iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49681

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

Exploring Deepfakes - Creation Techniques, Detection Strategies, and Emerging Challenges: A Survey

Jitesh Gadgilwar[1], Kunal Rahangdale[2], Om Jaiswal[3], Parag Asare[4], Pratik Adekar[5], Prof. Leela Bitla[6]
[1, 2, 3, 4, 5]*Student,* [6]*Mentor, Department of Information Technology, G H Raisoni College of Engineering, Nagpur*

Abstract: ***Deep learning, integrated with Artificial Intelligence algorithms, has brought about numerous beneficial practical technologies. However, it also brings up a problem that the world is facing today. Despite its innumerable suitable applications, it poses a danger to public personal privacy, democracy, and corporate credibility. One such use that has emerged is deepfake, which has caused chaos on the internet. Deepfake manipulates an individual's image and video, creating problems in differentiating the original from the fake. This requires a solution in today's period to counter and automatically detect such media. This study aims to explore the techniques for deepfake creation and detection, using various methods for algorithm analysis and image analysis to find the root of deepfake creation. This study examines image, audio, and ML algorithms to extract a possible sign to analyze deepfake. The research compares the performance of these methods in detecting deepfakes generated using different techniques and datasets. As deepfake is a rapidly evolving technology, we need avant-garde techniques to counter and detect its presence accurately.***

Keywords: Deepfakes, Deep Learning, Artificial Intelligence, Encoder, Decoder, GAN, Survey

I. INTRODUCTION

Deepfake has emerged as the threat to public's privacy, corporate reputation, political elections and reassemble a threat to the democracy. Deepfake technology has emerged as a emerging technology and a powerful tool to fabricate images and videos. Over 18 lakhs photographs and video clips are gets upload every day to web based products, including professional and social media sites. [1] Most of these videos and pictures appeared to be manipulated [2] for cordial reasons or for propaganda or misleading campaigns. Social networking sites is one of the areas where deepfakes technology is most frequently used to propagate rumours and false information quickly. People are more likely to believe information coming from their social networks, especially friends and family who have substantial personal connections, as the "infopocalypse" grows. This tendency to accept information without doing appropriate research can lead to the dissemination of false information, which can have an impact on social attitudes and behaviours. In reality, even when people are aware that something is false, they frequently decide to accept it if it is consistent with their values.

The development of deceptive operations is now simpler than ever due to the growth of realistic and high-quality deep counterfeiting manufacturing tools that are increasingly available as open-source software. With growing ease and accuracy, this trend has made it possible for people with little technical knowledge to edit films and control information by swapping out faces, creating conversation, and changing expressions. Such manipulations can be very difficult to spot and have a big impact on political campaigns, public debate, and people's reputations. Deep learning models often need a lot of picture and video data to produce lifelike images and films. Due of their extensive internet media libraries, public figures like politicians and celebrities are frequently targeted. Deepfakes can be produced to alter these people's appearances or put them in made-up situations thanks to the capacity to manipulate these photographs and videos. The effects of this technology go beyond entertainment and could pose severe risks to international security. [3] Deepfakes, for instance, might be used to create films of fake remarks made by world leaders, inflaming international tensions over politics. Furthermore, it is used to mislead voters, influence election outcomes, and disseminate false information about financial markets. [4] Deepfake techniques superimpose existing footage and can establish one person face to another body and manipulate existing footage to create entirely new scenes. Face manipulation includes the modify of facial attributes such as age, makeup, eyeglasses, skin color, hair color, morphing, mouth open or closed, color, and adding imperceptible perturbations. The technology poses serious threats and concerns for society even though it has numerous potential applications in industries like entertainment and education. As a result, there is a growing need for research to better understand the technology, its impact, and to develop effective countermeasures to mitigate its negative effects.

II. DEEPFAKE CREATION

Depfake technology has become increasingly popular due to its ability to create high-quality modified videos and its user-friendly interface, catering to both expert and beginner users. Unfortunately, this technology is being misused and circulated on social media in two prominent ways: for pornographic purposes and in political campaigns. The first involves making pornographic content by swapping out real people's faces for those of famous musicians or public figures with the intention of slandering their reputation. The second includes the use of Deepfakes in election campaigns intended to delegitimize particular people. In order to sway public opinion and make them appear less electable, political personalities' faces are substituted for these people's in videos of them making contentious utterances. [5]

A. Face Swap

In contemporary times, the process of face manipulation has become increasingly popular, with one of its most prevalent subcategories being face swap. Using this method, the facial features of one person are digitally swapped out for those of another, creating a new picture that incorporates the physical characteristics of both individuals. FakeApp - created by a Reddit user was the first attempt of face swapping. An autoencoder-decoder pairing structure is often used during the Deepfake development phase. The decoder reconstructs and decodes the face images once the autoencoder has first extracted the latent features from them. A dual encoder-decoder architecture is required to enable facial substitution between the source and destination images. The encoder network settings are shared between the two pairs, and each pair trains on a distinct collection of images. Both couples are able to consistently extract facial features thanks to the shared encoder network.[6].

By employing this approach, the shared encoder is able to identify and comprehend the similarities between two distinct sets of facial images. This task is relatively straightforward, as facial structures typically contain relative features like positioning of the mouth, nose, and eyes. Fig. 1 shows a step-by-step methodology for creating Deepfakes, where the feature set of a source face A is integrated with decoder B to generate a manipulated version of face B that resembles the original face A, this approach is employed in various works such as DeepFaceLab and Deep-Fake tensorflow framework. [7]

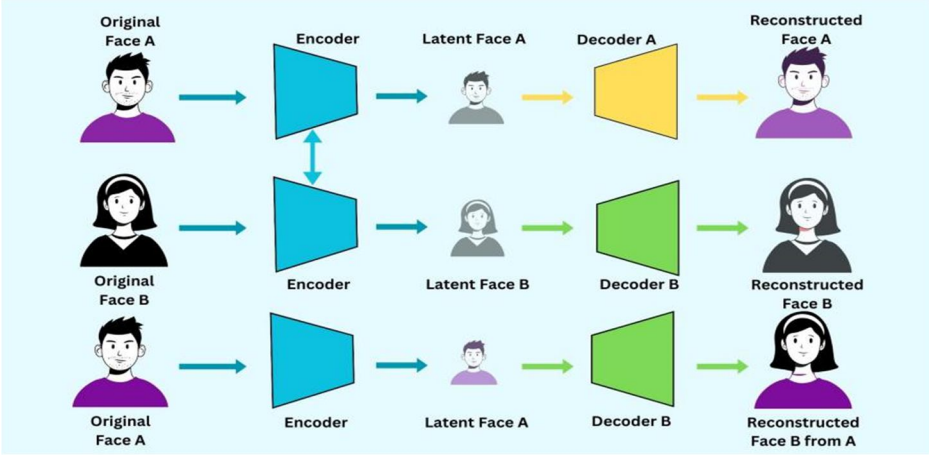


Fig. 1. Deepfake creation model

B. Generative Adversarial Networks (GAN)

A generator and discriminator make up the dual neural network design known as a GAN (Generative Adversarial Network). Although the discriminator identifies the difference between authentic content and fraudulent content, the generator generates bogus content using a random vector. Realistic deep fakes can be created with GANs, which can create nonexistent actual faces. Popular GAN-based techniques for creating deepfakes, such as the "this person does not exist" website, include STYLEGAN and VGGFace. The architecture of deepfake methods like GANs includes two additional layers, namely adversarial loss and perceptual loss layers. These layers captured latent facial features such as eye movements using an autoencoder-decoder approach, thereby enhancing the quality and authenticity of generated synthetic images.[9] CycleGAN is a deepfake method that uses the GAN architecture to extract distinctive features from one image and apply them to another image. This method makes use of a cycle loss function to speed up the discovery of latent features. Unsupervised method, CycleGAN that can execute image-to-image conversion without the necessity for matched samples, in contrast to supervised methods. In other words, even if the images are unrelated to one another, the model can learn the properties of numerous images from the source and target domains.

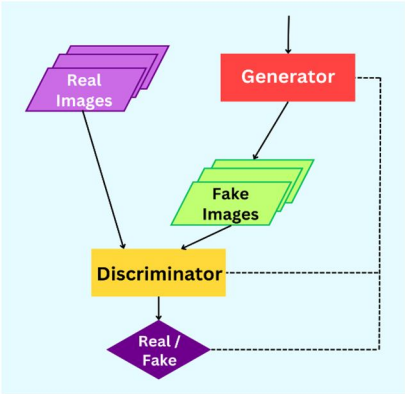


Fig. 2 Illustration of GAN

III. DEEPFAKE DETECTION

The detection of deepfakes is typically regarded as a binary classification task that involves the use of classifiers to differentiate between genuine and manipulated videos. This approach necessitates a vast collection of genuine and manipulated videos for training deep learning models for classification. Despite the growing quantity of manipulated images and videos available, it is still insufficient to establish a standardized measure for assessing the efficacy of different detection techniques. To tackle this concern, [10] created a noteworthy dataset that consisted of over 600 fake videos generated via GAN, utilizing the open-source software Faceswap-GANs. Videos with varying levels of realism produced through effectively imitating expressions, lip movements, and blinking, were generated using content from the publicly accessible VidTIMIT database. Subsequently, these videos aided in assessing several techniques to identify deepfake content. The outcomes indicate that commonly used facial identification systems such as FaceNet showed suboptimal effectiveness in detecting deepfakes. When employed for identifying deepfake content, other methods like audio-video synchronization approaches and visual quality metrics, displayed a significantly elevated rate of inaccuracies. These findings highlight the pressing necessity to develop more robust and efficient methods capable of detecting deepfakes from authentic videos. This section will provide a complete summary of the diverse techniques that have been suggested for identifying deepfake content.

A. Visual Feature-based Deepfake Detection

In 2018, a deepfake detection method based on visual features was proposed, which used blink detection [11]. This approach is based on the premise that there are distinguishable differences in the eye blinking patterns between the altered and real videos. In another approach that relies on visual features, involves detecting differences in head orientations to analyze potential modification of videos. The method involves calculating the differences between facial positioning and the adjacent regions like ears and shoulders. The measurement of head poses inconsistencies is illustrated in Figure 3. Specifically, inconsistency is calculated by contrasting the pose orientations of the central region of the face, represented by red key points, with those of the entire face, captured by blue key points. [12] proposed the use of techniques based on artificial neural networks for identifying deepfake content generated by GANs. Their approach involves employing techniques for analyzing the quantitative characteristics of video frames, which enhances the detection of deepfake videos. Similarly, [13] also presented an alternative approach that uses CNNs for detecting deepfake images generated by GANs. However, a critical issue that concerns the generalization capabilities of forensics models has been neglected in most previous research, as they tend to use identical datasets for training as well as testing their models. To address this issue, has proposed a forensics CNN that employs two image processing techniques, namely Gaussian Blur and Gaussian Noise, for detecting deepfake images.


Fig 3 Illustration of Measuring Head's Pose Inconsistency

This model is based on the application of preprocessing techniques that eliminate subtle, high-frequency interference signals found in GAN images and enhance pixel noise in local pixel characteristics. This helps the classifier can identify between real and false facial photos by better capturing the distinguishing characteristics of authentic and fraudulent photographs, proposed a hybrid methodology, (Fig. 4) beyond conventional deepfake detection models, to effectively detect fabricated images. It consisted of a two-stream network, as depicted in Figure 4, to identify instances of face modification. For training the model on authentic and manipulated images, the face classification pipeline utilizes GoogleNet. Subsequently, the patch triplet pipeline examines image characteristics by utilizing a steganalysis feature extractor. The results of this experiment indicate that this approach is efficient in identifying both fake and authentic images.

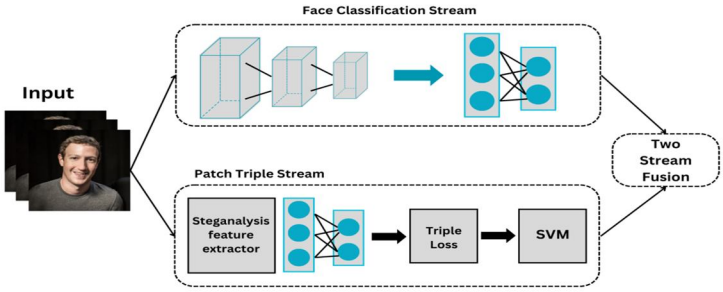


Fig. 4 Two-stream neural networks.

B. Spatiotemporal Feature Analysis

At present, a significant proportion of deepfake detection techniques consider a single frame of a video. It is crucial to understand that video manipulation can involve a variety of frame-level features. Recent research has shown that identifying real from fake videos can be accomplished by examining the temporal sequence between frames. Deepfake detection based on temporal features involves analyzing sequential frames from a video to identify patterns and inconsistencies. In simpler terms, a series of frames can be conceptualized as sequential data. This can be achieved using recurrent neural networks (RNNs). Fig 5. depicts the schematic of a basic RNN. The input data is represented as x, while the processing that occurs in RNN is represented by s. The output of the model is represented by o. "U" represents the input layer weights whereas "V" represents the output layer weights. The processing matrix W, represents the output matrix which will be get added to the input in the subsequent time step. As an example, input "x_0" will undergo processing and the generated output will be used for processing the subsequent data "x_1".

Apart from the approach that relies on Recurrent networks as its main methodology, another technique for detecting modifications in videos uses a CNN-based classification algorithm. This method is known as the Optical Flow approach. In 2019, a new approach was introduced for detecting fake videos using a model based on bidirectional Recurrent neural networks that used DenseNet frames for extracting features. This method represented a re-examination of the use of RNNs for deepfake detection. . Several models, such as ResNet50 and DenseNet, were outperformed by this particular method. In another study, G¨uera and Delp [14] drew attention to discrepancies within individual frames and between different frames of deepfake videos. To address this problem, they introduced a temporal-aware pipeline approach that utilizes CNN layers (Convolutional Neural Networks) and LSTM models for identifying deepfake videos. This method utilizes a convolutional layer to extract features from individual frames. Next, the features are fed into a Long Short-Term Memory network, which produces a descriptor for the temporal sequence. Then, a dense neural network classifies the videos into either fake or categories using this descriptor (Fig. 6). The methodology was evaluated on a dataset of about 500-600 videos, with 50% consisting of fake videos collected from multiple sources and the remaining 50% being real videos sourced from the Hollywood live dataset [15]. During the evaluation process, the methodology achieved an impressive accuracy rate of over 97%.

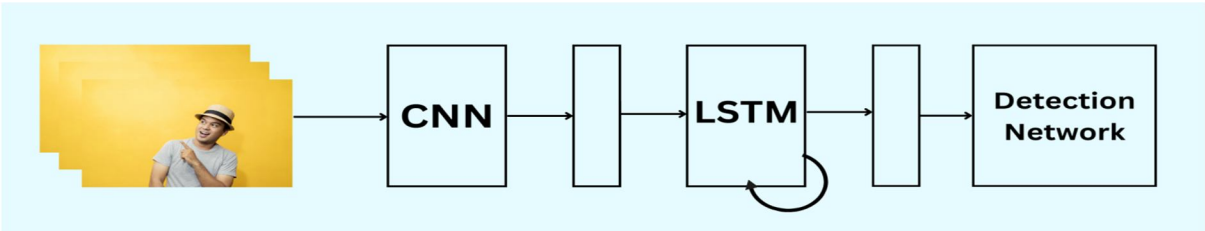


Fig 6. Spatiotemporal features analysis.

C. Biological Signals Analysis

Based on the findings of numerous research studies, it has been established that certain biological signals such as the heartbeat, can serve as a dependable predictor of real videos. In order to accomplish this objective, [17] mentioned a team of researchers has devised a model utilizing Generative Adversarial Networks (GANs), which is capable of analyzing the heartbeat signal in order to accurately identify manipulated videos created through the use of deepfake technology.

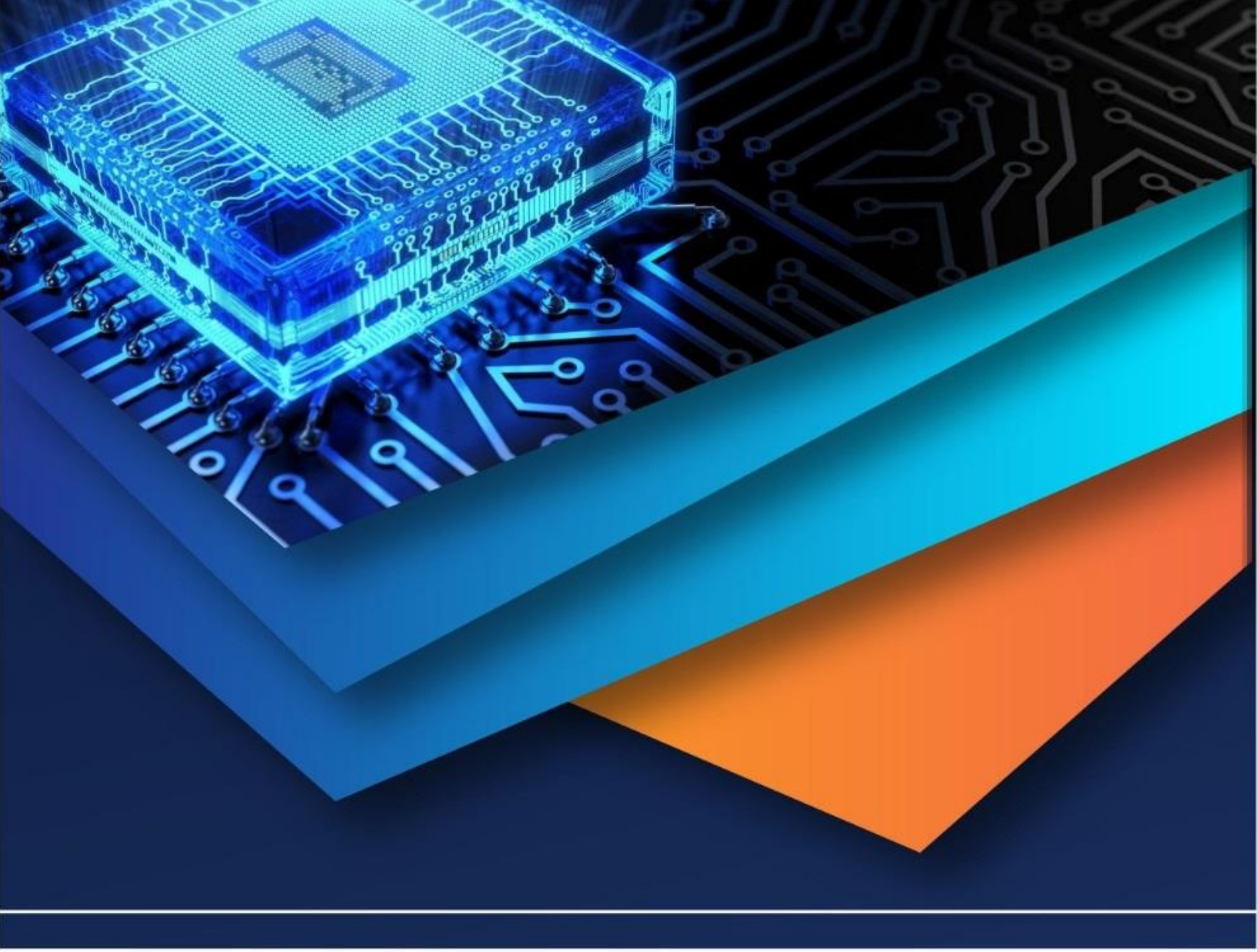
The model takes in authentic videos as inputs. A registration layer is also incorporated, which identifies the facial regions of significance and extracts the corresponding biological signals to produce spatiotemporal windows called PPG cells. These cells may contain multiple faces that were detected using a face detector. Then, a classification layer determines the authenticity of the video. The model's accuracy in identifying deepfakes was 97.3%, according to the authors' evaluation of it using a variety of publicly available datasets. To calculate the heart rate of the subjects in videos, Neural-Ordinary Differential Equations (Neural-ODEs) were used to trained on the original videos [16]. To draw attention to the heart rhythm signals in videos, the authors have suggested using a motion-enhanced spatiotemporal visualization technique. Using the results of the mentioned technique, a spatiotemporal attention network is then used to identify fake videos. Heart rate analysis has also been to detect DeepFake using remote Photoplethysmography (rPPG) [18]. To show that blood flow is present, this technique uses subtle colour changes in human skin.

IV. CONCLUSION

The emergence of deepfake content has led to a decreased trust in media content, as the act of seeing is no longer sufficient for believing. The consequences of this phenomenon are extensive and include the potential for distress and adverse effects on targeted individuals, the aggravation of disinformation and hate speech, and the potential of inciting political unrest and violence. The urgency of this problem is particularly pronounced in the present-day context, because technology has made it quite easy to create modified videos and create deep fake content. Furthermore, social media platforms possess the capability to swiftly circulate such content on a global scale. This paper offers an up-to-date overview of both deepfake creation and detection methods to meet this challenge. This paper examines the current developments and issues in the field of deepfakes, offering useful insights to researchers worldwide who are working on artificial intelligence. It will help them to develop effective strategies to detect deepfakes and protect individuals from the potential harm they can cause.

REFERENCES

[1] The Conversion Article: 32 billion images and 720000 hours of video are shared online daily online.
[2] CNBC Article : So its fine if you edit your selfies
[3] Foreign Affair Article: Deepfakes and the new disinformation war: The coming age of post-truth geopolitics.
[4] T. Hwang. Deepfakes: A grounded threat assessment. Technical report, Centre for Security and Emerging Technologies, Georgetown University, 2020.
[5] E. Meskys, A. Liaudanskas, J. Kalpokiene, and P. Jurcys, “Regulating deep fakes: legal and ethical considerations,” J.Intellect. Prop. Law Pract., vol. 15, no. 1, pp. 24–31, Jan. 2020, doi: 10.1093/jiplp/jpz167.
[6] Faceswap: Deepfakes Software for All. https://github.com/deepfakes/faceswap
[7] DeepFake tf: https://github.com/StromWine/DeepFake tf.
[8] L. Guarnera, O. Giudice, and S. Battiato, “Fighting deepfake by exposing the convolutional traces on images,” IEEE Access, vol. 8, pp. 165085–165098, 2020, doi: 10.1109/ACCESS.2020.3023037.
[9] Chopra P, Junath N, Singh SK, Khan S, Sugumar R, Bhowmick M. Cyclic GAN Model to Classify Breast Cancer Data for Pathological Healthcare Task. Biomed Res Int. 2022 Jul 21;2022:6336700. doi: 10.1155/2022/6336700. PMID: 35909482; PMCID: PMC9334078.
[10] Pavel Korshunov and S´ebastien Marcel. Vulnerability assessment and detection of deepfake videos. In 2019 International Conference on Biometrics (ICB), pages 1–6. IEEE, 2019.
[11] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, Dec. 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630787.
[12] Zhou, P., Han, X., Morariu, V.I. and Davis, L.S. (2017) Two-Stream Neural Networks for Tampered Face Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, 21-26 July 2017, 1831-1839. https://doi.org/10.1109/CVPRW.2017.229
[13] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, “Recurrent Convolutional Strategies for Face Manipulation Detection in Videos,” in IEEE 2019 Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 80–87.
[14] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In 15th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), pages 1–6. IEEE, 2018
[15] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008.
[16] Fernandes S, Raj S, Ortiz E, Vintila I, Salter M, Urosevic G, Jha S (2019) Predicting heart rate variations of deepfake videos using neural ode. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 0–0
[17] Almars, A. (2021) Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications, **9**, 20-35. doi: 10.4236/jcc.2021.95003.
[18] K. N. Ramadhani and R. Munir, "A Comparative Study of Deepfake Video Detection Method," 2020 3rd International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2020, pp. 394-399, doi: 10.1109/ICOIACT50329.2020.9331963.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY