Scientific studies on:"Forensic analysis of fake videos in large-scale events towards automated detection and mitigation"

Maha Sabri Altememe
College of Computer Science & Information Technology,
University of Karbala, Iraq
maha.sabri@uokerbala.edu.iq

Wael Mahdi Brich
Information &Telecommunications Public Company, Ministry of Communications
albdairy.waal@itpc.gov.iq

Abstract

Mass gatherings, especially the Arba'in pilgrimage in Karbala, are a fertile ground for the spread of deepfake videos. These gatherings rely heavily on visual content circulating on social media and crowd-sourced platforms. Adversaries can exploit this by creating fake videos that mislead people and confuse public opinion. Based on the study's recommendations, there are some ways to combat the spread of deepfake videos in the context of mass gatherings. In this work, we classify deepfake detection methods according to the applications they are used for: hybrid multimedia detection, image, audio and video detection. This study aims to provide the reader with an enhanced understanding of: how deepfakes are created and identified; the latest developments and discoveries in this field; the deficiencies in current security measures; and the four areas that need further research and thinking. According to the results, the most widely used deep learning technique in research is traditional neural network (CNN) methodology.

Introduction

The Artificial Intelligence and deep learning technology that is created by a fake image or video is called a deep fake. Although AI can achieve amazing results, detecting deepfakes using AI is still a difficult and complex process. Deep learning is an essential tool for identifying deepfakes. Promising deepfake detection techniques

Semi-Annual Scientific Journal
www.arbnj.com
March 2025 - Ramadan 1446

include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs). It can lead to a variety of issues, such as influencing public opinion or using false evidence in court. In light of these variables, we must possess the skills that enable us to distinguish between reality and illusion. This paper provides a comprehensive analysis of the literature on algorithm-based deepfake detection techniques. Deepfake is a term used to describe a variety of face alteration techniques that uses advanced technology including computer vision and deep learning (DL).

Four categories comprise face modification: identity switch, full-facial synthesis, attribute manipulation, and expression swap (Balaji et al., 2021). One of the most popular forms of deepfake video is identity swap, also referred to as face swap, in which the faces of the targeted individuals are substituted for the faces of the source people (Jafar et al., 2020). Users may find it challenging to differentiate between fabricated images and videos and deepfake news when they combine the two (Matern et al., 2019). Through its widespread distribution on social media, this kind of deepfake has the potential to have a major negative impact on people's lives (H. H. Nguyen, Yamagishi, et al., 2019). The use of deep learning in detecting deepfakes is crucial. Advanced algorithms like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) have shown promise in identifying deepfakes.

To address this threat, the research recommends several proactive measures:

1. Developing advanced detection algorithms and verification tools: These can be integrated into social media platforms and other content aggregators to identify and remove deepfake videos in real-time, preventing their widespread dissemination.

2. Implementing robust fact-checking and content moderation protocols: Event organizers, media outlets, and social media platforms should work collaboratively to quickly identify and debunk any manipulated or misleading content related to the gathering.

3. Empowering users with media literacy education: Equipping the public, especially attendees of these events, with the skills to recognize and critically analyze online content can help mitigate the impact of deepfake videos and other forms of disinformation.

4. Establishing clear communication channels and rapid response mechanisms: Organizers, authorities, and platforms should have well-defined processes to promptly address and correct any false narratives or manipulated content that emerges during or after the event.

By proactively implementing these strategies, the research suggests that the threats posed by deepfake videos and other forms of visual manipulation can be significantly reduced in the context of massive public gatherings, preserving the integrity of these

Semi-Annual Scientific Journal
www.arbnj.com
March 2025 - Ramadan 1446

important events and the broader public discourse.

Related Works

There are several varieties of deepfake detection techniques available today, and each has benefits and drawbacks of its own. In order to obtain more accurate findings, this work attempts to analyze these approaches from other papers and highlights how they might be integrated and altered in a new project.

This study (D. Guera¨ and E. J. Delp, 2018), proposes a temporal-aware process to automatically recognize deepfake videos. Understanding the production process of deepfake videos is the first step in identifying them. With this knowledge, we will be able to pinpoint the weak places in the production process of deepfakes and exploit them to recognize them when they occur. Frame-level scene inconsistency is the first feature used in the strategy presented in this study. There are 300 videos in this dataset, which is derived from the HOHA dataset. This technique could identify whether the section under study is from a deepfake video or not with an accuracy rate of above 97%.

In this paper (M. A. Younus and T. M. Hasan, 2020), This approach uses the limitation that the deepfake algorithm can only produce fake faces with a certain size and resolution while creating videos. An additional blur function needs to be added to the synthetic faces in order for them to fit and resemble the source's face arrangement in the original videos. The produced face and

the background of the deepfake films it creates are exclusively blurry as a result of this alteration. The technique compares the blurred synthesized ROI and the surrounding context using a specialized Haar wavelet transform function in order to identify such inconsistency. The accuracy percentage of this suggested model is 90.5.

In this study (Mousa Tayseer Jafar et. Al, 2020), a method for detecting deepfakes utilizing mouth features was proposed. This article describes the design and implementation of a deepfake detection model with mouth features (DFT-MF), which uses a deep learning technique to isolate, analyze, and verify lip/mouth movement in order to detect deepfake films. Next, a face's mouth region is cropped. For the face, there will be set coordinates. To estimate the location of 68 (X, Y) coordinates, a face landmark detector is utilized, working on a typical image frame. The next phase involves excluding all faces with closed mouths and tracking faces with only open mouths and teeth with a good degree of resolution. By computing three variables: words per phrase, speech rate, and frame rate, CNN is used to categorize videos as real or fraudulent based on a threshold number of bogus frames. If a video has more than fifty bogus frames, it is labeled as fake; otherwise, it is labeled as authentic.

Semi-Annual Scientific Journal
w w w . a r b n j . c o m
March 2025 - Ramadan 1446

3. Deepfake detection methods

3.1. Fake image detection

The process of detecting and marking images that have been altered or edited in some way is known as "fake image detection," and it usually involves the use of machine learning algorithms and image processing techniques. Numerous studies and projects deal with the topic of false picture identification, employing a variety of techniques, including deep learning models, to identify facial changes in videos—DeepFakes, Face2Face, and FaceSwap, for example. Texture analysis, especially for faces produced by generative adversarial networks (GANs), can be used to distinguish between real and synthetic faces.

Developing generalized representations to describe the artifacts produced by generation models are possible through tools and services available for fake image detection, such as the Fake Image Detector, which uses advanced techniques like Metadata Analysis and Error Level Analysis (ELA) to detect manipulated images. However, it's important to note that no method is 100% accurate, and there is always a risk of false positives or negatives. Fake face image detection is the most difficult challenge in the field of image forgery detection.

Fake images can be used to create false personas on social networking sites, which makes it possible for personal data to be stolen. For instance, the fake picture generator can be used to

produce potentially harmful images of celebrities with unsuitable material. Specifically, Deepfake replaces the face of an original image with a different person's face using GANs. The GAN models are more likely to generate realistic faces that can be precisely spliced into the main image because they were trained on 10 out of 100 images.

3.2. Fake Video detection

The process of identifying and marking videos that have been altered or created using artificial intelligence (AI) is known as fake video detection. With the advancement and accessibility of deepfake technology, this is becoming more and more crucial. Many techniques, including monitoring motion patterns, modeling the human vocal tract, and evaluating color changes in faces to infer blood flow, are being developed for the purpose of detecting false videos. These techniques seek to identify minute indications of video tampering and separate them from authentic footage. Researchers are also investigating the use of blockchain technology and watermarking to determine the source of information and stop deepfakes from spreading. However, it is important to note that technology alone cannot solve the problem of deepfakes, and education and media literacy are also crucial for combating the spread of misinformation (D. Guera and E. J. Delp, 2018).

There are several common methods for detecting deepfake videos:

Semi-Annual Scientific Journal
www.arbnj.com
March 2025 - Ramadan 1446

- Analysis of color changes in faces: This method analyzes the color changes in facial regions to infer blood flow and detect signs of manipulation.
- Motion pattern analysis: This method studies the motion patterns in videos to detect any abnormalities that may indicate manipulation.
- Modeling the human vocal tract: This method models the human vocal tract to detect any inconsistencies in the audio of a video, which may indicate manipulation.
- Watermarking and blockchain technology: These technologies can be used to establish the provenance of media and prevent the spread of deepfakes.
- Machine learning and artificial intelligence: These techniques can be used to train models to detect subtle signs of manipulation in videos.
- Training datasets: Creating and using large and diverse datasets to train machine learning models to detect deepfakes.
- Temporal sequence analysis: Analyzing the temporal sequence between frames to discriminate real videos from fake ones.
- Biological signals analysis: Analyzing biological signals such as eye blinking and heartbeat to detect deepfake videos.
- It's worth noting that the field of deepfake detection is constantly evolving, and new methods are being developed all the time. Additionally, deepfake detection methods should be used in

conjunction with education and media literacy to effectively combat the spread of misinformation.

3.3. Fake Audio detection

Audio analysis plays an important role in detecting deepfake videos as well as deepfake audio. In the case of deepfake videos, audio analysis can help identify inconsistencies between the audio and visual elements of the video. For example, if the lip movements of a person in a video do not match the words being spoken in the audio, it could be an indication that the video has been manipulated using deepfake techniques. In the context of deepfake audio, audio analysis can help differentiate between organic and synthetic speech. By analyzing the acoustic and fluid dynamic differences between voice samples, it is possible to identify whether a given audio sample was generated organically by a human or synthetically by a computer. So, audio analysis can help detect deepfakes by identifying inconsistencies between the audio and visual elements of a video, as well as by differentiating between organic and synthetic speech in the case of deepfake audio.

4. Deep Fake Detection Techniques by Deep Learning

Although deep learning models can directly extract or learn features from the data, their feature extraction and selection mechanisms have made them widely employed in computer vision. In deepfake detection studies, we found the following deep learning-

based models have been used: convolutional neural network (CNN) model (e.g., XceptionNet, GoogleNet, VGG, ResNet, EfficentNet, HRNet, InceptionResNetV2, MobileNet, InceptionV3, DenseNet, SuppressNet, StatsNet), recurrent neural network (RNN) model (e.g., LSTM, FaceNet), bidirectional RNN model, long-term recurrent Table 5. Distribution of used models. Deep Ensemble Learning (DEL), Hierarchical Memory Network (HMN), Multi-task Cascaded CNNs (MTCNN), Convolutional Neural Network (RCNN), and Faster RCNN models are among the models.

4.1.RNN-based neural network architecture for deepfake detection

In implementing a deepfake-detecting neural network architecture based on RNNs, a sigmoid activation function for binary classification (actual or fake) may be used in the final output layer, which is followed by fully connected layers and one or more RNN layers. To extract spatial or temporal characteristics from the video frames, it is also usual practice to employ other types of layers, such as convolutional layers or attention mechanisms. Using a suitable optimizer (like Adam), a loss function (like binary cross-entropy), and an evaluation measure (like accuracy), you can train the RNN model using the prepared dataset. Additionally, methods like early halting, batch normalization, and dropout can be used to enhance generalization and avoid overfitting [6]. Figure 1 presents the fundamental diagram of neural networks architectures.

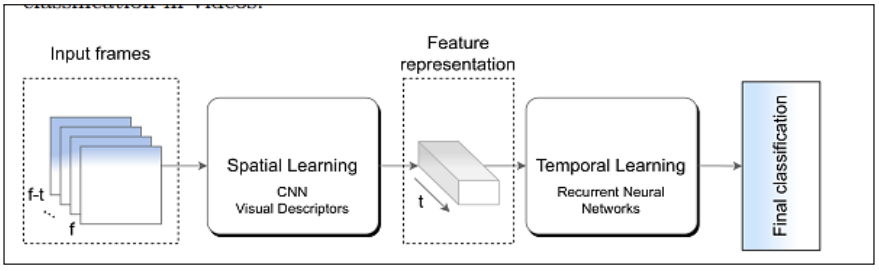

Figure (1) diagram of neural networks architectures [3].

CNN-based neural network architecture for deepfake detection

CNN-based neural network architectures, such as XceptionNet, GoogleNet, VGG, ResNet, EfficientNet, HRNet, InceptionResNetV2, MobileNet, and InceptionV3, are frequently employed for deepfake detection. These architectures work well for computer vision applications like deepfake detection since they are made to extract and learn features from data. This is an illustration of how to use Keras to create a CNN-based deepfake detection model. A collection of artificial neuron layers working together is called a convolutional neural network. Like their biological counterparts, artificial neurons are mathematical functions that analyze the weighted sum of all aggregate inputs and then provide an activation value. Each layer of a ConvNet produces many activation functions when an image is fed into it; these activation functions are then passed on to the following layer. In the first layer, critical features such as edges that are horizontal or diagonal are removed. The subsequent layer receives this information and uses it to identify more intricate features like edges and combinational edges. Based on the activation

Semi-Annual Scientific Journal
www.arbnj.com
March 2025 - Ramadan 1446

map of the preceding convolution layer, the layer categorization yields a range of confidence ratings (numbers between 0 and 1) that represent the likelihood that the image will fit into a particular "class" (Hsu, C.-C., Zhuang, Y.-X. and Lee, C.-Y, 2020) In Figure 2 workflow diagram for CNN deepfake detection.

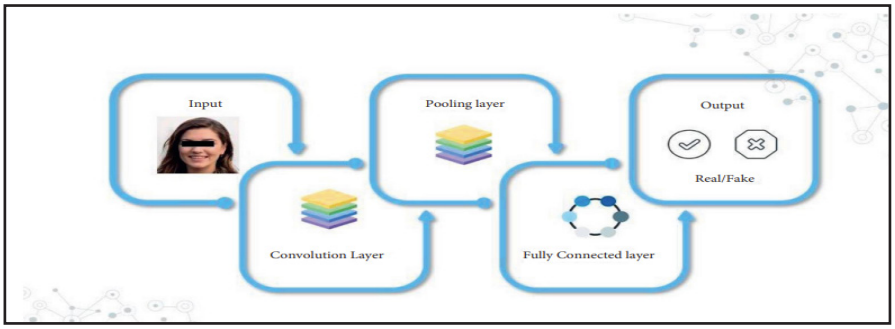

Figure (2) Workflow Diagram for CNN Deepfake Detection

4.2. LSTM-based neural network architecture for deepfake detection

To create a deepfake-detecting neural network architecture based on LSTMs, one or more LSTM layers, fully connected layers, and a final output layer with a sigmoid activation function for binary classification (actual or fake) can all be found in the architecture. Using a suitable optimizer (like Adam), a loss function (like binary cross-entropy), and an evaluation measure (like accuracy), train the LSTM model using the prepared dataset. Additionally, methods like early halting, batch normalization, and dropout can be used to enhance generalization and avoid overfitting (A. M. Almars, 2021). In Figure 3 workflow diagram for CNN-LSTM deepfake detection.

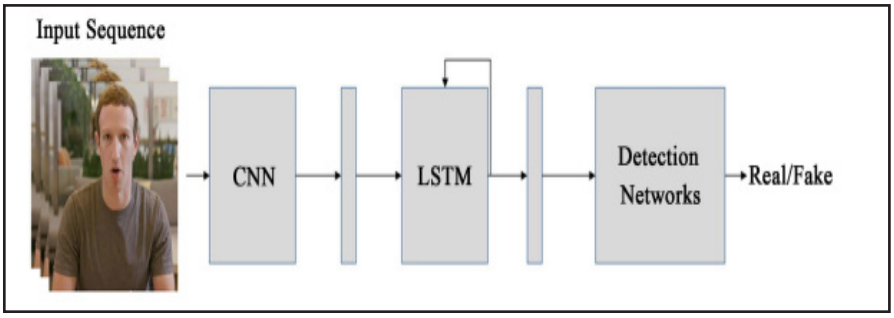

Figure (3) Workflow Diagram for CNN-LSTM Deepfake Detection

5. Deepfake Generation Technique

5.1.Generative Adversarial Networks (GANs)

Since recent GAN-based deepfake models acquire a great degree of realism, deepfake detection and prevention are extremely difficult undertakings. Using a pre-trained GAN to create deepfake samples is one way to approach this issue. Then, using different approaches such as ensemble methods, pixel-by-pixel comparison, or feature extraction, these samples can be compared with the original deepfake. Figure 4 explain GNA architecture.

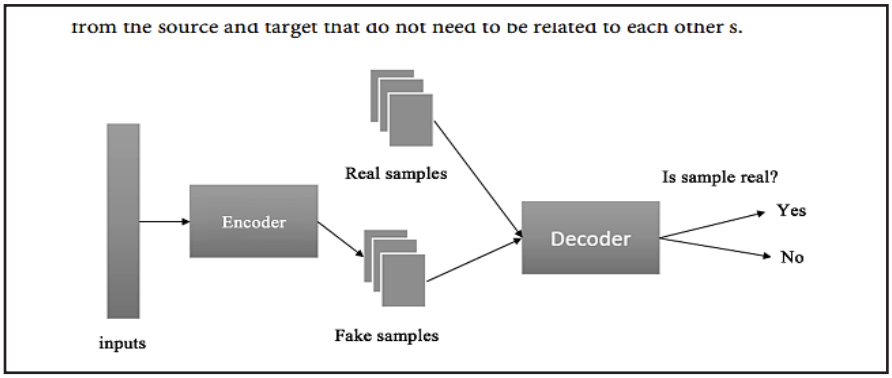

Figure (4) GNA architecture (Pouyanfar, S., et al., 2018)

Deepfake Generation: A type of deep neural network that is frequently used to produce deep fakes is called a generative adversarial network, or GAN. The ability of GNAs to learn from a set of training data sets and provide a sample of data with the same features and attributes is one of their advantages. GANs, for instance, can be used to replace a person's "fake" image or video with a "real" one (D. Guera¨ and E. J. Delp, 2018). An encoder and a decoder are the two neural network components that make up the architecture of GANs. To create fake data, the model first trains on a sizable data set using the encoder. The phony data is then learned from the actual data using the decoder. However, in order for this model to produce faces that look realistic, a lot of data—both photographs and videos—is needed. Figure 2: The architecture of the GNA. The encoder first gets random input seeds to create a fictitious sample, as seen in the picture. The decoder is trained using those fictitious samples. The decoder, which is essentially a binary classifier, receives inputs in the form of genuine and fake samples. It then uses the SoftMax function to separate the true data from the fake.

6. Methodology

This document includes a concise overview of several studies that detail several techniques for identifying fake images and videos. In order to obtain more accurate findings than current methods, we will also discuss how those methods might be merged or adjusted for our new project.

Here is an outline of a study that aims to investigate deep learning techniques for deepfake detection:

- Data Gathering and Preparation: A large dataset comprising both authentic and deepfake photos and videos is gathered. To guarantee consistency and quality, the dataset is cleansed and preprocessed.
- Feature Extraction: Relevant features are extracted from the dataset using the proper deep learning algorithms. For this, methods such as VGG, ResNet, or Efficient Net can be applied.
- Model Development: To address the issue of deepfake detection, a number of deep learning models are created. One can utilize CNNs, RNNs, and GANs separately or in combination. The models' performance can be raised by using strategies including ensemble techniques, fine-tuning, and transfer learning.
- Model Evaluation: A variety of metrics, including accuracy, precision, recall, and F1-score, are used to assess the models' performance. To accurately evaluate the model's performance, cross-validation methods such as k-fold cross-validation can

Semi-Annual Scientific Journal
www.arbnj.com
March 2025 - Ramadan 1446

be applied.

- Hyperparameter tuning: To maximize the performance of the model, hyperparameters are adjusted. For this, methods such as Grid Search, Random Search, or Bayesian Optimization can be applied.
- Final thoughts and Upcoming Projects: Conclusions regarding the efficacy of the suggested deep learning methods for deepfake detection can be made in light of the findings. Additionally, areas for further study and development can be noted.

7. Conclusion

Massive public gatherings, such as protests, rallies, or festivals, have become a prevalent feature of modern society. These events often attract large crowds and generate significant media attention, both through traditional news outlets and social media platforms. This makes them an attractive target for bad actors who may seek to spread disinformation or manipulate the narrative surrounding these events.

One of the key concerns raised in the research is the growing threat of deepfake videos - synthetic media that can convincingly depict people saying or doing things they never actually did. In the context of massive public gatherings, deepfake videos could be used to create false narratives, discredit protest movements, or sow discord among participants. The widespread availability of photos and videos in social media materials has led to the rise in

popularity of deepfakes. This is especially crucial today because people may readily disseminate and share these kinds of fake contents on social media platforms and have easier access to the tools needed to create deep fakes. There has been a lot of interest in deep learning techniques across many fields.

Several deep learning-based techniques have been put out recently to effectively detect phony photos and videos and solve this problem. In this paper, we first go over the programs and resources that are currently in widespread use for producing phony photos and videos. After that, we examined the state-of-the-art deepfake approaches and separated them into two main categories in this paper: image and video detection. We gave a thorough explanation of the architecture, tools, and performance of the existing deepfake methods. Lastly, we have talked about the difficulties that still exist and offered suggestions for further deep learning research on deepfake detection.

While deep learning has demonstrated impressive results in detecting deepfakes, the quality of deepfakes has been rising. For the purpose of effectively identifying phony movies and photos, the existing deep learning techniques must therefore also be improved. Furthermore, there is currently no reliable way to determine which architecture is best suited for deepfake detection or how many layers are required for deep learning algorithms. In order to increase social media platforms' ability to handle the widespread effects of deepfakes and lessen their effects, another field of research is the integration of deepfake identification techniques.

Semi-Annual Scientific Journal
www.arbnj.com
March 2025 - Ramadan 1446

References

- M. Almars, (2021). "Deepfakes Detection Techniques Using Deep Learning: A Survey," Journal of Computer and Communications, no. ISSN Online: 2327-5227,
- Do, N.-T., Na, I.-S. and Kim, S.-H. (2018) Forensics Face Detection from GANS Using Convolutional Neural Network. ISITC.
- Grekousis, G. (2019) Artificial Neural Networks and Deep Learning in Urban Geography: A Systematic Review and Meta-Analysis. Computers, Environment and Urban Systems, 74, 244-256. https://doi.org/10.1016/j.compenvurbsys.2018.10.008
- Güera, D. and Delp, E.J. (2018) Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, 27-30 November 2018, 1-6. https://doi.org/10.1109/AVSS.2018.8639163
- Guera and E. J. Delp, (2018). ""Deepfake video detection using recurrent neural networks," in Proceedings of the 2018 15th," in IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6, Auckland, New Zealand.
- D. Guera¨ and E. J. Delp, (2018) "Deepfake video detection using recurrent neural networks," in 15th IEEE International Conference on Advanced Video and Signal Based Surveillance

(AVSS), 2018, pp. 1-6.

- Hasin Shahed Shad ,Md. Mashfiq Rizvee,Nishat Tasnim Roza ,(2021), "Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network," Hindawi, Vols. Volume 2021, Article ID 3111676, 18 pages.
- Hopfield, J.J. (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences, 79, 2554-2558. https://doi.org/10.1073/pnas.79.8.255
- Hsu, C.-C., Lee, C.-Y. and Zhuang, Y.-X. (2018) Learning to Detect Fake Face Images in the Wild. 2018 IEEE International Symposium on Computer, Consumer and Control (IS3C), Taichung, 6-8 December 2018, 388-391. https://doi.org/10.1109/IS3C.2018.00104
- Hsu, C.-C., Zhuang, Y.-X. and Lee, C.-Y., (2020). " Deep Fake Image Detection Based on Pairwise Learning. Applied Sciences," MDPI Journals, pp. on Pairwise Learning. Applied Sciences, 10, 370.,
- Li, H., Li, B., Tan, S. and Huang, J. (2018) Detection of Deep Network Generated Images Using Disparities in Color Components.
- Li, Y. and Lyu, S. (2018) Exposing Deepfake Videos by Detecting Face Warping Artifacts. [10] Yang, X., Li, Y. and Lyu, S. (2019) Exposing Deep Fakes Using Inconsistent Head Poses.

2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, 12-17 May 2019, 8261-8265. https://doi.org/10.1109/ICASSP.2019.8683164

- Marra, F., Gragnaniello, D., Cozzolino, D. and Verdoliva, L. (2018) Detection of Gan-Generated Fake Images over Social Networks. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, 10-12 April 2018, 384-389. https://doi.org/10.1109/MIPR.2018.00084
- Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets. [6] Kwok, A.O. and Koh, S.G. (2020) Deepfake: A Social Construction of Technology Perspective. Current Issues in Tourism, 1-5. https://doi.org/10.1080/13683500.2020.1738357
- Mousa Tayseer Jafar; Mohammad Ababneh; Mohammad Al-Zoube; Ammar Elhassan, (2020) "Forensics and Analysis of Deepfake Videos," in 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 053-058, doi: 10.1109/ICICS49469.2020.239493.
- Nataraj, L., et al. (2019) Detecting GAN Generated Fake Images Using Co-Occurrence Matrices. Electronic Imaging, 2019, 532-1-532-7. https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532
- Pouyanfar, S., et al., (2018). " A Survey on Deep Learning: Algorithms, Techniques,," and Applications. ACM Computing

Surveys (CSUR), 51, 1-36.

- Tariq, S., Lee, S., Kim, H., Shin, Y. and Woo, S.S. (2018) Detecting Both Machine and Human Created Fake Face Images in the Wild. Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, Toronto, 15 October 2018, 81-87. https://doi.org/10.1145/3267357.3267367
- Vaccari, C. and Chadwick, A. (2020) Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society, 6, 1-13. https://doi.org/10.1177/2056305120903408
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A. (2020) CNN-Generated Images Are Surprisingly Easy to Spot... for Now. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 13-19 June 2020, 8695-8704. https://doi.org/10.1109/CVPR42600.2020.00872
- Westerlund, M. (2019) The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9, 40-53. https://doi.org/10.22215/timreview/1282
- Xuan, X., Peng, B., Wang, W. and Dong, J. (2019) On the Generalization of GAN Image Forensics. In: Chinese Conference on Biometric Recognition, Springer, Berlin, 134-141. https://doi.org/10.1007/978-3-030-31456-9_

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