Deepfakes: Current and Future Trends

Ángel Fernández Gambín, Anis Yazidi, Athanasios Vasilakos, Hårek Haugerud

***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.**

***Index Terms*—Artificial Intelligence, Deep Learning, Deepfake, Digital Deception, Blockchain, GAN**

I. Impact Statement

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 just 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.

II. 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 and media content platforms, has entirely changed the information paradigm and increased massively the amount of online data. The availability of affordable digital devices, including smart phones, tablets, laptops, and digital cameras has resulted in the exponential growth of multimedia content in cyberspace. Additionally, the evolution of social media over the last decade has allowed people to share captured multimedia content rapidly, leading to a significant increase in content generation and ease of access to it [1].

In this context, where information is rapidly spread worldwide, it has become increasingly difficult to know the truth and trust the information, which may result in extremely harmful consequences. Indeed, reports indicate that the human ability to detect deception without special assistance is only 54% [2]. Today we live in a "post-truth" age, where disinformation is utilized by malicious actors to manipulate public opinion. This is known as *fake news*, and constitutes one of the greatest threats to democracy, journalism, and freedom of expression nowadays [3]. Disinformation can cause severe damage: election manipulation, creation of warmongering situations, defaming any person, etc [1]. The majority of individuals in developed economies will consume more false than true information by 2022 [4]. 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 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 [4].

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

Angel Fernández Gambín, Anis Yazidi and Hårek Haugerud are with the Department of Computer Science, Oslo Metropolitan University, Oslo, Norway. Athanasios Vasilakos is with the Department of Computer Science, University of Agder, Grimstad, Norway.
Corresponding author: angelfer@oslomet.no

into the following categories: (i) entire face synthesis, (ii) attribute manipulation, (iii) identity swap, and (iv) expression swap (i.e., reenactment) [5].

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 [6]. It has been used to create digital avatars or virtual assistance to improve the quality of experience in video conferencing [7]. For instance, the authors in [8] leverage deepfake algorithms to extract an accurate model of an individual and generate new content especially designed for benign use. Specifically, they create an interactive Digital Twin of a subject that can serve as a replacement for in-person or virtual presence. The purpose of the proposed application is to provide users with easy-to-use tools that enable them to produce their own digital replica for future use, so that it can be featured in re-enactments, interactive stories, memorials, and simulations. Another example is the virtual concert that the mythical band ABBA is preparing for 2022, that will feature digital versions of the band members [9]. Moreover, it has been used for creating facial visual effects in movie and TV show production in order to re-create a role appearance for some celebrities that may have passed away, or for paying tribute to the lost ones in a memorial concert. Besides, it has gained popularity in smartphone applications for entertainment purposes, especially targeted for making viral videos on social media platforms [10] [11]. Another case can be found in [12], where potential benefits in the tourism industry and related marketing are addressed.

However, the malicious uses largely dominate the positive ones. Deepfakes can be used to propagate online fake news, which can entail severe threats as aforementioned, cause political or religious tensions between countries, fool the public, create chaos in financial markets, sabotage, fraud, scams, obstruction of justice, and potentially many more. For instance, they can be even used to generate fake satellite earth images for military purposes [13]. And the largest concern is that it provides any user, with technology know-how, the ability to create videos that undermine the truth, combined thus with the advent of social networks, the proliferation of such content might be unstoppable [14].

Because of this, it is imperative to develop techniques to prevent, detect, and stop the spreading of deepfake content. Deepfakes should be combated through: (i) legislation and regulation [15], (ii) corporate policies and voluntary action, (iii) education, and (iv) countermeasures technology [6]. This entails a challenging task even if there is a credible, secure, and trusted way to trace the history of digital content. In this regard, the research community, big tech corporates and governments are focusing their efforts on launching proposals and regulations to stop digital deception.

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. 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.

The rest of the manuscript is organized as follows. A DL general overview is provided in Section III. A discussion about the battleground between generation and detection is presented in Section IV. Leveraging blockchain technology as a way to guarantee digital content authentication is addressed in Section V. In Section VI, scenarios of deepfake application are presented, focusing on social and communication networks. Future trends and potential challenges are proposed in Section VII. Finally, Section VIII summarizes our conclusions.

III. Deep Learning Outline

Machine Learning has revolutionized the way of understanding data, creating endless opportunities. Its aim is to give machines the ability to learn without being strictly programmed [21]. As an important breakthrough in Machine Learning (ML), Deep Learning has witnessed a strong burst into its application domains, including computer vision, speech recognition, and natural language processing among many other fields. In simple words, a DL model is learning automatically the features and decision making in contrast to a classical human-crafted ML system, thanks to its multiple-level representation.

The goal of this section is to provide the reader with a DL outline to better understand the rest of the paper. In this sense, an overview about main DL models used in the generation, detection and prevention of deepfakes is presented in next section.

Reference	Generation	Detection	Authentication	Scenarios of Application
[16]	x	✓	x	x
[17]	✓	✓	x	x
[5]	✓	✓	x	x
[18]	✓	✓	x	x
[19]	x	✓	x	x
[20]	x	x	✓	x
[20]	x	x	✓	x
This work	✓	✓	✓	✓

TABLE I: State of the art comparison.

A. 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, 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.

1) Artificial Neural Networks: The structure of a feed-forward (the input goes only one way within the network) Artificial Neural Network (ANN) is an input layer, some intermediate hidden layers, and an output layer, able to learn linear and nonlinear relationships between input and output pairs, through extracted features. Each layer comprises at least one neuron. These neurons run specific *activation* functions and are connected to each other with some weights, mapping its input to an output. Every neuron within a layer usually runs the same activation function, defining the layer type. The combination of used layers and the structure of the network, i.e., how neurons are inter-connected, defines the network type [22]. In order to find the weights for each neuron that minimize a certain error objective function, a common training procedure is the Backpropagation (BP) algorithm [23].

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 [23]. 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 [24].

3) Recurrent Neural Networks: A RNN is a recursive ANN, storing information within the network. Neurons within a recurrent layer can also be connected to each other, i.e., the output of a neuron is connected both to the next neuron within the same layer and the neuron(s) of next layer [24]. A specific type of RNNs is LSTM networks. The LSTM neurons are called Memory Cells (MCs). A MC is able to store information about past network states by using *gates*. A gate consists of a neuron with sigmoid activation function and a multiplication block. Thanks to this structure, the MC output relies on the sequence of past states, making LSTMs suitable for processing time series with long-term dependencies [25]. Gate Recurrent Unit (GRU) is another RNN model. A GRU cell is composed of a reset gate and an update one. The first is used to decide how much past information to forget. The latter decides what new information to add in every iteration, helping the model to determine how much of the past information needs to be utilized in the future [26]. RNNs are good at handling temporal and predictive problems.

4) Autoencoders: An Autoencoder (AE) is an unsupervised ANN trained to reproduce its input to its output [27]. It is composed of an encoder and a decoder. Each part contains some hidden layers. The encoder transforms the input into a feature-based representation, reducing the data dimensionality. Subsequently, the decoder tries to rebuild the original input from that representation. The training process should be accomplished by minimizing the reconstruction error, while prioritizing which characteristics of the inputs should be learned. The 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 [18]. CNNs and RNNs can be used as AEs, increasing the model complexity to solve certain problems [28] and [29].

5) General Adversarial Networks: The concept of a General Adversarial Network was first introduced in 2014 [30], inspired by the zero-sum game from game theory. A GAN consists of two (deep) ANNs pitting one against the other: a

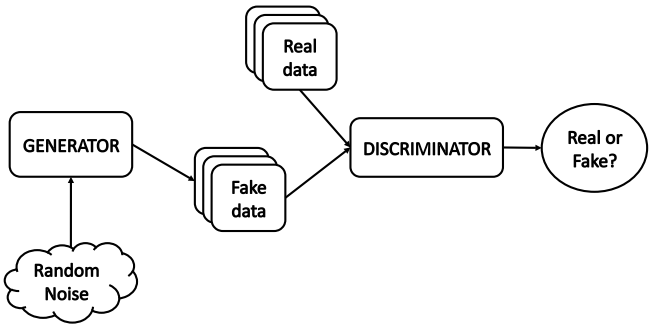

Fig. 1: GAN architecture.

generator and a discriminator, learning at the same time. The GAN optimization process tries to reach Nash equilibrium, where both generative and discriminative models work as adversaries. The generator attempts deception through the generation of samples using random noise. The discriminator, usually a binary classifier, attempts to authenticate real training data samples from deceptive samples generated by the generative model. A graphical explanation can be found in Fig. 1. They work well with images, videos and voice generation. Numerous variations and improvements of GANs have been proposed over the years. Regarding deepfakes, two popular GAN-based image translation frameworks are pix2pix and CycleGAN [18].

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 [31] and [32]. 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 [32] discuss about how GANs can benefit multiple aspects of computer and communication networks, including mobile networks, IoT, and cybersecurity.

6) Transformers: Transformers are a type of ANN first introduced in 2017 by A. Vaswani et al. [33]. They were developed to solve the problem of sequence transduction, i.e., any task that transforms an input sequence to an output one. This includes speech recognition, text-to-speech transformation, etc. They have been gaining popularity within natural language processing field. For instance, OpenAI used them in their language models [34] and DeepMind for AlphaStar [35].

For models to perform sequence transduction, it is necessary to understand dependencies and connections within a certain input, e.g., a sentence. RNNs and CNNs have been used to deal with this because of their inherent properties. However, they present some drawbacks. RNNs do not perform well when input sentences are too long, because the probability of keeping the context from a word that is far away from the current word being processed decreases exponentially with the distance from it. Moreover, CNNs are not good at capturing dependencies. To solve these issues, attention models were developed, which their focus is on a subset of the given input. The idea behind is that there might be relevant information in every word in a sentence. So in order for the decoding to be precise, it needs to take into account every word of the input, using attention.

Transformers usually combine CNNs with attention models, to provide parallelization. Attention boosts the speed of how fast the model can translate from one sequence to another. A transformer consists of two parts: encoder and decoder. Both are composed of modules that can be stacked on top of each other multiple times. The modules consist mainly of attention and feed-forward layers [33].

In this way, transformers present a demonstrated performance in modeling dependencies between pixels for a variety of recognition tasks in computer vision and therefore are good candidates for combating deepfakes. Authors in [36] propose a multi-scale transformer that detects local inconsistencies at different spatial levels. To improve the detection results and enhance the robustness of their method to image compression, frequency information is combined with RGB features. A video transformer with incremental learning is employed in [37]. Another model considering a video transformer is presented in [38]. They design a distillation methodology where a patch-based positioning CNN model learns to interact with all positions to find the artifact region for solving false negative problem. Finally, the joint model based on CNN and vision transformer is discussed in [39]. The CNN extracts features while the transformer categorizes them using an attention mechanism [40].

IV. 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 Section IV-A. Moreover, a summary of the reviewed surveys is provided in Table II with key ideas and scope. Then, main insights regarding the reviewed literature are distilled in Section IV-B, where we elaborate on both generation and detection sides.

A. State of the Art

A thorough review is presented in [16] with a focus on deepfake video detection, especially, generation process, several detection methods and existing benchmarks. According to the goal behind facial image manipulation, algorithms are divided into two categories: face swapping and face reenactment. Depending on the technique, the authors provide the following classification: general ANN-based, temporal consistency features, visual artifacts, camera fingerprints and biological signals. According to the main outcomes of the latter study, current detection methods are still not ready yet to be applied in real-world scenes, and further research should pay more attention to the generalization and robustness.

Deepfake creation and detection techniques using DL are surveyed in [17]. Regarding detection, the authors distinguish between whether the content is images or videos. Within video detection, two categories are proposed: biological signals analysis, and spatio-temporal features analysis. Besides, they provide access to several public datasets. Their main conclusion is that current DL methods are facing scalability issues, and thus more robust models are needed to be applied to any large and high-quality dataset.

A similar idea is pursued in [5] where a comprehensive overview and detailed analysis on deepfake generation and detection is presented. The authors discuss the taxonomy of various generation methods and the categorization of several detection models, putting emphasis on the battleground between the two sides. Interesting interactive diagrams are provided as add-ons for further exploration. Regarding generation methods, four categories are presented: entire face synthesis, attribute manipulation, identity swap, and expression swap. As for detection, they classify the works based on spatial features, frequency features, and biological signals. Moreover, the concept of evasion of deepfake detection is introduced, i.e., techniques to evade the fake faces being detected. Three types are discussed: adversarial attack, removing the fake traces in the frequency domain and use of advanced image filtering or generative models.

Following the trend, the purpose of [18] is to provide a deeper understanding of how deepfakes are created and detected, map the shortcomings of the current defense solutions, and several areas that require further research. As countermeasures for deepfakes generation, they delve into prevention and mitigation. To prevent deepfakes, data provenance should be tracked through distributed ledgers. We will discuss more about this in Section V. Within the context of combating deepfakes, they review some works dealing with adversarial ML as a way to disrupt and corrupt deepfake networks. Finally, the authors claim that deepfakes extend beyond human visuals, and have spread to many other domains, including healthcare, social media and finances among others, demonstrating that deepfakes are not just attack tools for misinformation, defamation, and propaganda, but also sabotage, fraud, scams, obstruction of justice, and potentially many more.

Another extensive survey is presented in [13] and [41]. In this work, deepfake detection methods are grouped into two major categories: image and video. The latter is distinguished into two smaller groups: visual artifacts within single video frame-based methods and temporal features across frames-based ones. Whilst most of the methods based on temporal features use deep recurrent classification models, visual artifacts methods are usually implemented by either deep or shallow classifiers. Their main conclusions are that detection methods shall be integrated with social media platforms, where distributed ledger technology can be a viable solution, as well as, explainable AI has to be promoted to facilitate the understanding and efficient use of the information that this technology itself provides.

The work in [42] reviews techniques for manipulating face images including deepfake methods, and tools to detect such manipulations. Four types of facial manipulation are reviewed: i) entire face synthesis, ii) identity swap, iii) attribute manipulation, and iv) expression swap. For each group, details regarding techniques, existing databases, and key benchmarks are discussed. Their concluding remarks include the need for further research on the generalization ability of the fake detectors against unseen conditions, where a viable solution could be architectures that do not require fake videos for training. Fusion techniques, at a feature or score level, could provide a better adaptation of the fake detectors to the different scenarios. In addition, novel schemes, not only based on image/video information, should be studied in order to provide more robust tools.

An analysis of existing tools and ML-based approaches for generation and detection of both audio and video deepfakes is presented in [1]. For each category of deepfake, the authors of the latter paper discuss information related to manipulation approaches, public datasets, and performance evaluation of deepfake detection techniques along with their results.

Video forgery detection using passive techniques is reviewed by Shelke et al. [19]. They survey the existing literature based on the features, forgery identified, used datasets, and performance parameters. Although their main scope is not DL, their survey can be an interesting reference to the reader for further insights regarding video counterfeiting. Furthermore, anti-forensics strategies, i.e., to deceive forensic investigation by removing or hiding traces left after the forgery, are also discussed.

The authors in [43] examine the manipulations of images and videos produced with editing tools, reporting DL approaches adopted to counter these attacks. Next, they analyze issues related to source camera model and device identification, as well as monitoring image and video sharing on social media.

B. 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 [1]. 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.

Reference	Scope	Media content	Featured sections
[16]	Detection	Video	Face swapping, face reenactment
[17]	Generation and Detection	Video, Image	Biological signals, spatio-temporal features
[5]	Generation and Detection	Video, Image, Audio	Face synthesis, attribute manipulation, identity and expression swap
[18]	Generation and Detection	Video, Image, Audio	Reenactment, replacement, editing, and synthesis
[13], [41]	Generation and Detection	Video, Image	Visual artifacts and temporal features
[42]	Generation and Detection	Image	Face synthesis, attribute manipulation, identity and expression swap
[1]	Generation and Detection	Video, Audio	Face swapping, lip syncing, face reenactment, face synthesis, attribute editing
[19]	Detection	Video	Compression and noise artifacts, motion and statistical features
[43]	Generation and Detection	Video, Image	Multiple compression, anomaly-based architectures, device and social media identification

TABLE II: Summary of reviewed surveys.

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 [5], 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 [1].

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 [16] indicate generalization performance of existing detection algorithms is still insufficient and an urgent problem to be addressed.

Datasets: there is a need of public 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 [5] 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 specially 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 [1].

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 [16].

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 [16], 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 [1].

V. Authentication

Although advances in combating deepfakes are improving, current solutions are limited. As we discussed in Section IV, there are huge efforts on how to deal with malicious deepfake applications, from research community to big technological corporates. The key problem is that the better the defense, the smarter the offense [14], and detecting, fact-checking and

disproving deepfakes on real time is posing an enormous challenge given the speed of technological development and content volume uploaded every day. Nowadays, there are no established methods for checking the originality of an online published digital media content. It is extremely difficult to determine in a trusted way the true origin of a posted digital item [14]. In this way, most research is focused on the development of AI-based detection techniques, but however, there is one missing aspect, *authentication*. Instead of attempting to detect what content is fake, techniques to provide tamper-proof evidence of what content is real are a powerful solution. Therefore, there exists a need for a Proof of Authenticity (PoA) system regarding online digital content to identify trusted published sources.

Distributed ledgers, and specifically Blockchain (BC), present excellent opportunities as potential technologies that can help to combat digital deception. They enable privacy, security, and trust in a decentralized peer-to-peer network without any central managing authority [4]. Blockchain is an emerging technology that uses cryptography to secure transactions within a network. A blockchain delivers a decentralized database (known as *digital ledger*) of transactions, of which each node on the network is aware [44]. The network is a chain of devices (e.g., computers) that all need to endorse a transaction before it can be verified and recorded. Moreover, transactions are easily auditable by all the involved stakeholders. Namely, a BC is simply a data structure that allows the production and distribution of a "tamper-proof digital ledger" of exchanges. Therefore, BC systems can render transactions relatively more secure and transparent than those in centralized systems. BC ability to combat digital deception is focused on controlling the traceability of the media, the communications architecture, and the transactions. However, problems involved in developing effective ways to identify, test, transmit, and audit information are still open [4].

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 Section V-A 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 Section V-B.

A. State of the Art

An overview of techniques to provide tamper-proof evidence of what content is real is presented in [14]. They discuss potential use cases and solutions to tackle deepfakes via BC functionalities and features. The authors claim that current research in BC authentication of digital content for the deepfakes is still in infancy and provide interesting future research lanes.

The same goal is pursued in [4]. The authors explore the potential of BC to combat digital deception, describing the most relevant applications and identifying their main open challenges. Among the most promising solutions figure: (i) decentralized content moderation; (ii) fact-checking incentivized applications, where reliable fact-checkers can validate content for financial rewards (e.g., tokens), while the received rewards increase as the fact-checker improves its reputation; and (iii) decentralized social media platforms.

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 [20]. The smart contract utilizes hashes that store the digital content and its metadata. The metadata contains information related to the device capturing the video, date and time, as well as logs and manually added information that the video creator can add, such as a trust stamp. Their solution relies on the principle that if the content can be credibly traced to a trusted or reputable source, the content can then be real and authentic. Moreover, security analysis on how their BC-based proposal ensures key security goals such as integrity, accountability, authorization, availability and non repudiation is addressed. They evaluate also if their solution is resilient against popular attacks, including Man In the Middle and Distributed Denial of Service (DDoS).

A similar idea is discussed in [45], where permissioned BC is coupled with LSTM. This means media content would require the original artist attestation of untampered data. The smart contract combines multiple LSTM networks into a process that allows for tracing of a digital content historical provenance. The result is a theoretical framework that enables PoA for digital media using a decentralized BC, where LSTMs are used as a deep encoder for creating unique discriminative features, which are then compressed and hashed into a transaction.

Bitcoin, the most well-known BC network, contains its entire legal transaction history, thereby providing convenience for tracking money. However, mixing services, usually offered by third-party companies, are used as an effective means to hide the identity of a transaction address by combining several transfers from different users. 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, and to regulate the cryptocurrency market and avoid financial crime in general, many governments and third parties are seeking to find ways to prevent or identify mixing services. 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 Section VII. The authors in [46] demonstrate that Bitcoin transaction graphs possess community properties and that a mixing service can be regarded as a cluster outlier. They leverage on a deep AE to identify mixing services in a real Bitcoin ledger.

Ethereum smart contracts are immutable and the attackers or developers cannot modify them. However, they can be terminated and new contracts can be created. Therefore, they are vulnerable to attacks and financial fraud within this BC. Moreover, identifying anomalies in this massive network is challenging because of anonymity. In [47], an AD method

Reference	Scope	Highlights
[14]	Leveraging BC	Use cases: private keys & smart contracts; Challenges: product integration by industry
[4]	Leveraging BC	Apps: content moderation & decentralized social media; Challenges: distributed ledger technology
[20]	PoA	Content traceability; Challenges: decentralized app for automatic PoA
[45]	PoA	Content traceability based on smart contracts; LSTM encoder fingerprinting
[46]	Anomaly Detection	BC mixing services; AD within Bitcoin network
[47]	Anomaly Detection	AD within Ethereum smart contracts
[48]	Leveraging BC	Cryptocurrency malware detection

TABLE III: Summary of reviewed works.

based on one-class graph ANN is proposed and evaluated on the publicly available Ethereum data.

The authors in [48] propose a deep RNN learning model for hunting cryptocurrency malware threats. Their approach analyzes Windows applications operation codes as a case study. The proposed model trains with five different LSTM structures.

B. 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 [14]. 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 greatest impact in the short term will come from traceability and tracking services implemented by big media platforms. Nevertheless, more disruptive solutions like decentralized social media platforms cannot be neglected [4]. Regarding open challenges, the following shall 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 [14]. 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. Moreover, there is no one size fits all solution for the general intervention mechanisms [4].

Integration with AI: BC technology alone is not able to fully solve the problem. In order to detect falsification attacks, contextual knowledge to corroborate the media integrity (e.g., social context features, domain location, and temporal patterns) shall be considered. 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 [4].

VI. 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 Section VI-A, 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 Section VI-B.

A. Online Social Networks

Online social media networks play an essential role in making communication between humans more accessible. However, sensitive information may be available through them and other online services that lack the security measures to protect this data. Communication systems can be penetrated by malicious users through social engineering attacks [57]. These attacks are psychological techniques and fraudulent methods with the aim of obtaining confidential information, e.g. passwords, personal intimate data, incriminating evidence, bank card numbers among others, through tricking individuals or enterprises into accomplishing several actions that benefit attackers. Currently, social engineering attacks are the biggest threats facing cybersecurity. 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 [57].

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 [58]. (ii)

Reference	Scope	Highlights
[49]	Misinformation spreading	Role played by political interest within deepfake sharing
[50]	Disinformation spreading	Deepfake sharing through Twitter social network
[51]	Data collection	Twitter data for deepfake detection
[2]	Disinformation spreading	DL-based fake news detection mechanism
[52]	Disinformation spreading	DL-based fake news detection mechanism
[53]	Misinformation spreading	Political and pornographic deepfakes analysis
[54]	Disinformation spreading	DL-based fake news detection mechanism
[55]	Disinformation spreading	Opinion mining-based fake news detection mechanism
[56]	Disinformation spreading	ML-based fake news detection mechanism

TABLE IV: Summary of reviewed literature.

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 [59] and [57].

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 improvements that impersonation can achieve including video and text context generation.

In this way, deepfakes postulate to be as one of the greatest challenges for social media networks in the upcoming years. Facebook and Adobe already raised policies to detect and fight deepfakes. The latest was Twitter, which recently announced a new policy to combat the impact of manipulated content. Moreover, Google has also decided to take action to limit their reach by creating an algorithm to detect and automatically delete deepfakes uploaded to YouTube and other Google services. A tool called Assemble was created to helps journalists to identify manipulated images. Although big tech corporates are making big efforts, academic research has just recently begun addressing digital disinformation on social media [50].

In this section, we review literature related to the generation of online text, mainly fake news, and spread through 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 [49] offers insights into the inadvertent sharing of deepfakes and highlights the role played by political interest, a key motivation for political engagement, which is also positively associated with the sharing of deepfakes. The core findings suggest politically interested and low-cognitive abled users are more likely to share deepfakes inadvertently. Moreover, network size moderates the relationship between political interest and sharing.

The authors in [50] analyze the deepfake phenomenon on Twitter. NodeXL was used to identify main actors and their connections. In addition, the semantic networks of the tweets were analyzed to discover hidden patterns and predominant content. Results show that half of the actors involved in the deepfake spreading are journalists and media companies, which is a sign of the concern that this sophisticated form of manipulation generates in this collective. Moreover, although most of the deepfakes that spread over the Internet are pornographic in nature, public attention is focused above all on political deepfakes because of their ability to generate instability in many aspects, including inside a country and among countries.

It is crucial to develop deepfake social media messages detection systems. With this goal, a publicly available dataset of deepfake tweets is provided in [51]. Every collected tweet was actually posted on Twitter, including tweets from a total of 23 bots, imitating 17 human accounts. The bots are based on various generation techniques, i.e., Markov Chains, RNN, LSTM. Moreover, some randomly tweets from the humans imitated by the bots were selected to have an overall balanced dataset of $25,572$ tweets. Lastly, they evaluate several state-of-the-art text detection approaches. Their results suggest that a wide variety of detectors (text representation-based using ML or DL methods and transformer-based using transfer learning) have greater difficulties in detecting correctly a deepfake tweet rather than a human-written one.

A classifier that can predict whether a piece of news is fake or not based only its content is built in [2], thereby approaching the problem from a purely deep learning perspective by RNN models, i.e., vanilla, GRU and LSTM. They leverage on a public benchmark dataset called LIAR. Collected a decade-long, $12.8k$ 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 [52]. 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 source material from Twitter, the work in [53] explores the relationship between political and pornographic deepfakes, finding that they operate in similar ways to silence critical speech. Authors claim that policy makers should consider the reasons why people create and consume fake porn and seek to challenge the inequalities that lead this technology to disproportionately target women.

Fake news generally spread exponentially and more rapid than real news, due to fact that they are usually more dramatic. Fake tweets also tend to have more rumor path propagation hops, known as retweets. Tweets of real news on the other hand, tend to expand at a constant and slow pace, with usually lower people reach [54]. The thesis in [54] 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 [55] 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. Their method employs SentiWordNet to consider the cognitive cues of the text to facilitate opinion mining. It further leverages a bidirectional Gated RNN incorporating objective factors, such as sentiment and a credibility score to provide efficient decisions. 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 [56], 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 [60], 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.

B. 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 [22]. IoT is a distributed network of embedded systems communicating through wired or wireless communication technologies. It is composed of physical objects empowered with limited computation, storage, and communication capabilities as well as embedded with electronics (such as sensors and actuators), software, and network connectivity that enables these objects to collect, sometimes process, and exchange data. The things in IoT refer to the objects from our daily life ranging from smart house-hold devices to more sophisticated ones such as Radio Frequency IDentification (RFID) devices, heartbeat detectors, accelerometers, and every type of sensor [62].

Due to the complexity of IoT systems, guaranteeing security is challenging. IoT devices mostly work in an unattended and sometimes unpredictable environment, where an attacker may physically access the device by eavesdropping. IoT devices cannot support complex security structures given their limited computation and power resources. Moreover, due to the interdependency and interconnectivity between the IoT device and the rest of the cyberphysical system, new ways of attack can easily arise [27]. From all these reasons, IoT systems must have a transition, from merely facilitating secure communication amongst devices, to intelligence enabled by DL techniques to build strong holistic security solutions [27]. Indeed, DL has shown improvements over traditional signature-based and rule-based systems as well as classic ML solutions [65]. Several security aspects shall be considered in every IoT system [27]:

Integrity: ensuring an effective checking mechanism to detect any modification during communication over an insecure wireless network is key. A deficiency in integrity inspection can allow for modification of the data stored in the IoT memory device.

Authentication: entities/agents identification should be perfectly established prior to performing any other process. However, due to the nature of IoT systems, authentication requirements differ from system to system and trade-offs are a major challenge in developing an effective scheme.

Authorization: refers to granting users access rights to the IoT system. The main challenge is how to grant access successfully in an environment where users may be not only humans but also physical devices or services.

Availability: services delivered by IoT systems must always be available to authorized entities. However, IoT systems can

Reference	Scope	Highlights
[27]	IoT security systems	ML apps for intrusion detection & DL for malware and anomalies detection
[61]	IoT security systems	Intrusion detection models; ML & DL detection tools; Open datasets
[62]	IoT security systems	Security requirements and solutions within IoT systems; main attack vectors
[63]	IoT framework	Deep ANN-based intrusion detection system
[64]	IoT framework	5G-enabled solution combining DL and BC technology

TABLE V: Summary of reviewed literature.

still be rendered unavailable by many threats, such as Denial of Service (DoS) or active jamming. Therefore, ensuring the continuous availability it is a critical point.

Non-repudiation: providing access logs that serve as evidence in situations where IoT users cannot refuse an action. Non-repudiation is not considered a key security aspect for many IoT systems, but it can be in specific contexts, such as payment systems where both parties cannot repudiate a transaction.

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. If an attack path exists, active intruders can attempt to partially or fully impersonate an IoT entity; (ii) data tampering is the act of intentionally changing (deleting or editing) information via unauthorized operations. Most attack detection systems have a common structure: (i) a data gathering module collects data, which possibly contains evidence of an attack, (ii) an analysis module detects attacks after data processing, and (iii) a mechanism 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 [61]. 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 ML methods and recent advances in DL used to develop enhanced security within IoT systems is presented in [27]. 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 [61] review IoT technologies, protocols, architectures and threats emerging from compromised IoT devices along with providing an overview of intrusion detection models. 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 [62]. The authors shed light on the gaps in these security solutions that call for ML and DL approaches. Their findings suggest that the theoretical foundations of DL models need to be strengthened so that the performances can be quantified based on parameters such as computational complexity, learning efficiency, as well as parameter tuning strategies. Furthermore, new hybrid learning strategies and novel data visualization techniques will be required for intuitive and efficient data interpretation.

The authors in [63] propose a DL-based intrusion detection system. The multi-class classifier comprises a feed-forward ANN with embedding layers to identify four categories of attacks, namely denial of service (DoS), DDoS, data gathering, and data theft, while differentiating traffic of these attack types from routine network traffic. In addition, the encoding of high-dimensional categorical features is extracted through the concept of network embedding and subsequently applied to a binary classifier via a transfer learning-based approach. 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 Section V-B, a 5G-enabled IoT security framework combining DL and BC technology is proposed in [64]. Their hierarchical architecture is described across the four layers of cloud, fog, edge, and user. The framework is evaluated employing various standard measures of latency, accuracy, and security to demonstrate its validity in practical applications.

VII. 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 Section VII-A. Then, we elaborate on some research opportunities in Section VII-B, aiming at encouraging work in those together with the main agents and potential technologies involved.

A. 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 [22].

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. There exist open issues related to the compliance with national and international laws, such as General Data Protection Regulation (GDPR) [66], 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 [67], where a review of the existing anonymization techniques for privacy preserving publishing of social network data is presented. 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 [68] [69].

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 [49]. 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 [70]. The authors analyze the fake bot classifier *Botometer* [71]. This classifier was successfully introduced as a way to estimate the number of bots in a given list of accounts and, as a consequence, has been frequently used in academic publications. Their conclusions show that Botometer scores are imprecise when it comes to estimating bots; especially in a different language. They further show that Botometer's thresholds, even when used very conservatively, are prone to variance, which, in turn, will lead to false negatives (i.e., bots being classified as humans) and false positives (i.e., humans being classified as bots). This has immediate consequences for academic research as most studies in social sciences using the tool will unknowingly count a high number of human users as bots and vice versa. Moreover, the authors in [72] point out fundamental theoretical flaws of social bot research. They inspect hundreds of accounts that had been counted or even presented as social bots in peer-reviewed studies. Their results show that they were unable to find a single social bot. They conclude that studies claiming to investigate the prevalence or influence of social bots have, in reality, just investigated false positives and artifacts of the flawed detection methods employed.

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 [70].

Furthermore, future research has to focus on group behavior and user networks [73] [74]. 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) [75]. 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.

B. Future Trends

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 [76]. 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 [76] and [77].

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 [78] 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 [79] and [80].

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 [81]. In this sense, *Explainable Artificial Intelligence (XAI)* was first mentioned in 2004 by Van Lent [82], 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 [83] and [84].

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 [85] propose a CNN-based deepfake detection framework tested on various XAI techniques, evaluating its applicability within real-life scenarios. The authors in [86] discuss both practical and novel ideas for leveraging XAI to improve the efficacy of digital forensic analysis, usually employed in deepfake detection mechanisms.

3) Knowledge Fusion: With the aim of achieving models that maximally learn from data, one promising solution is complementing data-driven with theory-driven models. The first are highly flexible in adapting to data and finding hidden patterns, while the latter are easier to interpret. This concept is known as *knowledge fusion* [87], where information discovered from different areas of expertise can reinforce each other to derive more meaningful understanding. DL techniques can be potential candidates in this regard [88]. The authors in [89] study the applicability and limitations of different knowledge fusion techniques. The leverage data fusion to identify reliable information among several analyzed data sources, enabling efficient decision making. The authors in [90] survey data fusion for IoT, focusing on AI, probabilistic methods and theory of belief.

The work in [42] suggests that fusion techniques, at a feature level, could provide a better adaptation for deepfake detection in different scenarios. In fact, they discuss some examples, where different fake detection approaches are already based on the combination of different sources of information, such as steganalysis and DL features, or spatial and spectral features. Another two interesting fusion approaches are pointed out, combining RGB, Depth, and InfraRed information to detect physical face attacks. Moreover, fusion of other sources of information such as the text, keystroke, or audio that accompanies the videos when uploading them to social networks could be very valuable to improve deception detectors.

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 [22].

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 [91]. 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 [92]. 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.

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 [1], existing deepfake detectors have mainly relied on the fixed features of existing cyberattacks by using ML techniques, including unsupervised clustering and supervised classification methods, and therefore they are less likely to detect unknown deepfakes. Hence, Reinforcement Learning (RL) techniques could play a key role on the detection side. A step beyond, deep RL, could offer great potential for not only deepfake detection but also to counter antiforensic attacks on the detectors. Since RL can model an autonomous agent to take sequential actions optimally with limited or without prior knowledge of the environment, it could be used to meet a need for developing algorithms to capture traces of anti-forensic processing, and to design attack-aware deepfake detectors. A review of deep RL approaches developed for solving cyber-security problems can be found in [13], including autonomous intrusion detection techniques and multiagent game theory simulations for defense strategies.

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 [93]. 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 [94].

Edge Intelligence (EI) [94] 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 [95] 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 [96] 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 [97]. Their aim is to detect highly sophisticated GAN generated deepfake images at the edge, reducing training and inference time while achieving a certain accuracy.

7) Blockchain Technology and AI: The combination of DL with BC technology reveal a powerful symbiosis that can provide a fully functional security system. Firstly, DL may assist BC technology in realizing smart decision-making, improved evaluation, filtering and comprehension of data and devices within a network to facilitate the effective implementation of BC for enhanced trust and security services. Secondly, BC may assist AI by providing a large volume of data, since its inherent decentralization database stresses the importance of data distribution among several nodes on a specific network.

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.

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.

VIII. 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.

REFERENCES

[1] M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *arXiv preprint arXiv:2103.00484*, Feb 2021.

[2] S. Girgis, E. Amer, and M. Gadallah, "Deep learning algorithms for detecting fake news in online text," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. Cairo, Egypt: IEEE, Dec 2018, pp. 93–97.

[3] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, Oct 2020.

[4] P. Fraga-Lamas and T. M. Fernández-Caramés, "Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality," *IT Professional*, vol. 22, no. 2, pp. 53–59, March 2020.

[5] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *arXiv preprint arXiv:2103.00218*, Feb 2021.

[6] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, Nov 2019.

[7] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 10 039–10 049.

[8] N. Caporusso, "Deepfakes for the good: A beneficial application of contentious artificial intelligence technology," in *International Conference on Applied Human Factors and Ergonomics*. Orlando, USA: Springer, July 2020, pp. 235–241.

[9] Abba, "Abba Voyage," https://abbavoyage.com/, 2021.

[10] FaceApp, "FaceApp," https://www.faceapp.com/, 2021.

[11] Facebrity, "Facebrity," https://apps.apple.com/us/app/facebrity-face-swap-morph-app/id1449734851, 2021.

[12] A. O. Kwok and S. G. Koh, "Deepfake: a social construction of technology perspective," *Current Issues in Tourism*, vol. 24, no. 13, pp. 1798–1802, March 2021.

[13] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection: A survey," *arXiv preprint arXiv:1909.11573*, Sept 2019.

[14] A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, "Making Sense of Blockchain for AI Deepfakes Technology," in *2020 IEEE Globecom Workshops (GC Wkshps*. Taipei, Taiwan: IEEE, Dec 2020, pp. 1–6.

[15] J. Langguth, K. Pogorelov, S. Brenner, P. Filkuková, and D. T. Schroeder, "Don't trust your eyes: Image manipulation in the age of deepfakes," *Frontiers in Communication*, vol. 6, p. 26, 2021.

[16] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A Survey on Deepfake Video Detection," *IET Biometrics*, pp. 1–18, April 2021.

[17] A. M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey," *Journal of Computer and Communications*, vol. 9, no. 5, pp. 20–35, May 2021.

[18] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, April 2021.

[19] N. A. Shelke and S. S. Kasana, "A comprehensive survey on passive techniques for digital video forgery detection," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 6247–6310, Oct 2020.

[20] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *Ieee Access*, vol. 7, pp. 41 596–41 606, March 2019.

[21] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[22] A. F. Gambín, E. Angelats, J. S. González, M. Miozzo, and P. Dini, "Sustainable Marine Ecosystems: Deep Learning for Water Quality Assessment and Forecasting," *IEEE Access*, 2021.

[23] H. D. Trinh, A. F. Gambin, L. Giupponi, M. Rossi, and P. Dini, "Mobile traffic classification through physical control channel fingerprinting: a deep learning approach," *IEEE Transactions on Network and Service Management*, Oct 2020.

[24] M. Sit, B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir, "A comprehensive review of deep learning applications in hydrology and water resources," *Water Science and Technology*, vol. 82, no. 12, pp. 2635–2670, Aug 2020.

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

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

[27] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for internet of things (IoT) security," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1646–1685, May 2020.

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

[29] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, June 2014.

[31] A. Arora and Shantanu, "A Review on Application of GANs in Cybersecurity Domain," *IETE Technical Review*, pp. 1–9, Dec 2020.

[32] H. Navidan, P. F. Moshiri, M. Nabati, R. Shahbazian, S. A. Ghorashi, V. Shah-Mansouri, and D. Windridge, "Generative Adversarial Networks (GANs) in networking: A comprehensive survey & evaluation," *Computer Networks*, p. 108149, July 2021.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, Long Beach, CA, USA, Dec 2017, pp. 5998–6008.

[34] OpenAI, "OpenAI," https://openai.com/blog/better-language-models/, 2021.

[35] DeepMind, "DeepMind - AlphaStar," https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii, 2021.

[36] J. Wang, Z. Wu, J. Chen, and Y.-G. Jiang, "M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection," *arXiv preprint arXiv:2104.09770*, 2021.

[37] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," in *Proceedings of the 29th ACM International Conference on Multimedia*, Lisbon, Portugal, Oct 2021, pp. 1821–1828.

[38] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake detection scheme based on vision transformer and distillation," *arXiv preprint arXiv:2104.01353*, 2021.

[39] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[40] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[41] A. Deshmukh and S. B. Wankhade, "Deepfake detection approaches using deep learning: A systematic review," *Intelligent Computing and Networking*, vol. 146, pp. 293–302, Oct 2020.

[42] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, Dec 2020.

[43] I. Amerini, A. Anagnostopoulos, L. Maiano, L. R. Celsi *et al.*, "Deep learning for multimedia forensics," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 4, pp. 309–457, Aug 2021.

[44] M. Nofer, P. Gomber, O. Hinz, and D. Schiereck, "Blockchain," *Business & Information Systems Engineering*, vol. 59, no. 3, pp. 183–187, Mar 2017.

[45] C. C. K. Chan, V. Kumar, S. Delaney, and M. Gochoo, "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, Switzerland: IEEE, Sept 2020, pp. 55–62.

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

[47] V. Patel, L. Pan, and S. Rajasegarar, "Graph Deep Learning Based Anomaly Detection in Ethereum Blockchain Network," in *International Conference on Network and System Security*. Melbourne, Australia: Springer, Nov 2020, pp. 132–148.

[48] A. Yazdinejad, H. HaddadPajouh, A. Dehghantanha, R. M. Parizi, G. Srivastava, and M.-Y. Chen, "Cryptocurrency malware hunting: A deep recurrent neural network approach," *Applied Soft Computing*, vol. 96, p. 106630, Nov 2020.

[49] S. Ahmed, "Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size," *Telematics and Informatics*, vol. 57, p. 101508, March 2021.

[50] J. Á. Pérez Dasilva, K. Meso Ayerdi, and T. Mendiguren Galdospin, "Deepfakes on Twitter: Which Actors Control Their Spread?" *Media and Communication*, vol. 9, no. 1, pp. 301–312, March 2021.

[51] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "Tweepfake: About detecting deepfake tweets," *Plos one*, vol. 16, no. 5, p. e0251415, May 2021.

[52] R. K. Kaliyar, A. Goswami, and P. Narang, "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network," *The Journal of Supercomputing*, vol. 77, no. 2, pp. 1015–1037, May 2020.

[53] S. Maddocks, "'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes," *Porn Studies*, vol. 7, no. 4, pp. 415–423, June 2020.

[54] H. Mjaaland, "Detecting Fake News and Rumors in Twitter Using Deep Neural Networks," Master's thesis, University of Stavanger, Norway, June 2020.

[55] V. Sabeeh, M. Zohdy, A. Mollah, and R. Al Bashaireh, "Fake News Detection on Social Media using Deep learning and Semantic Knowledge Sources," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, no. 2, Feb 2020.

[56] M. M. Sadr *et al.*, "The Use of LSTM Neural Network to Detect Fake News on Persian Twitter," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 11, pp. 6658–6668, May 2021.

[57] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, p. 89, April 2019.

[58] N. Ryabchuk, "Artificial Intelligence Technologies Using in Social Engineering Attacks," in *CEUR Workshop Proceedings*, vol. 2654, Kiev, Ukraine, Aug 2020, pp. 546–555.

[59] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks," *Journal of Information Security and applications*, vol. 22, pp. 113–122, June 2015.

[60] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, p. 106983, March 2021.

[61] J. Asharf, N. Moustafa, H. Khurshid, E. Debie, W. Haider, and A. Wahab, "A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions," *Electronics*, vol. 9, no. 7, p. 1177, July 2020.

[62] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in IoT security: Current solutions and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1686–1721, April 2020.

[63] M. Ge, N. F. Syed, X. Fu, Z. Baig, and A. Robles-Kelly, "Towards a deep learning-driven intrusion detection approach for Internet of Things," *Computer Networks*, vol. 186, p. 107784, Feb 2021.

[64] S. Rathore, J. H. Park, and H. Chang, "Deep Learning and Blockchain-empowered Security Framework for Intelligent 5G-enabled IoT," *IEEE Access*, vol. 9, pp. 90 075–90 083, May 2021.

[65] D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A survey of deep learning methods for cyber security," *Information*, vol. 10, no. 4, p. 122, April 2019.

[66] E. Commission, "GDPR," https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, 2016.

[67] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM Sigkdd Explorations Newsletter*, vol. 10, no. 2, pp. 12–22, 2008.

[68] R. Mayer and H.-A. Jacobsen, "Scalable deep learning on distributed infrastructures: Challenges, techniques, and tools," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.

[69] C. S. Wickramasinghe, D. L. Marino, K. Amarasinghe, and M. Manic, "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, Dec 2018, pp. 745–751.

[70] A. Rauchfleisch and J. Kaiser, "The false positive problem of automatic bot detection in social science research," *PloS one*, vol. 15, no. 10, p. e0241045, 2020.

[71] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th international conference companion on world wide web*, New York, USA, April 2016, pp. 273–274.

[72] F. Gallwitz and M. Kreil, "The rise and fall of 'social bot'research," *SSRN: https://ssrn. com/abstract*, vol. 3814191, 2021.

[73] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach, "Aggregate characterization of user behavior in twitter and analysis of the retweet graph," *ACM Transactions on Internet Technology (TOIT)*, vol. 15, no. 1, pp. 1–24, 2015.

[74] D. Ediger, K. Jiang, J. Riedy, D. A. Bader, C. Corley, R. Farber, and W. N. Reynolds, "Massive social network analysis: Mining twitter for social good," in *2010 39th International Conference on Parallel Processing*. San Diego, CA, USA: IEEE, Oct 2010, pp. 583–593.

[75] M. Kanakaraj and R. M. R. Guddeti, "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, March 2015, pp. 169–170.

[76] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, july 2019.

[77] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, Dec 2017.

[78] M. Kim, S. Tariq, and S. S. Woo, "FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Online, June 2021, pp. 1001–1012.

[79] S. Suratkar, E. Johnson, K. Variyambat, M. Panchal, and F. Kazi, "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)*. Kharagpur, India: IEEE, July 2020, pp. 1–9.

[80] S. Suratkar, F. Kazi, M. Sakhalkar, N. Abhyankar, and M. Kshirsagar, "Exposing deepfakes using convolutional neural networks and transfer learning approaches," in *2020 IEEE 17th India Council International Conference (INDICON)*. New Delhi, India: IEEE, Dec 2020, pp. 1–8.

[81] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *Journal of Applied Remote Sensing*, vol. 11, no. 4, p. 042609, Sept 2017.

[82] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*. San Jose, California, USA: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, July 2004, pp. 900–907.

[83] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. Opatija, Croatia: IEEE, July 2018, pp. 0210–0215.

[84] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, Sept 2018.

[85] B. Malolan, A. Parekh, and F. Kazi, "Explainable deep-fake detection using visual interpretability methods," in *2020 3rd International Conference on Information and Computer Technologies (ICICT)*. San Jose, CA, USA: IEEE, March 2020, pp. 289–293.

[86] S. W. Hall, A. Sakzad, and K.-K. R. Choo, "Explainable artificial intelligence for digital forensics," *Wiley Interdisciplinary Reviews: Forensic Science*, p. e1434, Sept 2020.

[87] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb 2019.

[88] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, "A survey on an emerging area: Deep learning for smart city data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 392–410, May 2019.

[89] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang, "From data fusion to knowledge fusion," *arXiv preprint arXiv:1503.00302*, 2015.

[90] F. Alam, R. Mehmood, I. Katib, N. N. Albogami, and A. Albeshri, "Data fusion and IoT for smart ubiquitous environments: A survey," *IEEE Access*, vol. 5, pp. 9533–9554, April 2017.

[91] O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting GAN DCT anomalies," *arXiv preprint arXiv:2101.09781*, 2021.

[92] J. Pu, N. Mangaokar, B. Wang, C. K. Reddy, and B. Viswanath, "Noisescope: Detecting deepfake images in a blind setting," in *Annual Computer Security Applications Conference*, Austin, USA, Dec 2020, pp. 913–927.

[93] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, May 2016.

[94] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge

computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, June 2019.

[95] M. A. Ferrag and L. Maglaras, "DeepCoin: A novel deep learning and blockchain-based energy exchange framework for smart grids," *IEEE Transactions on Engineering Management*, vol. 67, no. 4, pp. 1285–1297, July 2019.

[96] E. Hasanaj, A. Aveler, and W. Söder, "Cooperative edge deepfake detection," Master's thesis, Jönköping University, School of Engineering, Aug 2021.

[97] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "EasyDeep: An IoT Friendly Robust Detection Method for GAN Generated Deepfake Images in Social Media."