Artificial Intelligence Review (2024) 57:159
https://doi.org/10.1007/s10462-024-10810-6

Deepfake video detection: challenges and opportunities

Achhardeep Kaur[1] · Azadeh Noori Hoshyar[2] · Vidya Saikrishna[3] · Selena Firmin[1] · Feng Xia[4]

Accepted: 16 May 2024 / Published online: 29 May 2024

Abstract
Deepfake videos are a growing social issue. These videos are manipulated by artificial intelligence (AI) techniques (especially deep learning), an emerging societal issue. Malicious individuals misuse deepfake technologies to spread false information, such as fake images, videos, and audio. The development of convincing fake content threatens politics, security, and privacy. The majority of deepfake video detection methods are data-driven. This survey paper aims to thoroughly analyse deepfake video generation and detection. The paper's main contribution is the classification of the many challenges encountered while detecting deepfake videos. The paper discusses data challenges such as unbalanced datasets and inadequate labelled training data. Training challenges include the need for many computational resources. It also addresses reliability challenges, including overconfidence in detection methods and emerging manipulation approaches. The research emphasises the dominance of deep learning-based methods in detecting deepfakes despite their computational efficiency and generalisation limitations. However, it also acknowledges the drawbacks of these approaches, such as their limited computing efficiency and generalisation. The research also critically evaluates deepfake datasets, emphasising the necessity for good-quality datasets to improve detection methods. The study also indicates major research gaps, guiding future deepfake detection research. This entails developing robust models for real-time detection.

Keywords Deepfake detection · Fake video · Deep learning · Efficiency · Generalisation · Computational time

1 Introduction

Deepfake is a technical term for fake content on social platforms (Guo et al. 2020). This mainly includes fake images and videos. Fake images and videos are an old tradition. Since the advent of digital visual media, there has been a desire to manipulate them. Manipulation technologies have been widely used to forge images and videos for deception and entertainment. Using professional software like Adobe Photoshop to edit an image takes knowledge, time, and work. Instead of editing software like Adobe Shop, fake videos and images can be made by machines that don't require domain knowledge. In these new images and videos, an individual's face is transformed to mimic that of a target subject,

Extended author information available on the last page of the article

Table 1 Evolution of deepfake fraud over the last 5 years

Year	Key events and trends
2019	Emergence and proliferation
	– Introduction of deepfake technology exploitation for fraudulent activities
	– Limited awareness and detection capabilities
2020	Accessibility and widespread use
	– Increased accessibility of user-friendly deepfake creation tools
	– Improvement in detection approaches, but deepfakes become more sophisticated
2021	Technological advancements
	– Deepfake scam gains mainstream attention
	– Regional variations in the frequency of deepfake fraud
2022	Realism and sophistication
	– Significant rise in deepfake fraud incidents across various industries
	– Offenders adopt more advanced techniques, making detection challenging
	– First move towards interdisciplinary teamwork to fight deepfake fraud
2023	Present: technological duel between offenders and defenders
	– There is a continuous technological arms race between fraudsters and security measures
	– The never-ending arms race between cybercriminals and defence systems
	– Deepfake risks are now more widely recognised around the world
	– There is a tenfold surge in the worldwide detection of deepfakes across all sectors from 2022 to 2023

resulting in an amazingly realistic image or video of events that never occurred (Tolosana et al. 2020). For example, deepfake may modify a person's appearance while preserving their facial expression (Xu et al. 2022).

Deepfakes, made up of images, audio, and videos, seem to be the most common type of fake media. The very first "deepfake" video was released in 2017, in which a celebrity's face was replaced with that of a porn actor. Deepfakes received attention and began to become widespread when a Reddit user known as "Deepfake" demonstrated how a renowned person's face could be modified to give them a featured part in a pornographic video clip (Güera and Delp 2018).

Deepfake is among the top five identity fraud types in 2023. According to DeepMedia, a startup developing tools to identify fake media, the number of video deepfakes of all types has tripled, and the number of speech deepfakes has increased eightfold in 2023 compared to the same period in 2022. They have estimated that about 500,000 video and audio deepfakes will be uploaded on social media sites worldwide by the end of 2023 (Ulmer and Tong 2023). We have listed some key trends in the evolution of deepfake frauds over the last 5 years in Table 1.

Deepfake media can be of different types based on the content that has been manipulated. These manipulations include visual, audio, and textual modifications (Tolosana et al. 2020). Figure 1 shows types of deepfake content. Among visual, text-based, and audio, visual deepfakes are most common. They mainly include fake images and videos. As we know, today is the era of social media. These fake images and videos are used on social media platforms to spread false information about events that have never happened (Zhou and Zafarani 2020). "Face swapping", involves replacing the target's face with that of the original image, is a common method for creating deepfake images. On the other hand,

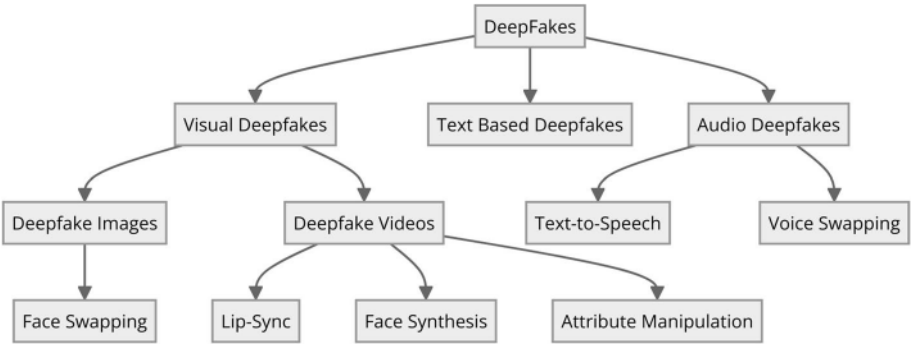

Fig. 1 Hierarchical classification of deepfake content available on social media platforms. The image also shows the methods used to create visual and audio deepfakes. Deepfake images and videos are frequently used on social media platforms

deepfake videos may be created using three techniques: lip-sync, face synthesis, and attribute manipulation (Nguyen et al. 2019b; Masood et al. 2023). The second type of deepfake is text-based deepfake. These textual deepfakes are mostly used on social media for fake comments and reviews on e-commerce websites. The third kind of deepfake is known as an audio deepfake. Such deepfakes involve using AI to create synthetic, realistic-sounding human speech. These deepfakes can be created using text-to-speech or voice-swapping methods.

Although deepfake technology is seen from a detrimental perspective, it can also be used in some productive projects. Deepfake can potentially improve multimedia, movies, educational media, digital communications, gaming and entertainment, social media, healthcare delivery, material science, and many commercial and content development industries. Furthermore, deepfake has the potential to be used in medical technology. We will consider some examples to understand the positive application of deepfake technologies.

Deepfake technology allows for automated and realistic voice dubbing for films and educational media in any language (Mahmud and Sharmin 2021). Companies that use digital video characters can make high-quality visual effects by re-synthesising music and video. Deepfake is also widely used in gaming to ensure realistic voice narration. Game characters' mouth motions are coordinated with the actors' voices. A deepfake video conferencing system may also be used to cross the language barrier. This technology may increase eye contact and make everyone appear to speak the same language during video conferences. Furthermore, technology may be used to digitally restore an amputee's leg or to help transgender people perceive themselves in a more favourable way as their desired gender. Deepfake technology can potentially assist people suffering from Alzheimer's disease in interacting with a younger face they may recall (Westerlund 2019). Scientists are currently looking into using Generative adversarial networks (GANs) to detect anomalies in X-rays and their potential to create virtual chemical molecules to speed up material research and medical discoveries. You may construct digital clones of yourself and have them go with you throughout e-stores, so you can also put on a bridal gown or suit digitally and then virtually experience a wedding site.

Although there are various advantages, there is also potential for misuse. The negative uses of deepfakes outnumber the favourable ones by a wide margin (Westerlund 2019). Deepfake has had a significant impact on today's social and virtual worlds. For example,

images and videos used as evidence in court proceedings or police investigations were widely regarded as legitimate. But deepfake technology makes it hard to believe in such evidence now. Deepfake poses risks like identity theft, computer fraud, blackmail, voice or image manipulation during authentication, and making fake evidence (Rao et al. 2021). Deepfakes are intended for use on social media platforms, where conspiracies, rumours, and misinformation spread quickly because users tend to follow what is trending (Masood et al. 2023). Recent advancements in AI-powered deepfakes have even amplified the issue (Liu et al. 2021b). Most GAN-generated faces do not even exist in the real world. Additionally, GAN may make realistic face changes in a video, such as identity swapping (Rao et al. 2021). This type of false information may be easily transmitted to millions of people on the internet via easy access to technology (Westerlund 2019).

With these advancements, the volume of fake content on the internet is increasing significantly. According to a survey by Deeptrace in 2020, there were 7964 deepfake videos online at the start of 2019. Nine months later, that number had risen to 14,678 (Toews 2020). They point out the possibility of using deepfake technology in political campaigns, which should be considered (Cellan-Jones 2019). Deeptrace again claimed in 2021, and they reported that the number of deepfakes on the web surged by 330%, reaching over 50,000 at their peak between October 2019 and June 2020 (Toews 2020). It has continued to expand since then. Video-sharing websites like YouTube and Facebook are the source of news for one in five internet users.

Deepfake technology has made it possible to make these videos look real; therefore, it is necessary to assess the videos' authenticity (Westerlund 2019; Karras et al. 2019). The difficulty of distinguishing between authentic and manufactured content has sparked widespread concern. As a result, research aimed at identifying fake media is critical for public safety and privacy. In addition to being a major threat to the privacy of personal information and national security, they could also be used in cyber warfare. This is likely to generate fear and distrust of digital content.

1.1 Previous surveys

Deepfake creation and detection is a new area of study in computer vision. Several survey papers on detecting deepfakes have been published in the past. 90% of these surveys focus on image or video deepfakes. The rest of the surveys explored deepfakes related to audio or a combination of audio, video, and other media formats (Stroebel et al. 2023). For example, Tolosana et al. (2020) covers facial image alteration methods, including deepfake and detection methods. However, this survey has only considered fake images.

Mirsky and Lee (2021) focuses on reenactment approaches for deepfake generation and provides model architecture charts for each deep neural network (DNN) used for deepfake generation methods. The survey lacks discussion on the technical challenges associated with generation and detection systems.

Verdoliva (2020) focuses on visual media integrity verification or the detection of manipulated images and videos. Deepfakes created by deep learning are featured alongside new data-driven forensic ways to combat them. They categorise detection methods into traditional approaches and deep learning-based methods. The analysis also shows the problems with current forensic methods and the challenges and opportunities ahead.

Nguyen et al. (2022) gave a complete overview of deepfake strategies and encouraged the development of more reliable approaches to fighting the challenges of deepfakes.

Another survey paper, Masood et al. (2023) reviews deepfake generation tools and machine learning (ML)-based techniques for detecting audio and video manipulations. The authors discuss available datasets and accuracy as the most important criteria for evaluating deepfake detection strategies.

Xu et al. (2022) evaluates research on deepfake generation, detection, and evasion of detection methods. They also illustrate the battlefield between the two sides, including the adversaries (DeepFake creations) and the defenders (DeepFake detection). This is an extensive survey with an analysis of more than 300 references; despite this, they have not addressed the issue of the computational complexity of detection methods.

Patil et al. (2023) has outlined the importance of biological classifiers in deepfake detection. They have discussed how these procedures can make it harder to identify facial features. Thus, these algorithms may misidentify deepfake videos as fakes.

Rana et al. (2022) examined deepfake detection methods by categorising them into four distinct groups: approaches based on deep learning, traditional machine learning methods, statistical techniques, and blockchain-based techniques.

Yu et al. (2021) have thoroughly reviewed the literature on detecting deepfake videos. They covered the generation of deepfakes, methods for detecting them, and benchmarks for evaluating the performance of detection models. The research indicates that current detection approaches are insufficient for real-world scenarios. The survey highlights the need for detection methods that are efficient, adaptable, and resistant to deepfake manipulation techniques. The study concluded that current detection methods are inappropriate for real-time use and should focus on time efficiency, generalisation and reliability.

Gambín et al. (2024) emphasises that collaboration among researchers, governments, and business organisations is essential to create and implement successful deepfake detection and prevention strategies. They discussed the potential of distributed ledgers and blockchain technology in improving cybersecurity measures against deepfakes.

A recent survey by Gong and Li (2024) has classified deepfake detection methods as conventional CNN-based detection, CNN with semi-supervised detection, transformer-based detection, and biological signal detection. The survey compares deepfake detection datasets and methodologies, highlighting their pros and cons. The authors discuss the challenges of obtaining accurate findings across datasets and suggest future directions to increase detection reliability.

Table 2 compares several surveys from the literature, including their strengths and weaknesses. This table summarises cutting-edge research in deepfake detection and sets a foundation for future advances in this crucial topic.

1.2 Motivation

Most existing surveys work on similar grounds, ensuring they can stand up to attacks, unclear how reliable existing detection technologies perform in terms of computational complexity and robustness. Only a few surveys examine the application of detection methods in real-world scenarios. While most of these surveys are concerned with detecting fake images, only a few discuss deepfake video detection. Detection results in terms of accuracy provided in most articles are over-confidence. These detectors do not perform similarly in real-time applications.

Following the discussion on research conducted on deepfake detection algorithms and datasets, there are some noticeable findings below:

Table 2 Comparative analysis of existing surveys on deepfake detection techniques, focusing on their core areas of research, key strengths, and identified limitations

Survey	Focus area	Strengths	Limitations
Tolosana et al. (2020)	Facial image alteration methods Deepfake detection methods	Detailed on facial alteration methods	Only considered fake images
Mirsky and Lee (2021)	Reenactment approaches for deepfake generation DNN architectures	In-depth look at generation methods	Lacks discussion on detection challenges
Verdoliva (2020)	Visual media integrity verification Traditional and deep learning methods	Categorisation of detection methods	Limited discussion on emerging threats
Nguyen et al. (2019b)	Overview of deepfake strategies Various deepfake strategies	Comprehensive overview	Less focus on real-world application
Masood et al. (2023)	ML-based audio and video manipulations	Reviews generation tools and detection methods	Narrow scope, overlooking future methods
Xu et al. (2022)	Deepfake generation, detection, and evasion Research evaluation on deepfakes	Extensive survey with 300+ references	Does not address computational complexity
Patil et al. (2023)	Biological classifiers in deepfake detection	Focus on novel detection methods	May misidentify videos as fakes
Rana et al. (2022)	Categorisation of deepfake detection methods Four groups of detection methods	Comprehensive categorisation	Assumes consistency in Deepfake detection methods that may not exist.
Yu et al. (2021)	Detecting deepfake videos Benchmarks for detection model evaluation	Thorough literature review	Inappropriate for real-time use
Gambín et al. (2024)	Collaboration in deepfake detection and prevention Cybersecurity measures against deepfakes	Discussion on cybersecurity measures	Limited historical perspective
Gong and Li (2024)	Classification of deepfake detection methods CNN, transformers, biological signals	In-depth comparison and classification	Challenges in obtaining accurate findings
Akhtar (2023)	Similar focus on deepfake detection techniques and challenges Various methods in deepfake detection	Likely offers unique insights into detection methodologies and challenges	Specific limitations and challenges in adaptation to real-world scenarios
Our survey	Deepfake video generation and detection Deep Learning, ML, Statistical methods	Focus on real-time application Addresses computational complexity Emphasis on the practical performance of existing detectors	–

- *Predominant focus on image-based detection* Researchers have conducted significantly more detection experiments on deepfake images than deepfake videos. Even if it is conducted on deepfake videos, they have mostly looked at spatial inconsistencies instead of temporal ones.
- *Insufficient real-world testing* A significant percentage of researchers have not tested their techniques in the real world. This includes testing against new and different deepfake technologies, looking at how efficient the technology is when used in real life, and ensuring it can stand up to attacks meant to get around detection systems.
- *Gap in dataset quality and relevance* Among the detectors that have attempted both image and video detection, most experimented on the existing high-fidelity image-based datasets rather than the latest video-based datasets. There is an urgent need for substantial effort towards developing effective deepfake video detectors and high-quality video datasets.

These findings reflect the genuine performance of existing detection methods for detecting deepfake videos, which are still unclear regarding reliability, generalisation, and computing complexity.

1.3 Survey contributions

This paper evaluates the reliability and data efficiency of state-of-the-art deepfake detectors in real-time, focusing on deepfake video detection. We aim to offer valuable insights to enhance the performance of current deepfake detection systems. No surveys have addressed some of the difficulties and potential future opportunities discussed in this paper.

This paper's particular contribution revolves around tackling the special issues associated with detecting deepfake videos, which sets it apart from current surveys that mostly concentrate on assessing the detection of deepfake still images. Furthermore, there is currently limited study on the computational time required for deepfake video detectors. While both fake images and video detection pose challenges, video deepfakes are particularly demanding in terms of computational resources. This increased requirement comes from the temporal dimension, the larger volume of data involved, and the complex nature of the models used to generate deepfakes, making video deepfake detection more challenging than identifying fake images. Our survey covers this potential aspect as one of the major challenges. The contributions of this survey paper are as follows:

- *Consolidation of existing knowledge* Our study consolidates existing deepfake detection research, comparing methods' effectiveness, efficiency, and scalability. It focuses on video dynamics, data requirements for model training, and deep learning applications.
- *Comprehensive taxonomy of detection challenges* This work goes beyond the scope of existing surveys by offering a taxonomy for deepfake detection challenges that categorises the broad spectrum of challenges in deepfake video detection. The taxonomy will guide future studies on developing more resilient detection algorithms.
- *Insight into deepfake datasets* We comprehensively analysed deepfake datasets and assessed them based on their diversity, realism, and availability. This analysis is crucial in the creation of more representative and challenging datasets.
- *Exploring new trends and future directions* New trends and strategies to increase deepfake detection reliability, computing complexity, and real performance have been explored in this survey.

- *Practical observations and applications in the real world* The survey connects academic research to practice by merging practical observations and theoretical results. It highlights the significance of detection methods' ability to be used in real-time to improve security, privacy, and media integrity.

1.4 Survey structure

The remainder of the paper is structured as follows: Sect. 2 describes the systematic review methodology. Section 3 focuses on the deepfake generation algorithms. Section 4 reviews the most used datasets in deepfake generation and detection methods. Section 5 explains how deepfake video detection differs from image detection. Section 6 provides a concise summary of deepfake detection methods. Section 7 presents a taxonomy of deepfake video detection challenges and existing solutions. Section 8 discusses various open issues in this research domain. In Sect. 9, future opportunities for improving deepfake video detection have been thoroughly discussed. Finally, Sect. 10 concludes the paper.

2 Systematic literature review methodology

The main objective of our systematic literature review (SLR) is to explore and analyse the existing research on deepfake video detection.

2.1 Survey scope

We are focusing on data-driven methods of deepfake video detection. We are working towards understanding the current challenges, proposed solutions in literature to these challenges, and potential for future research. To comprehensively address the challenges and opportunities in deepfake video detection, we have divided our main objective into the following sub-objectives to thoroughly address the challenges and potentials in deepfake video detection.

- Explore the evolution of deepfake generation techniques
- Investigate existing deepfake datasets
- Explore how deepfake video detection differs from image detection
- Identify and categorise various methods used to detect deepfakes and assess the state-of-the-art in the detection of video deepfakes
- Analysis of current challenges in deepfake video detection
- Discussion of open issues in deepfake detection
- Explore emerging trends and future opportunities

2.2 Paper collection strategy

Conducting your search is crucial for gathering literature on deepfake video creation and detection. We searched several electronic databases, such as IEEE Xplore, Google Scholar, ACM Digital Library, SpringerLink, PubMed, arXiv, CVPR, Scopus, ScienceDirect(ELSEVIER) and Web of Science. The keywords we used are as follows:

- "deepfake"/"fake content"/"video manipulation"/"deepfake generation"
- "deepfake detection"/ "fake video detection"/"deepfake detection challenges
- "deep learning for deepfake."

To capture the most research studies on this fast-developing topic, we also reviewed the reference lists of all papers to find more relevant literature.

2.3 Selection criteria

We used predefined inclusion criteria to select significant papers: (1) literature from only peer-reviewed journals (includes articles, editorials, and commentaries) that describe deepfake detection approaches, (2) literature that highlights challenges in deepfake detection, and (3) that explain future research potential, (4) we filtered out the research that focuses on data-driven detection techniques, (5) the studies directly answer one or more of the research questions of this study, (6) if research has been published in multiple journals or conferences, the latest version is included. Non-peer-reviewed journals and studies that are not written in English were excluded.

2.4 Data extraction and quality assessment

A standardised approach captures data from selected research about the study's aims, methodology, major findings, and deepfake detection contributions. Data is extracted from the 132 papers. The selected research is assessed using criteria based on established guidelines. We evaluated the selected papers' contributions, methodologies, results, implications, and future research directions.

2.5 Findings

We found a literature emphasis on using deep learning techniques, namely convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to detect deepfake videos. Furthermore, common concerns in detection are limited data availability, unbalanced datasets, and the requirement of models. The review also grew interested in investigating alternate methodologies, such as blockchain technology and statistical analysis, to improve detection capabilities.

3 Evolution of deepfake generation techniques

This section discusses the history and current state of deepfake-generating methods, stressing the importance of artificial intelligence (AI). This study covers the complex technological processes of deepfake creation. By studying these aspects, we prepare for a thorough comprehension of datasets needed to design efficient detection methods. This is crucial in the fight against deepfakes from theory to practice.

Computer graphics have utilised video modification techniques for many years. They frequently employ 3D reconstructions of the video's face geometry. There are two aspects to focus on in the deepfake generation research domain: generation methods and datasets. Deepfake is a technology that uses generative adversarial networks to produce

fake content. All research teams working on GAN aim to improve the quality of their applications in terms of image and video quality. It has been demonstrated that GAN-based synthesis approaches may create unexpectedly high-quality false videos (Karras et al. 2019). According to the study by Afchar et al. (2018), video and image manipulation are becoming more prevalent, primarily due to technological advancements in machine learning and deep learning.

3.1 Use of artificial intelligence

Deepfake leverages the power of artificial intelligence (AI) (Liu et al. 2018; Xia et al. 2019) to manipulate or generate visual and audio content with a high potential to deceive (Kietzmann et al. 2020). GAN framework was developed by Goodfellow et al. (2020) in 2014. Several researchers have investigated computer vision methods in areas linked with the creation of deepfakes, which use a variety of neural network models and architectures, primarily GAN. The name GAN, as shown in Fig. 2, signifies the combination of two networks (Rana et al. 2022; Nguyen et al. 2022; Xia et al. 2021). These networks are named generator (G) and discriminator (D). The generator uses an encoder and a decoder to generate fake videos with the intent of tricking D, and the discriminator develops the ability to distinguish between genuine and fake video samples using a training set. GAN has seen several changes and enhancements throughout the years since its introduction in 2014. It is becoming increasingly simple to use pre-trained GAN to instantly replace one person's face in an image or video with the face of another person (Liu et al. 2021b). Moreover, GUI-based applications such as FakeApp have made it easy for non-technical individuals to create these deepfakes. Anyone with sufficient desire, time, and computing power can use the technology now. Figure 3 shows how deepfake technology can create authentic-looking images or videos. We can compare the top and lower rows to differentiate between real and fake frames (Shahzad et al. 2022).

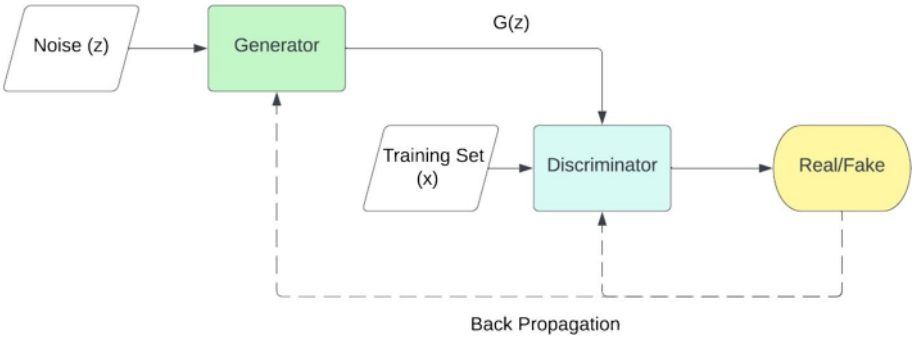

Fig. 2 Generative Neural Network (GAN). The diagram represents the GAN framework in which a Generator (G) generates data from noise (z), and a Discriminator (D) assesses it against a real dataset (x) for authenticity

Fig. 3 Comparison of original and deepfake video frames. The top row shows frames from original videos, while the bottom row showcases deepfake frames, proving the technology's sophistication to replicate actual video with high precision (Shahzad et al. 2022)

3.2 Classification of fake media generation techniques

Fake media generation techniques are classified as traditional methods and deep learning-based methods. Both of these methods require media editing tools to create convincing fake media.

3.2.1 Traditional fake media generation methods

Traditional approaches to developing fake images and videos rely on computer vision and image processing algorithms. Because these methods were developed before the advent of deep learning, the fake media they produce may not be as realistic as those produced by methods based on deep learning. These traditional fake media generation algorithms are classified into four types based on their target use: entire face synthesis, attribute modification, identity swapping, and expression swapping/face reenactment (Xu et al. 2022).

Entire face synthesis This involves creating complete digital representations of faces that do not exist. Image warping and morphing are two approaches that can be used to accomplish this goal (Zhao et al. 2016; Berthouzoz et al. 2011; Xu et al. 2022).

Attribute modification "Attribute modification" is modifying certain aspects of an image or video. Deepfakes use specific attribute modifications to create realistic fakes. Modifications may impact behaviour, appearance, or content (Berthouzoz et al. 2011).

Identity swapping A video or still image can be altered by using this method by putting one person's face onto another person's body. Korshunova et al. (2017) utilised CNN to create a face-swapping system. Moreover, Wang et al. (2018) developed a real-time face-swapping method.

Face reenactment/expression swapping To "reenact" a face means to imitate another person's expression. A person's facial expression in the target image or video is swapped out with that of a different person in the source image or video (Xu et al. 2022; Akhtar 2023). This technique is also known as puppet master (Tolosana et al. 2020).

Kim et al. (2018) has developed a method for reanimating portrait videos using only an input video. Unlike prior methods that manipulate facial expression, they are the first

to transfer head position, rotation, and eye blinking from a source to a target video. Figure 4 depicts frames from the video clip of Barack Obama, including a lip-sync deepfake, a comic impersonator, a face-swap deepfake, and a puppet master deepfake (Agarwal et al. 2019). Figure 4 originates from OpenFace2, a free software suite for analysing facial behaviour.

3.2.2 Deep learning based methods

Generation techniques based on deep learning have revolutionised the area of deepfakes. These techniques generate fake content using sophisticated neural networks and extensive datasets, making creating more realistic and believable content simpler.

Autoencoders These neural networks try to reconstruct the data they were fed in the first place. In the context of deep fakes, they can encode and decode facial characteristics, allowing for the swapping and manipulation of faces (Khalid and Woo 2020).

Variational Autoencoders (VAEs) This is an extension of autoencoders; they apply a framework of probabilities to the encoding method. VAEs combine the best features of autoencoders and neural networks (Child 2020).

Generative Adversarial Networks (GANs) GANs consist of two separate neural networks known as generators and discriminators (Goodfellow et al. 2020). The generator tries to create convincing fake content, while the discriminator attempts to identify the difference between real and fake content (Brock et al. 2018).

Transformers Transformers are well known for natural language processing (NLP). Recently, they have advanced in deepfake generation. The transformer model uses an encoder–decoder design that incorporates self-attention techniques. Deepfake images or videos can be created by fine-tuning a pre-trained transformer model on a specific dataset (Mubarak et al. 2023). Transformer models can be used to create human-like content

Fig. 4 Example frames from a video clip of US president Obama demonstrating deepfake techniques, including lip-sync, face swap, impersonation, and puppet master manipulations using a free software suite OpenFace2 (Agarwal et al. 2019)

with contextually relevant replies. OpenAI's Generative pretrained transformer (GPT) model is remarkable (Brown et al. 2020).

Diffusion models Diffusion models repeatedly modify an initial noise distribution to match the intended data distribution to produce realistic fake images with less blurriness and more distinguishing features (Ho et al. 2020). Diffusion models can produce more realistic images compared to GANs and VAEs (Aghasanli et al. 2023; Dhariwal and Nichol 2021). The DeepFakeFace (DFF) dataset is an open-sourced comprehensive collection of artificial celebrity images generated using diffusion models.

The development of these more complex deep learning models has led to a remarkable development in the sophistication of deepfakes. New techniques allow for creating more realistic and convincing false media, which has both beneficial and concerning applications. When technology is used ethically and responsibly, it lessens the likelihood of unintended consequences. Figure 5 compares Variational Autoencoders, Generative Adversarial Networks, and Diffusion Models across four metrics: data diversity, realism, stability, and ease of training. All three models have their strengths and weaknesses. GANs can generate samples that closely resemble the real data, exhibiting high fidelity. However, GANs are prone to mode collapse, where they fail to capture the full variety of the data. On the other hand, VAEs generate samples with lower fidelity, but they provide a wider range of diversity. Diffusion models are not straightforward to train. However, the generated samples may possess the same level of realism as those produced by GANs or VAEs. Table 3 shows a quick summary of deepfake generation classification.

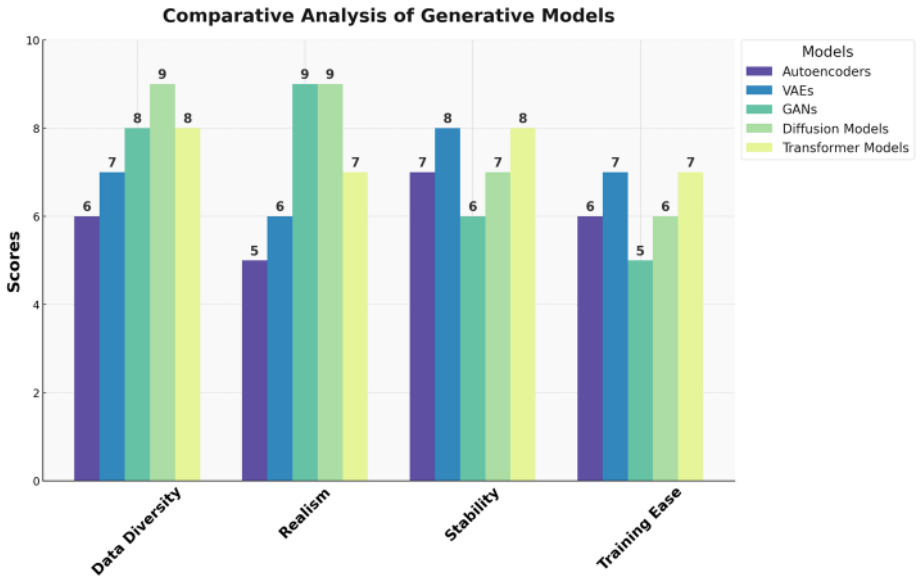

Fig. 5 The figure compares autoencoders, VAEs, GANs, transformers and diffusion models based on their data diversity, realism, stability, and ease of training while generating deepfake content. From 0 to 10, VAEs are balanced, GANs are realistic, and Diffusion Models are promising

Table 3 Summary of fake media generation techniques

Classes of generation techniques	Description	Types	Comments
Traditional methods	These methods rely on image processing and computer vision algorithms	Entire face synthesis Attribute Modification Identity Swapping Face Reenactment	Content generation is not as realistic as deep learning-based methods
Deep learning-based methods	DL-based methods use neural networks and extensive datasets	Autoencoders Variational Autoencoders (VAEs) Transformers Generative Adversarial Networks (GANs) Diffusion models	Create more realistic and convincing fake media content

3.3 Technological evolution of deepfake creation

In this section, we will examine the origins, progress, and problems of deepfake technology by following its historical trajectory. There have been significant advancements in machine learning models, methods for modifying audio and video, and other approaches along the technological progression of deep fake production (Khanjani et al. 2023). Deepfake technology has been an exciting and eventful rise from a small-scale pastime to an effective weapon with far-reaching implications for many sectors of the economy and beyond. There is growing concern about the possible misuse of deepfakes, images or videos created by artificial intelligence that replace a person's likeness with another. These deepfakes are becoming more lifelike and harder to detect (Kingra et al. 2023). In this section, we will examine the origins, progress, and problems of deepfake technology by following its historical trajectory.

Figure 6 shows the timeline of the evolution of fake media creation over the past few years. The beginning of the technology used to create deep fakes can be traced back to the early 2010 s. This period was characterised by more fundamental picture manipulation and the development of deep learning frameworks. The trend continued to advance with the introduction of face-swapping applications in 2017, which led to deepfake videos in 2018, in which GAN played an important role. In the following years, developments such as realistic lip-syncing, voice cloning, and the creation of full-body deepfakes were undertaken (Masood et al. 2023). Data dependencies were decreased through few-shot learning techniques, and the technology was found to have creative uses in the entertainment industry. Deepfakes, on the other hand, provide several substantial challenges, such as ethical concerns, difficulties in detection, and the requirement for legal frameworks to handle privacy threats and misuse. Examining the future reveals that ongoing research, industry cooperation, and regulatory activities will shape the future of deepfake technology. These efforts will emphasise ethical use, increased detection, and appropriate restrictions.

4 Existing deepfake datasets

Another critical component of deepfake detection is the dataset. Various datasets for deepfake-related study and experimentation have been made public. For DeepFake detection algorithms to be effective, they must be trained and tested. The lack of deepfake datasets

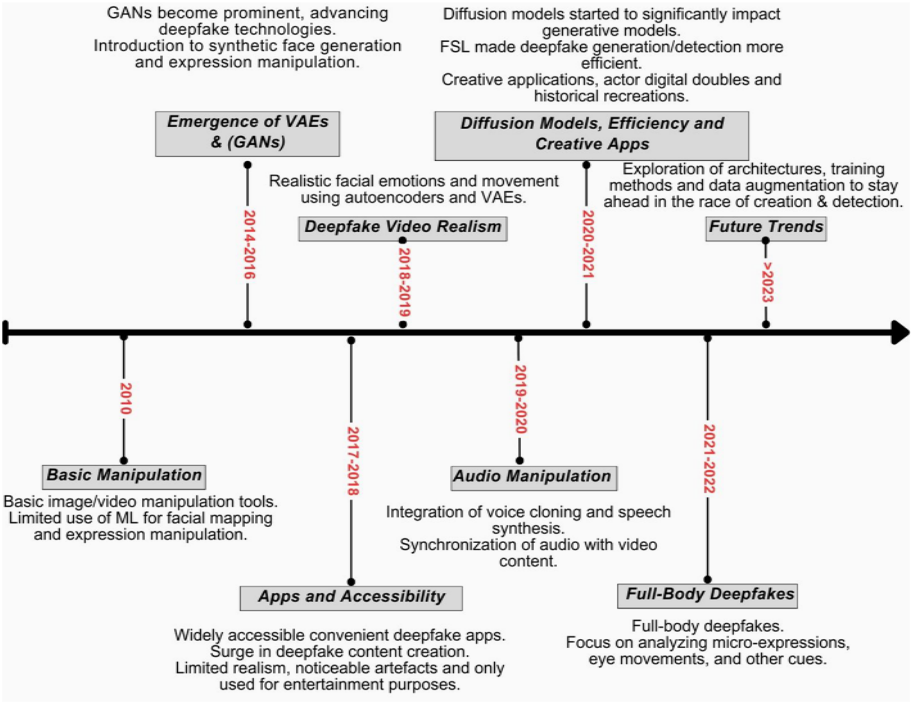

Fig. 6 The timeline of the evolution and advancements in deepfake technology: showing the progress from simple manipulation in the early 2010s to full body deepfake videos in the year 2022, along with predictions for future advancements (Dang et al. 2020; Masood et al. 2023)

or fragmented deepfake datasets is a significant obstacle to deepfake detection (Sohan et al. 2023). There is a growing need for large-scale deepfake video datasets for appropriate training. Table 4 describes the statistics of the most frequently used deepfake datasets for research in this field. We have listed datasets with their launch dates and the number of videos and images (both real and fake).

4.1 Classification of deepfake datasets

Li et al. (2018) have classified the datasets launched before 2019 into two categories: first-generation and second-generation. Now, most recent deepfake datasets are considered third-generation datasets (Dolhansky et al. 2020). They proposed the Deep Fake dataset (DFDC). DFDC dataset is the largest Deepfake dataset currently accessible and one of the only datasets to include video recorded to be used in machine learning applications (Dolhansky et al. 2020). First-generation datasets typically comprise fewer than 1000 videos. Also, these databases often do not claim they own the underlying content or have individual consent. Second-generation datasets have increased the number of videos to around 10 thousand videos. Moreover, the second-generation dataset contains high-quality videos as compared to first-generation datasets. Preceding datasets usually suffer from overfitting because of the number of videos. The new generation improves on the preceding one by increasing the number of frames.

Table 4 The table summarises deepfake datasets from 2018 to 2023, highlighting their introduction dates, content types, real vs fake numbers, sources of real and fake, and usage rates

Dataset name	Launch date	Gen	Images/videos	Real/fake	Source of real	Source of fake	Usage rate
UADFV	2018	1	Videos	49/49	YouTube	Artificial created	5%
Deepfake TIMIT	2018	1	Videos	320/640	Vid-TIMIT dataset	Artificial created	11%
FF++	2019	1	Videos	1000/4000	YouTube	Artificial created	51%
Google-DFD	2019	1	Videos	363/3068	Volunteer actors	Artificial created	23%
DFDC	2020	2	Videos	23,654/104,500	Volunteer actors	Artificial created	20%
Celeb-Df	2020	2	Videos	590/5639	YouTube	Artificial created	34%
Df-1.0	2020	2	Videos	50,000/10,000	YouTube	Artificial created	2%
DFFD	2020	2	Images and videos	58,703/240,336	YouTube	Internet	1%
Wild-Deepfake	2020	2	Videos	3805/3509	Internet	Internet	1%
DF-W	2021	3	Videos	1869/1869	YouTube	Artificial created	–
Open-Forensics	2021	3	Images	45,473/70,325	Google open images	Artificial created	–
DFFD	2023	3	Images	30,000/90,000	IMDB-WIKI dataset	Diffusion models	–

UADFV UADFV stands for the Uncompressed and Authentic Deepfake Video Dataset. The videos in this dataset are created using the Face2Face and NeuralTextures approaches and a unique combination of lighting and background. Moreover, the videos in UADFV are uncompressed, which makes them more suitable for research purposes. UADFV is used in research to develop deepfake detection methods and improve the robustness of existing methods (Yang et al. 2019).

Deepfake TIMIT The Deepfake TIMIT dataset supports deepfake detection and forgery localization research. The original videos in the dataset are of different individuals speaking different sentences. In contrast, deepfake videos were created using various deepfake techniques such as face swapping, reenactment, and face generation. The dataset contains videos with various visual artefacts and modifications, making it appropriate for testing deepfake detection algorithms (Korshunov and Marcel 2018).

Face Forensics++ (FF++) It is an extension of the original Face Forensics dataset. The dataset includes manipulated videos created using four different manipulation methods: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. The manipulated videos were created using different levels of manipulation strength, making it possible to evaluate the performance of deepfake detection methods under different scenarios. Face Forensics++ is frequently used in research for deepfake detection and facial forensics (Rossler et al. 2019).

Google-DFD Google-DFD stands for "Google Deepfake Detection" dataset. The dataset includes binary labels indicating whether each video is genuine or manipulated. The research community uses the Google-DFD dataset to create and test deepfake detection algorithms (Dufour and Gully 2019).

Celeb-Df It was created in November 2019 and is named after the CelebA dataset, a popular face recognition dataset. The dataset also includes a set of spatial and temporal annotations, providing ground-truth information on the manipulated regions and the frame-level manipulation. It is one of the widely used datasets in deepfake detection research (Li et al. 2020).

DeeperForensics-1.0 DF-1.0, also known as DeepFake 1.0, is a dataset of manipulated videos. The manipulation levels in the videos vary, from subtle manipulations to more severe ones, making it possible to evaluate the effectiveness of deepfake detection methods under different scenarios (Jiang et al. 2020).

DeepFake Detection (DFD) challenge It is a large-scale dataset of manipulated videos and images created for the DeepFake Detection Challenge (DFDC) hosted by Facebook in 2020 (Dang et al. 2020).

Wild-Deepfake Wild-Deepfake is a deepfake detection dataset created in 2020. Wild-Deepfake is widely used in the research community for developing deepfake detection algorithms (Zi et al. 2020).

DF-W DF-W is part of the Face Recognition Vendor Test (FRVT) 1:N Identification and Vendor Masking track, which aims to evaluate the performance of face recognition systems in the presence of deepfake manipulations (Pu et al. 2021).

OpenForensics It is an open-source dataset. OpenForensics is freely available to the research community and is intended to be used for developing and testing deepfake detection algorithms (Le et al. 2021).

DeepFakeFace (DFF) The DFF dataset consists of 120,000 images-30,000 real and 90,000 fake. The dataset uses genuine images from the IMDB-WIKI dataset to test detection methods in various looks and settings.

5 Understanding the specific challenges in detecting deepfake videos

This section aims to understand the reasons behind the significant rise in analytical and computational requirements triggered by fake video content. We will explore the consequences of video compression methods and the impact of temporal sequences and audio-visual synchronisation. This analysis will emphasise the significance of obtaining an in-depth knowledge of the theoretical and technical advancements required to detect deepfake videos successfully. These differences in video and image detection areas will serve as a foundation for a more focused analysis in this survey.

5.1 Deepfake image vs deepfake video detection

The process of determining whether or not an image has been manipulated to deceive or mislead viewers is known as deepfake image detection. The manipulation may involve modifying the content or context of the image, such as altering a person's appearance, adding or deleting objects, or changing the lighting or background. Another possibility is that the image may be flipped horizontally or vertically. Common methods for detecting deepfake images include analysing the image's metadata, searching for inconsistencies in the image's pixels or patterns, and comparing the image in question to real and fake image datasets.

The process of detecting deepfake videos, on the other hand, entails determining whether or not a video has been altered to trick or mislead viewers. The modification may involve changing the content or context of the video in some way by adding or deleting objects, changing the facial expressions or movements of the people in the video, or making adjustments to the audio or visual effects (Sabir et al. 2019). Due to the greater volume of data and the temporal nature of the video, detecting deepfake videos is typically more difficult than detecting deepfake images (Tolosana et al. 2020). Analysing the video metadata, searching for abnormalities in the video frames or optical flow, and applying machine learning algorithms to identify the video as real or false are all common techniques for detecting deepfake videos. The presence of temporal aspects in videos adds more complexity, requiring the creation of increasingly sophisticated detection methods that can precisely recognise deepfake content in a constantly evolving setting.

Detecting deepfake images and deepfake videos are two distinct yet interconnected problems, each presenting its unique set of obstacles and opportunities. Below are some of the main differences between deepfake image detection and video detection.

5.1.1 Temporal and continuity

- *Images* Fake image detection focuses solely on static properties and does not include any temporal processing. So, the detection algorithms aim to detect visual anomalies such as irregular texturing, lighting discrepancies, pixel-level characteristics, colour histograms, and digital artefacts that may suggest tampering. Techniques such as 2D CNNs could be used (Ji et al. 2012).
- *Videos* On the other hand, fake video detection includes temporal data and maintains frame coherence. Deepfake video detection approaches utilise the time dimension to identify flaws and artefacts that may not be readily apparent in a single frame. Video detection employs inter-frame comparison, motion analysis, and temporal coherence

(Patel et al. 2023). Video detection requires advanced methods like 3D CNNs or recurrent neural network architectures. Initially, 3D Convolutional Neural Networks (CNNs) were introduced for action recognition. Various video-based projects utilise the core concept of integrating learning frames within a given time frame (Ji et al. 2012). Liu et al. (2021a) proposes a lightweight 3D CNN with an outstanding ability to learn in integrating spatial information in the time dimension and employs a channel transformation (CT) module to minimise parameters while learning deeper extracted features. Their experiments demonstrate the proposed network outperforms previous DeepFake detection approaches.

5.1.2 Computational complexity

- *Images* Fake image detection often requires less computing power than video detection because it involves analysing static, single-frame input. Since real-time analysis is unnecessary, more complicated models can be used per frame (Tyagi and Yadav 2023).
- *Videos* In the video detection process, processing many frames, often in real-time, makes video analysis computationally costly (Bansal et al. 2023; Kumar et al. 2016). This requires more advanced computational resources and efficient algorithms to analyse the information across an entire video quickly (Anjum et al. 2016).

5.1.3 Real-time detection requirements on social platforms

- *Images* Image detection needs less rapid recognition than video feeds on social platforms. The flexibility in terms of urgency enables the use of more complex and time-consuming detection methodologies.
- *Videos* However, detecting deepfakes in video often requires fast detection. Videos need immediate analysis to ensure the dependability of each frame on the previous frame. So, video detection algorithms must be accurate and fast. These two restrictions need developing and refining detection methods to meet live content filtering requirements (Mezaris et al. 2019). Real-time detection requires developments that balance speed and accuracy, which are being sought (Mitra et al. 2021).

5.1.4 Diverse sources and manipulation techniques

- *Images* Image manipulations focus only on aspects like face swapping or object insertion.
- *Videos* Deepfake videos might involve sophisticated voice cloning with synchronised facial expressions (Tyagi and Yadav 2023; Mittal et al. 2023). So, the complexities and variation of these techniques can be more evident in videos because of the inclusion of movement, audio, and sequential editing.

Although deepfake video and image detection have the same objective, each process's approaches, challenges, and factors differ. The detection of deepfake videos has several challenges, including the necessity for real-time analysis and the additional complexity of temporal information that must be managed. Table 5 summarises the key differences in the fake image and video detection approaches.

5.2 Deepfake video detection process

Detecting deepfake images and videos shares certain methodologies, yet these tasks diverge significantly in complexity and necessitate different approaches. The deepfake video detection system involves all the steps from the image detection system. However, the video detection process has a few additional steps in input processing, like converting the video into frames before inputting it to the detection system. Other phases, like applying deep learning strategies, model training and testing, result determination, and accuracy calculation, are the same as image detection.

Figure 7 shows the steps involved in the deepfake video system. The details of these detection system steps are as follows:

- *Input* The process starts with a video as input to the system. This video could be a real or fake video created using deepfake techniques.
- *Pre-processing* Before the video undergoes analysis, pre-processing improves video quality and prepares data for analysis. This may require resizing, normalising, or other processes to prepare the input for deep learning algorithms.
- *Create a model* A detection model is then created using Deep Learning (DL). DL methods like CNNs and RNNs are popular for feature extraction and pattern identification.
- *Model training* The detection model is trained using datasets. The training entails exposing the model to labelled instances of real and deepfake movies, enabling the model to learn the distinguishing features between the two categories.
- *Model testing* Testing the model after training uses data not used during training. This test assesses the model's ability to apply learning to new examples.
- *Result determination* The final step is using the trained model's predictions to verify a video's authenticity. This phase involves deciding if videos are real or deepfake.
- *Accuracy calculation* The system compares model predictions against testing data ground truth labels to determine model accuracy. Accuracy measures the model's ability to classify actual and deepfake videos.

5.3 Feature extraction techniques used by deep learning models

In high-dimensional data analysis, visualisation, and modelling, dimensional reduction is widespread preprocessing. Feature selection is one of the simplest approaches to minimising dimensionality. This method involves selecting only those input dimensions that have the information necessary to solve the specific problem at hand. Feature extraction is a broader technique in which one attempts to build a transformation of the input space onto the low-dimensional subspace in such a way that the majority of the pertinent information is maintained. The detection of deepfakes can be accomplished by applying several feature extraction strategies. Each method has a distinct set of benefits and drawbacks; selecting the appropriate method is based on the particular demands of any specific task.

Face landmarks and texture information extraction from the video frames using methods such as Scale-Invariant Feature Transform (SIFT), Active Appearance Models (AAMs), and Local Binary Patterns (LBP) have been implemented in several different deep fake detection models (Li and Lyu 2018). 3D CNNs and RNNs analyse spatiotemporal patterns in input video frames. This approach has been used in several

Table 5 A summary of comparative analysis of detecting deepfake images and videos

Aspects	Deepfake image detection	Deepfake video detection	Findings
Data	Involves analysing a single image	Involves analysing a sequence of images	This indicates that the detection of deepfake videos involves the processing of more data than the detection of deepfake images does, which can make it more challenging
Compression	Typically, uncompressed or minimally compressed images	Must consider the compression used in the video	Compression changes the data and makes it more difficult to detect
Temporal information	It does not require temporal information since the image is static	Requires the analysis of temporal information, such as motion and movement between frames	It is more challenging to analyse temporal information
Audio information	Does not involve audio analysis	May involve analysing audio information, in addition to visual information	It is a more complex task, including visual and audio computations
Performance trade-offs	Less complex techniques could be used	More complex and computationally expensive methods than deepfake image detection are required due to videos' larger data and temporal nature	There are higher computational requirements for deepfake video detection than for deepfake image detection methods
Training data	Relies on image datasets with known labels	May require video datasets with both real and fake videos	There is less labelled data available for deepfake videos

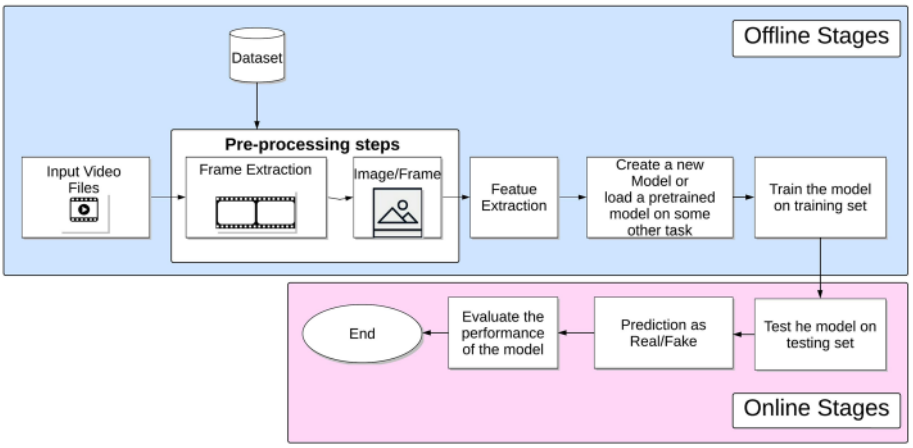

Fig. 7 Workflow of a fake video detection system. Offline stages include frame extraction, image processing, feature extraction, and model creation, followed by training. Online stages consist of model testing, specific video prediction, and accuracy calculation

deepfake detection models, such as the model proposed by Güera and Delp (2018). The Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT) are other approaches to extracting features. These methods analyse the frequency domain features of the input video frames using methods (Zhao et al. 2019). Some methods analyse the input video frames in search of signs of tampering. Using techniques such as copy-move forgery detection, JPEG compression analysis, and image splicing detection. Eye-tracking algorithms, known as gaze tracking, are used to analyse the direction in which the subjects of the input video frames are looking. Analysing the anomalies in the gaze direction throughout the frames is one method utilised in the quest to identify deep fakes (Ciftci et al. 2020). Another method involves analysing the quality of the video frames by extracting attributes such as sharpness, contrast, and noise level. It is a quality-based technique used (Nguyen and Derakhshani 2020).

There are several different deep learning approaches, each of which has demonstrated great performance in feature extraction for deepfake detection. Table 6 summarises the various methods used for feature extraction for deepfake detection. CNNs are frequently employed for various image and video analysis tasks, such as detecting deepfakes. It has been demonstrated that CNNs can successfully collect high-level features from the video frames fed into them. These features include facial expressions, poses, and motions. CNNs can be used in conjunction with other methods to improve performance. RNNs are utilised for tasks involving the analysis of sequential data, such as the analysis of video. RNNs make it easier to see the minute shifts and inconsistencies frequently found in deepfake videos. These deep learning algorithms have demonstrated great performance in feature extraction for deepfake detection, and they can be combined with other approaches to achieve higher levels of accuracy. However, It is essential to remember that the performance of these approaches might change based on the dataset and the particular deepfake detection task being performed.

Table 6 Feature extraction techniques used by deep learning models for deepfake detection

Feature extraction technique	Description
Face landmarks and texture information	Extraction of face landmarks and texture information from video frames using techniques like Scale-Invariant Feature Transform (SIFT), Active Appearance Models (AAMs), and Local Binary Patterns (LBP)
3D CNNs and RNNs	Analysis of spatiotemporal patterns in input video frames using 3D Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)
Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT)	Analysis of frequency domain features in input video frames
Tampering detection	Detection of signs of tampering in input video frames using techniques like copy-move forgery detection, JPEG compression analysis, and image splicing detection
Eye-tracking algorithms	Analysis of the gaze direction of subjects in the input video frames to identify anomalies indicative of deep fakes
Quality-based techniques	Analysis of video frame attributes such as sharpness, contrast, and noise level
CNNs	Convolutional Neural Networks capture high-level features from video frames, including facial expressions, poses, and motions
RNNs	Recurrent Neural Networks are used for analysing sequential data, making detecting shifts and inconsistencies in deepfake videos easier

6 Existing deepfake detection techniques

Researchers have developed wide-range deepfake detection methods using various factors (Lyu 2020). This section will cover some very important aspects of existing deepfake video detectors. By analysing these approaches, we understand the present level of advancement and equip ourselves to explore the classification of challenges, ready to examine the open issues and possibilities.

6.1 Classification of deepfake detection methods

Deepfake video detection methods include ML-based, DL-based, blockchain-based, statistical measurement-based, and frequency domain feature methods. We have summarised different types of detection methods in Table 7.

Most existing deepfake detection algorithms are based on DNN because of their capability in feature extraction and selection processes (Afchar et al. 2018; Li et al. 2018; Rossler et al. 2019).

6.1.1 ML-based methods

ML-based detection approaches incorporate conventional machine learning techniques. Identifying patterns, anomalies, or inconsistencies in media information is typically accomplished using statistical models and algorithms (Dolhansky et al. 2020). Based on statistical characteristics, ML-based algorithms have the potential to be useful in detecting small

Table 7 Classification of deepfake detection methods: an overview of their qualities, associated challenges, and illustrative example methods

Classes	Qualities	Challenges	Example methods
ML based methods	Explainability: tree-based ML techniques offer explainability	May struggle with complex patterns	Random Forest
	Model adjustments are straightforward	Performance depends on data quality	Decision Tree
	Understands human logic	Prone to overfitting	SVM
	High accuracy (up to 98%)	Some methods are limited by feature extraction.	
DL based methods	High accuracy over 99%	Tendency towards overfitting.	MesoNet
	Identifies specific artefacts	Significant computational resources required.	CNNs
	Utilise spatial and temporal features	Needs large, diverse datasets	GAN based detection
	Combines audio and visual for accuracy	Lacks explainability in decisions	
Blockchain-based methods	Decentralisation: provides decentralised verification of digital content in a trusted manner	Adoption and integration: widespread adoption and seamless integration with existing systems may be challenging	Blockchain for provenance tracking
	Provenance tracking: enables tracking the historical provenance of digital content	Computational overhead: blockchain processes introduce computational overhead	Blockchain for content authenticity
	Tamper-proof records: ensures tamper-proof records and logs in the public blockchain	Susceptibility to attacks: vulnerabilities to attacks in blockchain-based systems	
Statistical measurements methods	Objective measures: employ statistical measures such as normalised cross-correlation scores	Sensitivity to noise: statistical measures may be sensitive to noise and variations	PRNU analysis (Koopman et al. 2018)
	Non-image analysis: PRNU analysis focuses on photo response non-uniformity	Limited feature coverage: reliance on specific statistical features may limit detection in certain cases	Normalised cross-correlation
	Hypothesis testing: evaluate the statistical significance between original and GAN-created images	Statistical framework design: designing effective statistical frameworks for detection can be complex	
Frequency domain feature methods	Frequency analysis: explores deepfake detection by analysing frequency domain features	Interpretability: ensuring interpretability of features extracted from the frequency domain	Frequency analysis (Zhang et al. 2020)

Table 7 (continued)

Classes	Qualities	Challenges	Example methods
	Rich latent patterns: exploits the richness of latent patterns in the frequency domain	Noise sensitivity: addressing sensitivity to noise and variations in frequency-based features	ID-revelation (Qian et al. 2020)
	ID-revelation: learns temporal facial features based on a person's speech movement	Computational complexity: requires more computational resources	

anomalies in deepfake videos. In most cases, these methods extract relevant features from the videos. In addition to statistical characteristics, colour distributions, texture patterns, and other observable characteristics may also comprise part of the extraction. Examples of these models, such as Decision Trees, Support Vector Machines (SVM), and Random Forests, are frequently used. The models are trained using labelled datasets incorporating real and fake content characteristics.

Although ML-based methods can be successful, they may encounter challenges in managing the complexity of deepfake videos, particularly as generative models advance in sophistication. Machine learning algorithms may encounter difficulties in capturing the complex and nonlinear connections that exist within deepfake content (Rana et al. 2022). The efficacy of machine learning-based detection is highly contingent upon the calibre and variety of the training data. An ML model must be exposed to diverse, genuine, and manipulated content to acquire strong distinguishing characteristics. ML models frequently provide interpretability, enabling practitioners to comprehend the specific features contributing to the model's decision-making process. This level of transparency can facilitate the identification of the cues that the model depends on to differentiate between authentic and counterfeit content (Maksutov et al. 2020).

6.1.2 DL models for deepfake detection

In this section, we will cover the most successful existing deepfake video detection methods. Rossler et al. (2019) used a CNN-based method to find content that had been changed. They trained the neural network with a mix of datasets in a supervised way. This deep convolutional neural network, known as XceptionNet, has demonstrated high accuracy in detecting deepfake videos. It was submitted to the DeepFake Detection Challenge (DFDC), receiving a score of 0.9965 for its AUC-ROC. Afchar et al. (2018) proposed a deep-learning method called MesoNet for detecting fake content using the Deepfake and Face2Face techniques with two network architectures. MesoNet is a compact convolutional neural network developed to identify manipulated facial expressions. It can detect deepfakes and other facial modifications with high accuracy. Convolutional neural network (CNN) models are the most extensively used deepfake detection classifiers due to their outstanding performance (Xu et al. 2022). These DL-based detection approaches are entirely data-driven and employ the extraction of spatial characteristics to enhance detection efficacy. EfficientNet is a deep convolutional neural network that has demonstrated exceptional performance in the image classification tasks it has been given. It was applied in the DFDC, and the AUC-ROC score that it received was 0.9974 (Tan and Le 2019). ResNet is a deep convolutional neural network that has achieved high performance in image classification tasks. It has also been used in deepfake detection and achieved high accuracy (He et al. 2016). As per (Agarwal et al. 2020), ResNet is a deep convolutional neural network that has achieved high performance in image classification tasks. It has also been used in deepfake detection and achieved high accuracy. Another DL method is the transformer, which has significantly progressed in several vision classification tasks. Zhao et al. (2023) proposed a video transformer which analyses spatial and temporal information in fake videos and improves deepfake detection performance and generalisation. Video transformer is a video-based detection approach that processes numerous frames simultaneously and applies self-attention to distinct token dimensions. The transformer with spatial-temporal inconsistency detection demonstrates better generalisation in unseen data than earlier video-based detection

approaches. Coccomini et al. (2022) compare CNNs and Vision Transformers (ViTs) in the context of deepfake image detection. They used the ForgeryNet dataset to assess the efficiency of their cross-forgery performance. EfficientNetV2 performs better in training techniques, but ViTs are more proficient in generalisation, making them superior in detecting deepfakes. This difference demonstrates the adaptability of ViTs to the evolving field of deepfake detection.

The mean squared error MSE between the actual and predicted labels is used as the loss function for network training. Other earlier methods exploited the inconsistencies in deepfake videos. Nguyen et al. (2019a) employs a capsule network to detect spoofs from printed images or recorded videos to computer-generated videos using deep convolutional neural networks. Amerini et al. (2019) have used CNN with optical flow to differentiate between fake and real videos. Güera and Delp (2018) claims they can take advantage of the time differences by building a pipeline that starts with a CNN and ends with a recurrent neural network (RNN). Their approach extracts frame-level information using a CNN. These features are then used to train an RNN to detect video manipulation. Most published detection methods consider deep fake detection a binary classification problem (real vs. fake). Pu et al. (2021) proposed a transfer learning system to improve the performance detection system and used a Support Vector Machine (SVM) as a classifier for training.

Another type of deepfake detectors uses diffusion models to detect fake content. Song et al. (2023) explores the increasing concerns about deepfake images, specifically involving prominent individuals and their influence on the spread of genuine information. They also introduced the DeepFakeFace (DFF) dataset, which was created using sophisticated diffusion models to improve the training and testing of deepfake detection systems. Ivanovska and Struc (2024) discusses how denoising diffusion models (DDMs) can target deepfake detectors. It shows how even minor DDM adjustments may damage synthetic media detectors. Detection methods can be deceived by small modifications made by DDMs that humans cannot see, making detection systems vulnerable. The findings emphasise the need for more robust detection approaches to survive diffusion model changes.

6.1.3 Blockchain based methods

Integrating blockchain technology into deepfake detection methods adds security and traceability, utilising the unchangeable and transparent features intrinsic to blockchain systems. Blockchain-based detection methods utilise blockchain technology to improve authenticity and traceability (Narayan et al. 2022; George and George 2023). Blockchain guarantees data integrity, offering a secure and transparent record for verifying the origin of media (Chan et al. 2020). These strategies are especially valuable when verifying the genuineness and source of media content is vital (Rana et al. 2022).

Deepfake detection approaches that use blockchain technology can improve authentication. Media content can be timestamped and recorded on the blockchain for tamper-proof authenticity. A record added to the blockchain cannot be changed or erased because of its immutability. This functionality is useful for immutable media provenance records. An unforgettable audit trail of an image or video's production and alterations can be stored on the blockchain for deepfake detection. Blockchain's decentralised consensus process prevents a single entity from controlling the network. The decentralised nature of deepfake detection improves security. The possibility of malevolent actors changing or compromising blockchain provenance data is reduced.

6.1.4 Statistical measurement-based methods

Quantitative analysis is used in statistical measurement-based strategies to find media content anomalies. Pixel distribution and colour patterns are common statistical features assessed by these approaches.

Statistical measurements can quantify deviations from natural changes to find video discrepancies. Pixel distribution deviations may suggest content manipulation. Unnatural sharpness, artefacts, and pixel-intensity irregularities are anomalies. Colour patterns are often analysed using histograms or other statistical methods. Natural lighting and environmental circumstances affect authentic content's colour distribution and variance. These expected colour patterns are compared to the analysed content using statistical measurements. Deepfake generation may create non-natural textures. Another statistical method for dimensionality reduction and feature extraction is Principal Component Analysis (PCA). In deepfake detection, PCA may analyse statistical changes in pixel values and find anomaly-causing components. Statistical measurement-based approaches can handle some situations, but deepfake content is complicated. As generative models improve, deepfake statistical variations may more accurately approximate natural patterns (Ciftci et al. 2020).

6.1.5 Frequency domain feature methods

Frequency domain feature methods analyse media content frequency components for deepfake identification. Content frequency distribution features are often extracted using the Fourier transform or wavelet analysis (Malik et al. 2023).

A Fourier transform converts image or video pixel values from the spatial to the frequency domain. The frequency spectrum shows content frequency component intensity. Another approach is wavelet analysis, used alongside the Fourier transform to collect high- and low-frequency components with localised information (Kohli and Gupta 2021). Multi-resolution wavelet transformations can reveal frequency-domain anomalies at different scales. Frequency domain feature approaches are great at finding deepfake creation artefacts. Imbalances in the frequency distribution, especially high-frequency components, suggest tampering. Hybrid approaches combine frequency-domain information with spatial and temporal analysis. This integration uses complimentary feature extraction methods to strengthen deepfake detection models (Frank et al. 2020).

6.1.6 Note on deep learning (reason for dominance)

Every class has advantages and disadvantages; an integrative approach may provide a more robust solution. However, DL-based approaches are widely used because they are highly effective at extracting and selecting features, making them particularly adept at detecting fake media content (Li et al. 2018). Deepfake generation involves the use of advanced generative models to imitate genuine content. This poses difficulties for standard methods that may have trouble adjusting to the complex patterns present in synthetic media (Naitali et al. 2023). Deep learning architectures, such as Convolutional Neural Networks (CNNs) and GANs, can find complex and subtle features in deepfake content because they are deep and do not work in a straight line (Rossler et al. 2019).

Table 8 is the summary of the top existing deepfake detectors. DL-based deepfake detection poses some difficulties that have not yet been adequately resolved. The

Table 8 List of top deepfake detection methods with the used dataset, classifier, type of content in the dataset, and their accuracy rate

Description	Dataset	Classifier	Content type	Accuracy
MesoNet: Afchar et al. (2018) proposed two deep models Meso-4 and MesoInception-4 are used to analyse deepfake videos	DFDC	MesoNet (shallow CNN)	Videos	Often above 0.90
XceptionNet: It is a variant of the Inception architecture, which consists of depthwise separable convolutions Rossler et al. (2019)	FF++	XceptionNet (DNN architecture)	Videos	0.91
Agarwal et al. (2019) combines a convolutional neural network (CNN) and a long short-term memory (LSTM) network as a two-stage detection system	Custom from YouTube	LSTM and CNN	Videos	0.91
The model proposed by Amerini et al. (2019) aims to detect deepfake videos by analysing the optical flow patterns between frames using a convolutional neural network (CNN)	DFDC	CNN	Videos	0.913
Güera and Delp (2018) uses a temporal-aware pipeline to automatically detect deepfake videos	Custom Dataset	RNN	Videos	0.971
Li and Lyu (2018) proposed a method based on the fact that the current DeepFake algorithm can only make low-resolution pictures that must be further distorted to look like the faces in the original video. Such changes leave unique artefacts in the DeepFake videos that result, and we show that convolutional neural networks (CNNs) can be used to record them	UADFV, DeepFake-TIMIT	CNN	Videos	0.999
Another method is based on detecting eye blinking in the videos proposed by Li et al. (2018), a physiological signal not well presented in manipulated fake videos	UADFV	CNN	Videos	0.991
Matern et al. (2019) have used simple visual artefacts to expose manipulations like Deepfakes and Face2Face	FF	KNN, MLP, LR	Images and videos	0.866
Capsule Network is a method to detect replay assaults utilising printed images or recorded videos and deep convolutional neural network-generated videos (Nguyen et al. 2019a)	FF	CNN	Videos	0.993
Sabir et al. (2019) provides an optimal approach for combining variants in deep models with domain-specific face preprocessing techniques to achieve state-of-the-art performance	FF++	RNN	Videos	0.995

Table 8 (continued)

Description	Dataset	Classifier	Content type	Accuracy
Yang et al. (2019) used the observation that deepfakes are formed by splicing synthesised face regions into the original image, producing faults that can be detected when 3D head postures are computed from face photographs	MFC, UADFV	SVM	Images and videos	0.891

above-mentioned DNN-based detection algorithms are vulnerable to adversarial noise attacks, and none of the research has evaluated their performance against adversarial noise attacks. Furthermore, DL-based deepfake video detection has focused on improving model performance regarding accurate classification (such as precision and recall). Table 8 depicts that most detection approaches can achieve superior performance with an accuracy rate greater than 90%. However, the study has revealed that these methods do not consider other important performance parameters for a model, such as time and cost complexity.

7 Challenges to deepfake video detection: a taxonomy

Although GANs have improved the efficiency of deepfake technology, the generator algorithms remain vulnerable and could be exploited to detect deepfakes. Most of the current detection methods are supervised in nature (Zotov et al. 2020). Despite the theoretical promises of DL-based deepfake detectors, practically, they are constrained by many aspects, like a lack of data (specifically, deepfake video datasets), generalisation, vulnerability to adversarial attacks, and computational capacity. This section investigates deep learning-based fake detection challenges and analyses current research to address these challenges. Figure 8 depicts a taxonomy of challenges that data-driven techniques for finding deepfakes face. Table 9 describes the challenges of detecting deepfake videos and the present approaches that are attempting to overcome these challenges. The remaining part of this section will present these challenges in detail.

7.1 Data-related challenges

Deep learning-based detection approaches are entirely data-driven and employ the extraction of spatial characteristics to enhance detection efficacy. According to Dimensional Research, 96% of organisations face data quality and labelling issues in DL initiatives

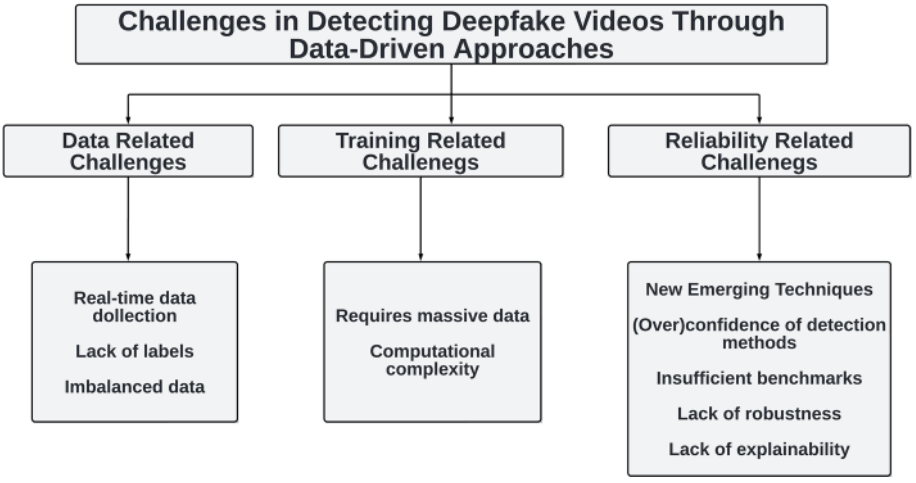

Fig. 8 Taxonomy of challenges in data-driven deepfake video detection. The image summarises three main kinds of challenges: data-related challenges, training-related challenges, and reliability-related challenges

Table 9 The challenges of detecting deepfake videos and the present approaches that are attempting to overcome these challenges

Challenges	Sub-categories	Methods addressing the challenges
Data-related challenges	Lack of labels	Semi-supervised learning
		Transfer learning
		Domain adaptation
	Imbalanced labels	Random oversampling and undersampling
Training-related challenges	Need of massive training data	Unpaired self-supervised training
		One-shot and few-shot learning
		Data-efficient models
	Computational complexity	Reduced computations with neural networks.
Reliability challenges	(Over)confidence of detection methods	Evaluation against unseen deepfakes and perturbation attacks
	New emerging manipulation techniques	Transfer learning
	Insufficient benchmarks	Development of publicly available benchmark datasets
	Lack of robustness	Robust detection against adversarial attacks. Generalisation approaches
	Lack of explainability	Improved explainability of detection results

(Silver et al. 2016). Consequently, if deepfake video detection is mostly based on DL techniques, they would face the same issues. It has been observed that DL techniques are popular but extremely data-hungry, and efficiency often gets reduced when the data set is small.

7.1.1 Lack of labels

As discussed in previous sections, for deep neural networks to achieve human-level performance, millions of labelled images are required for training (Qi and Luo 2020). However, Litjens et al. (2017) mentioned that a lack of labelled data is a common issue when applying machine learning to medical images, even when enormous volumes of unlabelled data are available (Zhang et al. 2020). Since getting sufficient labelled data in many scenarios can be challenging, researchers are increasingly interested in utilising unstructured data for training sophisticated learning models (Ren et al. 2022).

Impact on fake video detection This is one of the biggest challenges in deepfake video detection, as in numerous other DNN applications. It is due to the rapid development of generation techniques that surpass the creation of annotated datasets that accurately reflect the latest advancements.

Methods addressing lack of labels The quantity of labelled data available is restricted, and each detection technique should be able to cope with this constraint, regardless of its implementation. Due to insufficient labelled data, researchers may switch to semi-supervised learning instead. Semi-supervised/self-supervised learning allows models to learn meaningful representations without labelled input (Zhao et al. 2022). The methodologies that go beyond typical supervised learning are still being developed. These approaches include multiple-instance, reinforcement, semi-supervised, and transfer learning. However, given that labels are unknown during training, unsupervised learning presents a greater challenge than its supervised counterpart (Fung et al. 2021). To the best of our knowledge, research on using unsupervised learning for deepfake detection is extremely limited.

The scarcity of labelled data can be made up for by transferring knowledge from other labelled data sets (Lu et al. 2015). Transfer learning increases the model's performance on the target task by incorporating additional information from a different task. As per Adadi (2021), inspired by human beings' abilities, transfer learning aims to transfer information from one activity to another. It reduces the number of labelled samples required for a target job by acquiring knowledge from the source job. The degree of similarity between the tasks and the domains in which they exist determines how beneficial it is.

Domain adaptation is another word frequently heard in the transfer learning field. Cozzolino et al. (2018) and Tariq et al. (2021) have worked to apply transfer learning to deepfake detection tasks. As per the researchers, convolutional neural networks are the best deep learning strategies for deepfake video detection with a high accuracy rate. Transfer learning will become increasingly important in areas where annotated data is scarce. In areas with lots of annotated data, the concept of transfer learning can help improve learning performance (Liang et al. 2019; Zhou et al. 2018; Suratkar et al. 2020).

7.1.2 Imbalanced labels

Supervised learning approaches have certain drawbacks, such as the need for human labelling, data imbalance challenges, and expensive computations (Ren et al. 2021). Most publicly available datasets have a significant normal/abnormal data imbalance.

Impact on fake video detection The imbalance between real and fake videos in training datasets is more significant in deepfake video detection than in other areas. This leads to model biases due to the amount of real video content compared to the comparatively few examples of high-quality deepfakes. A model bias occurs when systems are better at recognising real videos than detecting deepfakes.

Methods addressing imbalanced labels To solve this issue, researchers are using sophisticated data augmentation techniques and investigating the development of synthetic data to increase the resilience of our models and achieve a balance between our datasets. Furthermore, deep learning models use random oversampling and undersampling to deal with unbalanced classes. The goal of oversampling is to improve the representation of disadvantaged minorities and get statistically significant results when comparing network attacks to background traffic. To make the minority and majority groups more comparable in size, undersampling eliminates samples from the larger group. Minority oversampling randomly duplicates a minority training example. During imbalanced learning, this might cause overfitting and prolonged training time (Sui et al. 2019).

7.2 Training-related challenges

Remembering that training data can impact how well data-driven models perform is important. As a result, most of these deep learning-based detection approaches are computationally intensive. The need for data for training detection models increases the computational time and number of resources required. Researchers are looking for more data-efficient models that exploit the capabilities of artificial learners without requiring a large amount of training data. There is limited work in this area. In this section, we'll examine a few research papers that address the above-mentioned challenge of deepfake detection (Mitra et al. 2021).

7.2.1 Need of massive training data

Deep learning techniques are popular but require a lot of data and frequently slow down when the data set is small. In many situations, gathering sufficient training data is costly, time-consuming or even impossible due to a lack of available resources.

Impact on fake video detection Similarly, detecting deepfake videos efficiently requires huge training and test datasets mostly based on deep learning approaches. However, in real-world scenarios like detection on social media platforms, we cannot afford to use a huge amount of data to make these deep-learning models work. So, many real-world applications want to use just a few data points because it costs less or takes less time. This has prompted discussion in academia and industry on creating models that fully use artificial learners' potential with less training data and less human supervision.

Methods addressing the requirement of massive training data One of the notable advancements includes unpaired self-supervised training techniques to reduce the amount of initial training data (Mirsky and Lee 2021). In 2019 and 2020, academics began exploring one-shot and few-shot learning to reduce training data. However, it is true that when a model is trained with a limited data set, the resulting model is over-specific to the training data and has trouble generalising (Adadi 2021). The neural network-based approach proposed by Mitra et al. (2021) can identify deepfake videos in social media regardless of the level of compression used. Only the most important and necessary frames are taken from

each video. This cuts down on the number of frames that must be checked for authenticity without lowering the quality of the results.

7.2.2 Computational complexity

The focus of machine learning and deep learning research in deepfake detection has been on improving model performance in terms of accurate classification (such as precision and recall) while not paying attention to other performance parameters that are important for a model, such as time and cost complexity. Social media platforms require fast and robust detection. Also, the results of methods to find deepfakes can be used in court as video evidence. The current detection methods are impractical due to their high computational cost.

Impact on fake video detection Deepfake video detection methods require significant computational resources due to the video's high resolution and temporal complexity. Detecting deepfakes necessitates prompt identification and mitigation of harmful information, making this computational necessity crucial.

Methods addressing computational complexity Research aims to use limited data to train the model to lower the computing complexity in the proposed research on deepfake detection. To determine whether a video is fake or real, the Mitra et al. (2021) proposed a method to reduce computations. This method brings deepfake detection closer to being deployed at the edge, as detection requires fewer computations. Afchar et al. (2018) presents a shallow architecture that can train and validate fake videos with significantly reduced computational complexity and fewer resources, but at the expense of accuracy, resulting in a total accuracy of just 0.66. Kawa and Syga (2020) presents a technique for detecting deepfakes that does not require high computational power. Specifically, they enhanced MesoNet by swapping out the default activation functions, resulting in an almost 1% improvement and increased decision consistency. Patel et al. (2020) describes transfer learning and its benefits when computational resources are constrained and a deep learning model does not need to be trained for days. By incorporating global texture features, GramNet, as proposed by Liu et al. (2020), increases the stability and generalisation of CNNs. Researchers must optimise network architectures and model pruning techniques to reduce computational burden while maintaining detection accuracy.

7.3 Reliability challenges

Current detection methods' reliability and efficiency are insufficient, especially in the case of deepfake video detection (Zhang 2022).

7.3.1 (Over)confidence of fake video detection methods

The current studies emphasise high confidence in detecting deepfakes with high accuracy and a low error rate. However, most have not evaluated their performance against unseen deepfakes or at least against perturbation attacks. To make a practical deepfake detector, we need to improve its ability to generalise, lower its cost to compute, and make it resistant to evasion attacks like adversarial attacks and simple transformations. According to Xu et al. (2022), they reviewed more than 100 peer-reviewed or arXiv papers on deepfake detection and found that only a few papers had tested their method from all three points of view above.

7.3.2 New emerging manipulation techniques

Our survey found that most deepfake detection methods assume a static game (Mirsky and Lee 2021; Wang and Gupta 2015). In practice, most deepfake detection algorithms do not work well because they are trained to look for certain types of synthetic videos. Most deepfake detection techniques are data-driven and, therefore, cannot be applied to unknown datasets. Moreover, developing supervised classifiers with one tampering technique works effectively, and keeping the baseline training up to date with the latest forging techniques is challenging. Constantly updating the supervised training is not feasible, considering new manipulation techniques may arise without notice.

Impact on fake video detection Deepfakes are an emerging technology than any other DNN applications. The continual advancement of generating technologies makes deepfake video detection very challenging. Detection algorithms must accommodate new patterns and artefacts. This arms race requires continual research and development to upgrade detection models, making it a more dynamic challenge than many other DNN applications.

Methods addressing new emerging manipulation techniques The majority of suggested deep learning-based deepfake detection algorithms are unsuitable for generalisation, and there is much to accomplish in this area. Ranjan et al. (2020) have improved the generalisation capabilities of deepfake detection with transfer learning. Suratkar et al. (2020) combine CNN with transfer learning. This will generalise the method in certain contexts using what has been previously learned in another context. Since transfer learning uses existing knowledge, it frequently produces better outcomes, even when data is scarce for training. In areas with lots of annotated data, transfer learning can help improve learning performance (Liang et al. 2019). To identify synthetic fake faces, the OC-FakeDect system (Khalid and Woo 2020) learns from real-world examples. In contrast to fake face detectors based on binary classifiers, OC-FakeDect takes a more balanced approach. Although its resistance to perturbation attacks is debatable, the methodology generalises well among DeepFake approaches. Zhang et al. (2019) claim that artefacts created using GANs have the potential to generalise to other synthetic methods. However, they have not evaluated how well their method can withstand perturbation attacks.

7.3.3 Insufficient benchmarks

Benchmarks play an important role in DNN research since they offer standardised datasets and assessment methodologies, enabling researchers to objectively evaluate the performance of their models and reproducible.

Impact on fake video detection Despite many deepfake video detection-related works published in recent years, publicly available benchmark datasets are still scarce. Deepfake video datasets are as important as detection algorithms. However, there is a lack of reliable standards because collecting fake videos is a complex and time-consuming task (Guo et al. 2020). According to Zhang (2022), standard benchmark datasets are needed for deepfake detection because current datasets have varied resolutions (for images and videos), small video lengths, and a lack of variety. The training and benchmark datasets should include gender, age, race, and scenario diversity.

Methods addressing insufficient benchmarks Many publicly available datasets in this area can be used to test the efficacy of various approaches to deepfake video detection. The present size of the deepfake video collection is sufficient for detection algorithms. However, videos in these datasets still have certain obvious visual artefacts of low quality.

It's important to note that the Deepfake Detection Challenge (DFDC) dataset collects data more randomly to use the deepfake detection algorithm in the real world. This causes more visual fluctuations. This is something that should be kept in mind (Dolhansky et al. 2020). Li et al. (2020) presented the Celeb-Df dataset, which enhanced the flickering and low-resolution generated faces of early deepfake videos. There are 590 real movies and 5639 fakes in the training set. When compared to other datasets, Celeb-DF has the lowest detection accuracy.

7.3.4 Lack of robustness

DNNs are vulnerable to performance decline outside their training environment due to their lack of resilience. This problem is crucial for implementing DNNs with innovative or hostile inputs in real-world applications. DNNs lack robustness for several reasons, and overcoming these issues is crucial for their development; a robust, adversary-proof, deepfake detection system is necessary for maintaining public trust in Media.

Impact on fake video detection However, Hulzebosch et al. (2020) recently concluded that these DNNs-based deepfake detectors are not robust enough to be used in real-world scenarios. Deepfake generators employ a wide variety of evasion methods as well as adversarial machine learning (AML) approaches to trick deepfake detectors. Cybercriminals can employ AML to corrupt a machine learning model. According to Neekhara et al. (2021), Carlini and Farid (2020), CNN-based deepfake detection systems have recently been exposed to adversarial strategies using gradient-based adversarial attacks. These attacks cause the classifier's accuracy to degrade to a near-0% level. A video created to fool an open-source deepfake detection system could also consistently fool other unknown CNN-based detection methods, posing a serious security risk to the production of CNN-based detectors (Hulzebosch et al. 2020).

Methods addressing the lack of robustness Detection procedures must be resistant to intentional, and incidental countermeasures must be reliable. Several studies have examined how changes to an adversarial white-box surrogate source model can be sent to an unknown target network (Hussain et al. 2021; Cheng et al. 2019). Because of this, robust detectors must be built and tested against a wide range of attack scenarios and attacker abilities. Experiments indicated that Gram-Net resists common image degradations such as JPEG compression, blur, and noise (Liu et al. 2020). Wang et al. (2020) describes a binary classifier with impressive generalisation for recognising GAN-synthesised still images. Their data augmentation strategy has proven to be resistant to perturbation attacks.

7.3.5 Lack of explainability

DNNs are hard to explain because of their decision-making processes. DNNs are termed "black boxes" since their underlying workings and logic for predictions and judgements are not immediately accessible or intelligible by humans despite their excellent performance across a wide range of activities.

Impact on fake video detection Another key aspect of a practical deepfake detector is its ability to explain why it believes a video is fake. Current video detection approaches are unsuccessful in generating evidence to support the results. As a result, the explainability of current investigations is restricted. This lack of explainability poses several challenges, including trust and adoption, debugging and improvement, regulatory compliance, ethical and fair decision-making, and human collaboration.

8 Open issues

Despite the significant progress in deepfake video detection, several crucial challenges remain unsolved for present deepfake video detection methods.

Real-time and high-quality data collection To detect deepfakes in real time, acquiring and analysing a large and unbiased dataset is necessary. Collecting real-time data is one of the DL-based method's primary limitations. Unfortunately, many real-time application areas cannot access large amounts of new data.

High computation time/cost In a real-world scenario, the time required to detect deepfake is critical. Due to their significant time consumption, current detection algorithms are not widely used in practical applications. Unfortunately, in the existing literature on deepfake detection, detection accuracy is regarded as the only criterion, with only a few studies paying attention to the amount of time required for deepfake detection to be performed.

No strong bench-marking to evaluate Ddtector's performance In 2020, Facebook hosted a DeepFake Detection Challenge (DFDC), which attracted over 2000 teams. On the public dataset, the top-performing model attained an accuracy of 82.56%. However, when the entries were compared to the black box dataset, the top-performing models' scores shifted considerably. Selim Seferbekov's model was the most successful, scoring 65.18% accuracy against the black box dataset (Dolhansky et al. 2020). On the other hand, many existing deepfake detectors claim accuracy. These results show that it's still unclear how well-existing detection methods work. The genuine performance of current and future deepfake detectors cannot be evaluated without a platform with competitive baselines and challenging datasets.

Adversarial attacks on deepfake detectors Gaussian noise, blurring, image or video compression, and other factors can all degrade deepfakes. Additionally, rival researchers are starting to pay attention to designing strategies to avoid deepfake detectors recognising fake faces. DNNs are used in more than 90% of ways to classify genuine from fake. Adversarial noise attacks using undetectable additive noises are effective against DNNs. The current study has not evaluated their resistance to adverse noise attacks.

Lack of generalised deepfake detectors A key performance indicator for algorithms is generalisation. One of the most difficult issues in the battle against deepfakes is dealing with unknown deepfakes. Most of the current methods for detecting deepfakes are troubled by the problem of overfitting the training data and a lack of generalisation across different datasets and generative models. Generalisation is frequently used to evaluate the performance of algorithms on unknown datasets. Many proposed detection techniques are built around supervised learning, which will likely work better on their datasets. Research on existing detection algorithms showed that their generalisation ability is still not good enough for cross-dataset detection. Due to this research gap, existing deepfake detection algorithms cannot generalise well across datasets and new types of deepfakes. Several studies focus on this objective of developing more generic detection approaches.

Quality of deepfake video datasets Developing deepfake detection algorithms largely depends on available datasets of deepfake videos. Most existing deepfake detection algorithms require extensive training datasets. The higher the quality of the datasets, the better the detection. Unfortunately, most of the available datasets contain very low-quality videos. Figure 9 shows examples of low-quality modified faces from the DFDC dataset. The low-quality examples include colour mismatches, evident splicing boundaries, and inconsistent synthetic face orientations. Most DeepFake detectors can confidently identify low-quality deepfakes with observable artefacts, but the problematic high-quality deepfakes that can

Fig. 9 Examples of low-quality deepfakes from DFDC dataset. These examples show how defects in deepfakes, such as colour mismatches, visible splicing lines, and mismatched face layouts, make them easily detectable. (Color figure online)

mislead our eyes can only be rarely detected by detectors. Moreover, we discovered that many publicly available datasets did not guarantee that their individuals were willing participants or had consented to altering their faces.

9 Future opportunities

The "opponents" are the DeepFake generating techniques, while the "defenders" are the DeepFake detection methods. We believe the conflict between opponents and defenders may result in gradual but persistent scientific progress and discoveries. We anticipate important directions for deepfake detection systems that will gain more attention in the coming years. We have attempted to link existing open challenges with potential future opportunities in Fig. 10.

Data-efficient learning to reduce computation time Research should aim to use limited data to train the model to lower the computing cost in the proposed research on deepfake detection. In addition to limiting the amount of data used, the new detection techniques should also minimise the processing time and cost, which is regarded as an essential aspect when discussing the adoption of deepfake detection algorithms in real-world applications. An essential requirement for a sound learning system is that new types of tasks must be learned quickly, which most existing methods have been unable to do. Deepfake detection algorithms will be widely utilised on streaming media platforms to limit the negative impact of deepfake videos on social security. In the future, greater emphasis should be placed on the study of how to create a detection method that is both efficient and accurate.

Use of unsupervised/semi-supervised learning Most of these detection methods are supervised, and they have difficulty generalising across domains and datasets (Zotov et al. 2020). Semi-supervised learning improves the network's generalisation. Kumar et al. (2018) confirm the CNN model achieves 91.7% accuracy, but only in a set environment. There have been a lot of ideas about how to use machine learning in the last few years, like meta-learning, embedding learning, and generative modelling. A few-shot learning approach, transfer learning, and adversarial machine learning are examples of these learning strategies (Sun et al. 2019).

Hybrid models for improved generalisation Despite their popularity, existing detection models can't be generalised based on a few samples. On the other hand, humans can quickly

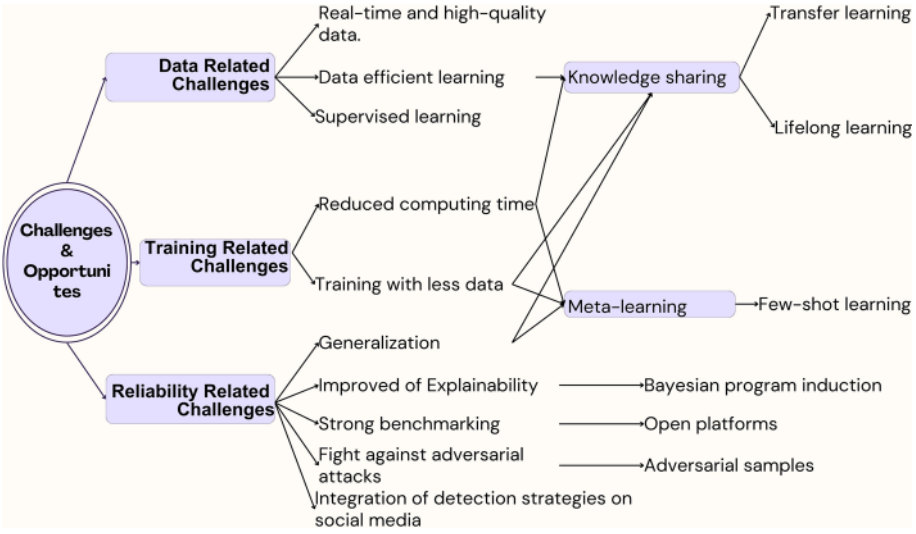

Fig. 10 A taxonomy of potential solutions to challenges in deepfake video detection. This image connects current open challenges with future opportunities, outlining strategies for reducing data requirements, training efficiency, and reliability in deepfake detection systems

acquire new skills by applying previous knowledge. The ability to transfer knowledge that a model has learned without having to do more target-specific supervised learning is a new way to deal with the problem of overfitting (Lampert et al. 2009). Another approach to achieving generalisation may be a hybrid approach like physics-guided machine learning.

Academic researchers are looking for more data-efficient models that exploit artificial learners' capabilities without much supervision and with a reduced amount of training data. Addressing these research challenges is essential for ML/DL-based detection models to be applied to real-world cases. Hybrid models are easier to scale up and use less computing resources (Ren et al. 2021; Peng et al. 2022). The hybrid strategies could be useful to:

- achieves generalisation by embedding "knowledge" into your model, so it can anticipate previously unseen data and perform well.
- achieves explainability because the physical formula is predictable, so you can add insights and consistency to otherwise "black box" machine learning models.

Strong bench-marking There is a strong need for standardised benchmarks, comprising protocols and tools for deepfake generation and detection, common criteria, or open platforms to transparently compare detection models against benchmarks. Moreover, developing deepfake detection algorithms largely depends on the available datasets of deepfake videos. Most existing research uses GANs to produce their image dataset to test the deepfake detection methods. Nobody knows the quality of these fake images or whether they contain noticeable flaws. Public availability of high-quality video datasets will aid in developing more efficient detection models.

Defending against adversarial attacks DNNs are used in more than 90% of the methods for classifying real videos from fake ones. Studies have demonstrated that adversarial noise attacks are effective against DNNs. The above-mentioned DNN-based detection algorithms

are vulnerable to adversarial noise attacks, and limited research has evaluated their performance against adversarial noise attacks. Most existing models have not been evaluated for their resistance to adverse noise attacks. So, there is an opportunity to develop deepfake detectors that are more resistant to changes made by adversaries (Hou et al. 2021; Rao et al. 2021; Neekhara et al. 2021).

Integration of deepfake detection methods into social media Current deepfake video detection methods are less productive in real-time scenarios (Yu et al. 2021). Therefore, another research direction is to integrate detection methods into distribution platforms such as social media to increase their effectiveness in dealing with the widespread impact of deepfakes.

9.1 Summary of future opportunities

This section summarises potential future directions that researchers who are already working in the field of deepfake technology or who aspire to work there in the future can investigate:

- Researchers haven't paid much attention to computational complexity, which offers another area with the potential for efficient deepfake detection. Specifically, computational time has been neglected in the literature. Improvements in computational time can be studied for real-time applications.
- Hybrid methods have not been substantially explored or employed in deepfake video detection. Hybrid methods have the potential to provide excellent classification accuracy for real-time fake video detection.
- An existing deepfake detection model could be very useful if the results could be reproducible. This could be done by giving the research community access to large public datasets, experimental setups, and open-source tools and codes. It will help show real progress in the field by keeping people from overestimating how well things are going.

10 Conclusion

Deepfake technology and social media make it easier to spread fake content. Addressing this problem is very important because people's confidence in media content is decreasing due to deepfakes, as seeing them does not guarantee trust. The progression of deepfake technology increases the risk of spreading false information, which directly compromises the trustworthiness of news, information, and interpersonal exchanges.

We presented a brief overview of deepfake generation techniques and a detailed analysis of current deepfake video detection methods and their vulnerabilities. As a new research topic, there is a battleground between the two sides of deepfake technologies: opponents (deepfake generation methods) and defenders (deepfake detection methods). The competition between these two parties provides new opportunities that can help identify research questions, research trends, and directions in deepfake video detection. Since there is no indication that the development of deepfake technology will be slowed down, academics and government officials should discuss and resist this destructive technology. Researchers, policymakers, and technology experts should come forward to devise comprehensive strategies for mitigating the impact of deepfakes. To help researchers and practitioners working

in this rapidly growing and expanding field, this survey paper has highlighted many important research issues that need to be examined.

Author contributions AK: conceptualisation, investigation, writing—original draft, editing, FX: conceptualisation, review and editing, ANH: supervision, writing—review and editing, SF: supervision, review and editing, VS: review and editing.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Adadi A (2021) A survey on data-efficient algorithms in big data era. J Big Data 8(1):1–54
Afchar D, Nozick V, Yamagishi J et al (2018) MesoNet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–7
Agarwal S, Farid H, Gu Y et al (2019) Protecting world leaders against deep fakes. In: CVPR workshops. pp 38–45
Agarwal S, Farid H, Fried O et al (2020) Detecting deep-fake videos from phoneme-viseme mismatches. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp 660–661
Aghasanli A, Kangin D, Angelov P (2023) Interpretable-through-prototypes deepfake detection for diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 467–474
Akhtar Z (2023) Deepfakes generation and detection: a short survey. J Imaging 9(1):18
Amerini I, Galteri L, Caldelli R et al (2019) Deepfake video detection through optical flow based CNN. In: Proceedings of the IEEE/CVF international conference on computer vision workshops
Anjum A, Abdullah T, Tariq MF et al (2016) Video stream analysis in clouds: an object detection and classification framework for high performance video analytics. IEEE Trans Cloud Comput 7(4):1152–1167
Bansal N, Aljrees T, Yadav DP et al (2023) Real-time advanced computational intelligence for deep fake video detection. Appl Sci 13(5):3095
Berthouzoz F, Li W, Dontcheva M et al (2011) A framework for content-adaptive photo manipulation macros: application to face, landscape, and global manipulations. ACM Trans Graph 30(5):120–1
Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. arXiv Preprint http://arxiv.org/abs/1809.11096
Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
Carlini N, Farid H (2020) Evading deepfake-image detectors with white-and black-box attacks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp 658–659
Cellan-Jones R (2019) Deepfake videos 'double in nine months'. https://www.bbc.com/news/technology-49961089

Chan CCK, Kumar V, Delaney S et al (2020) Combating deepfakes: multi-LSTM and blockchain as proof of authenticity for digital media. In: 2020 IEEE/ITU international conference on artificial intelligence for good (AI4G). IEEE, pp 55–62
Cheng S, Dong Y, Pang T et al (2019) Improving black-box adversarial attacks with a transfer-based prior. In: Advances in neural information processing systems, vol 32
Child R (2020) Very deep VAEs generalize autoregressive models and can outperform them on images. arXiv Preprint http://arxiv.org/abs/2011.10650
Ciftci UA, Demir I, Yin L (2020) FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.3009287
Coccomini DA, Caldelli R, Falchi F et al (2022) Cross-forgery analysis of vision transformers and CNNs for deepfake image detection. In: Proceedings of the 1st international workshop on multimedia AI against disinformation. pp 52–58
Cozzolino D, Thies J, Rössler A et al (2018) ForensicTransfer: weakly-supervised domain adaptation for forgery detection. arXiv Preprint http://arxiv.org/abs/1812.02510
Dang H, Liu F, Stehouwer J et al (2020) On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 5781–5790
Dhariwal P, Nichol A (2021) Diffusion models beat GANs on image synthesis. Adv Neural Inf Process Syst 34:8780–8794
Dolhansky B, Bitton J, Pflaum B et al (2020) The deepfake detection challenge (DFDC) dataset. arXiv Preprint http://arxiv.org/abs/2006.07397
Dufour N, Gully A (2019) Contributing data to deepfake detection research. Google AI Blog 1(2):3
Frank J, Eisenhofer T, Schönherr L et al (2020) Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning. PMLR, pp 3247–3258
Fung S, Lu X, Zhang C et al (2021) DeepfakeUCL: deepfake detection via unsupervised contrastive learning. In: 2021 international joint conference on neural networks (IJCNN). pp 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534089
Gambín ÁF, Yazidi A, Vasilakos A et al (2024) Deepfakes: current and future trends. Artif Intell Rev 57(3):64
George AS, George AH (2023) Deepfakes: the evolution of hyper realistic media manipulation. Partn Univers Innov Res Publ 1(2):58–74
Gong LY, Li XJ (2024) A contemporary survey on deepfake detection: datasets, algorithms, and challenges. Electronics 13(3):585
Goodfellow I, Pouget-Abadie J, Mirza M et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144
Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6
Guo B, Ding Y, Yao L et al (2020) The future of false information detection on social media: new perspectives and trends. ACM Comput Surv (CSUR) 53(4):1–36
He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851
Hou M, Wang L, Liu J et al (2021) A3Graph: adversarial attributed autoencoder for graph representation learning. In: Proceedings of the 36th annual ACM symposium on applied computing. pp 1697–1704
Hulzebosch N, Ibrahimi S, Worring M (2020) Detecting CNN-generated facial images in real-world scenarios. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp 642–643
Hussain S, Neekhara P, Jere M et al (2021) Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 3348–3357
Ivanovska M, Struc V (2024) On the vulnerability of deepfake detectors to attacks generated by denoising diffusion models. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 1051–1060
Ji S, Xu W, Yang M et al (2012) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
Jiang L, Li R, Wu W et al (2020) DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 2889–2898

Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 4401–4410

Kawa P, Syga P (2020) A note on deepfake detection with low-resources. arXiv Preprint http://arxiv.org/abs/2006.05183

Khalid H, Woo SS (2020) OC-FakeDect: classifying deepfakes using one-class variational autoencoder. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp 656–657

Khanjani Z, Watson G, Janeja VP (2023) Audio deepfakes: a survey. Front Big Data 5:1001063

Kietzmann J, Lee LW, McCarthy IP et al (2020) Deepfakes: trick or treat? Bus Horiz 63(2):135–146

Kim H, Garrido P, Tewari A et al (2018) Deep video portraits. ACM Trans Graph (TOG) 37(4):1–14

Kingra S, Aggarwal N, Kaur N (2023) Emergence of deepfakes and video tampering detection approaches: a survey. Multimed Tools Appl 82(7):10165–10209

Kohli A, Gupta A (2021) Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN. Multimed Tools Appl 80:18461–18478

Koopman M, Rodriguez AM, Geradts Z (2018) Detection of deepfake video manipulation. In: The 20th Irish machine vision and image processing conference (IMVIP). pp 133–136

Korshunov P, Marcel S (2018) DeepFakes: a new threat to face recognition? Assessment and detection. arXiv Preprint http://arxiv.org/abs/1812.08685

Korshunova I, Shi W, Dambre J et al (2017) Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision. pp 3677–3685

Kumar P, Singhal A, Mehta S et al (2016) Real-time moving object detection algorithm on high-resolution videos using GPUs. J Real Time Image Proc 11:93–109

Kumar AD, Soman KP et al (2018) DeepImageSpam: deep learning based image spam detection. arXiv Preprint http://arxiv.org/abs/1810.03977

Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 951–958

Le TN, Nguyen HH, Yamagishi J et al (2021) OpenForensics: large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 10117–10127

Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. arXiv Preprint http://arxiv.org/abs/1811.00656

Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–7

Li Y, Yang X, Sun P et al (2020) Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 3207–3216

Liang H, Fu W, Yi F (2019) A survey of recent advances in transfer learning. In: 2019 IEEE 19th international conference on communication technology (ICCT). IEEE, pp 1516–1523

Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88

Liu J, Kong X, Xia F et al (2018) Artificial intelligence in the 21st century. IEEE Access 6:34403–34421

Liu Z, Qi X, Torr PH (2020) Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 8060–8069

Liu J, Zhu K, Lu W et al (2021a) A lightweight 3D convolutional neural network for deepfake detection. Int J Intell Syst 36(9):4990–5004

Liu MY, Huang X, Yu J et al (2021b) Generative adversarial networks for image and video synthesis: algorithms and applications. Proc IEEE 109(5):839–862

Lu J, Behbood V, Hao P et al (2015) Transfer learning using computational intelligence: a survey. Knowl Based Syst 80:14–23

Lyu S (2020) DeepFake detection: current challenges and next steps. In: 2020 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, pp 1–6

Mahmud BU, Sharmin A (2021) Deep insights of deepfake technology: a review. arXiv Preprint http://arxiv.org/abs/2105.00192

Maksutov AA, Morozov VO, Lavrenov AA et al (2020) Methods of deepfake detection based on machine learning. In: 2020 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus). pp 408–441. https://doi.org/10.1109/EIConRus49466.2020.9039057

Malik MH, Ghous H, Qadri S et al (2023) Frequency-based deep-fake video detection using deep learning methods. J Comput Biomed Inform 4(02):41–48

Masood M, Nawaz M, Malik KM et al (2023) Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 53(4):3974–4026
Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE winter applications of computer vision workshops (WACVW). IEEE, pp 83–92
Mezaris V, Nixon L, Papadopoulos S et al (2019) Video verification in the fake news era, vol 4. Springer. https://doi.org/10.1007/978-3-030-26752-0
Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Comput Surv (CSUR) 54(1):1–41
Mitra A, Mohanty SP, Corcoran P et al (2021) A machine learning based approach for deepfake detection in social media through key video frame extraction. SN Comput Sci 2(2):1–18
Mittal T, Sinha R, Swaminathan V et al (2023) Video manipulations beyond faces: a dataset with human-machine analysis. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 643–652
Mubarak R, Alsboui T, Alshaikh O et al (2023) A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access 11:144497–14452. https://doi.org/10.1109/ACCESS.2023.3344653
Naitali A, Ridouani M, Salahdine F et al (2023) Deepfake attacks: generation, detection, datasets, challenges, and research directions. Computers 12(10):216
Narayan K, Agarwal H, Mittal S et al (2022) DeSI: deepfake source identifier for social media. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 2858–2867
Neekhara P, Dolhansky B, Bitton J et al (2021) Adversarial threats to deepfake detection: a practical perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 923–932
Nguyen HM, Derakhshani R (2020) Eyebrow recognition for identifying deepfake videos. In: 2020 international conference of the biometrics special interest group (BIOSIG). IEEE, pp 1–5
Nguyen HH, Yamagishi J, Echizen I (2019a) Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2307–2311
Nguyen TT, Nguyen CM, Nguyen DT et al (2019b) Deep learning for deepfakes creation and detection. arXiv Preprint http://arxiv.org/abs/1909.11573 1:2
Nguyen TT, Nguyen QVH, Nguyen DT et al (2022) Deep learning for deepfakes creation and detection: a survey. Comput Vis Image Underst 223:103525
Patel M, Gupta A, Tanwar S et al (2020) Trans-DF: a transfer learning-based end-to-end deepfake detector. In: 2020 IEEE 5th international conference on computing communication and automation (ICCCA). IEEE, pp 796–801
Patel Y, Tanwar S, Gupta R et al (2023) Deepfake generation and detection: case study and challenges. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3342107
Patil K, Kale S, Dhokey J et al (2023) Deepfake detection using biological features: a survey. arXiv Preprint http://arxiv.org/abs/2301.05819
Peng C, Xia F, Saikrishna V et al (2022) Physics-informed graph learning: a survey. arXiv Preprint http://arxiv.org/abs/2202.10679
Pu J, Mangaokar N, Kelly L et al (2021) Deepfake videos in the wild: analysis and detection. In: Proceedings of the web conference 2021. pp 981–992
Qi GJ, Luo J (2020) Small data challenges in big data era: a survey of recent progress on unsupervised and semi-supervised methods. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.3031898
Qian Y, Yin G, Sheng L et al (2020) Thinking in frequency: face forgery detection by mining frequency-aware clues. In: European conference on computer vision. Springer, pp 86–103
Rana MS, Nobi MN, Murali B et al (2022) Deepfake detection: a systematic literature review. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3154404
Ranjan P, Patil S, Kazi F (2020) Improved generalizability of deep-fakes detection using transfer learning based CNN framework. In: 2020 3rd international conference on information and computer technologies (ICICT). IEEE, pp 86–90
Rao S, Verma AK, Bhatia T (2021) A review on social spam detection: challenges, open issues, and future directions. Expert Syst Appl 186:115742
Ren J, Xia F, Liu Y et al (2021) Deep video anomaly detection: opportunities and challenges. In: 2021 international conference on data mining workshops (ICDMW). IEEE, pp 959–966
Ren J, Xia F, Lee I et al (2022) Graph learning for anomaly analytics: algorithms, applications, and challenges. ACM Trans Intell Syst Technol. https://doi.org/10.1145/3570906

Rossler A, Cozzolino D, Verdoliva L et al (2019) FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 1–11
Sabir E, Cheng J, Jaiswal A et al (2019) Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3(1):80–87
Shahzad HF, Rustam F, Flores ES et al (2022) A review of image processing techniques for deepfakes. Sensors 22(12):4556
Silver D, Huang A, Maddison CJ et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489
Sohan MF, Solaiman M, Hasan MA (2023) A survey on deepfake video detection datasets. Indones J Electr Eng Comput Sci 32(2):1168–1176
Song H, Huang S, Dong Y et al (2023) Robustness and generalizability of deepfake detection: a study with diffusion models. arXiv Preprint http://arxiv.org/abs/2309.02218
Stroebel L, Llewellyn M, Hartley T et al (2023) A systematic literature review on the effectiveness of deepfake detection techniques. J Cyber Secur Technol 7(2):83–113. https://doi.org/10.1080/23742917.2023.2192888
Sui Y, Yu M, Hong H et al (2019) Learning from imbalanced data: a comparative study. In: International symposium on security and privacy in social networks and big data. Springer, pp 264–274
Sun Q, Liu Y, Chua TS et al (2019) Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 403–412
Suratkar S, Kazi F, Sakhalkar M et al (2020) Exposing deepfakes using convolutional neural networks and transfer learning approaches. In: 2020 IEEE 17th India council international conference (INDICON). IEEE, pp 1–8
Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, pp 6105–6114
Tariq S, Lee S, Woo S (2021) One detector to rule them all: towards a general deepfake attack detection framework. In: Proceedings of the web conference 2021. pp 3625–3637
Toews R (2020) Deepfakes are going to wreak havoc on society. We are not prepared. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/?sh=17d2da4f7494
Tolosana R, Vera-Rodriguez R, Fierrez J et al (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148. https://doi.org/10.1016/j.inffus.2020.06.014
Tyagi S, Yadav D (2023) A detailed analysis of image and video forgery detection techniques. Vis Comput 39(3):813–833
Ulmer A, Tong A (2023) Deepfaking it: America's 2024 election collides with AI boom. https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-AI-boom-2023-05-30/
Verdoliva L (2020) Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process 14(5):910–932
Wang X, Gupta A (2015) Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE international conference on computer vision. pp 2794–2802
Wang H, Xie D, Wei L (2018) Robust and real-time face swapping based on face segmentation and CANDIDE-3. In: Pacific Rim International Conference on Artificial Intelligence. Springer, pp 335–342
Wang Y, Yao Q, Kwok JT et al (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv (CSUR) 53(3):1–34
Westerlund M (2019) The emergence of deepfake technology: a review. Technol Innov Manag Rev 9(11):39–52
Xia F, Liu J, Nie H et al (2019) Random walks: a review of algorithms and applications. IEEE Trans Emerg Top Comput Intell 4(2):95–107
Xia F, Sun K, Yu S et al (2021) Graph learning: a survey. IEEE Trans Artif Intell 2(2):109–112. https://doi.org/10.1109/TAI.2021.3076021
Xu FJ, Wang R, Huang Y et al (2022) Countering malicious deepfakes: survey, battleground, and horizon. Int J Comput Vis. https://doi.org/10.1007/s11263-022-01606-8
Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 8261–8265
Yu P, Xia Z, Fei J et al (2021) A survey on deepfake video detection. IET Biom 10(6):607–624. https://doi.org/10.1049/bme2.12031
Zhang T (2022) Deepfake generation and detection, a survey. Multimed Tools Appl 81(5):6259–6276
Zhang X, Karaman S, Chang SF (2019) Detecting and simulating artifacts in GAN fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–6

Zhang J, Wang W, Xia F et al (2020) Data-driven computational social science: a survey. Big Data Res 21:100145
Zhao J, Mathieu M, LeCun Y (2016) Energy-based generative adversarial network. arXiv Preprint http://arxiv.org/abs/1609.03126
Zhao Y, Ge W, Li W et al (2019) Capturing the persistence of facial expression features for deepfake video detection. In: International conference on information and communications security. Springer, pp 630–645
Zhao H, Zhou W, Chen D et al (2022) Self-supervised transformer for deepfake detection. arXiv Preprint http://arxiv.org/abs/2203.01265
Zhao C, Wang C, Hu G et al (2023) ISTVT: interpretable spatial-temporal video transformer for deepfake detection. IEEE Trans Inf Forensics Secur 18:1335–1348
Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv (CSUR) 53(5):1–40
Zhou P, Han X, Morariu VI et al (2018) Learning rich features for image manipulation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1053–1061
Zi B, Chang M, Chen J et al (2020) WildDeepfake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM international conference on multimedia. pp 2382–2390
Zotov S, Dremliuga R, Borshevnikov A et al (2020) Deepfake detection algorithms: a meta-analysis. In: 2020 2nd symposium on signal processing systems. pp 43–48

Authors and Affiliations

Achhardeep Kaur[1] · Azadeh Noori Hoshyar[2] · Vidya Saikrishna[3] · Selena Firmin[1] · Feng Xia[4]

✉ Achhardeep Kaur
achhardeepkaur@students.federation.edu.au

Azadeh Noori Hoshyar
a.noorihoshyar@federation.edu.au

Vidya Saikrishna
v.saikrishna@federation.edu.au

Selena Firmin
s.firmin@federation.edu.au

Feng Xia
f.xia@ieee.org

[1] Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, VIC 3353, Australia

[2] Institute of Innovation, Science and Sustainability, Federation University Australia, Berwick, VIC 3806, Australia

[3] Global Professional School, Federation University Australia, Ballarat, VIC 3353, Australia

[4] School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia