D. Lees, and D. Golding, "The digital face and deepfakes on screen," *Converg., Int. J. Res. New Media Technol.*, vol. 27, no. 4, pp. 849–854, Aug. 2021.

[5] *Make Your Own Deepfakes*. Accessed: Oct. 17, 2023. [Online]. Available: https://deepfakesweb.com/

[6] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, ''DeepFaceLab: Integrated, flexible and extensible face-swapping framework,'' 2020, *arXiv:2005.05535*.

[7] *FaceApp: Face Editor*. Accessed: Oct. 17, 2023. [Online]. Available: https://www.faceapp.com/

[8] *ChatGPT*. Accessed: Oct. 17, 2023. [Online]. Available: https://openai.com/chatgpt

[9] *DALL·E 2*. Accessed: Oct. 17, 2023. [Online]. Available: https://openai.com/dall-e-2

[10] *DALL·E 2*. Accessed: Oct. 17, 2023. [Online]. Available: https://www.mid
journey.com

[11] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, ''DeepFakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,'' *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023.

[12] A. Boutadjine, F. Harrag, K. Shaalan, and S. Karboua, ''A comprehensive study on multimedia DeepFakes,'' in *Proc. Int. Conf. Adv. Electron., Control Commun. Syst. (ICAECCS)*, Mar. 2023, pp. 1–6.

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

[14] 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.

[15] Á. Figueira and L. Oliveira, ''The current state of fake news: Challenges and opportunities,'' *Proc. Comput. Sci.*, vol. 121, pp. 817–825, Jan. 2017.

[16] M. Caldwell, J. T. A. Andrews, T. Tanay, and L. D. Griffin, ''AI-enabled future crime,'' *Crime Sci.*, vol. 9, no. 1, pp. 1–13, Dec. 2020.

[17] T. Weikmann and S. Lecheler, ''Cutting through the hype: Understanding the implications of DeepFakes for the fact-checking actor-network,'' *Digit. Journalism*, vol. 2023, pp. 1–18, Mar. 2023.

[18] M. S. Raval, M. Roy, and M. Kuribayashi, ''Survey on vision based fake news detection and its impact analysis,'' in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 1837–1841.

[19] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, ''Incremental learning for the detection and classification of GAN-generated images,'' in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2019, pp. 1–6.

[20] C. Vaccari and A. Chadwick, ''DeepFakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news,'' *Social Media Soc.*, vol. 6, no. 1, Jan. 2020, Art. no. 205630512090340.

[21] A. de Ruiter, ''The distinct wrong of DeepFakes,'' *Philosophy Technol.*, vol. 34, no. 4, pp. 1311–1332, Dec. 2021.

[22] J. Langa, ''DeepFakes, real consequences: Crafting legislation to combat threats posed by deepfakes,'' *BUL Rev.*, vol. 101, p. 761, Jan. 2021.

[23] T. Hanitzsch, A. Van Dalen, and N. Steindl, ''Caught in the nexus: A comparative and longitudinal analysis of public trust in the press,'' *Int. J. Press/Politics*, vol. 23, no. 1, pp. 3–23, Jan. 2018.

[24] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, ''DeepFake detection: A systematic literature review,'' *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[25] T. Zhang, ''Deepfake generation and detection, a survey,'' *Multimedia Tools Appl.*, vol. 81, no. 5, pp. 6259–6276, Feb. 2022.

[26] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, ''Deep learning for deepfakes creation and detection: A survey,'' *Comput. Vis. Image Understand.*, vol. 223, Oct. 2022, Art. no. 103525.

[27] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, and P.-S. Group, ''PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews,'' *J. Med. Library Assoc.*, vol. 109, no. 2, pp. 1–19, Jul. 2021.

[28] P. Wang, R. Angarita, and I. Renna, ''Is this the era of misinformation yet: Combining social bots and fake news to deceive the masses,'' in *Proc. Companion Web Conf.*, 2018, pp. 1557–1561.

[29] 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.

[30] *FakeApp 2.2—Download for PC Free*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.malavida.com/en/soft/fakeapp/

[31] M. Zendran and A. Rusiecki, ''Swapping face images with generative neural networks for deepfake technology—Experimental study,'' *Proc. Comput. Sci.*, vol. 192, pp. 834–843, Jan. 2021.

[32] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, ''Protecting world leaders against deep fakes,'' in *Proc. CVPR Workshops*, vol. 1, 2019, p. 38.

[33] M. Koopman, A. M. Rodriguez, and Z. Geradts, ''Detection of deepfake video manipulation,'' in *Proc. 20th Irish Mach. Vis. Image Process. Conf. (IMVIP)*, 2018, pp. 133–136.

[34] 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.

[35] X. Yang, Y. Li, and S. Lyu, ''Exposing deep fakes using inconsistent head poses,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.

[36] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, ''Generative adversarial networks: An overview,'' *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[37] D. P. Kingma and M. Welling, ''Auto-encoding variational Bayes,'' 2013, *arXiv:1312.6114*.

[38] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, ''Deep unsupervised learning using nonequilibrium thermodynamics,'' in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[39] *GANs vs. VAEs: What is the Best Generative AI Approach? | TechTarget*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/feature/GANs-vs-VAEs-What-is-the-best-generative-AI-approach

[40] H. Khalid and S. S. Woo, ''OC-FakeDect: Classifying deepfakes using one-class variational autoencoder,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2794–2803.

[41] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, ''Combining efficientnet and vision transformers for video deepfake detection,'' in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2022, pp. 219–229.

[42] P. Dhariwal and A. Nichol, ''Diffusion models beat GANs on image synthesis,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8780–8794.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.

[44] P. Semaan, ''Natural language generation: An overview,'' *J. Comput. Sci. Res.*, vol. 1, no. 3, pp. 50–57, 2012.

[45] W. Yin, K. Kann, M. Yu, and H. Schütze, ''Comparative study of CNN and RNN for natural language processing,'' 2017, *arXiv:1702.01923*.

[46] X. Liu, P. He, W. Chen, and J. Gao, ''Multi-task deep neural networks for natural language understanding,'' 2019, *arXiv:1901.11504*.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, ''Language models are unsupervised multitask learners,'' *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[49] K. A. Pantserev, ''The malicious use of ai-based deepfake technology as the new threat to psychological security and political stability,'' in *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*. Cham, Switzerland: Springer, 2020, pp. 37–55.

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

[51] I. Korshunova, W. Shi, J. Dambre, and L. Theis, ''Fast face-swap using convolutional neural networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3697–3705.

[52] *GitHub—DeepFakes/Faceswap: Deepfakes Software for All*. Accessed: Oct. 18, 2023. [Online]. Available: https://github.com/deepfakes/faceswap

[53] R. Natsume, T. Yatagawa, and S. Morishima, ''RSGAN: Face swapping and editing using face and hair representation in latent spaces,'' 2018, *arXiv:1804.03447*.

[54] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "DeepFakes: Deceptions, mitigations, and opportunities," *J. Bus. Res.*, vol. 154, Jan. 2023, Art. no. 113368.

[55] 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)*, Jun. 2019, pp. 4396–4405.

[56] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, May 2023.

[57] L. Whittaker, K. Letheren, and R. Mulcahy, "The rise of deepfakes: A conceptual framework and research agenda for marketing," *Australas. Marketing J.*, vol. 29, no. 3, pp. 204–214, Aug. 2021.

[58] *ThisPersonDoesNotExist—Random AI Generated Photos of Fake Persons*. Accessed: Oct. 18, 2023. [Online]. Available: https://this-person-does-not-exist.com/en

[59] D. Patel, H. Zouaghi, S. Mudur, E. Paquette, S. Laforest, M. Rouillard, and T. Popa, "Visual dubbing pipeline with localized lip-sync and two-pass identity transfer," *Comput. Graph.*, vol. 110, pp. 19–27, Feb. 2023.

[60] J. Ice, "Defamatory political deepfakes and the first amendment," *Case W. Res. L. Rev.*, vol. 70, p. 417, Jan. 2019.

[61] 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, Aug. 2017.

[62] C. Hu, X. Xie, and L. Wu, "Face reenactment via generative landmark guidance," *Image Vis. Comput.*, vol. 130, Feb. 2023, Art. no. 104611.

[63] J. Thies, M. Zollhöfer, 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)*, Jun. 2016, pp. 2387–2395.

[64] Y. Nirkin, Y. Keller, and T. Hassner, "FSGANv2: Improved subject agnostic face swapping and reenactment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 560–575, Jan. 2023.

[65] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," in *Proc. AAAI*, vol. 34, 2020, pp. 10893–10900.

[66] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5932–5941.

[67] UCL. *'Deepfakes' Ranked as Most Serious AI Crime*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat

[68] 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.

[69] R. P. J. Wirth and P. Puchtler, "Neural speech synthesis in German," in *Proc. 14th Int. Conf. Adv. Hum.-Oriented Personalized Mech., Technol., Services*, 2021, pp. 26–34.

[70] J. Damiani. *A Voice Deepfake Was Used To Scam A CEO Out Of 243,000*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=469532a72241

[71] A. de Rancourt-Raymond and N. Smaili, "The unethical use of deepfakes," *J. Financial Crime*, vol. 30, no. 4, pp. 1066–1077, May 2023.

[72] R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar, "Towards automatic face-to-face translation," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1428–1436.

[73] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[74] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2017, *arXiv:1710.07654*.

[75] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*.

[76] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, F. Soong, T. Qin, S. Zhao, and T.-Y. Liu, "NaturalSpeech: End-to-end text to speech synthesis with human-level quality," 2022, *arXiv:2205.04421*.

[77] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 132–157, 2021.

[78] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Aug. 2020, pp. 716–731.

[79] *Advertising Standards Authority | Committee of Advertising Practice*. Accessed: Oct. 17, 2023. [Online]. Available: https://www.asa.org.uk/

[80] D. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1985, pp. 748–751.

[81] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[82] A. Godulla, C. P. Hoffmann, and D. Seibert, "Dealing with deepfakes—An interdisciplinary examination of the state of research and implications for communication studies," *Stud. Commun. Media*, vol. 10, no. 1, pp. 72–96, 2021.

[83] *Towards Personalised Synthesised Voices for Individuals With Vocal Disabilities: Voice Banking and Reconstruction*. Accessed: Oct. 18, 2023. [Online]. Available: https://aclanthology.org/W13-3917

[84] *AI Voice Generator With Text to Speech and Speech to Speech—Resemble AI*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.resemble.ai/

[85] *Voice Cloning: Realistic Text to speech Voice Cloning Online | Murf*. Accessed: Oct. 18, 2023. [Online]. Available: https://murf.ai/voice-cloning?lmref=oYsoew

[86] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting LLM-generated texts," 2023, *arXiv:2303.07205*.

[87] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Social Netw. Anal. Mining*, vol. 13, no. 1, p. 30, Feb. 2023.

[88] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.

[89] Á. Vizoso, M. Vaz-Álvarez, and X. López-García, "Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation," *Media Commun.*, vol. 9, no. 1, pp. 291–300, Mar. 2021.

[90] J. Mink, L. Luo, N. M. Barbosa, O. Figueira, Y. Wang, and G. Wang, "DeepPhish: Understanding user trust towards artificially generated profiles in online social networks," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 1669–1686.

[91] J. Tourille, B. Sow, and A. Popescu, "Automatic detection of bot-generated tweets," in *Proc. 1st Int. Workshop Multimedia AI against Disinformation*, Jun. 2022, pp. 44–51.

[92] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci.*, vol. 467, pp. 312–322, Oct. 2018.

[93] J. Lovato, L. Hébert-Dufresne, J. St-Onge, R. Harp, G. S. Lopez, S. P. Rogers, I. U. Haq, and J. Onaolapo, "Diverse misinformation: Impacts of human biases on detection of deepfakes on networks," 2022, *arXiv:2210.10026*.

[94] P. Regulation, "General data protection regulation," *Intouch*, vol. 25, pp. 1–5, Jan. 2018.

[95] A. Tong and A. Ulmer. *Deepfaking it: America's 2024 Election Collides With AI Boom*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/

[96] T. Zhang, L. Deng, L. Zhang, and X. Dang, "Deep learning in face synthesis: A survey on deepfakes," in *Proc. IEEE 3rd Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, Aug. 2020, pp. 67–70.

[97] M. Appel and F. Prietzel, "The detection of political deepfakes," *J. Comput.-Mediated Commun.*, vol. 27, no. 4, Jul. 2022, Art. no. zmac008.

[98] D. K. Citron and R. Chesney, "Deepfakes and the new disinformation war," *Foreign Affairs*, vol. 10, pp. 1–6, Jan. 2019.

[99] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. presidential election online discussion," *1st Monday*, vol. 21, p. 11, Nov. 2016.

[100] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature Commun.*, vol. 9, no. 1, pp. 1–9, Nov. 2018.

[101] M. Pavis, "Rebalancing our regulatory response to deepfakes with performers' rights," *Converg., Int. J. Res. New Media Technol.*, vol. 27, no. 4, pp. 974–998, Aug. 2021.

[102] T. Dobber, N. Metoui, D. Trilling, N. Helberger, and C. de Vreese, "Do (Microtargeted) deepfakes have real effects on political attitudes?" *Int. J. Press/Politics*, vol. 26, no. 1, pp. 69–91, Jan. 2021.

[103] K. Wahl-Jorgensen and M. Carlson, "Conjecturing fearful futures: Journalistic discourses on deepfakes," *Journalism Pract.*, vol. 15, no. 6, pp. 803–820, Jul. 2021.

[104] M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innov. Manage. Rev.*, vol. 9, no. 11, pp. 39–52, Jan. 2019.

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

[106] *Martin Lewis Felt 'Sick' Seeing Deepfake Scam ad on Facebook*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.bbc.com/news/uk-66130785

[107] T. Kirchengast, "Deepfakes and image manipulation: Criminalisation and control," *Inf. Commun. Technol. Law*, vol. 29, no. 3, pp. 308–323, Sep. 2020.

[108] R. Rini, "Deepfakes and the epistemic backstop," *Philosopher's Imprint*, vol. 20, no. 24, 2020.

[109] K. R. Harris, "Video on demand: What deepfakes do and how they harm," *Synthese*, vol. 199, nos. 5–6, pp. 13373–13391, Dec. 2021.

[110] A. Yadlin-Segal and Y. Oppenheim, "Whose dystopia is it anyway? Deepfakes and social media regulation," *Converg., Int. J. Res. New Media Technol.*, vol. 27, no. 1, pp. 36–51, Feb. 2021.

[111] M. Hameleers, T. G. van der Meer, and T. Dobber, "You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media," *Social Media Soc.*, vol. 8, no. 3, 2022, Art. no. 20563051221116346.

[112] J. Langguth, K. Pogorelov, S. Brenner, P. Filkuková, and D. T. Schroeder, "Don't trust your eyes: Image manipulation in the age of DeepFakes," *Frontiers Commun.*, vol. 6, May 2021, Art. no. 632317.

[113] M. Albahar and J. Almalki, "DeepFakes: Threats and countermeasures systematic review," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 22, pp. 3242–3250, 2019.

[114] H. Etienne, "The future of online trust (and why deepfake is advancing it)," *AI Ethics*, vol. 1, no. 4, pp. 553–562, Nov. 2021.

[115] R. Pfefferkorn, "'DeepFakes' in the courtroom," *BU Pub. Int. LJ*, vol. 29, p. 245, Jan. 2019.

[116] J. Botha and H. Pieterse, "Fake news and DeepFakes: A dangerous threat for 21st century information security," in *Proc. 15th Int. Conf. Cyber Warfare Secur.*, 2020, p. 57.

[117] J. Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Washington, DC, USA: Carnegie Endowment for International Peace, 2020.

[118] *How Can We Combat the Worrying Rise in Deepfake Content?* Accessed: Oct. 17, 2023. [Online]. Available: https. https://www.weforum.org/agenda/2023/05/how-can-we-combat-the-worrying-rise-in-deepfake-content/

[119] *Fears Grow of Deepfake ID Scams Following Progress Hack*. Accessed: Oct. 17, 2023. [Online]. Available: https://www.ft.com/content/167befa0-123f-4384-a37e-c8a5b78604b2

[120] A. K. Ghazi-Tehrani and H. N. Pontell, "Phishing evolves: Analyzing the enduring cybercrime," *Victims Offenders*, vol. 16, no. 3, pp. 316–342, Apr. 2021.

[121] *Risk Governance and The Rise of Deepfakes*. Accessed: Oct. 19, 2023. [Online]. Available: https://www.epfl.ch/research/domains/irgc/spotlight-on-risk-series/risk-governance-and-the-rise-of-deepfakes/

[122] 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.

[123] R. Metz. *Facebook and YouTube Say They Removed Zelensky Deepfake | CNN Business*. Accessed: Oct. 19, 2023. [Online]. Available: https://edition.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html

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

[125] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, May 2017.

[126] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[127] *Ethics Guidelines for Trustworthy AI—FUTURIUM—European Commission*. Accessed: Oct. 17, 2023. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[128] *Opinion | In India, Journalists Face Slut-Shaming and Rape Threats (Published 2018)*. Accessed: Oct. 18, 2023. [Online]. Available: https://www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shaming-rape.html

[129] *My Blonde GF—Tyke Films*. Accessed: Oct. 17, 2023. [Online]. Available: https://tykefilms.com/my-blonde-gf

[130] *DeepFake Porn Documentary Explores Its 'Life-Shattering' Impact*. Accessed: Oct. 19, 2023. [Online]. Available: https://www.bbc.com/news/entertainment-arts-65854112

[131] S. Greengard, "Will deepfakes do deep damage?" *Commun. ACM*, vol. 63, no. 1, pp. 17–19, Dec. 2019.

[132] *Sharing Deepfake Intimate Images to be Criminalised in England and Wales*. Accessed: Oct. 17, 2023. [Online]. Available: https://www.theguardian.com/society/2023/jun/27/sharing-deepfake-intimate-images-to-be-criminalised-in-england-and-wales

[133] J. Hong, "The state of phishing attacks," *Commun. ACM*, vol. 55, no. 1, pp. 74–81, Jan. 2012.

[134] R. Montasari, R. Hill, S. Parkinson, P. Peltola, A. Hosseinian-Far, and A. Daneshkhah, "Digital forensics: Challenges and opportunities for future studies," *Int. J. Organizational Collective Intell.*, vol. 10, no. 2, pp. 37–53, 2020.

[135] R. Montasari, H. Jahankhani, R. Hill, and S. Parkinson, *Digital Forensic Investigation of Internet of Things (IoT) Devices*. Cham, Switzerland: Springer, 2021.

[136] M. K. Johnson and H. Farid, "Exposing digital forgeries in complex lighting environments," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 450–461, Sep. 2007.

[137] E. Kee, J. F. O'Brien, and H. Farid, "Exposing photo manipulation with inconsistent shadows," *ACM Trans. Graph.*, vol. 32, no. 3, pp. 1–12, Jun. 2013.

[138] T. J. de Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha, "Exposing digital image forgeries by illumination color classification," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 7, pp. 1182–1194, Jul. 2013.

[139] A. Piva, "An overview on image forensics," *ISRN Signal Process.*, vol. 2013, pp. 1–22, Jan. 2013.

[140] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the unseen: Current trends and challenges in digital image and video forensics," *ACM Comput. Surv.*, vol. 43, no. 4, pp. 1–42, Oct. 2011.

[141] J. F. Chen, Z. J. Fu, W. M. Zhang, X. U. Cheng, and X. M. Sun, "Review of image steganalysis based on deep learning," *J. Softw.*, vol. 32, no. 2, pp. 551–578, 2021.

[142] P. Singh, "Robust homomorphic video hashing," in *Proc. IEEE 4th Int. Conf. Multimedia Inf. Process. Retr. (MIPR)*, Sep. 2021, pp. 11–18.

[143] Y. Zheng, Y. Cao, and C.-H. Chang, "A PUF-based data-device hash for tampered image detection and source camera identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 620–634, 2020.

[144] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, vol. 7, pp. 41596–41606, 2019.

[145] J. A. Costales, S. Shiromani, and M. Devaraj, "The impact of blockchain technology to protect image and video integrity from identity theft using deepfake analyzer," in *Proc. Int. Conf. Innov. Data Commun. Technol. Appl. (ICIDCA)*, Mar. 2023, pp. 730–733.

[146] D. Boneh, A. J. Grotto, P. McDaniel, and N. Papernot, "How relevant is the Turing test in the age of sophisbots?" *IEEE Secur. Privacy*, vol. 17, no. 6, pp. 64–71, Nov. 2019.

[147] P. Korus and N. Memon, "Content authentication for neural imaging pipelines: End-to-end optimization of photo provenance in complex distribution channels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8613–8621.

[148] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated face swapping and its detection," in *Proc. IEEE 2nd Int. Conf. Signal Image Process. (ICSIP)*, Aug. 2017, pp. 15–19.

[149] 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)*, Jan. 2019, pp. 83–92.

[150] D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, ''We need no pixels: Video manipulation detection using stream descriptors,'' 2019, *arXiv:1906.08743*.

[151] U. A. Ciftci, I. Demir, and L. Yin, ''FakeCatcher: Detection of synthetic portrait videos using biological signals,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/TPAMI.2020.3009287.

[152] T. Jung, S. Kim, and K. Kim, ''DeepVision: Deepfakes detection using human eye blinking pattern,'' *IEEE Access*, vol. 8, pp. 83144–83154, 2020.

[153] 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.

[154] T. Baltrušaitis, P. Robinson, and L.-P. Morency, ''OpenFace: An open source facial behavior analysis toolkit,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[155] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, ''Detecting deep-fake videos from phoneme-Viseme mismatches,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2814–2822.

[156] P. Korshunov and S. Marcel, ''Speaker inconsistency detection in tampered video,'' in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2375–2379.

[157] S. A. Shahzad, A. Hashmi, S. Khan, Y.-T. Peng, Y. Tsao, and H.-M. Wang, ''Lip sync matters: A novel multimodal forgery detector,'' in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 1885–1892.

[158] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, ''A lip sync expert is all you need for speech to lip generation in the wild,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 484–492.

[159] L. Guarnera, O. Giudice, and S. Battiato, ''DeepFake detection by analyzing convolutional traces,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2841–2850.

[160] S. McCloskey and M. Albright, ''Detecting GAN-generated imagery using saturation cues,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4584–4588.

[161] Z. Akhtar and D. Dasgupta, ''A comparative evaluation of local feature descriptors for DeepFakes detection,'' in *Proc. IEEE Int. Symp. Technol. Homeland Secur. (HST)*, Nov. 2019, pp. 1–5.

[162] 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.

[163] 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*.

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

[165] 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.

[166] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, ''Deepfake detection using spatiotemporal convolutional networks,'' 2020, *arXiv:2006.14749*.

[167] 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.

[168] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, ''Detecting deep-fake videos from appearance and behavior,'' in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2020, pp. 1–6.

[169] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, ''MTD-Net: Learning to detect deepfakes images by multi-scale texture difference,'' *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4234–4245, 2021.

[170] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, and Y. Li, ''MSTA-Net: Forgery detection by generating manipulation trace based on multi-scale self-texture attention,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4854–4866, Jul. 2022.

[171] 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*, Oct. 2020, pp. 2823–2832.

[172] 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.

[173] D. Güera 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.

[174] D. M. 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,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2851–2859.

[175] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, ''FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces,'' 2019, *arXiv:1909.06122*.

[176] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, ''Lips don't lie: A generalisable and robust approach to face forgery detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5037–5047.

[177] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, ''Interpretable and trustworthy deepfake detection via dynamic prototypes,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1972–1982.

[178] Z. Ma, X. Mei, and J. Shen, ''3D attention network for face forgery detection,'' in *Proc. 4th Inf. Commun. Technol. Conf. (ICTC)*, May 2023, pp. 396–401.

[179] 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.

[180] L. Nataraj, T. Manhar Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. S. Manjunath, ''Detecting GAN generated fake images using co-occurrence matrices,'' 2019, *arXiv:1903.06836*.

[181] N. Yu, L. Davis, and M. Fritz, ''Attributing fake images to GANs: Learning and analyzing GAN fingerprints,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7555–7565.

[182] 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)*, Jun. 2020, pp. 5780–5789.

[183] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, ''DeepFakesON-phys: DeepFakes detection based on heart rate estimation,'' 2020, *arXiv:2010.00400*.

[184] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, ''Face forgery detection by 3D decomposition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2928–2938.

[185] A. Rössler, 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.

[186] A. Jain, R. Singh, and M. Vatsa, ''On detecting GANs and retouching based synthetic alterations,'' in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–7.

[187] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[188] S. Khan, S. Parkinson, L. Grant, N. Liu, and S. Mcguire, ''Biometric systems utilising health data from wearable devices: Applications and future challenges in computer security,'' *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1–29, Jul. 2021.

[189] S. Parkinson, S. Khan, A. Crampton, Q. Xu, W. Xie, N. Liu, and K. Dakin, ''Password policy characteristics and keystroke biometric authentication,'' *IET Biometrics*, vol. 10, no. 2, pp. 163–178, Mar. 2021.

[190] S. Parkinson, S. Khan, A. Badea, A. Crampton, N. Liu, and Q. Xu, ''An empirical analysis of keystroke dynamics in passwords: A longitudinal study,'' *IET Biometrics*, vol. 12, no. 1, pp. 25–37, Jan. 2023.

[191] S. Parkinson, S. Khan, N. Liu, and Q. Xu, ''Repetition and template generalisability for instance-based keystroke biometric systems,'' in *Proc. IEEE 3rd Int. Conf. Comput. Commun. Artif. Intell. (CCAI)*, May 2023, pp. 272–277.

[192] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, ''DeepSonar: Towards effective and robust detection of AI-synthesized fake voices,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1207–1216.

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

[194] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[195] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio DeepFake detection," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Tokyo, 2020, pp. 132–137.

[196] P. R Aravind, U. Nechiyil, and N. Paramparambath, "Audio spoofing verification using deep convolutional neural networks by transfer learning," 2020, *arXiv:2008.03464*.

[197] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with Res2Net architecture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6354–6358.

[198] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," 2019, *arXiv:1904.01120*.

[199] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting AI-synthesized speech using bispectral analysis," in *Proc. CVPR Workshops*, 2019, pp. 104–109.

[200] Z. Wu, R. Kumar Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," 2020, *arXiv:2009.09637*.

[201] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Comput. Speech Lang.*, vol. 63, Mar. 2020, Art. no. 101096.

[202] R. K. Das, J. Yang, and H. Li, "Data augmentation with signal companding for detection of logical access attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6349–6353.

[203] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP J. Inf. Secur.*, vol. 2021, no. 1, pp. 1–14, Dec. 2021.

[204] M. Aljasem, A. Irtaza, H. Malik, N. Saba, A. Javed, K. M. Malik, and M. Meharmohammadi, "Secure automatic speaker verification (SASV) system through sm-ALTP features and asymmetric bagging," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3524–3537, 2021.

[205] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj, and R. Singh, "Generalized spoofing detection inspired from audio generation artifacts," 2021, *arXiv:2104.04111*.

[206] J. Yi, Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," 2021, *arXiv:2104.03617*.

[207] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "DeepFake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134018–134028, 2022.

[208] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, "DeepFake audio detection by speaker verification," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2022, pp. 1–6.

[209] L. Blue, K. Warren, H. Abdullah, C. Gibson, L. Vargas, J. O'Dell, K. Butler, and P. Traynor, "Who are you (I really Wanna know)? Detecting audio DeepFakes through vocal tract reconstruction," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 2691–2708.

[210] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting deepfake voice using explainable deep learning techniques," *Appl. Sci.*, vol. 12, no. 8, p. 3926, Apr. 2022.

[211] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "BTS-E: Audio DeepFake detection using breathing-talking-silence encoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[212] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6316–6320.

[213] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. Interspeech*, Sep. 2019, pp. 1068–1072.

[214] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 5859–5866.

[215] Y. Ma, Z. Ren, and S. Xu, "RW-ResNet: A novel speech anti-spoofing model using raw waveform," 2021, *arXiv:2108.05684*.

[216] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2021, pp. 13–22.

[217] J.-W. Jung, S.-B. Kim, H.-J. Shim, J.-H. Kim, and H.-J. Yu, "Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms," 2020, *arXiv:2004.00526*.

[218] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6369–6373.

[219] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.

[220] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro, "Deepfake speech detection through emotion recognition: A semantic approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8962–8966.

[221] D. Salvi, P. Bestagini, and S. Tubaro, "Synthetic speech detection through audio folding," in *Proc.2nd ACM Int. Workshop Multimedia AI Against Disinformation*, Jun. 2023, pp. 3–9.

[222] C. Papastergiopoulos, A. Vafeiadis, I. Papadimitriou, K. Votis, and D. Tzovaras, "On the generalizability of two-dimensional convolutional neural networks for fake speech detection," in *Proc. 1st Int. Workshop Multimedia AI Against Disinformation*, Jun. 2022, pp. 3–9.

[223] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 1985–1999, Dec. 2019.

[224] Y. Mo and S. Wang, "Multi-task learning improves synthetic speech detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6392–6396.

[225] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.

[226] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[227] S. Najari, M. Salehi, and R. Farahbakhsh, "GANBOT: A GAN-based framework for social bot detection," *Social Netw. Anal. Mining*, vol. 12, no. 1, pp. 1–11, Dec. 2022.

[228] M. Orabi, D. Mouheb, Z. A. Aghbari, and I. Kamel, "Detection of bots in social media: A systematic review," *Inf. Process. Manage.*, vol. 57, no. 4, Jul. 2020, Art. no. 102250.

[229] G. Jawahar, M. Abdul-Mageed, and L. V. S. Lakshmanan, "Automatic detection of machine generated text: A critical survey," 2020, *arXiv:2011.01314*.

[230] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bull. Rev.*, vol. 21, no. 5, pp. 1112–1130, Oct. 2014.

[231] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, J. C. Lai, and R. L. Mercer, "An estimate of an upper bound for the entropy of English," *Comput. Linguistics*, vol. 18, no. 1, pp. 31–40, 1983.

[232] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," 2019, *arXiv:1906.04043*.

[233] P. Bhatt and A. Rios, "Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions," 2021, *arXiv:2106.01170*.

[234] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," 2023, *arXiv:2301.07597*.

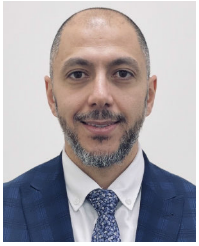
[235] L. Fröhling and A. Zubiaga, "Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover," *PeerJ Comput. Sci.*, vol. 7, p. e443, Apr. 2021.

[236] H.-Q. Nguyen-Son, N. T. Tieu, H. H. Nguyen, J. Yamagishi, and I. E. Zen, "Identifying computer-generated text using statistical analysis," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1504–1511.

[237] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. Wook Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release strategies and the social impacts of language models," 2019, *arXiv:1908.09203*.

[238] M. Gallé, J. Rozen, G. Kruszewski, and H. Elsahar, "Unsupervised and distributional detection of machine-generated text," 2021, *arXiv:2111.02878*.

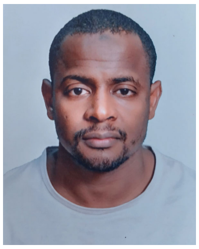
[239] E. N. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," *IEEE Access*, vol. 11, pp. 70977–71002, 2023.
[240] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–15.
[241] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," 2023, *arXiv:2301.11305*.
[242] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[243] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, and A. Szlam, "Real or fake? Learning to discriminate machine from human generated text," 2019, *arXiv:1906.03351*.
[244] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang, "Energy-based models," in *Predicting Structured Data*. Cambridge, MA, USA: MIT Press, 2007.
[245] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," 2019, *arXiv:1911.00650*.
[246] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "TweepFake: About detecting deepfake tweets," *PLoS ONE*, vol. 16, no. 5, May 2021, Art. no. e0251415.
[247] S. G. Tesfagergish and R. Damaševičius, "Deep fake recognition in tweets using text augmentation, word embeddings and deep learning," in *Proc. ICCSA*. Cham, Switzerland: Springer, 2021, pp. 523–538.
[248] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[249] P. Kowalczyk, M. Röder, A. Dürr, and F. Thiesse, "Detecting and understanding textual deepfakes in online reviews," Tech. Rep., 2022.
[250] S. Chakraborty, A. Singh Bedi, S. Zhu, B. An, D. Manocha, and F. Huang, "On the possibilities of AI-generated text detection," 2023, *arXiv:2304.04736*.
[251] S. Wadhera, D. Kamra, A. Rajpal, A. Jain, and V. Jain, "A comprehensive review on digital image watermarking," 2022, *arXiv:2207.06909*.
[252] G. C. Langelaar, I. Setyawan, and R. L. Lagendijk, "Watermarking digital image and video data. A state-of-the-art overview," *IEEE Signal Process. Mag.*, vol. 17, no. 5, pp. 20–46, Feb. 2000.
[253] M. J. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation," in *Information Hiding*. Cham, Switzerland: Springer, 2001, pp. 185–200.
[254] H. M. Meral, B. Sankur, A. Sumru Özsoy, T. Güngör, and E. Sevinç, "Natural language watermarking via morphosyntactic alterations," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 107–125, Jan. 2009.
[255] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," 2023, *arXiv:2301.10226*.
[256] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?" 2023, *arXiv:2303.11156*.
[257] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," 2023, *arXiv:2303.13408*.
[258] M. Roopak, S. Khan, S. Parkinson, and R. Armitage, "Comparison of deep learning classification models for facial image age estimation in digital forensic investigations," *Forensic Sci. Int., Digit. Invest.*, vol. 47, Dec. 2023, Art. no. 301637.
[259] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "BTS-E: Audio deepfake detection using breathing-talking-silence encoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5. [250] This is a preprint article and the requested information is not avaialble: https://arxiv.org/abs/2304.04736



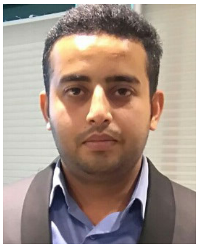
RAMI MUBARAK received the B.Sc. degree in information systems from the Arab Academy for Science, Technology and Maritime Transport, in 2006, and the M.Sc. degree in cyber security and digital forensics from the University of Huddersfield, U.K. He is currently the Head of the Department of Information Technology, Ministry of Interior, Bahrain. He has more than 16 years of experience in researching, evaluating, and deploying new technologies within the public sector.

TARIQ ALSBOUI received the B.Sc. degree in internet computing from Manchester Metropolitan University, U.K., in 2010, and the Ph.D. degree in computer science from the University of Huddersfield, U.K., in 2021. He is currently a Lecturer in computing with the School of Computing and Engineering, University of Huddersfield. He has authored several peer-reviewed international journals and conference papers. He is a fellow of the Higher Education Academy (FHEA). He is a Reviewer of high-impact-factor journals, such as IEEE Access and IEEE Internet of Things Journal.

OMAR ALSHAIKH received the M.Sc. degree in operations management to obtain the core knowledge in strategy, planning, and operations management. He is currently pursuing the Ph.D. degree with the University of Huddersfield. He is a Police Officer with the Ministry of Interior, Bahrain. He is also an artificial intelligence and leadership enthusiast; one of the roles in the career as a member of Bahrain law enforcement is to integrate cutting-edge technology to enhance the pedagogical methods in police sciences by carrying out simulations of virtual reality to meet the renewed security challenges. He is a fellow of the Higher Education Academy, U.K.



ISA INUWA-DUTSE received the Ph.D. degree from the University of St Andrews, with a focus on explainable artificial intelligence (XAI) project aimed at making ML systems more transparent and improving end-users engagements through an interactive argumentative framework. He is currently a Lecturer with the Department of Computer Science, University of Huddersfield, U.K. He is also a Visiting Lecturer with the University of Hertfordshire. His research interests include natural language processing (NLP) and machine learning (ML) aimed at developing useful tools with applications across various domains.



SAAD KHAN is currently a Senior Lecturer with the School of Computing and Engineering, University of Huddersfield, and a part of the Centre for Cyber Security. His research interests include developing and utilizing artificial intelligence and machine learning techniques for cyber security in various domains, such as SIEM systems, vulnerability, and anomaly detection, learning domain knowledge, mitigation planning, and access control.

SIMON PARKINSON is currently a Professor with the Department of Computer Science, University of Huddersfield, and leading the Centre for Cyber Security. His research interests include computer security and artificial intelligence. This includes undertaking research in biometric authentication systems and access control policy analysis. He has authored numerous articles on these topics as well as other cross-discipline applications of artificial intelligence, where he has a special interest in automated planning.