Artificial Intelligence Review (2024) 57:64
https://doi.org/10.1007/s10462-023-10679-x

# Deepfakes: current and future trends

Ángel Fernández Gambín[1] · Anis Yazidi[1] · Athanasios Vasilakos[2] · Hårek Haugerud[1] · Youcef Djenouri[3,4,5]

Published online: 19 February 2024

**Abstract**
Advances in Deep Learning (DL), Big Data and image processing have facilitated online disinformation spreading through Deepfakes. This entails severe threats including public opinion manipulation, geopolitical tensions, chaos in financial markets, scams, defamation and identity theft among others. Therefore, it is imperative to develop techniques to prevent, detect, and stop the spreading of deepfake content. Along these lines, the goal of this paper is to present a big picture perspective of the deepfake paradigm, by reviewing current and future trends. First, a compact summary of DL techniques used for deepfakes is presented. Then, a review of the fight between generation and detection techniques is elaborated. Moreover, we delve into the potential that new technologies, such as distributed ledgers and blockchain, can offer with regard to cybersecurity and the fight against digital deception. Two scenarios of application, including online social networks engineering attacks and Internet of Things, are reviewed where main insights and open challenges are tackled. Finally, future trends and research lines are discussed, pointing out potential key agents and technologies.

**Keywords** Artificial intelligence · Deep learning · Deepfake · Digital deception · Blockchain · GAN

## 1 Introduction

We live in the digital era. The exponential evolution of the Information and Communications Technologies (ICT) sector has transformed society, from the way we do daily things such as purchasing goods, e.g., e-commerce, to the way we communicate among each other. The worldwide adoption of Internet, together with the irruption of social networks

✉ Youcef Djenouri
youcef.djenouri@usn.no

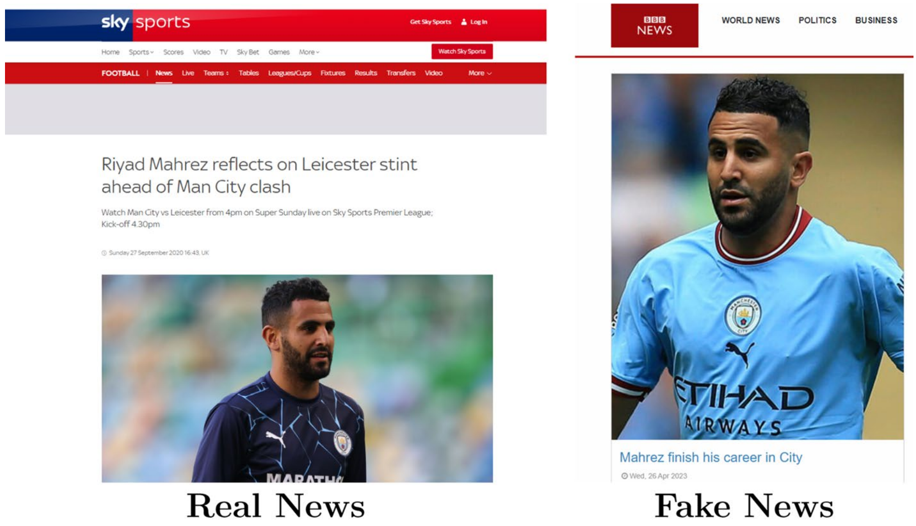
[1] Department of Computer Science, Oslo Metropolitan University, Oslo, Norway

[2] Department of Computer Science, University of Agder, Grimstad, Norway

[3] Department of Microsystems, University of South-Eastern Norway, Kongsberg, Norway

[4] Norwegian Research Center, Oslo, Norway

[5] IDEAS NCBR, Warsaw, Poland



Riyad Mahrez reflects on Leicester stint ahead of Man City clash

Watch Man City vs Leicester from 4pm on Super Sunday live on Sky Sports Premier League; Kick-off 4.30pm

Sunday 27 September 2020 16:43, UK

Mahrez finish his career in City

Wed, 26 Apr 2023

Real News

Fake News

**Fig. 1** Example of the real and fake news data collected

and media content platforms, has entirely changed the information paradigm and increased massively the amount of online data. The proliferation of budget-friendly digital devices such as smartphones, tablets, laptops, and digital cameras has led to an explosive surge in multimedia content across the online realm. Moreover, the advancement of social media in the past decade has enabled individuals to swiftly share their captured multimedia content, resulting in a substantial rise in content creation and convenient accessibility to it Masood et al. (2021). In this rapidly expanding global information landscape, the task of discerning truth and establishing trust in the veracity of information has grown increasingly challenging, potentially leading to severe repercussions (Girgis et al. 2018). Reports indicate that the unaided human capacity to detect deception stands at a mere 54% (Girgis et al. 2018). Presently, we find ourselves residing in an era characterized as "post-truth," where the deliberate dissemination of disinformation is employed by malicious entities to manipulate public opinion. This prevalent phenomenon, commonly referred to as "fake news," poses a substantial threat to democracy, journalism, and the fundamental principles of freedom of expression (Zhou and Zafarani 2020). Figure 1 outlines an example of the fake news data. It shows a real news and a fake news about the famous Algerian soccer player *Riyad Mahrez*. In fact, *Riyad Mahrez* played with his former team Leicester City in 2017. However, as far as I know, he still plays at Man City. Note that the fake news was generated by the authors using *worldgreynews* website.[1]

Disinformation can cause severe damage: election manipulation, creation of warmongering situations, defaming any person, etc Masood et al. (2021). The majority of individuals in developed economies will consume more false than true information by 2022 Fraga-Lamas and Fernández-Caramés (2020). Digital deception is commonly recognized as deceptive or misleading content created and disseminated to cause public or personal harm (e.g., post-truth, populism, and satire) or to obtain a profit (e.g., clickbaits, cloaking, ad

[1] https://www.worldgreynews.com/add-news.

farms, and identity theft). In the context of mass media, digital deception originates usually either from political institutions, governments or non-state actors, including media corporates and fraudsters, that publish content without economic or educational entrance barriers. As a consequence, these horizontal and decentralized communications cannot be controlled with traditional tools. In addition, this lack of supervision allows for security attacks (e.g., social engineering). Moreover, the veracity of information seems to be sometimes negotiable for the sake of profit, as the competition is increasingly tough (Fraga-Lamas and Fernández-Caramés 2020). At the same time, we have witnessed tremendous advancements in the field of Artificial Intelligence (AI) (especially Deep Learning (DL)), Big Data and cloud computing. This powerful technology combination is able to provide real-time data-driven intelligence, leveraging large amounts of collected data into useful information.

Thanks to these advances, together with those in image processing, the concept of *Deepfake* has appeared. It can be defined as the generation of fake digital content or manipulation of genuine one through the use of DL techniques.The content includes video, image, audio, and text among other sources. Its popularity comes mainly from the manipulation of facial appearance (attributes, identity, expression), usually classified into the following categories: (i) entire face synthesis, (ii) attribute manipulation, (iii) identity swap, and (iv) expression swap (i.e., reenactment) (Juefei-Xu et al. 2021).

Deepfake technology itself is neutral, and can be applied for good purposes in many fields including education, entertainment, online social media, healthcare, fashion, and marketing (Westerlund 2019). It has been used to create digital avatars or virtual assistance to improve the quality of experience in video conferencing (Wang et al. 2021a). For instance, in Caporusso (2020), the authors utilize deepfake algorithms to extract an accurate model of an individual and generate new content specifically designed for benign purposes. They create an interactive Digital Twin of a subject, serving as a substitute for in-person or virtual presence. The proposed application aims to provide users with user-friendly tools to create their own digital replicas for various uses, such as re-enactments, interactive stories, memorials, and simulations. Another example is the upcoming virtual concert by the legendary band ABBA in 2022, which will feature digital versions of the band members (Abba 2021). Furthermore, deepfake technology has been employed in movie and TV show production to recreate the appearance of deceased celebrities or pay tribute to them in memorial concerts through facial visual effects. Additionally, it has gained popularity in smartphone applications for entertainment purposes, particularly in creating viral videos for social media platforms (FaceApp 2021; Facebrity 2021). Another case discussed in Kwok and Koh (2021) explores the potential benefits of deepfakes in the tourism industry and related marketing.

However, the malicious uses largely dominate the positive ones. Deepfakes, enabled by advanced technologies, pose significant threats by facilitating the propagation of online fake news. The consequences of this are far-reaching, including the potential to ignite political or religious tensions between nations, deceive the public, disrupt financial markets, perpetrate acts of sabotage, fraud, scams, obstruct justice, and much more. Remarkably, deepfakes can even be employed to generate counterfeit satellite earth images with military implications (Nguyen et al. 2019). The most concerning aspect is that these technologies grant individuals with technical expertise the ability to create videos that undermine the very concept of truth. When combined with the widespread adoption of social networks, the proliferation of such manipulative content becomes immensely challenging to control or curtail (Yazdinejad et al. 2020b). Due to the gravity of the situation, it is crucial to develop techniques aimed at preventing, detecting, and mitigating the spread of deepfake content. Effective measures to combat deepfakes include: (i) implementing legislation and regulations (Langguth et al. 2021), (ii) adopting corporate policies and voluntary initiatives, (iii) promoting education and awareness, and (iv) advancing countermeasures technology (Westerlund 2019). This undertaking presents a formidable challenge, even if there exists a credible, secure, and trusted method to trace the origins of digital content. In response, the research community, major technology corporations, and governments are directing their efforts towards proposing and implementing regulations to curtail digital deception.

After a thorough review of the literature, we could not find any similar article dealing with the addressed topics in our contribution. Specifically, there exists vast literature dealing with the generation and detection of deepfakes. Therefore, our goal in this regard is to provide main insights for the reader to quickly delve into the topic. Regarding authentication, blockchain technology in the field is explored, gathering some related published articles. As far as we know, this is one of the first publications surveying this specific topic. We also provide scenarios of application for better understating the tackled issues and the potential research opportunities that the addressed technologies can offer. Finally, our extracted conclusions and lesson learned constitute an important contribution to the guidance for further research, and we believe they are the key point of this paper. All in all, the main impact of this paper is to position the reader in a perfect spot to further investigate all the aforementioned items with a big picture glimpse. Along these lines, we present a big picture perspective of the deepfake paradigm, by reviewing current and future trends. This is supported by surveying the state of the art and providing insightful references for guidance and further research. Our main contributions in this work are the following:

- A compact summary of DL techniques used for deepfakes is presented to facilitate a non-familiar reader with the topic and to get involved with technical terms.
- A review of the fight between generation and detection techniques is elaborated, focusing on the highest-impact literature to extract the current status and possible research spots.
- A discussion about the potential that new technologies, such as distributed ledgers and blockchain, can offer with regard to cybersecurity and the fight against digital deception.
- Two scenarios of application, including social media engineering attacks and Internet of Things (IoT) networks, are reviewed where main insights and open challenges are tackled.
- Future trends and research lines are discussed, mentioning potential involved agents and technologies that can play an essential role.

To the best of our knowledge, we could not find any similar paper in the literature tackling all the topics that we discuss and comprising useful information that can help to easily understand this ecosystem and the potential opportunities that it offers. Specifically, there exists a vast literature dealing with the generation and detection of deepfakes. Therefore, we provide a big picture overview in this matter, with main insights and focusing on hot-topic challenges. Regarding authentication, we delve into the opportunities that blockchain technology could provide, tackling several topics such as use cases and applications, content proof mechanisms and anomaly detection. Finally, our extracted conclusions and lesson learned constitute an important contribution to the guidance for further research (Table 1).

**Table 1** State of the art comparison

| References | Generation | Detection | Authentication | Scenarios of application |
|---|---|---|---|---|
| Yu et al. (2021) | x | ✓ | x | x |
| Almars (2021) | ✓ | ✓ | x | x |
| Juefei-Xu et al. (2021) | ✓ | ✓ | x | x |
| Mirsky and Lee (2021) | ✓ | ✓ | x | x |
| Shelke and Kasana (2020) | x | ✓ | x | x |
| Hasan and Salah (2019) | x | x | ✓ | x |
| Hasan and Salah (2019) | x | x | ✓ | x |
| This work | ✓ | ✓ | ✓ | ✓ |

The rest of the manuscript is organized as follows. A DL general overview is provided in Sect. 2. A discussion about the battleground between generation and detection is presented in Sect. 3. Leveraging blockchain technology as a way to guarantee digital content authentication is addressed in Sect. 4. In Sect. 5, scenarios of deepfake application are presented, focusing on social and communication networks. Future trends and potential challenges are proposed in Sect. 6. Finally, Sect. 7 summarizes our conclusions.

## 2 Deep learning outline

Machine Learning has revolutionized the way we comprehend data, opening up endless possibilities. Its objective is to empower machines with the ability to learn autonomously, without rigid programming (Goodfellow et al. 2016). Deep Learning, a significant breakthrough in ML, has rapidly gained traction across various application domains such as computer vision, speech recognition, and natural language processing, among many others. Unlike traditional human-crafted ML systems, a DL model automatically learns features and decision-making processes through its multi-level representation.

The purpose of this section is to provide the reader with an overview of DL, laying the foundation for a better understanding of the subsequent content in this paper. Consequently, the next section presents an outline of the key DL models utilized in the generation, detection, and prevention of deepfakes.

### 2.1 Deep learning architectures

In the following, some of the most common DL architectures used within deepfakes topic are presented. General Adversarial Networks (GANs) were the first used to build up deepfakes. In this way, architectures combining GANs with other models dominate the literature for generation purposes. Regarding detection, approaches based on Convolutional Neural Networks (CNNs) are the most common strategy, due to the nature of the used data, i.e., image and video. As for authentication, Recurrent Neural Networks (RNNs), and specifically Long Short-Term Memory (LSTM) networks, networks, are the most used models for content traceability. Nonetheless, these are general trends where endless problem-tailored solutions can be found mixing any of the following architectures.

### 2.1.1 Artificial neural networks

The structure of a feed-forward (where input only flows in one direction within the network) Artificial Neural Network (ANN) consists of an input layer, intermediate hidden layers, and an output layer. This network is capable of learning linear and nonlinear relationships between input and output pairs by utilizing extracted features. Each layer consists of at least one neuron. These neurons employ specific activation functions and are interconnected with weights, which map their input to an output. Typically, neurons within a layer employ the same activation function, thereby defining the layer type. The network type is determined by the combination of utilized layers and the structure of interconnections between neurons (Gambín et al. 2021). The Backpropagation (BP) algorithm is commonly used for training the network by finding weights for each neuron that minimize a specific error objective function (Trinh et al. 2020).

### 2.1.2 Convolutional neural networks

A CNN is a feed-forward ANN that comprises one or more convolutional layers. A number of kernels is defined per layer, with a certain number of weights. These are convolved across the whole input. Thanks to this weights reuse, the network becomes sparse, providing reduced computational complexity with respect to fully-connected feed-forward neural networks (Trinh et al. 2020). Rectified Linear Unit (ReLU) model is usually utilized as an activation layer to recognize nonlinear correlations, whereas Max Pooling is used to reduce the input size (maintaining the positional information). CNNs work well with images as inputs, with relevant contributions within image classification, object and computer vision in general (Sit et al. 2020).

### 2.1.3 Recurrent neural networks

A Recurrent Neural Network (RNN) is a type of ANN that possesses a recursive structure, allowing it to store information within the network. Neurons in a recurrent layer can be interconnected, where the output of a neuron is connected to both the next neuron within the same layer and the neuron(s) in the subsequent layer (Sit et al. 2020). A specific variant of RNN is the Long Short-Term Memory (LSTM) network. The neurons in an LSTM network are referred to as Memory Cells (MCs). MCs have the ability to retain information from past network states by utilizing gates. A gate consists of a neuron with a sigmoid activation function and a multiplication block. This unique structure enables the MCs to incorporate the sequence of past states, making LSTM networks suitable for processing time series with long-term dependencies (Hochreiter and Schmidhuber 1997). Another variant of RNN is the Gated Recurrent Unit (GRU). A GRU cell consists of a reset gate and an update gate. The reset gate determines how much past information should be forgotten, while the update gate determines what new information to incorporate in each iteration. This mechanism allows the model to decide the amount of relevant past information to be utilized in the future (Cho et al. 2014). RNNs are particularly effective in handling temporal and predictive problems.

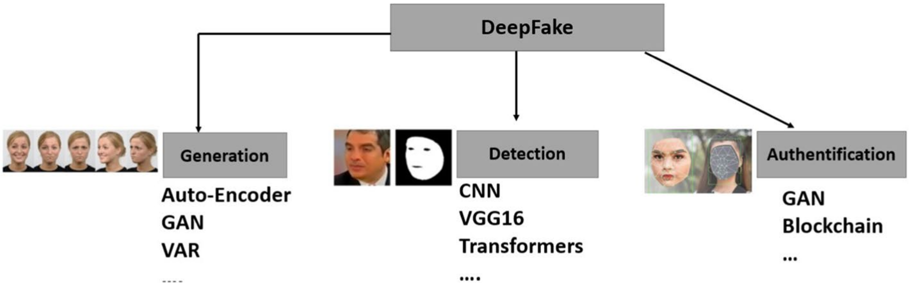


**Fig. 2** GAN architecture

### 2.1.4 Autoencoders

An Autoencoder (AE) is an unsupervised Artificial Neural Network (ANN) designed to replicate its input at its output (Al-Garadi et al. 2020). It consists of two main components: an encoder and a decoder. Each component includes one or more hidden layers. The encoder takes the input and transforms it into a feature-based representation, thereby reducing the dimensionality of the data. The decoder then attempts to reconstruct the original input from this representation. During the training process, the objective is to minimize the reconstruction error while prioritizing the learning of essential input characteristics. The Backpropagation (BP) algorithm is used in this regard. AEs are potentially important for automatic feature extraction and dimensionality reduction.

If encoder and decoder are not symmetrical, then other applications can be achieved and the ANN is called encoder-decoder network. Variational AE is another type of AE, where the encoder learns the posterior distribution of the decoder given a certain input. Variational AEs are usually better at generating content than standard ones, due to the fact that the concepts in the latent space, i.e., feature representation, are disentangled, and, thus, encodings respond better to interpolation and modification (Mirsky and Lee 2021). CNNs and RNNs can be used as AEs, increasing the model complexity to solve certain problems (Pu et al. 2016) and (Chung et al. 2016).

### 2.1.5 General adversarial networks

The concept of a General Adversarial Network was first introduced in 2014 Goodfellow et al. (2014), inspired by the zero-sum game from game theory. A GAN consists of two (deep) ANNs pitting one against the other: a generator and a discriminator, learning at the same time. The optimization process of Generative Adversarial Networks (GANs) aims to achieve a Nash equilibrium, where both the generative and discriminative models act as adversaries. The generator strives to deceive by generating samples using random noise, while the discriminator, typically a binary classifier, attempts to distinguish real training data samples from deceptive samples generated by the generator. A graphical representation of this process is shown in Fig. 2. GANs are particularly effective in tasks such as image, video, and voice generation. Over the years, numerous variations and enhancements of GANs have been proposed. In the context of deepfakes, two popular GAN-based image translation frameworks are pix2pix and CycleGAN (Mirsky and Lee 2021).

GANs play an essential role in the development of deepfakes. In this regard, cybersecurity stakeholders are employing them with outstanding results in fields such as intrusion detection, steganography, password cracking, and Anomaly Detection (AD). Two interesting papers for further research are Arora and Shantanu (2020) and Navidan et al. (2021). In the first, a systematic literature review of GANs applications in the cybersecurity domain is elaborated, including analysis of specific extended GAN frameworks, such as deep convolutional, bidirectional and cycle GANs. Moreover, several cybersecurity datasets are presented. The authors in Navidan et al. (2021) discuss about how GANs can benefit multiple aspects of computer and communication networks, including mobile networks, IoT, and cybersecurity.

### 2.1.6 Transformers

Transformers are a type of Artificial Neural Network (ANN) first introduced in 2017 by Vaswani et al. (2017). They were developed to address the challenge of sequence transduction, which involves transforming an input sequence into an output sequence. This encompasses tasks such as speech recognition, text-to-speech conversion, and more. Transformers have gained significant popularity in the field of natural language processing. For instance, OpenAI utilized transformers in their language models (OpenAI 2021), and DeepMind employed them in the development of AlphaStar (DeepMind 2021).

In order for models to effectively perform sequence transduction, it is crucial to understand the dependencies and connections within a given input, such as a sentence. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have been commonly used for this purpose due to their inherent properties. However, they also have limitations. RNNs struggle with longer input sentences as they tend to lose context for words that are distant from the current word being processed. On the other hand, CNNs are not adept at capturing dependencies effectively. To overcome these challenges, attention models were introduced, which focus on relevant subsets of the input. The idea is that every word in a sentence may contain valuable information, and attention allows the decoding process to consider each word accurately.

Transformers often combine CNNs with attention models to enable parallelization and expedite the translation between sequences. A transformer consists of two main components: an encoder and a decoder. Both components are composed of modules that can be stacked multiple times. These modules primarily consist of attention and feed-forward layers (Vaswani et al. 2017).

Transformers have demonstrated exceptional performance in modeling dependencies for various recognition tasks in computer vision and therefore hold promise in combating deepfakes. In the paper by Wang et al. (2021b), they propose a multi-scale transformer that detects local inconsistencies at different spatial levels. To enhance the detection results and improve the method's robustness against image compression, they combine frequency information with RGB features. Khan and Dai (2021) employ a video transformer with incremental learning. Another model that considers a video transformer is presented in Heo et al. (2021), where they introduce a distillation methodology using a patch-based positioning CNN model to effectively address false negative issues. Finally, a joint model based on CNN and vision transformer is discussed in Wodajo and Atnafu (2021). The CNN extracts features while the transformer categorizes them using an attention mechanism (Niu et al. 2021).

## 3 Generation vs detection

Advances in deepfake generation and detection methods are growing at a fast pace. Both sides naturally form a battleground, where the attackers operate the generation, and the defenders perform the detection. Indeed, this incessant dispute is what pushes the topic forward and enhances its remarkable progress. The study of deepfakes has gained a lot of attention in recent years and the number of publications is increasing exponentially since the first works dating back to 2016. Moreover, the field itself is getting broader, not only including media content but also other topics. Due to this, and the scope of this work, we do not intend to provide a comprehensive survey in this section comprising every available work in the topic. Instead, a discussion including the highest-impact reviews found in the literature is presented, focusing on image, video and audio related works. In this way, we also support the reader with key references for further research.

The section is organized as follows. First, a state-of-the-art analysis is presented in Sect. 3.1. Moreover, a summary of the reviewed surveys is provided in Table 2 with key ideas and scope. Then, main insights regarding the reviewed literature are distilled in Sect. 3.2, where we elaborate on both generation and detection sides.

### 3.1 State of the art

A thorough review is presented in Yu et al. (2021) with a focus on deepfake video detection, specifically addressing the generation process, various detection methods, and existing benchmarks. The authors classify algorithms into two categories based on the goal of facial image manipulation: face swapping and face reenactment. They further classify the techniques into the following categories: general Artificial Neural Network (ANN)-based methods, temporal consistency features, visual artifacts, camera fingerprints, and biological signals. The study concludes that current detection methods are not yet ready for real-world applications, and future research should prioritize generalization and robustness.

In Almars (2021), a survey on deepfake creation and detection techniques using Deep Learning (DL) is provided. The authors differentiate between image and video content for detection purposes. Within video detection, they propose two categories: analysis of biological signals and analysis of spatio-temporal features. The paper also provides access to several public datasets. The main conclusion is that current DL methods face scalability issues, necessitating the development of more robust models applicable to large, high-quality datasets.

A comprehensive overview and detailed analysis of deepfake generation and detection is presented in Juefei-Xu et al. (2021). The authors discuss the taxonomy of various generation methods and the categorization of detection models, focusing on the battle between these two sides. Interactive diagrams are provided for further exploration. The generation methods are classified into four categories: entire face synthesis, attribute manipulation, identity swap, and expression swap. For detection, the works are classified based on spatial features, frequency features, and biological signals. The paper also introduces the concept of evading deepfake detection, discussing adversarial attacks, removing fake traces in the frequency domain, and the use of advanced image filtering or generative models.

The purpose of Mirsky and Lee (2021) is to provide a deeper understanding of deepfake creation and detection, identify the shortcomings of current defense solutions, and highlight areas requiring further research. The paper discusses prevention and mitigation

**Table 2** Summary of reviewed surveys

| References | Scope | Media content | Featured sections |
|---|---|---|---|
| Yu et al. (2021) | Detection | Video | Face swapping, face reenactment |
| Almars (2021) | Generation and Detection | Video, Image | Biological signals, spatio-temporal features |
| Juefei-Xu et al. (2021) | Generation and Detection | Video, Image, Audio | Face synthesis, attribute manipulation, identity and expression swap |
| Mirsky and Lee (2021) | Generation and Detection | Video, Image, Audio | Reenactment, replacement, editing, and synthesis |
| Nguyen et al. (2019), Deshmukh and Wankhade (2020) | Generation and Detection | Video, Image | Visual artifacts and temporal features |
| Tolosana et al. (2020) | Generation and Detection | Image | Face synthesis, attribute manipulation, identity and expression swap |
| Masood et al. (2021) | Generation and Detection | Video, Audio | Face swapping, lip syncing, face reenactment, face synthesis, attribute editing |
| Shelke and Kasana (2020) | Detection | Video | Compression and noise artifacts, motion and statistical features |
| Amerini et al. (2021) | Generation and Detection | Video, Image | Multiple compression, anomaly-based architectures, device and social media identification |

as countermeasures for deepfake generation. Data provenance tracking through distributed ledgers is proposed for prevention. The paper also reviews works on adversarial Machine Learning (ML) as a means to disrupt and corrupt deepfake networks. The authors emphasize that deepfakes extend beyond human visuals, impacting domains such as healthcare, social media, and finances, highlighting the need to address a wide range of potential risks.

Another extensive survey is presented in Nguyen et al. (2019) and Deshmukh and Wankhade (2020). The detection methods are grouped into two major categories: image and video. Within video detection, two subgroups are distinguished: visual artifacts within single video frame-based methods and temporal features across frames-based methods. Temporal feature-based methods often utilize deep recurrent classification models, while visual artifacts methods are implemented using deep or shallow classifiers. The main conclusions emphasize the integration of detection methods with social media platforms, the potential of distributed ledger technology, and the promotion of explainable AI for effective understanding and utilization of information.

The work in Tolosana et al. (2020) reviews techniques for manipulating face images, including deepfake methods, as well as tools for detecting such manipulations. Four types of facial manipulation are discussed: entire face synthesis, identity swap, attribute manipulation, and expression swap. For each group, the paper provides details on techniques, existing databases, and key benchmarks. The concluding remarks highlight the need for research on the generalization ability of fake detectors against unseen conditions and suggest fusion techniques at a feature or score level to improve adaptability. The exploration of novel schemes beyond image/video information is also suggested for more robust tools.

An analysis of existing tools and ML-based approaches for the generation and detection of both audio and video deepfakes is presented in Masood et al. (2021). The authors discuss manipulation approaches, public datasets, and the performance evaluation of deepfake detection techniques along with their results for each category of deepfake.

Shelke and Kasana (2020) review video forgery detection using passive techniques. Although their main focus is not DL, their survey provides insights into video counterfeiting based on features, identified forgeries, used datasets, and performance parameters. The paper also discusses anti-forensics strategies aimed at deceiving forensic investigation by removing or hiding traces left after the forgery.

In Amerini et al. (2021), the authors examine image and video manipulations created with editing tools and discuss DL approaches employed to counter these attacks. They also analyze issues related to source camera model and device identification, as well as monitoring image and video sharing on social media platforms.

## 3.2 Main insights

Regarding generation, several open challenges can be highlighted:

*Generalization* DL models are data-driven, and therefore they reflect the learned features during training (Masood et al. 2021). To generate high-quality deepfakes, large data volumes are required, and obtaining this is a challenging task in most cases. Due to this, generalized models that adapt properly to unseen data are needed to enable the execution of a trained model for multiple target identities.

*Datasets* There is a need for large-scale diversified datasets. Most of the existing ones only expand the diversity of the content-related factors such as gender, age, or location. According to Juefei-Xu et al. (2021), the diversity regarding video, such as several resolutions and compression degrees among others, have not been fully taken into account. Moreover, they claim there is a lack of ultra-high-resolution images to work with.

*Image/Video conditions* Existing deepfake techniques generate good results in controlled environments with suitable conditions. However, several elements can compromise the final output. First, *pose variations*: the quality of manipulated content degrades significantly for scenarios where a person is looking off-camera. Moreover, another big challenge is the facial distance of the target from the camera, as an increase in distance from capturing devices results in low-quality face synthesis. Second, *illumination*: an abrupt change in illumination conditions such as in indoor/outdoor scenes results in color inconsistencies and strange artifacts in the resultant videos. Third, *occlusions*: when the face region of the source and victim are obscured with a hand, hair, glasses, or any other items, which eventually causes inconsistent facial features in the manipulated content. Finally, *temporal coherence:* the presence of evident artifacts like flickering and jitter among frames is another important drawback. These effects occur because generation frameworks work on each frame without taking into account the temporal consistency (Masood et al. 2021).

*Synthetic audio* There exists still a lack of realism in synthetic audio, including the lack of natural emotions, pauses, and speaking pace.

Regarding detection, the following issues need further attention:

*Generalization* Often adopted to evaluate a DL algorithm on unseen datasets, generalization is an important factor regarding performance and the ability to adapt to real-world scenarios. Authors in Yu et al. (2021) indicate generalization performance of existing detection algorithms is still insufficient and an urgent problem to be addressed.

*Datasets* There is a need of publicly available datasets and consensus on which benchmarks should be used for evaluation purposes. Furthermore, current DL methods are facing scalability issues. Most of the works use fragmented datasets, which translates into unacceptable results when applied to large-scale datasets. In this regard, high-quality and bigger datasets are required. Moreover, the authors in Juefei-Xu et al. (2021) suggest there is a lack of competitive baselines for comparison. Existing studies employ simple baselines rather than strong state-of-the-art to demonstrate that their DL models improve on prior studies.

*Interpretability* has been an inherent problem for ANN-based algorithms, i.e., mainly all DL architectures. Due to the black-box nature of DL models, their outputs are often difficult or in some cases even impossible to understand by human expertise. This is especially critical in practical forensic scenarios, such as those that the deepfake detection schemes are developed for. Although there has been some progress in other fields, interpretability within deepfakes is still an open issue.

*Architecture evaluation* Current deepfake detection approaches are formulated as a binary classification problem, where each sample can be either real or fake. However, for real-world scenarios, videos can be altered in ways other than deepfakes, so content not detected as manipulated does not guarantee the video is an original one. Furthermore, deepfake content can be the subject of multiple types of alteration, i.e. audio/visual, and therefore a single label may not be completely accurate. Therefore, the classification shall be enhanced to multi-class/multi-label (Masood et al. 2021).

*Time efficiency* The final goal of deepfake detection algorithms will be to widely use them on streaming media platforms. However, current models are far from this due to their high time consumption (Yu et al. 2021).

*Robustness* Assesses the ability of DL algorithms to maintain its performance when random noise or informed perturbations are present. Compared with original videos, compressed ones are more difficult to detect because they do not contain a lot of image information. According to Yu et al. (2021), an effective way to improve robustness is to add noise within the detection networks.

*Social media networks* In order to save network bandwidth or to secure users' privacy, some manipulations are performed by social media networks before uploading any content. This is known as social media laundering and removes clues with respect to underlying forgeries, and eventually increases false positive detection rates. A measure to increase the accuracy of deepfake identification approaches over social media laundering is to include simulations of these effects in training data (Masood et al. 2021).

## 4 Authentication

Although advances in combating deepfakes are improving, current solutions are limited. As we discussed in Sect. 3, there are huge efforts on how to deal with malicious deepfake applications, from research community to big technological cooperates. The main issue is that while defenses get better, offenses get smarter, and real-time detection, fact-checking, and debunking of deepfakes is becoming increasingly difficult given the rate of technical advancement and the amount of content that is posted every day (Yazdinejad et al. 2020b). There are currently no recognized techniques for determining the originality of digital media content that has been released online. It is extremely difficult to determine in a trusted way the true origin of a posted digital item (Yazdinejad et al. 2020b). In this approach, the majority of research focuses on the creation of AI-based detection techniques, although *authentication* is one component that is absent. Techniques to give tamper-proof evidence of which data is authentic are a potent alternative to trying to identify what content is phony. Therefore, a Proof of Authenticity (PoA) system is required for online digital content in order to recognize reliable published sources. Distributed ledgers, and in particular Blockchain (BC) have great potential as solutions that can aid in thwarting digital deceit. They allow for anonymity, security, and confidence in a decentralized peer-to-peer network that lacks a centralized controlling body. Blockchain is a new technology that secures network transactions using encryption. A blockchain provides a distributed database of transactions that every node on the network can access, also referred to as a *digital ledger* (Nofer et al. 2017). The network is a collection of hardware (such as computers) that must all approve a transaction for it to be validated and recorded. Transactions can also be easily audited by all the relevant parties. A BC is merely a data format that enables the creation and distribution of an exchanges "tamper-proof digital ledger". Therefore, compared to centralized systems, BC systems can make transactions more safe and transparent. Managing the ability to track of the media, the transmission infrastructure, and the transactions is the key to BC's capacity to prevent digital deceit. However, there are still issues to be solved in terms of creating efficient methods for information identification, testing, transmission, and auditing (Fraga-Lamas and Fernández-Caramés 2020).

Therefore, the goal of this section is to leverage on BC technology as a way to guarantee digital content authentication. In this regard, we review in Sect. 4.1 several interesting works that can help the reader to better understand the potential of this technology as an open research spot and its future applications. Moreover, some insights are discussed in Sect. 4.2.

## 4.1 State of the art

A list of methods for demonstrating tamper-proof verification that content is authentic is offered (Yazdinejad et al. 2020b). They talk about possible use cases for BC's functions and features as well as ways to combat deepfakes. The authors assert that present research in BC about the authentication of digital content for deepfakes is still in its early stages and offer promising directions for further study. Fraga-Lamas and Fernández-Caramés (2020) is pursuing the same objective. The authors examine BC's ability to counteract digital fraud, outlining the most pertinent applications and outlining their biggest unresolved issues. The three most promising solutions are (i) decentralized content moderation (ii) fact-checking applications with financial incentives (e.g., tokens), (iii) and decentralized social media platforms. Reliable fact-checkers can validate content for these rewards while the received rewards rise as the fact-checkers‘s reputation improves (Table 3).

A PoA of digital media looks a promising way of helping to eradicate the epidemic of forged content. A scheme using BC-based Ethereum technology, the second largest BC network, specifically Ethereum smart contracts, to track the provenance and history of digital content to its original source, even if it is copied multiple times, is presented in Hasan and Salah (2019). The smart contract utilizes hashes that store the digital content and its metadata. The metadata includes details on the camera used to record the video, the time and date, as well as any logs or manually added data that the producer of the video may have added, like a trust stamp. Their answer is based on the idea that content can be real and authentic if it can be reliably linked to a reliable or trustworthy source. Additionally, a security analysis of their BC-based solution shows how it addresses important security objectives like integrity, accountability, authorization, availability, and non-repudiation. They assess whether their solution is resistant to well-known attacks like Man in the Middle and ac DDoS.

A similar idea is discussed in Chan et al. (2020), where permissioned BC is coupled with LSTM. This indicates that original artist attestation of unaltered data would be required for media content. The smart contract merges many LSTM systems into a procedure that enables the historical provenance of digital content to be tracked. LSTMs are utilized as a deep encoder to provide distinctive biased features, that are subsequently reduced and hashed into a transaction. The end result is a theoretical framework that permits PoA for digital media via a decentralized BC.

The most popular BC network, Bitcoin, keeps all of its legal transaction history, making it simple to track money. However, by combining many transfers from various users, mixing services-which are typically provided by third-party businesses which are utilized as an efficient way to conceal the true nature of a transaction address. This type of services are against the key idea of using BC as a way of authentication. They are not illegal but they provide an extra layer of anonymity, which can be suitable for privacy purposes but also can entails security concerns. Due to this, many authorities and third parties are looking for solutions to stop or detect mixing services in order to control the cryptocurrency marketplace and prevent financial crime in general. Detecting the original user of a Bitcoin address within a mixing service goes into the AD field. Some insights about the opportunities that AD can provide in combating deepfakes are discussed in Sect. 6. The authors in Nan and Tao (2018) show that mixing services can be viewed as cluster outliers and that Bitcoin event graphs have community features. They leverage on a deep Autoencoder (AE) to identify mixing services in a real Bitcoin ledger.

**Table 3** Summary of reviewed works

| References | Scope | Highlights |
|---|---|---|
| Yazdinejad et al. (2020b) | Leveraging BC | Use cases: private keys & smart contracts; Challenges: product integration by industry |
| Fraga-Lamas and Fernández-Caramés (2020) | Leveraging BC | Apps: content moderation & decentralized social media; Challenges: distributed ledger technology |
| Hasan and Salah (2019) | PoA | Content traceability; Challenges: decentralized app for automatic PoA |
| Chan et al. (2020) | PoA | Content traceability based on smart contracts; LSTM encoder fingerprinting |
| Nan and Tao (2018) | Anomaly Detection | BC mixing services; AD within Bitcoin network |
| Patel et al. (2020) | Anomaly Detection | AD within Ethereum smart contracts |
| Yazdinejad et al. (2020a) | Leveraging BC | Cryptocurrency malware detection |

Because Ethereum smart contracts are immutable, neither developers nor attackers can alter them. They can, however, be canceled and replaced by new agreements. They are therefore at risk of assault and financial fraud inside of this BC. Furthermore, anonymity makes it difficult to spot irregularities in this vast network. In Patel et al. (2020), an AD method based on one-class graph Artificial Neural Network (ANN) is proposed and evaluated on the publicly available Ethereum data. For the purpose of identifying bitcoin malware risks, Yazdinejad et al. (2020a) suggest a deep RNN learning model. Their method uses the operation codes of Windows apps as an illustration study. The suggested model uses five different LSTM architectures to train.

## 4.2 Main insights

It is undeniable that the adoption of BC can help in combating pernicious deepfakes, thanks to its inherent features including scalability, decentralization, and transaction transparency (Yazdinejad et al. 2020b). On the other hand, BC is still in its infancy and, indeed, most of the current proposals lack practical implementations as they are based on customized assumptions. However, and based on its growth speed, we believe it will shortly be massively adopted covering the entire digital ecosystem. Furthermore, the implementation of traceability and tracking services by major media platforms will have the biggest influence in the immediate future. Decentralized social media sites, for example, cannot be disregarded, according to Fraga-Lamas and Fernández-Caramés (2020). The following issues with regard to open challenges will be highlighted:

*Detection is not enough* Research community is mainly focused on detecting verifiable false content, while other malicious uses within the digital deception field are barely investigated. Besides, strategies for guaranteeing trustworthy content sources are needed to be addressed.

*BC-based solutions* Vast majority of digital deception detection proposals are based on cryptographic hashes, which are sensitive to noise. Slight changes in a certain hash can imply the lost of information and/or content traceability. Moreover, cryptography schemes are vulnerable to certain quantum computing attacks. Therefore, solutions optimized for a better noise sensitivity, and post-quantum BC architectures must be further investigated.

*Social media networks* The integration of BC within common social media networks, e.g., Twitter, Whatsapp, Instagram, etc. is essential towards preventing the release of counterfeit videos by deepfake technology (Yazdinejad et al. 2020b). In this matter, big giant tech companies such as Google, Facebook, etc play an essential role, since they have the potential and resources to develop, test and implement countermeasure against digital deception.

*Web browsing* The implementation of a BC-based web extension that is able to trace the video origin source is another powerful solution still to be addressed. In this way, decentralized applications that are able to automate the establishment of PoA for any content shall be investigated and developed.

*Cross-disciplinary partnerships* The rapid evolution of digital deception requires multidisciplinary collaborations including corporates, academia, media and governments. Furthermore, there is not a standard intervention technique that works for everyone (Fraga-Lamas and Fernández-Caramés 2020).

*Integration with AI* BC technology alone is not able to fully solve the problem. Contextual knowledge that supports media integrity, such as social context traits, domain setting, and temporal patterns, must be taken into account in order to recognize falsification attacks.

Therefore, the combination of BC and AI looks a promising solution that can be enhanced by the huge amount of available information and complex data interactions which social media platforms can provide. The ultimate goal in this sense would be to devote strategies to prevent counterfeit reality before its spreading (Fraga-Lamas and Fernández-Caramés 2020).

## 5 Scenarios of application

The goal of this section is to analyze deepfakes in specific contexts, evaluating possible impacts and problems that can generate, and highlighting challenges and potential opportunities. In this way, we focus on two scenarios of application where digital deception is a hot topic by virtue of its devastating implications. First, in Sect. 5.1, social engineering attacks are presented where fake news generation within social media networks is addressed. Then, cybersecurity issues found in IoT networks are tackled in Sect. 5.2.

### 5.1 Online social networks

Online social media platforms are crucial for facilitating human interaction and communication. However, they and other websites which do not have the security safeguards to safeguard this data may make sensitive information accessible. Social engineering attacks can be used by hostile persons to access communication networks (Salahdine and Kaabouch 2019). These attacks use psychological tricks and dishonest tactics with the goal of obtaining sensitive data, such as passwords, private information, incriminating evidence, and bank card numbers, among others, by deceiving people or businesses into taking actions that are advantageous to the attackers. The greatest threats to cybersecurity at the moment are social engineering assaults. With the Big Data advent, attackers use the vast amount of collected data for businesses purposes, selling it in bulk as goods within black markets (Salahdine and Kaabouch 2019).

There exist many types of social engineering attacks. Among them, the following are relevant to the scope of this paper: (i) Carding, where a malicious agent/bot performs device fingerprinting and ML-based behavioral analysis to commit fraud related to bank cards and accounts (Ryabchuk 2020). (ii) Phising, where the main goal is capturing access credentials, such as usernames and passwords, from relevant websites and accounts, by sending emails or instant messaging with fraudulent information. These attacks are moving towards spear phishing attacks, i.e., more sophisticated phising where highly targeted messages are sent after initial data mining on target users. (iii) Pharming, based on the words "farming" and "phishing", is intended to redirect a website's traffic to another deceptive site. This can be done by installing a malicious program on the victim computer or by exploitation of a Domain Name System (DNS) server vulnerability. Further, insights on these matters can be found in Krombholz et al. (2015) and Salahdine and Kaabouch (2019).

In recent years, these attacks have been combined to perform online identity theft. This is the ultimate and more sophisticated attack, where the scammer, after collecting confidential information from the victim through the aforementioned methods, is able to impersonate the victim and act online on his behalf without consent. Considering social networks credentials and bank details subtraction, the potential harm to the victim can be extremely severe, specially nowadays that our lives are based on online digital services. Regarding this, deepfakes can exponentially increase the potential damage, considering the

**Table 4** Summary of reviewed literature

| References | Scope | Highlights |
|---|---|---|
| Ahmed (2021) | Misinformation spreading | Role played by political interest within deepfake sharing |
| Pérez Dasilva et al. (2021) | Disinformation spreading | Deepfake sharing through Twitter social network |
| Fagni et al. (2021) | Data collection | Twitter data for deepfake detection |
| Girgis et al. (2018) | Disinformation spreading | DL-based fake news detection mechanism |
| Kaliyar et al. (2020) | Disinformation spreading | DL-based fake news detection mechanism |
| Maddocks (2020) | Misinformation spreading | Political and pornographic deepfakes analysis |
| Mjaaland (2020) | Disinformation spreading | DL-based fake news detection mechanism |
| Sabeeh et al. (2020) | Disinformation spreading | Opinion mining-based fake news detection mechanism |
| Sadr (2021) | Disinformation spreading | ML-based fake news detection mechanism |

improvements that impersonation can achieve including video and text context generation (Table 4).

Deepfakes are thus predicted to be one of the biggest problems for social media platforms in the next years. To recognize and combat deepfakes, Facebook and Adobe have already developed regulations. The most recent being Twitter, which just recently declared a new set of guidelines to lessen the impact of altered information. Google has additionally made the decision to take action in order to curtail their influence by developing an algorithm to identify and instantly remove deepfakes posted to YouTube and other Google sites. For the purpose of assisting journalists in spotting faked photos, the Assemble tool was developed. Academic research on digital misinformation on social media has just lately started, despite the significant efforts of huge tech corporations (Pérez Dasilva et al. 2021).

In this section, we examine the literature on the creation and dissemination of online text, particularly fake news on social media networks. Relevant state of the art is presented in the following, where main findings are also highlighted for each reviewed contribution.

The study in Ahmed (2021) provides insights into the unintentional sharing of deep fakes and emphasizes the importance of political interest, a major driver of political involvement and a factor that is positively correlated with deep fake sharing. The main findings indicate that individuals who are politically interested and poor cognitively ability are more likely to unintentionally spread deepfakes. Furthermore, the association between political motivation and sharing is moderated by network size.

The authors in Pérez Dasilva et al. (2021) investigate the deepfake trend on Twitter. To pinpoint key players and their relationships, NodeXL was employed. To find hidden patterns and dominant content, the underlying semantic connections of the messages were also examined. The fact that journalists and media organizations make up half of those participating in disseminating deep fakes is evidence of the anxiety that this sophisticated kind of deception incites among this group, according to the results. Furthermore, despite the fact that most deepfakes that circulate on the web are pornographic in nature, political deepfakes are the ones that have the most public attention due to their capacity to produce instability in a variety of contexts, including both within and across countries.

Systems for detecting deepfake social media posts must be developed. In order to achieve this, a publicly accessible collection of deepfake messages is offered in the reference (Fagni et al. 2021). Every tweet that was gathered-including those from a whopping 23 in bots that were impersonating 17 in human accounts-was truly posted on Twitter. The bots are built using a variety of generational methods, including Markov Chains, RNN, and LSTM. Additionally, a small sample of randomly chosen tweets from the individuals that the bots imitated were chosen to provide an equal number of 25,572 tweets. Finally, they assess a number of cutting-edge text detection techniques. According to their findings, a wide range of detectors based on ML or DL approaches and transformer-based utilizing transfer learning) had more trouble successfully identifying a deepfake tweet than a tweet that was actually produced by a human. A classifier that can predict whether a piece of news is fake or not based only its content is built in Girgis et al. (2018), thereby approaching the problem from a purely deep learning perspective by RNN models, i.e., vanilla, Gate Recurrent Unit (GRU) and LSTM. They leverage on a public benchmark dataset called LIAR. Collected a decade-long, 12.8*k* manually labeled short statements in various contexts from Politifact, which provides a detailed analytical report and a link to its source level for each case.

News content and the existence of communities sharing the same opinions in the social network are taken into account for fake news detection in Kaliyar et al. (2020). The news-user engagement (relation between user profiles on social media and news articles) is captured and combined with user community information (users having the same perception about a news article) to form a 3-mode (content, context and user-community) tensor. A tensor is a multidimensional array that gives a higher dimensional generalization of matrices. The proposed technique is tested on real-world datasets, including BuzzFeed and Politifact. An ensemble machine learning classifier (XGBoost) and a deep neural network are employed for classification tasks. Results show the combined content and context approach gives better results. As future work, they propose real-time text-based classification of news articles by utilizing these content and context-based features.

Using Twitter as a source, the piece in Maddocks (2020) investigates the connection between sociopolitical and pornographic deepfakes and discovers that they both attempt to restrict dissenting expression. The authors contend that in order to address the injustices that cause this kind of technology to disproportionately harm women, policymakers should take into account the motivations behind the production and consumption of fake porn.

Because fake news is typically more dramatic than actual news, it spreads exponentially and more quickly. Retweets, also known as gossip path propagation hops, are more common in fake tweets. On the other hand, tweets about actual news typically grow slowly and steadily, reaching fewer people (Mjaaland 2020). The thesis in Mjaaland (2020) proposes a hybrid fake-news detection model that combines metadata with article content and rumor path propagation. These are represented as temporal patterns, used as inputs for a bidirectional LSTM network. Some other DL architectures are also implemented for comparison. The dataset comes from Politifact website.

To improve the identification of fake news, the authors in Sabeeh et al. (2020) suggest it is necessary to explore the interaction between the user and the news. In this regard, they say credibility analysis is essential to verify the trustworthiness of news to improve the detection accuracy. The comments of the users on social networks are the most reliable signals of the user intent. The authors propose a model to detect fake news that incorporates opinion mining on user comment, and credibility analysis of Twitter metadata. SentiWordNet is used in their approach to take into account the text's cognitive clues in order to help with opinion mining. In order to produce effective decisions, it also makes use of a bidirectional Gated RNN that incorporates objective criteria like emotion and a credibility score. As future work, they point the need to improve feature selection, and to consider also media content and specific writing styles for improving the detection.

Another example of fake news identification can be found in Sadr (2021), where a hybrid model of LSTM and bidirectional LSTM has been used on Persian texts and tweets. Word2Vec was employed for the embedding phase. Rumors are extracted from DataHeart database, upgraded and combined with own collected data.

An automatic approach embedded in Chrome web browser is presented in Sahoo and Gupta (2021), with the goal of detecting fake news on Facebook. Specifically, Sahoo et al. leverage on Facebook account features combined with news content to analyze the account behavior through LSTM. Other ML methods are also available in the add-on framework. As main limitations found on literature, authors from the latter article claim that further research on feature selection shall be conducted in order to reduce detection time. Moreover, online systems are needed for real-world scenarios.

It should be noted again the need for accurate, diverse and large enough datasets in accomplishing these tasks. Thus, further work on this should be conducted. Moreover, most of surveyed literature is focused on Twitter, because of its inherent characteristics as text sharing network. Therefore, additional research focused on other platforms, such as Whatsapp, Telegram, Instagram and Facebook, is still an open spot that shall be addressed (Table 5).

## 5.2 IoT networks

Wireless Sensor Networks (WSN) collect large volumes of data through the rollout of a vast number of self-organized agents, including sensors, actuators and computers among others. Furthermore, IoT provides interconnectivity within the different involved *things*, with the goal of intelligently monitoring and controlling them (Gambín et al. 2021). The Internet of Things (IoT) is an interconnected system of embedded systems that communicate via wired or wireless technologies. It is made up of physical items that are integrated with electronics (such sensing and actuation), software, and network connectivity that allow them to collect, occasionally process, and transmit data. These devices also have some limited calculation, storage, and communication capabilities. The term “things” refers to various items from our daily lives, including intelligent household gadgets and more advanced ones like RFID gadgets, pulse sensors, accelerometers, and every form of sensor (Hussain et al. 2020). As a result of the intricate nature of IoT systems, ensuring security is difficult. The majority of IoT devices operate in an uncontrolled, occasionally unpredictable environment where an attacker might physically control the device by listening in. IoT devices’ low computational and power capabilities prevent them from supporting elaborate security mechanisms. Additionally, new attack strategies might rapidly emerge because to the interdependence and connectivity between the Internet of Things (IoT) devices and the remaining components of the cyberphysical system (Al-Garadi et al. 2020). For all of these explanations, IoT systems need to go from simply enabling secure communication between devices to intelligence provided by DL approaches to create robust, all-encompassing security solutions, as noted by the al2020survey. Indeed, DL has demonstrated advantages over conventional signature-based, rule-based, and classic ML solutions (Berman et al. 2019). Several security aspects shall be considered in every IoT system (Al-Garadi et al. 2020):

*Integrity* The idea is to make sure that there is a reliable checking system in place to find any modifications made during communication across an unsecured wireless network. The

**Table 5** Summary of reviewed literature

| References | Scope | Highlights |
|---|---|---|
| Al-Garadi et al. (2020) | IoT security systems | ML apps for intrusion detection & DL for malware and anomalies detection |
| Asharf et al. (2020) | IoT security systems | Intrusion detection models; ML & DL detection tools; Open datasets |
| Hussain et al. (2020) | IoT security systems | Security requirements and solutions within IoT systems; main attack vectors |
| Ge et al. (2021) | IoT framework | Deep ANN-based intrusion detection system |
| Rathore et al. (2021) | IoT framework | 5 G-enabled solution combining DL and BC technology |

data saved in the IoT storage devices may be modified if there is a problem with integrity checks.

*Authentication* Before beginning any other step, entities/agents identity must be perfectly established. The nature of IoT systems, however, means that trade-offs are a significant challenge in creating an efficient scheme because authentication requirements vary from system to system.

*Authorization* It is used to describe giving consumers access to the IoT system. The major issue is figuring out how to provide access effectively in a situation where users might include not only actual people but also actual objects or services.

*Availability* IoT system services must always be accessible to authorized entities. IoT systems can still be deemed inoperable by a variety of threats, including active jamming and ac DoS. Therefore, maintaining continual availability is essential.

*Non-repudiation* It produces access logs that can be used as proof in circumstances where IoT users are unable to object to a certain behavior. For many IoT systems, non-repudiation is not seen as a critical security feature, although it can be in some circumstances, such as payment methods where a transaction cannot be revoked by either party.

These security properties can be threatened by numerous attacks, such as passive and active hazards. Deepfakes and deception falls within the active ones, where the two main potential deceptive attacks are the following: (i) impersonation (e.g., spoofing, man-in-the-middle) pretends to be/act as an authorized IoT device or user. Active intruders may try to completely or partially pass as an IoT entity if a malicious path exists; (ii) data manipulation is the act of purposefully altering (deleting or altering) information through unauthorized actions. The structure of the majority of attack detection systems is similar: (i) a data gathering module gathers data that may contain evidence of an attack; (ii) an analysis module identifies attacks after data processing; and (iii) an interface for reporting an attack.

The analysis module can be implemented using various methods, however, DL techniques are the most suitable and dominant due to its powerful features regarding data examination and pattern learning, including AD based on IoT devices interactions. Furthermore, DL methods are good at prediction of new attacks, which are often different from previous ones (Asharf et al. 2020). In this section, we focus on reviewing literature related to digital deception within the IoT ecosystem. Due to the wideness of the topic, we do not intend to elaborate a comprehensive survey in this section, but just to review relevant state of the art, mainly high-impact surveys, where key findings are highlighted for each reviewed contribution.

A comprehensive survey of Machine Learning (ML) methods and recent advances in DL used to develop enhanced security within IoT systems is presented in Al-Garadi et al. (2020). IoT security threats and attack surfaces are discussed. Among the potential surveyed applications, ML has been used mainly for intrusion detection, while DL for malware and anomalies detection. As main conclusions, the authors claim the need for diverse datasets within IoT security. The success of AI models depends merely on this. This is still an open issue due to the wide diversity of IoT devices in the ecosystem, as well as the privacy concerns related to critical information stored, such as industrial and medical data. Indeed, the heterogeneity present in IoT systems arises the need for multi-modal DL architectures, able to handle large-scale streaming, heterogeneous and high-noise data.

The authors in Asharf et al. (2020) provides a summary of intrusion detection methods and reviews IoT technologies, protocols, topologies, and hazards arising from compromised IoT devices. Besides, they analyze various ML and DL techniques suitable to detect cyberattacks. Several IoT security datasets are presented. Among the open challenges related to AD in IoT networks, they highlight that the security system may generate false alarms in order to improve the attack detection. Moreover, they claim completely avoiding or minimizing false-positive and false-negative is another research challenge.

Security requirements and solutions within IoT systems, together with attack vectors are discussed in Hussain et al. (2020). The authors highlighted the security solutions' shortcomings and the need for ML and DL methods. According to their research, DL models' theoretical underpinnings need to be reinforced in order to enable quantification of performance based on factors including computational complexity, learning effectiveness, and parameter adjustment techniques. For intuitive and effective data interpretation, innovative hybrid methods of learning and cutting-edge data visualization approaches will also be necessary. The authors in Ge et al. (2021) suggests an intrusion detection system based on DL. In order to distinguish between the traffic of these attack types and regular network traffic, the multi-class classifier uses a feed-forward ANN with embedding layers to recognize four categories of attacks: denial of service (DoS), DDoS, data gathering, and data theft. Additionally, a transfer learning-based approach is used to extract the encoding of high-dimensional categorical features, which are then used to a binary classifier. The authors consider as future work the use of GANs for data augmentation purposes, in order to generate synthetic data to carry additional experiments. Furthermore, they plan to improve the classifier to operate in real time, and to investigate feature ranking techniques for time-series feature-based classifiers.

Along the main insights from Sect. 4.2, a 5 G-enabled IoT security framework combining DL and BC technology is proposed in Rathore et al. (2021). The four levels of cloud, fog, edge, and user are described as their hierarchical architecture. To prove the framework's validity in real-world applications, it is assessed using a variety of common measurements of delay, accuracy, and security.

## 6 Open challenges

In this section, we distill the most relevant lessons learned throughout the reviewed literature and discuss open challenges. Our concluding remarks are presented in Sect. 6.1. Then, we elaborate on some research opportunities in Sect. 6.2, aiming at encouraging work in those together with the main agents and potential technologies involved.

### 6.1 Concluding remarks

DL represents a cutting-edge technology, that combined with Big Data, cloud computing, IoT and image processing is revolutionizing a vast number of fields. The automatic feature extraction is a major advantage, providing proven high performances in complex scenarios, in contrast with the traditional ML based on human expertise for feature engineering. On the other hand, DL requires more computing power and time, as well as the need for larger well-balanced input data. Besides, DL models rely on sample data and suffer from low interpretability, which translates into specific gained experience from the addressed dataset (Gambín et al. 2021).

One of the clearest conclusions we can extract after reviewing the literature, is the need for useful datasets. Data collection must be diverse and large enough in order to provide DL architectures with the right amount of information to learn from. In this way, output models will be able to adapt better to every possible scenario. Probably, one of the main reasons behind this lack of useful data is privacy implications. Regarding the observance of domestic and international legislation, there are unresolved questions, including General Data Protection Regulation (GDPR) (Commission 2016), especially when dealing with feasibility of data anonymization, and the ease of subject rights. In this sense, the development of mechanisms able to collect data and process it without the need of storing and/or analyzing critical information must be further investigated. Some related insights can be found in Zhou et al. (2008), where an evaluation of the available anonymization methods for releasing social network data is offered for privacy protection. Aligned with the need of data, performance evaluation within not ideal environments and contexts is mandatory to improve robustness in DL strategies. Therefore, generalization, scalability and robustness in DL schemes are still open challenges to be tackled. Further information can be found in Mayer and Jacobsen (2020), Wickramasinghe et al. (2018).

As for deepfakes, it can be concluded that, even being a neutral technology, malicious applications can be very harmful and disruptive in society. Therefore, techniques advocated for prevention, detection and detention of spreading are essential, where huge coordinated efforts from research community, governments and private institutions should be promoted. Moreover, the detection of digital deception content is not enough, due to the fast development on the generation side. Therefore, further research has to be carried out regarding authentication. Guaranteeing a trusted content origin source in combination with a tracked history of it are powerful tools to combat online fake content.

Regarding its spreading, we can state that online social networks are the cornerstones, where content related to politics, finance and porn are the main topics. Besides, the main misinformation spreading actors within social media networks are people with low education and/or strong affiliation to certain institutions or school of thought, easily manipulated (Ahmed 2021). As for disinformation spreading, the usual origin source are third-party companies specialized in social media, hired by a private corporate, public agency or government with the aim of inferring public opinion manipulation in order to attain certain financial and/or geopolitical benefits. In this sense, it seems to constitute a demagogic paradox, the fact that worldwide governments support and fund the fight against digital deception and disinformation spreading publicly, while they also leverage this technology behind the scenes in favor of their own interest.

As for its generation and detection, it has been and still is a hot-topic where the number of publications is massively increasing every year. However, some concerns arise when referring to detection research. These are related to the fact that most of the available research is focused on how to detect fake content and/or fake spreading accounts, i.e., spambot accounts that spread fake content, based on supervised learning techniques, that require a labeled dataset. And this is the crux of the question: the labeling process is not trivial. Specially when bot accounts are very sophisticated, it becomes almost impossible to distinguish between human and machine. Hence, this is generating a lot of research that can be biased by the available datasets used as ground-truth, and its labeling accuracy can be questioned. In this regard, interesting insights can be found in Rauchfleisch and Kaiser (2020). The authors analyze the fake bot classifier *Botometer* (Davis et al. 2016).

This classifier has been widely utilized in academic articles as a result of its successful introduction as a method to calculate the amount of bots in an identified set of accounts. Their findings demonstrate that Botometer ratings are inaccurate for estimating bots, particularly when used in an alternate language. They also demonstrate that Botometer's limits, even when applied very conservatively, are subject to fluctuation, which can result in both false positives and false negatives (i.e., the classification of humans as bots) and false negatives (i.e., the classification of bots as humans). The majority of social science studies employing the program will inadvertently count a significant portion of human consumers as bots and vice versa, which has direct implications for academic research. Moreover, the authors in Gallwitz and Kreil (2021) identifies critical theoretical shortcomings in social bot research. In peer-reviewed studies, numerous profiles that had been tallied or even identified as social bots were examined. Their findings indicate that they were able to locate even one social bot. They come to the conclusion that research purporting to look into the presence or impact of online bots have, in fact, just looked into false positives and byproducts of the inadequate detection techniques used.

In this way, the focus for long-term research has to be put on unsupervised, semi-supervised and reinforcement learning strategies, that do not depend on prior information and are able to adapt to the context. Anomaly detection and pattern recognition are thus the main potential actors in solving these issues. We continue the discussion of these topics in the following subsection. These tools could also be enhanced through descriptive digital forensic analysis (Rauchfleisch and Kaiser 2020).

Furthermore, future research has to focus on group behavior and user networks (Bild et al. 2015; Ediger et al. 2010). Some elements such as neighborhood properties, user account metadata, content trends and relationships among accounts sharing the same type of information are key for improving detection mechanisms. Another strategy to improve the detection would be to concentrate in a small portion of data. This can be referring to a specific area or country, language, or topic. In this sense, meaningful information can be obtained more easily, especially if network interactions and group characterization are expected to be analyzed.

Finally, vast research has been carried out over the Twitter platform, due to its inherent features, and with strong focus on text content analysis through Natural Language Processing (NLP) (Kanakaraj and Guddeti 2015). However, we believe there is a lack of research considering other major platform such as Telegram, Youtube and Instagram among others, where video and image play an essential role.

## 6.2 Future trends

### 6.2.1 Transfer learning & data augmentation

The integration and processing of the massive amount of data that is available nowadays from different sources poses an open challenge. Further research is needed to extract the optimal valuable information from the measured data. *Transfer learning* is a good candidate in this matter. In general, traditional ML/DL models are designed to solve specific problems, with the consequent drawback that they have to be rebuilt from scratch if the problem context changes. Transfer learning overcomes this by leveraging knowledge acquired for one task to solve related ones, even if the learning crosses domains. It is specially popular in DL due to the need of large datasets.

Together with transfer learning, new ways of obtaining additional data would be highly beneficial. *Data augmentation* is used to expand limited data by generating new samples from existing ones, i.e., synthetic data, and can be a powerful strategy to reduce overfitting and therefore improving the performance of DL models. It encompasses a suite of techniques that enhance the size and quality of training datasets (Shorten and Khoshgoftaar 2019). In this way, ANNs are incredibly powerful at mapping high-dimensional inputs into lower-dimensional representations, and thus several DL-based methods have been proposed for data augmentation. Feature space augmentation based on CNNs and AEs, adversarial training and especially GAN-based models are among the key techniques. Further information can be found in Shorten and Khoshgoftaar (2019) and Perez and Wang (2017).

Currently, mature literature is scarce on this matter. Some examples where deepfake detection performance is enhanced through transfer learning and data augmentation are the following. Representation learning and knowledge distillation paradigms are employed in Kim et al. (2021) to introduce a transfer learning-based feature representation model. The authors perform domain adaptation tasks on new deepfake datasets while minimizing losses regarding prior knowledge about deepfakes. Moreover, a CNN architecture combined with transfer learning for video fake detection is proposed in Suratkar et al. (2020a, b).

### 6.2.2 Explainable AI

DL architectures are complex systems, usually seen as black boxes. *Visualization tools* able to provide insights about what actually the system is doing and how the network is learning are still an open research opportunity (Ball et al. 2017). In this sense, *Explainable Artificial Intelligence (XAI)* was first mentioned in 2004 by Van Lent et al. (2004), to describe the ability of their system to explain the behavior of AI-controlled entities in simulation games application. It has surged as a new research arena that promotes the interpretability of the output performances in AI, and specially in DL, facilitating the understanding and efficient use of the information that this technology itself provides. Further information can be found in Došilović et al. (2018) and Adadi and Berrada (2018).

Current deepfake detection methods also fail to convince of its reliability. Since the fundamental issue revolves around earning the trust of human agents, the construction of interpretable and also easily explainable models is imperative. The authors in Malolan et al. (2020) propose a CNN-based deepfake detection framework tested on various XAI techniques, evaluating its applicability within real-life scenarios. The authors in Hall et al. (2020) discuss both practical and novel ideas for leveraging XAI to improve the efficacy of digital forensic analysis, usually employed in deepfake detection mechanisms.

### 6.2.3 Knowledge fusion

Combining data-driven models with theory-driven models is a potential approach for achieving models that learn as much as possible from data. The former are much more adaptable to data and adept at seeing hidden patterns, whilst the latter are simpler to understand. The idea of knowledge fusion where information from several fields of expertise can support one other to provide more insightful understanding. In this sense, DL approaches may be contenders. Dong et al. (2015) examine the applicability and constraints of several knowledge fusion approaches. Utilizing data fusion to find trustworthy information among several sources of studied data allows for effective decision-making. The work in Tolosana et al. (2020) implies that feature-level fusion approaches might offer a superior adaptability for deepfake detection in various contexts. In fact, they highlight several instances where various fake detection techniques are already based on the synthesis of various information sources, such as steroids and DL features, spatial and spectral characteristics, or other combinations thereof. There are two further intriguing fusion methods that use RGB, depth, and infrared data to identify physical facial attacks. Additionally, combining data from other sources, such as the text, keystrokes, or audio that accompany films when they are uploaded to social networks, could be very beneficial to enhance deception detectors.

### 6.2.4 Anomaly detection

The identification of observations that differ from the majority of the data and do not follow an expected behavior is known as Anomaly Detection. These anomalies are usually classified into two types: (i) outliers, point-wise data; and (ii) anomaly patterns, fractions of data such as certain trends and fluctuations, that provide more information than outliers (Gambín et al. 2021).

AD can provide powerful insights with respect to deepfake prevention and detection thanks to its inherent capability to recognize patterns. These can be used as prior information, essential to build early-warning systems. Some literature can be found in this respect. For instance, a pipeline to detect GAN specific traces left during the deepfake creation is proposed in Giudice et al. (2021). The authors in the latter article employ discrete cosine transforms to detect anomalies. Moreover, an unsupervised fingerprint classification module based on anomaly detection to identify GAN images is presented in Pu et al. (2020). However, prevention strategies based on the obtained anomalies are still not addressed. This entails a research opportunity to be assessed. In this sense, integrated security systems should be designed involving prevention, detection and forecasting, where early-warning systems can be powered by AD, and intelligent control by reinforcement learning.

### 6.2.5 Decision making & reinforcement learning

One of the DL advantages is the ability to automate and speed up processes, such as management and *decision-making*, reducing the need for human intervention. Few DL works related to deepfakes tackle decision-making strategies as main contributions, representing an excellent opportunity for further research. The aim is not just to detect them but also to take intelligent actions to combat them. *Reinforcement learning* powered by DL models are the perfect combination in this matter, empowering the systems with foresighted control. This looks like a promising future research line to be addressed.

According to Masood et al. (2021), deepfake detectors that are now in use generally rely on the fixed characteristics of current cyberattacks by utilizing ML approaches, such as unsupervised clustering and supervised classification methods, and as a result, they are less likely to detect unidentified deepfakes. So, on the detecting side, RL approaches might be crucial. Deep RL, which goes one step farther, has enormous potential for both deepfake detection and defending against antiforensic assaults on the detectors. Since RL can simulate an autonomous agent to perform consecutive actions ideally with little to no prior understanding of the environment, it might be applied to the development of algorithms to identify anti-forensic processing and the creation of deepfake attack detectors. A review of deep RL approaches developed for solving cyber-security problems can be found in Nguyen et al. (2019), including autonomous intrusion detection techniques and multiagent game theory simulations for defense strategies.

### 6.2.6 Edge computing

Due to the increasing number of data sources, frequency, type and volume, a centralized system handling all this input may not be an optimal solution regarding scalability and efficiency. The *Edge Computing* paradigm tries to solve this by virtualizing network functions and deploying them at the network edge (Shi and Dustdar 2016). In this way, content, computation and even some control are moved "closer" to the end users. This entails some advantages such as low latency, energy and bandwidth efficiency, privacy protection, and context awareness (Zhou et al. 2019).

*Edge Intelligence (EI)* Zhou et al. (2019) proposes a new paradigm combining AI and edge computing, with the goal of performing distributed computing of DL models. This technology could be used to improve DL computing time, reducing drastically the training phase within the algorithms development, among other aforementioned benefits. Some examples are in the following. Authors in Ferrag and Maglaras (2019) present an energy framework for smart grids, combining BC technology and EI. It provides a peer-to-peer energy trading system, that is complemented with an intrusion detection block based on RNNs. The work in Hasanaj et al. (2021) proposes a solution to train deepfake detection models cooperatively on the edge, with the goal of evaluating time-computing efficiency. Finally, a memory-efficient DL-based deepfake detection method deployed in the IoT is explained in Mitra et al. (2021). Their aim is to detect highly sophisticated GAN generated deepfake images at the edge, reducing training and inference time while achieving a certain accuracy.

### 6.2.7 Blockchain technology and AI

A remarkable symbiosis between DL and BC technology is revealed, capable of producing a fully effective security system. In order to effectively use BC for increased trust and security services, DL may first help BC technology realize intelligent decision-making, improved evaluation, filtering, and understanding of data and devices within a network. Second, since its built-in decentralized database emphasizes the significance of data dispersion among numerous nodes on a particular network, BC may help AI by offering a big volume of data. Therefore, this a powerful tool chain still on its first steps. Further research has to be accomplished where numerous opportunities are still to be evaluated.

### 6.2.8 Real-time systems

The ultimate goal in order to combat deepfakes is to develop real-time frameworks. Due to the complexity of the challenge, regarding training times and efficiency issues, it is still an open issue of the topic. These online systems should leverage AD-based early-warning forecasting blocks to prevent and predict deepfakes, on DL architectures to detect digital deception, and on RL-based decision making to stop the spreading. This would provide a complex cybersecurity platform, where its adoption within social networks and Internet applications would be the ideal integrated solution.

## 7 Conclusions

Advances in Deep Learning, Big Data and image processing have facilitated online disinformation spreading through Deepfakes. This entails severe threats including public opinion manipulation, geopolitical tensions, chaos in financial markets, scams, defamation and identity theft among others. Therefore, it is imperative to develop techniques to prevent, detect, and stop the spreading of deepfake content. In this paper, we have conducted a review targeting the entire deepfake paradigm, by reviewing current and future trends. First, a compact summary of DL techniques used for deepfakes has been presented. Then, a review of the fight between generation and detection techniques has been elaborated. Moreover, we have discussed about the potential that new technologies, such as distributed ledgers and blockchain, can offer with regard to cybersecurity and the fight against digital deception. Two scenarios of application, including online social networks engineering attacks and Internet of Things, have been reviewed providing main insights and open challenges. Finally, future trends and research lines have been examined, mentioning potential key agents and technologies.

**Author contributions** AFG: Reading papers, paper analysis, and review organization. AY: Reading papers, paper analysis, and review organization. AV: Reading papers, paper analysis, and review organization. HH: Reading papers, paper analysis, and review organization. YD: Reading papers, paper analysis, and review organization.

**Funding** Open access funding provided by University Of South-Eastern Norway.

## Declarations

**Competing interests** The authors declare no competing interests .

## References

Abba (2021) Abba Voyage. https://abbavoyage.com/

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160

Ahmed S (2021) Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. Telematics Inform 57:101508

Al-Garadi MA, Mohamed A, Al-Ali AK, Du X, Ali I, Guizani M (2020) A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun Surv Tutor 22(3):1646–1685

Almars AM (2021) Deepfakes detection techniques using deep learning: a survey. J Comput Commun 9(5):20–35

Amerini I, Anagnostopoulos A, Maiano L, Celsi LR et al (2021) Deep learning for multimedia forensics. Found Trends Comput Graphics Vis 12(4):309–457

Arora A, Shantanu (2020) A review on application of GANs in cybersecurity domain. IETE Techn Rev 1–9

Asharf J, Moustafa N, Khurshid H, Debie E, Haider W, Wahab A (2020) A review of intrusion detection systems using machine and deep learning in internet of things: challenges, solutions and future directions. Electronics 9(7):1177

Ball JE, Anderson DT, Chan CS (2017) Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. J Appl Remote Sens 11(4):042609

Berman DS, Buczak AL, Chavis JS, Corbett CL (2019) A survey of deep learning methods for cyber security. Information 10(4):122

Bild DR, Liu Y, Dick RP, Mao ZM, Wallach DS (2015) Aggregate characterization of user behavior in twitter and analysis of the retweet graph. ACM Trans Internet Technol 15(1):1–24

Caporusso N (2020) Deepfakes for the good: a beneficial application of contentious artificial intelligence technology. In: International conference on applied human factors and ergonomics. Springer, Orlando, pp 235–241

Chan CCK, Kumar V, Delaney S, Gochoo M (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). Geneva, IEEE, pp 55–62

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078

Chung Y-A, Wu C-C, Shen C-H, Lee H-Y, Lee L-S (2016) Audio word2vec: unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. arXiv preprint arXiv:1603.00982

Commission (2016) GDPR. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web, New York, USA, pp 273–274

DeepMind (2021) DeepMind- AlphaStar. https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

Deshmukh A, Wankhade SB (2020) Deepfake detection approaches using deep learning: a systematic review. Intell Comput Netw 146:293–302

Dong XL, Gabrilovich E, Heitz G, Horn W, Murphy K, Sun S, Zhang W (2015) From data fusion to knowledge fusion. arXiv preprint arXiv:1503.00302

Došilović FK, Brčić M, Hlupić N (2018) Explainable artificial intelligence: a survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). Opatija, Croatia. IEEE, pp 0210–0215

Ediger D, Jiang K, Riedy J, Bader DA, Corley C, Farber R, N W (2010) Reynolds, Massive social network analysis: mining twitter for social good. In: 2010 39th international conference on parallel processing. San Diego, CA, USA. IEEE, pp 583–593

FaceApp (2021) FaceApp. https://www.faceapp.com/

Facebrity (2021) Facebrity. https://apps.apple.com/us/app/facebrity-face-swap-morph-app/id1449734851

Fagni T, Falchi F, Gambini M, Martella A, Tesconi M (2021) Tweepfake: about detecting deepfake tweets. PLoS ONE 16(5):e0251415

Ferrag MA, Maglaras L (2019) DeepCoin: a novel deep learning and blockchain-based energy exchange framework for smart grids. IEEE Trans Eng Manag 67(4):1285–1297

Fraga-Lamas P, Fernández-Caramés TM (2020) Fake news, disinformation, and deepfakes: leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. IT Prof 22(2):53–59

Gallwitz F, Kreil M (2021) The rise and fall of 'social bot'research. SSRN: https://ssrn.com/abstract, vol 3814191

Gambín AF, Angelats E, González JS, Miozzo M, Dini P (2021) Sustainable marine ecosystems: deep learning for water quality assessment and forecasting,. IEEE Access

Ge M, Syed NF, Fu X, Baig Z, Robles-Kelly A (2021) Towards a deep learning-driven intrusion detection approach for Internet of Things. Comput Netw 186:107784

Girgis S, Amer E, Gadallah M (2018) Deep learning algorithms for detecting fake news in online text. In: 2018 13th international conference on computer engineering and systems (ICCES). Cairo, Egypt. IEEE, pp 93–97

Giudice O, Guarnera L, Battiato S (2021) Fighting deepfakes by detecting GAN DCT anomalies. arXiv preprint arXiv:2101.09781

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv preprint arXiv:1406.2661

Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning. MIT press, Cambridge, p 2

Hall SW, Sakzad A, Choo K-KR (2020) Explainable artificial intelligence for digital forensics. Wiley, New York, p e1434

Hasan HR, Salah K (2019) Combating deepfake videos using blockchain and smart contracts. IEEE Access 7:41596–41606

Hasanaj E, Aveler A, Söder W (2021) Cooperative edge deepfake detection. Master's thesis, Jönköping University, School of Engineering

Heo Y-J, Choi Y-J, Lee Y-W, Kim B-G (2021) Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hussain F, Hussain R, Hassan SA, Hossain E (2020) Machine learning in IoT security: current solutions and future challenges. IEEE Commun Surv Tutor 22(3):1686–1721

Juefei-Xu F, Wang R, Huang Y, Guo Q, Ma L, Liu Y (2021) Countering malicious deepfakes: survey, battleground, and horizon. arXiv preprint arXiv:2103.00218

Kaliyar RK, Goswami A, Narang P (2020) DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. J Supercomput 77(2):1015–1037

Kanakaraj M, Guddeti RMR (2015) Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In: Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015). Anaheim, CA, USA. IEEE, pp 169–170

Khan SA, Dai H (2021) Video transformer for deepfake detection with incremental learning. In: Proceedings of the 29th ACM international conference on multimedia. Lisbon, Portugal, pp 1821–1828

Kim M, Tariq S, Woo SS (2021) FReTAL: generalizing deepfake detection using knowledge distillation and representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1001–1012

Krombholz K, Hobel H, Huber M, Weippl E (2015) Advanced social engineering attacks. J Inf Secur Appl 22:113–122

Kwok AO, Koh SG (2021) Deepfake: a social construction of technology perspective. Curr Issue Tour 24(13):1798–1802

Langguth J, Pogorelov K, Brenner S, Filkuková P, Schroeder DT (2021) Don't trust your eyes: image manipulation in the age of deepfakes. Front Commun 6:26

Maddocks S (2020) A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. Porn Stud 7(4):415–423

Malolan B, Parekh A, Kazi F (2020) Explainable deep-fake detection using visual interpretability methods. In: 2020 3rd international conference on information and computer technologies (ICICT). San Jose, CA, USA. IEEE, pp 289–293

Masood M, Nawaz M, Malik KM, Javed A, Irtaza A (2021) Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. arXiv preprint arXiv:2103.00484

Mayer R, Jacobsen H-A (2020) Scalable deep learning on distributed infrastructures: challenges, techniques, and tools. ACM Comput Surv 53(1):1–37

Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Comput Surv 54(1):1–41

Mitra A, Mohanty SP, Corcoran P, Kougianos E (2021) EasyDeep: an IoT friendly robust detection method for GAN generated deepfake images in social media

Mjaaland H (2020) Detecting fake news and rumors in twitter using deep neural networks. Master's thesis, University of Stavanger, Norway

Nan L, Tao D (2018) Bitcoin mixing detection using deep autoencoder. In: 2018 IEEE Third international conference on data science in cyberspace (DSC). Guangzhou, China. IEEE, pp 280–287

Navidan H, Moshiri PF, Nabati M, Shahbazian R, Ghorashi SA, Shah-Mansouri V, Windridge D (2021) Generative Adversarial Networks (GANs) in networking: a comprehensive survey & evaluation. Comput Netw 194:108149

Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S (2019) Deep learning for deepfakes creation and detection: a survey. arXiv preprint arXiv:1909.11573

Niu Z, Zhong G, Yu H (2021) A review on the attention mechanism of deep learning. Neurocomputing 452:48–62

Nofer M, Gomber P, Hinz O, Schiereck D (2017) Blockchain. Bus Inf Syst Eng 59(3):183–187

OpenAI (2021) OpenAI. https://openai.com/blog/better-language-models/

Patel V, Pan L, Rajasegarar S (2020) Graph deep learning based anomaly detection in ethereum blockchain network. In: International conference on network and system security. Springer, Melbourne, pp 132–148

Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621

Pérez Dasilva JÁ, Meso Ayerdi K, Mendiguren Galdospin T (2021) Deepfakes on Twitter: which actors control their spread? Med Commun 9(1):301–312

Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A, Carin L (2016) Variational autoencoder for deep learning of images, labels and captions. arXiv preprint arXiv:1609.08976

Pu J, Mangaokar N, Wang B, Reddy CK, Viswanath B (2020) Noisescope: detecting deepfake images in a blind setting. In: Annual computer security applications conference, Austin, USA, pp 913–927

Rathore S, Park JH, Chang H (2021) Deep learning and blockchain-empowered security framework for intelligent 5G-enabled IoT. IEEE Access 9:90075–90083

Rauchfleisch A, Kaiser J (2020) The false positive problem of automatic bot detection in social science research. PLoS ONE 15(10):e0241045

Ryabchuk N (2020) Artificial intelligence technologies using in social engineering attacks. In: CEUR workshop proceedings, vol 2654. Kiev, Ukraine, pp 546–555

Sabeeh V, Zohdy M, Mollah A, Al Bashaireh R (2020) Fake news detection on social media using deep learning and semantic knowledge sources. Int J Comput Sci Inf Secur 18(2)

Sadr MM et al (2021) The use of LSTM neural network to detect fake news on persian Twitter. Tur J Comput Math Educ 12(11):6658–6668

Sahoo SR, Gupta BB (2021) Multiple features based approach for automatic fake news detection on social networks using deep learning. Appl Soft Comput 100:106983

Salahdine F, Kaabouch N (2019) Social engineering attacks: a survey. Future Internet 11(4):89

Shelke NA, Kasana SS (2020) A comprehensive survey on passive techniques for digital video forgery detection. Multimed Tools Appl 80(4):6247–6310

Shi W, Dustdar S (2016) The promise of edge computing. Computer 49(5):78–81

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):1–48

Sit M, Demiray BZ, Xiang Z, Ewing GJ, Sermet Y, Demir I (2020) A comprehensive review of deep learning applications in hydrology and water resources. Water Sci Technol 82(12):2635–2670

Suratkar S, Johnson E, Variyambat K, Panchal M, Kazi F (2020a) Employing Transfer-Learning based CNN architectures to enhance the generalizability of deepfake detection. In: 2020 11th international conference on computing, communication and networking technologies (ICCCNT).minus Kharagpur, India. IEEE, pp 1–9

Suratkar S, Kazi F, Sakhalkar M, Abhyankar N, Kshirsagar M (2020b) Exposing deepfakes using convolutional neural networks and transfer learning approaches. In: 2020 IEEE 17th India council international conference (INDICON). New Delhi, India. IEEE, pp 1–8

Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148

Trinh HD, Gambin AF, Giupponi L, Rossi M, Dini P (2020) Mobile traffic classification through physical control channel fingerprinting: a deep learning approach. IEEE Trans Netw Serv Manag 18:1946–1961

Van Lent M, Fisher W, Mancuso M (2004) An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the national conference on artificial intelligence. minus. Menlo Park, San Jose; MIT Press, Cambridge; AAAI Press, London, pp 900–907

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, Long Beach, CA, USA, pp 5998–6008

Wang T-C, Mallya A, Liu M-Y (2021a) One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10039–10049

Wang J, Wu Z, Chen J, Jiang Y-G (2021b) M2TR: multi-modal multi-scale transformers for deepfake detection. arXiv preprint arXiv:2104.09770

Westerlund M (2019) The emergence of deepfake technology: a review. Technol Innov Manag Rev 9:11

Wickramasinghe CS, Marino DL, Amarasinghe K, Manic M (2018) Generalization of deep learning for cyber-physical system security: a survey. In: IECON 2018-44th annual conference of the IEEE industrial electronics society. Washington, DC, USA. IEEE, pp 745–751

Wodajo D, Atnafu S (2021) Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126

Yazdinejad A, HaddadPajouh H, Dehghantanha A, Parizi RM, Srivastava G, Chen M-Y (2020a) Cryptocurrency malware hunting: a deep recurrent neural network approach. Appl Soft Comput 96:106630

Yazdinejad A, Parizi RM, Srivastava G, Dehghantanha A (2020b) Making sense of blockchain for AI deepfakes technology. In: 2020 IEEE globecom workshops (GC Wkshps). Taipei, Taiwan. IEEE, pp 1–6

Yu P, Xia Z, Fei J, Lu Y (2021) A survey on deepfake video detection. IET Biometrics 1–18

Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 53(5):1–40

Zhou B, Pei J, Luk W (2008) A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explor Newsl 10(2):12–22

Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J (2019) Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE 107(8):1738–1762