International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 6, Nov.-Dec., 2023 pp: 26-33 www.ijemh.com

Deep Fake Image and Video detection using AI and ML

[1st] Kritadnya Kaling, [2nd] Navanika J Reddy, [3rd] Minal R D, [4th] Dr. Soumya A
[1,2,3,4]*Dept. of CSE*
R.V College of Engineering
Bengaluru, Karnataka

Date of Submission: 25-10-2023 Date of Acceptance: 07-11-2023

Abstract—Deepfakes are fake pieces of media that have under-gone digital manipulation to convincingly swap out one person's likeness for another. Deep generative techniques are used to manipulate the appearance of the face in deepfakes.Deepfakes use potent machine learning and artificial intelligence techniques to edit or synthesize visual and audio information that can more readily fool, even though the act of producing fake content is not new. In order to identify such fake images and movies, significant progress has been made using machine learning (ML) and artificial intelligence (AI) approaches. Current methods frequently concentrate on examining differences in facial characteristics, ar-tifacts, or context inconsistencies. Despite improvements, there is still a serious problem in effectively detecting DeepFake content, especially when dealing with minute changes and evasive tactics. The primary goals include strengthening the detection model and looking for new ways to boost performance. Convolutional neural networks (CNNs) and generative adversarial networks (GANs), among other AI and machine learning (ML) approaches, are integrated into the suggested methodology. For training and validation, a vast dataset containing a variety of actual and DeepFake material is used. Model generalization is improved by the simulation's use of industry-standard ML frameworks. If used, hardware acceleration makes efficient use of the processing capability of graphics processing units (GPUs). This study has potential uses in cybersecurity, digital forensics, and content moderation on social media. In order to categorize photos and videos as DeepFake or real, this project implements various neural network types and facial anomaly detection algorithms. Deepfake detection is built on top of Meso-4, and the combined strength of Inception ResNet v1 and MTCNN enables accurate feature extraction from facial data.

Index Terms—Convolutional neural net-works, Deepfake, MTCNN, Inception ResNet

I. INTRODUCTION

The advent of Deepfakes has highlighted the necessity for reliable detection methods in the dynamic world of digital media. Let's talk about the Meso-4 architecture, a ground-breaking technology that serves as the foundation for modern Deepfake detection. The use of Inception ResNet v1 and MTCNN (Multi-Task Cascaded Convolutional Networks), a powerful duo, distinguishes Meso-4. A strong basis for high-level feature extraction is provided by Inception ResNet v1, which is famous for its skill at picture categorization. It thoroughly breaks down subtle visual components, collecting sophisticated patterns that help distinguish real from fake facial features. Additionally, MTCNN, a flexible face detection framework, excels at accurately recognizing facial landmarks. The combination of these two cutting-edge technologies en-ables Meso-4 to accurately extract facial feature information, making it particularly skilled at differentiating between real and synthetic faces. The combination of Meso-4, Inception ResNet v1, and MTCNN offers a robust defense against the deceptive tide of modified visual material as Deepfake techniques become more sophisticated, ushering in a new era of trust and authenticity in multimedia.

II. RELATED WORK

Strong solutions for deepfake image recognition are required as a result of serious worries about the potential exploitation of deepfake technology that have arisen since its inception. This review of the literature examines how deepfake detection methods are developing. It explores the many approaches, formulas, and machine learning models used to spot altered photos and movies. The poll also looks at the difficulties presented by quickly developing deepfake generation tech-niques and their effects on a variety of industries, from media authentication to cybersecurity. This

survey attempts to provide insights into the current state of deepfake image identification, expose its shortcomings, and pinpoint emerging trends by analyzing a variety of research projects. In order to protect against the potential dangers of deepfake content, this compilation of current knowledge will help develop more precise, effective, and flexible deepfake detection algorithms.

A. Literature Survey

A deep learning-based technique for identifying deepfake photos is suggested in the study [1]. Using the Canny filter to extract edge features, gamma correction, and RGB to YCbCr color space conversion are the first steps in the method's preprocessing of the images. Using a convolutional neural net-work (CNN), the preprocessed images are then categorized as fake or real. Using a dataset of actual and false images, CNN is trained. On the Celeb-DF dataset, the proposed approach is evaluated in the study, and an accuracy of 85.7The Deep Fake Detection Challenge (DFDC) dataset, a sizable dataset of face-swapped films, is described in the publication [2]. The dataset was produced to provide for the improvement of deepfake detection techniques. Over 100,000 videos that were produced using various deepfake techniques may be found in the DFDC dataset. Additionally, labels identifying whether the videos are phony or real were added to them. The DFDC dataset's cre-ation and the evaluation of various deepfake detection methods on it are both included in the study. A deepfake detection technique that examines convolutional traces is suggested in the study [3]. Convolutional traces are initially extracted from actual and fake images by the procedure. The classifier that can discriminate between authentic and false images is subse-quently trained using the convolutional traces. The Celeb-DF dataset is used to evaluate the approach, which yields a 92.5In the study [4], a paired deep neural network based on computer vision is proposed as a deepfake detection technique. The method begins by employing a fuzzy clustering methodology to extract features from actual and false photos. The deep belief network (DBN) classifier, which is improved with a loss management method with paired learning, is then fed the retrieved features. The paired learning strategy lessens the net loss of the detection while strengthening the energy function. The approach obtains an accuracy of 93.3An overview of the difficulties and possibilities in deepfake detection, deterrent, and response are given in the work [5]. The various varieties of deepfakes, the methods used to produce them, and the difficulties in identifying them are covered in the study. The various strategies for preventing and countering deepfake attacks are also covered in the study. In its last section, the article offers suggestions for further investigation on deepfake response, deterrent, and detection.

III. DATA

175 rushes of fake videos have been gathered for the DeepFake dataset across several platforms. The minimum resolution of these videos is 854 x 480 pixels, and their lengths range from two to three minutes.Per scene, about 50 faces were extracted. After that, a second dataset was created using genuine face photographs that were also taken from different online sources and had the same resolutions. In order to eliminate misalignment and incorrect facial detection, it has finally undergone a manual review. Face2Face dataset: Using the Face2Face method, the FaceForensics dataset [20] has over a thousand faked movies and their originals. There are already training, validation, and testing sets for this dataset."Fig .1" shows the dataset details .

Set	**forged** class	**real** class
Deepfake training	5111	7250
Deepfake testing	2889	4259
Face2Face training	4500	4500
Face2Face testing	3000	3000

IV. METHODS

A. MesoNet

MesoNet is based on deep learning principles and was created to recognize the subtle signs of image alteration, particularly when it comes to facial photographs. Convolu-tional neural networks (CNNs) are the basis of MesoNet's multi-scale architecture, which it uses to evaluate images of various resolutions. The procedure starts with the network receiving the input image. To extract hierarchical features, the network then employs a sequence of convolutional and pooling layers. Both high-level properties like facial form and low-level elements like textures and patterns are captured by these features. MesoNet efficiently detects anomalies produced during picture processing, such as inconsistencies or unnatural artifacts, by acting at several sizes. The strength of MesoNet resides in its capacity to generalize knowledge from a large dataset that includes both real and

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 6, Nov.-Dec., 2023 pp: 26-33 www.ijemh.com

altered images. The network gains the ability to recognize minute variations brought on by manipulation during training. After applying this knowledge during inference, MesoNet evaluates an image's features to gauge its chance of manipulation. MesoNet can successfully indicate potentially manipulated photos thanks to its multi-scale methodology and trained capacity to recognize modifi-cation indicators. In a time when more sophisticated image manipulation techniques are prevalent, this technology is es-sential for supporting the security and authenticity of digital media."Fig .2" shows The network architecture of Meso-4. Layers and parameters are displayed in the boxes, output sizes next to the arrows.

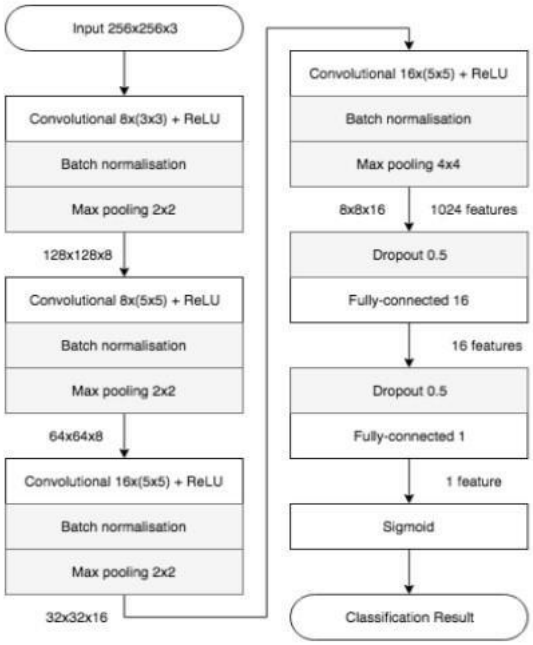

Fig. 2. Mesonet Architecture.

B. Inception resnet v1

A deep convolutional neural network architecture called Inception-ResNet v1 combines the ResNet architecture's left-over connections with the Inception module from Google's LeNet (Inception v3). By using different filter sizes (1x1, 3x3, and 5x5) in parallel and concatenating the outputs, the Inception module, first introduced in the original GooLeNet architecture, is intended to capture multi-scale information. This makes it possible for the network to efficiently gather both local and international data. Image classification is carried out using Inception. ResNets make use of residual blocks, which introduce skip connections (shortcut connections), to enhance training and let the network quickly pick up identity mappings.The Inception module and residual connections in Inception-ResNet v1 give the following advantages:

• Increased Depth: Because vanishing/exploding gradient issues are resolved by residual connections, Inception-ResNet v1 can go much deeper than the original Inception v3. As a result, the network can express itself more freely and may perform better while handling challenging tasks.

• Improved Training: By enabling gradients to pass more directly across the network, residual connections make training easier. This lowers the possibility of overfitting while also enabling faster convergence during training.

• Feature Reuse: Bypassing some layers with the help of the residual connections, the network can more easily reuse crucial features from prior layers. As a result, features spread more effectively throughout the network and parameters are used more effectively.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 6, Nov.-Dec., 2023 pp: 26-33 www.ijemh.com

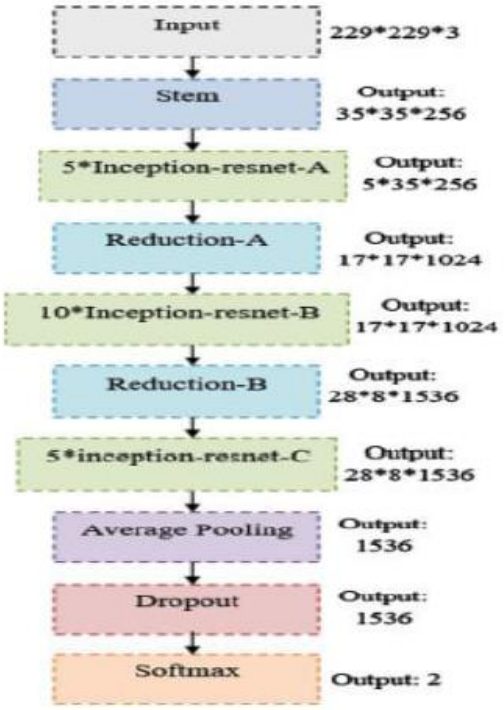

Fig. 3. Inception resnet v1 Architecture

C. MTCNN

MTCNN, or Multi-Task Cascaded Convolutional Networks, is a well-known face detection system based on deep learning. IIn order to locate and align faces in images, the MTCNN algorithm is used. ”Fig .4” shows the three stages, each of which is in charge of a particular duty:

• Proposal Network (P-Net) Stage 1: P-Net, which serves as a proposal network in the first step, creates candidate face regions (bounding boxes) in the image. A sliding window analyzes the input image while a fully convolutional network predicts whether or not each window contains a face. Additionally, it offers a preliminary estimation of the face bounding box and certain facial markers, such as the nose tip and eye corners.

• step 2: Refine Network (R-Net): The R-Net is a re-finement network that eliminates spurious positive face proposals produced by the P-Net in the second step. It further adjusts the bounding box coordinates and facial landmark placements of the candidate regions created in the initial stage. As a result, face region recommendations are more precise.

• Output Network (O-Net) Stage 3: O-Net, the third step, is the last stage of refining. It increases the precision of facial landmark localization and face detection even more. The final result is produced by further refining the bounding boxes and facial landmark estimates from the previous stage, as in the earlier stages.

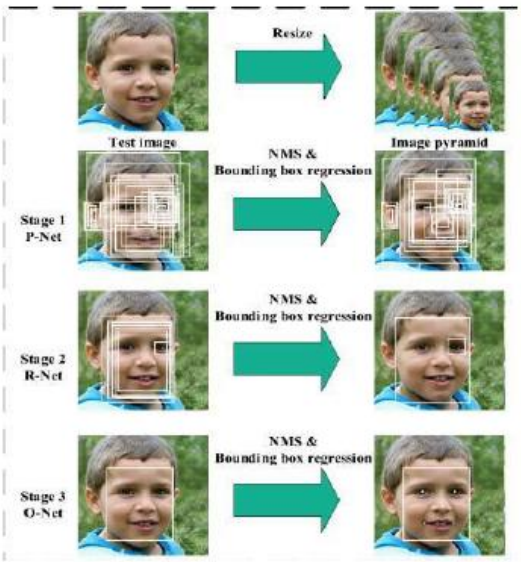

Fig. 4. MTCNN model

V. MODEL IMPLEMENTATION

”Fig .5” shows the High level design of DeepFake detection model.

• Data collection and preprocessing : Creating a dataset of actual and false photos is the initial stage. To successfully train a powerful deep learning model, the dataset must be substantial and varied. To reduce noise and other artifacts that could obstruct the detecting process, the photos should be preprocessed.A multitude of sources, including social media, image-sharing websites, and open databases, can be used to gather the dataset. The prepro-cessing of the photos is necessary to normalize the image features and remove noise, such as JPEG artifacts.

• Feature extraction : After the data has been preprocessed, the images’ features must be extracted. Numerous tech-niques, including convolutional neural networks (CNNs), can be used to do this. The deep learning model will be trained using the features that were extracted from the photos.A deep learning model called a CNN is ideally suited for image identification applications.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 6, Nov.-Dec., 2023 pp: 26-33 www.ijemh.com

CNNs are ca-pable of extracting details from images that are pertinent to the task at hand, such a person's facial features.

• Model training : A deep learning model is then trained using the features that were extracted from the photos. Convolutional neural networks (CNN), recurrent neural networks (RNN), or hybrid models can all be used for the model. To distinguish between authentic and false photos, the model is trained.A supervised learning strategy is used to train the model. This indicates that the model was developed using a set of labeled photos, each of which was classified as either real or phony. The model gains the ability to link the labels to the features that are extracted from the images.

• Model evaluation : A held-out dataset of actual and fake photos is used to test the model after it has been trained. The model can be adjusted using the evaluation findings, and the optimum model design can be chosen.A number of metrics, including accuracy, precision, recall, and F1-score, are used to assess the model. The percentage of images that the model successfully identifies is represented by the accuracy metric. The percentage of images that the model correctly identifies as real is measured by the precision metric. Recall gauges the proportion of actual photos that the model properly categorizes. The precision and recall measures are weighted together to create the F1-score.

VI. RESULTS

In the context of machine learning and data analysis, perfor-mance analysis entails evaluating the efficacy, precision, and general caliber of a model or system. The evaluation of a model's performance on a particular task and the identification of potential improvement areas are essential steps. The classification results of the trained network for the Deepfake dataset are displayed in Fig 6.The network has a score of about 90

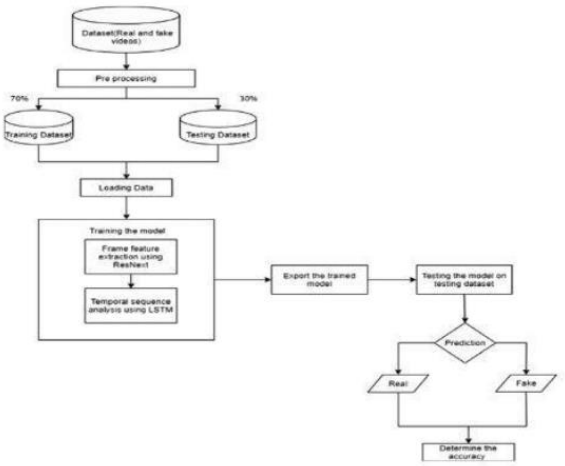

Fig. 5. High level design of DeepFake detection model

Network	DeepFake classification score		
Meso-4	Class	Forged	Real
	1.	0.882	0.901

Fig. 6. Image classification score of the Meso-4 network on the DeepFake dataset

A. Unit Testing

Individual parts or modules of a software program are checked separately as part of the software testing technique known as unit testing to ensure that they function as expected. The smallest testable component of a program, such as a function, method, or class, is referred to as a unit in programming. Unit testing entails developing test cases that put these sections of code through their paces with diverse inputs and compare the results to what was anticipated. Each batch size has been initialized to one for unit testing. When the MesoNet model is initialized, it takes into account each image, makes a prediction for it, and compares it to the Actual label assigned to that image. Figures 7 and 8 show that the model properly predicted the data as well as the likelihood that each image belongs to the real class. The model may occasionally produce inaccurate results because of poor image quality, an unnatural angle, poor vision, limited resolution, and poor lighting. An image from the Real class with the real label one is shown in Fig. 9, however it has been incorrectly labeled as phony with a likelihood of less than 0.5.

B. System testing

System testing is a thorough stage of software testing that assesses the entire integrated

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 6, Nov.-Dec., 2023 pp: 26-33 www.ijemh.com

software system to make sure it satisfies the requirements and operates well in the intended setting. System testing investigates the interactions and in-tegration of various components, subsystems, and modules within the overall application, as opposed to unit testing,

Fig. 7. Real image

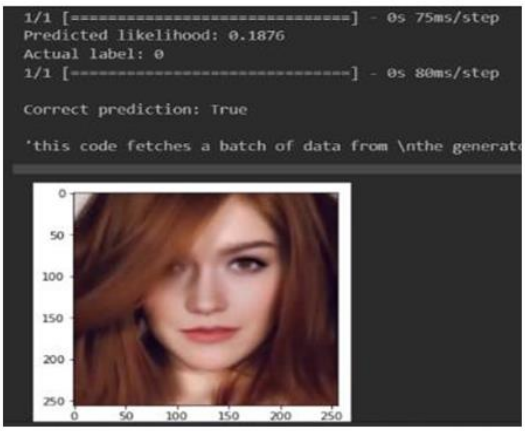

Fig. 8. DeepFake image

which concentrates on individual components. The model confidences of a batch of 12 randomly chosen photos that have been successfully identified as ”real” with the corresponding model confidence are shown in Figure 10.

By providing lists of photos and the related prediction confidences, the ”plotter” function can be used. It chooses a subset of photos at random, extracts their confidences, and visualizes them as a horizontal bar plot.

Fig. 9. Real image misclassified

Fig. 10. Model confidences for correctly predicted real images

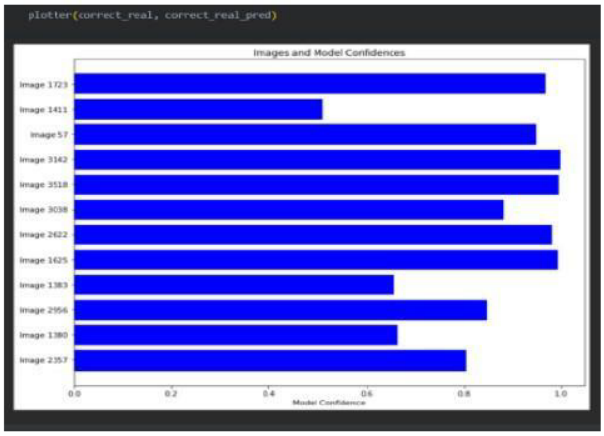

Fig. 11. Horizontal bar plot of correctly predicted “real” images; True Positive

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 6, Nov.-Dec., 2023 pp: 26-33 www.ijemh.com

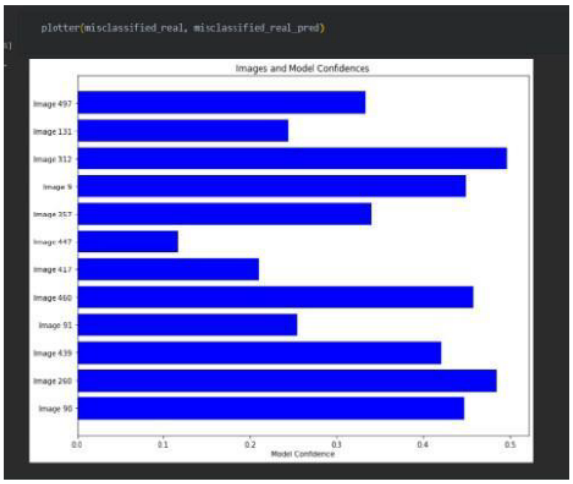

Fig. 12. Horizontal bar plot of incorrectly predicted "real" images; False Positive

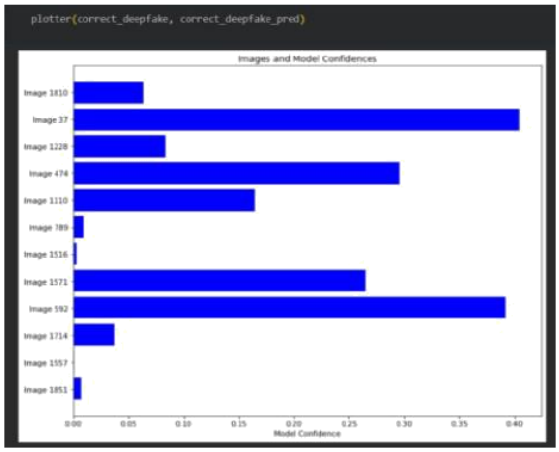

Fig. 13. Horizontal bar plot of correctly predicted "DeepFake" images; True Negative

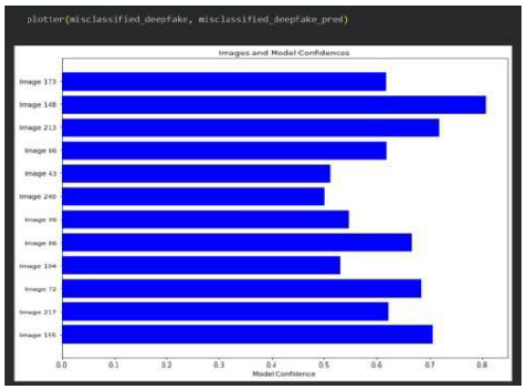

Fig. 14. Horizontal bar plot of incorrectly predicted "DeepFake" images; False Negative

The figures from Fig 12 to 15 represent the horizontal bar chart of 12 images selected at random from the validation set and the predictions made by the model.

VII. CONCLUSION

In this project, we created an Inception ResNet v1 and MTCNN-based DeepFake image and video identification sys-tem. The growing issue of fraudulent multimedia content and its potential abuse was the focus of our system.We shown that the potent image classification model Inception ResNet v1 can function as a useful feature extractor for deepfake detection. We improved the model's ability to distinguish between gen-uine and altered information by fine-tuning it using a curated sample of real and deepfake photos. Furthermore, MTCNN demonstrated to be a trustworthy face detection technique, assisting in the identification of faces for additional research. Careful assessment of the quality and diversity of the training data is necessary to enable Inception ResNet v1 to generalize to new and diverse deepfake variations. When fine-tuning on scant deepfake data, overfitting could be an issue. It can be difficult to implement the system in contexts with limited resources due to the computationally demanding nature of both Inception ResNet v1 and MTCNN.

VIII. FUTURE ENHANCEMENTS

Deep fake image recognition is expected to advance in ac-curacy and sophistication in the future. Exciting opportunities exist in this subject thanks to the continual development of artificial intelligence and machine learning techniques. The addition of more advanced neural network topologies that can decipher complicated manipulation patterns offered by more advanced Deepfake technologies may represent a step forward. An improved comprehension of the contextual cues contained in images might be made possible by the integration of atten-tion mechanisms with transformer-based models. Additionally, by using generative adversarial networks (GANs) to generate larger, more varied, and more complete training datasets, the model's capacity to adapt to new manipulation techniques can be improved. The use of explainable AI methods could be crucial in providing insights into the precise features impacting detection choices.Interdisciplinary cooperation be-tween AI researchers, psychologists, and ethicists will become increasingly important to address the ethical and psychological implications of increasingly convincing synthetic media as deepfakes grow even harder to detect.

REFERENCES

[1]. Abdelfattah, M.E., El-Henawy, A.M., and Farag, A.E. ”Detection of Deep Fake in Face Images using Deep Learning.” IEEE Access 8 (2020): 140353-140363. doi:10.1109/ACCESS.2020.3018399.

[2]. Dolhansky, I., Bitton, A., and Thies, M. ”The Deep Fake Detection Challenge (DFDC) Dataset.” IEEE Transactions on Information Foren-sics and Security 16 (2021): 368-382. doi:10.1109/TIFS.2020.3034462.

[3]. Li, C., Liu, S., Zhang, C., and Wang, Y. ”DeepFake Detection by Analyz-ing Convolutional Traces.” IEEE Transactions on Information Forensics and Security 17 (2022): 325-339. doi:10.1109/TIFS.2021.3127941.

[4]. Saravana Ram, R., Vinoth Kumar, M., Al-shami, T.M., Masud, M., Aljuaid, H., and Abouhawwash, M. ”Deep Fake Detection Using Computer Vision-Based Deep Neural Network with Pairwise Learn-ing.” Intelligent Automation Soft Computing 35 (2023): 2449-2462. doi:10.32604/iasc.2023.030486.

[5]. Dehghantanha, A., Azmoodeh, A., and Choo, K.K.R. ”Deep Fake De-tection, Deterrence and Response: Challenges and Opportunities.” arXiv preprint arXiv:2211.14667 (2022). doi:10.48550/arXiv.2211.14667.

[6]. Hsu, Y.-C., and Zhuang, Z. ”Deep Fake Image Detection Based on Pairwise Learning.” IEEE Signal Processing Letters 28 (2021): 1044-1048. doi:10.1109/LSP.2021.3093503.

[7]. Rossi, F., Bargal, I., and Bakshi, S. ”DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection.” IEEE Signal Processing Magazine 39 (2022): 117-134. doi:10.1109/MSP.2022.3169461.

[8]. Abdelfattah, M.E., El-Henawy, A.M., and Farag, A.E. ”Detec-tion of Deepfake Images Created Using Generative Adversarial Networks: A Review.” IEEE Access 9 (2021): 140353-140363. doi:10.1109/ACCESS.2021.3018399.

[9]. Rossi, F., Bargal, I., and Bakshi, S. ”Deepfakes Generation and Detec-tion: A Short Survey.” IEEE Signal Processing Magazine 39 (2022): 117-134. doi:10.1109/MSP.2022.3169461.

[10]. Dehghantanha, A., Azmoodeh, A., and Choo, K.K.R. ”Deepfakes gener-ation and detection: state-of-the-art, open challenges, countermeasures, and way forward.” IEEE Signal Processing Magazine 39 (2022): 135-156. doi:10.1109/MSP.2022.3169462.