DEEPFAKE DETECTION IN VIDEO USING INTEGRITY VERIFICATION METHOD

V. Premanand, V. Arulalan, G. Divya and S. Sandhya
Department of Information Technology, Madras Institute of Technology Campus, Anna University, Chennai, INDIA

Abstract— **The new advancement in deep generative networks has remarkably improved the quality and efficiency of generating realistically-looking fake face videos. Deepfakes or artificially generated visual renderings can be intended to accuse a public figure or influence a person's opinion. With the recent discovery of Generative Adversial Network(GAN), an aggressor employing a traditional PC can create renditions that are realistic enough to fool the observer. Disclosing the deepfakes is thus turning into a necessary precaution for reporters, social media platforms, and the general public. In this method, the eye blinking which is a physiological signal that is not well presented in the synthesized fake videos and head pose estimation in the videos which reveal facial landmarks specific to each person are analyzed. The eye blinking is detected using an improved aspect ratio algorithm and the General CNN-based classifier has been extended by incorporating the temporal relationship between consecutive frames which discloses the number of eye blinks, as eye blinking is a temporal process. The deepfakes are created by splicing the central part of the synthesized face and merging synthesized face regions into the original image. Head pose estimation is based on these observations and this introduces errors that can be revealed when head poses are estimated from the face images. Using features based on this condition and using a set of real face images and Deepfakes, an SVM classifier is evaluated. This method is tested over benchmarks of video detection dataset namely UADFV, Celeb-df(v2), and shows an accuracy rate of 82% in detecting videos that are deepfakes.**

Keywords: Temporal information, facial feature detection, eye blink detection, Alignment Error Classification, Support Vector Machine Classifier.

I. INTRODUCTION

The advancement of AI and technologies have led to the growth of synthesized videos.The front line of this pattern is the alleged Deep Fakes. The highly regarded "DeepFake" is referred to a deep learning based technique that is able to produce fake videos by swapping the face of an individual with the face of another person. The term "Deepfake" is a combination of "Deep Learning " and "Fake", which means the fake videos are created with the assistance of Deep Learning methods to create more realistic videos. These are generated by swapping synthesized faces into original images and videos using deep neural networks. Fake images and videos as well as facial data generated by digital manipulation, especially with DeepFake ways, became a good public concern recently. Due to this increased technology, there is a very high chance of spreading misinformation and also it becomes very difficult to differentiate false information from true ones. These Deepfakes matter more recently as they can be used to manipulate and threaten individuals and organizations. But by understanding this technology better, corporate organizations can protect themselves from manipulators. Historically, journalists have played an important role in the critical examination of information before publication and dissemination. However, with misleading information, it becomes harder to verify journalists thus creating fear among the public.

Lately, utilizing a normal PC with an inbuilt GPU an attacker can make realistic videos that are unable to differentiate from true information which can deceive the common people. To create a deepfake video, the person's face is swapped and replaced with another person, utilizing the facial recognition algorithm and deep learning network called Variational auto-encoder [VAE]. Autoencoders are one type of neural network that consists of an encoder and a decoder. Encoder decreases the image to a lower-dimensional latent space, whereas the image from the latent representation is regenerated by the decoder. This architecture is utilized by Deepfakes which uses a universal encoder to code a person into the latent space. The key features of facial features and body posture of the person are contained in the latent representation. This can be decoded with a model trained explicitly for the target.

II. RELATED WORK

Liu et al. [1] proposed an unsupervised image-to-image translation of 28 frameworks based on Coupled GANs, with the intent to learn the joint distribution of images in various areas. To address the problem, they make a shared-latent space assumption and propose an unsupervised image-to-image translation framework based on Coupled GANs. When analyzing the image translation problem from a probabilistic modeling perspective, the key challenge is to learn a joint distribution of images in different domains. In the unsupervised setting, the two sets consist of images from two marginal distributions in two different domains, and the task is to infer the joint distribution using these images. They compare the proposed framework with competing approaches and present high-quality image translation results on various challenging unsupervised image translation tasks, including street scene image translation, animal image translation, and face image translation. They also apply the proposed framework to domain adaptation and achieve state-of-the-art performance on benchmark datasets.

In [2] Goodfellow et al, introduced a novel approach to training generative models using a two-player minimax game. The proposed approach, known as Generative

Adversarial Networks (GANs), consists of a generator and a discriminator that are trained together in an adversarial setting. GANs have had a significant impact on the field of generative modeling, and have been successfully applied to a wide range of tasks such as image and video synthesis, natural language generation, and drug discovery.Despite their success, GANs also have some limitations. They can be difficult to train and require careful hyperparameter tuning to ensure convergence and also provide a significant contribution to the field of generative modeling and introduced a new paradigm for training generative models using adversarial learning

Building a good generative model of natural images has been a fundamental problem within computer vision. However, images are complex and highly dimensional, making them hard to model well, despite extensive efforts. Given the difficulties of modeling an entire scene at high resolution, most existing approaches instead generate image patches. So Denton et al. [3] propose an approach that can generate plausible looking scenes at 32×32 and 64×64. To do this, we exploit the multiscale structure of natural images, building a series of generative models, each of which captures image structure at a particular scale of a Laplacian pyramid. By training the conceptually simple generative model that can produce high-quality sample images that are qualitatively better than other deep generative modeling approaches. While they exhibit reasonable diversity, we cannot be sure that they cover the full data distribution. A key point in this work is giving up any "global" notion of fidelity, and instead breaking the generation into plausible successive refinements. The author states that many other signal modalities have a multiscale structure that may benefit from a similar approach.

The success of convolutional neural networks (CNNs) in supervised image classification tasks has been remarkable, but they face challenges in unsupervised learning. Alec Radford et al. [4] proposed a class of CNNs called Deep Convolutional Generative Adversarial Networks (DCGANs) that have architectural constraints and show promise in unsupervised learning. DCGANs bridge the gap between supervised and unsupervised learning and can be used for image classification with competitive performance. The DCGAN's algorithm evaluates a set of constraints on the architectural topology of Convolutional GANs that make them stable to train in most settings and give that adversarial networks learn good representations of images for supervised learning and generative modeling. Further research is needed to address instability. Extending this framework to other domains such as video and audio would be interesting.

The Wasserstein GAN (WGAN) paper by Arjovsky et al. [5] presents a novel technique for training Generative Adversarial Networks (GANs). The authors propose the Wasserstein distance as a metric for measuring the similarity between the true and generated probability distributions. They introduce the WGAN algorithm that uses the Wasserstein distance in the objective function, which provides several benefits over the traditional GANs.The paper highlights the limitations of traditional GANs, such as mode collapse, instability in training, and difficulty in assessing the quality of generated samples. The WGAN algorithm solves these problems by providing stable training, continuous estimation of the Wasserstein distance, and more informative learning curves.

Isola et al. [6] investigated the use of conditional adversarial networks as a general solution to problems with image-to-image translation, it also has some potential demerits, the cGAN approach proposed in the paper tends to overfit to the training data, resulting in limited generalization to new and unseen data.In some cases, the generator of the cGAN may collapse to produce a limited set of outputs, resulting in a lack of diversity in the generated image. Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. In addition to learning how to map input images to output images, these networks also pick up on how to train loss functions. This enables the use of a generic approach to situations that would ordinarily need very specific loss formulas. Among other tasks, the method is efficient at synthesizing photographs from label maps, reconstructing objects from edge maps, and coloring images

In [7] Kalicharan Jalui et al. designed a new method to detect and report the existence of deep fake content in digital video. This method achieves high detection accuracy, outperforming other methods in terms of both false positive and false negative rates. However, the paper also has some limitations. One potential limitation is the use of a single dataset to evaluate the performance of the proposed method. Additionally, the paper does not discuss the potential ethical or social implications of deepfake detection and its impact on privacy and freedom of speech. They have implemented the deep learning model by using a pre-trained ResNext CNN model to extract the frame level features and LSTM for training the model for video classification. The proposed method is capable of detecting the video as a Deepfake or Real with good accuracy on unseen data

III. METHODOLOGY

Our architecture is based on performing integrity verification by tracking significant changes in the eye blinking pattern of a subject along with the head pose estimated in the video. We are detecting faces in each frame of the video and regions corresponding to each eye are extracted to form a stable sequence. Features of the mouth and nose are detected for performing head pose estimation. After these pre-processing steps, eye blinking is detected by quantifying the degree of openness of an eye in each frame by sequence learning. Head poses are estimated using the difference in the facial landmarks from the whole face and those only from the central face region. The alignment error is revealed as differences within the head pose. This difference in the head poses and eye blinking pattern is then fed to an SVM classifier to differentiate the original one from the Deep Fake. The system architecture is shown in Fig 1.

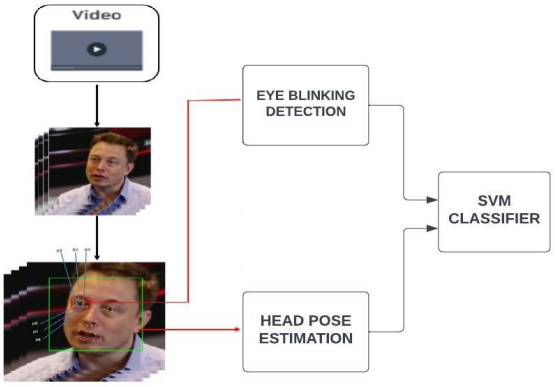


Fig 1 - System Architecture

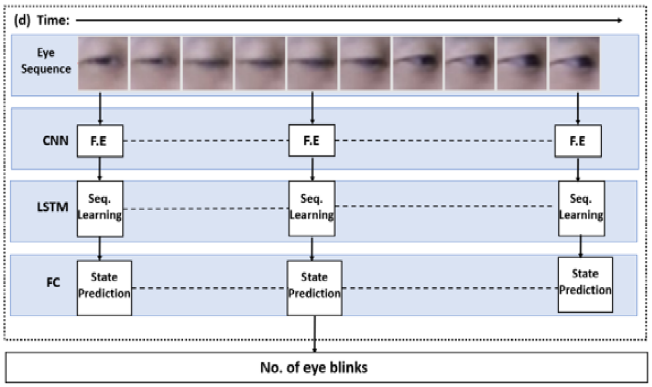


Fig 2 - Eye Blinking Detection

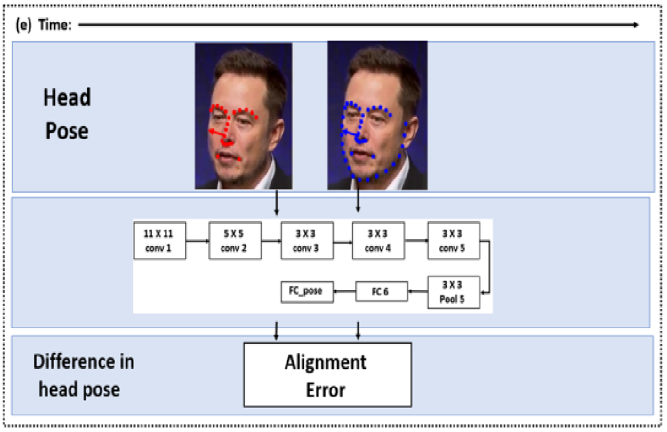


Fig 3 - Head Pose Estimation

A. Facial Features Detection:

The HyperFace algorithm is used for detecting the face region in the given image. In the Pre-Process, it performs data type conversion, image resizing, and normalization. The first module generates region -proposals that are independent of class from the given image and scales the images to 227 × 227 pixels. This is passed through the second module of HyperFace which is CNN. This module makes use of the candidate regions that are resized in the previous module and classifies them whether they are face or non-face regions. In case if the region gets categorized as the face, then the bounded box is drawn across the face and passed through the next module for extracting individual face features. The face features are plotted using 68 feature points which detect include the left and right eyes, left and right eyebrows, the nose, the mouth, and the Jawline of the face. Here for this method, we require only face regions such as the eyes and nose. The particular features from the face are extracted by splicing the array that contains those features. In R-CNN, the Selective Search algorithm is utilized for creating face region proposals in an image. If the Intersection over Union (IOU) overlap with the candidate region is more than 0.5 then it is a positive sample and it contains a face. If the IOU overlap with the candidate region is less than 0.35 then it is a negative sample and other regions are omitted. The softmax loss function is used for training the face detection task.

B. Eye Blink Detection

The Eye Tracker was implemented based on AR (Aspect-Ratio). AR takes six points around the eyes and calculates the area of the horizontal axis and vertical axis. In general, eye blinks occur at the same time in both eyes. The equation has been proposed that sums and divides in half the eye's ratio by using the value of the left eye and right eye. When eyes are closed, the eye's ratio is reduced and drops below the threshold in the consecutive video frames. Then, when the eyes are open, the eye's ratio is restored as before. That is when the eye is in the open state the aspect ratio of the eye is found to be a constant value, but the value abruptly falls approximately to 0 when the eye is in the closed state. In this process, the time required to blink, and blinking frequency is obtained. But the main drawback of the aspect ratio method is they are dependent on the eye landmarks found. Also in many cases, it is found that eye landmarks are not reliable as in some frames the eye area is too small to find the correct aspect ratio value which perplexes finding the eye state. By only depending on the image domain the eye state cannot be determined correctly. By using the temporal domain, the states of the eye can be memorized. Current strategies use Convolutional Neural Networks (CNN) as a binary classifier to differentiate the open and closed eye state of every frame.

However, CNN generates predictions depending on a single frame, that doesn't leverage the information in the temporal domain. As human eye blinking has a powerful temporal correlation with previous states, we tend to use Long- Recurrent Convolutional Neural Networks (LRCN) that distinguish open and closed eye states with the thought of previous temporal information. The Eye tracker algorithm is improved for detecting the eye blinking pattern by making the algorithm incorporate the temporal relationship between consecutive frames which memorizes the previous state to produce output based on sequence learning. If the eye blink has occurred in the previous frames, then the state of the eye in the next set of frames is likely to be in the open state. If the eye has been in the open state in previous frames, the state of the eye is most likely to be the open state for the next frames. This improvisation in the algorithm would result in less time in detecting the state with better accuracy.

C. Alignment Error Classification:

It is well-known that Deepfake takes only the central face region and swaps faces whereas the outer contour of the face will remain the same. This inconsistent match between the landmarks at the center region and the outer face region of faked faces is an alignment error. We consider the head orientation vector and the difference between the headpose of the central face region and the whole face region to be small in real images, but large in fake images. Ra is denoted as the rotation matrix calculated using facial landmarks from the entire face, and Rc as the one calculated solely using landmarks within the central region. we tend to acquire the 3D unit vectors Ta and Tc equivalent to the orientations of the head calculated. Then cosine distance between the 2 unit vectors Ta and Tc is compared, which takes value in [0, 2] with zero which means the 2 vectors accept as true with one another. The smaller this value is, the nearer the 2 vectors are to every alternative. The cosine distances of the 2 estimated head pose vectors for the real pictures concentrate on a considerably smaller range of values up to 0.02, whereas for Deepfakes the majority of the values are within the range between 0.02 and 0.08. The variation within the distribution of the cosine distances of the 2 head orientation vectors for real and Deepfakes counsel that they can be differentiated based on this cue.

D. Pose Estimation:

Head pose Estimation research in computer vision focuses on the prediction of the pose of an individual's head in a picture. In computer vision, the pose of an object refers to its relative direction and location with respect to a camera. The pose can be altered by either moving the object to the camera or the camera to the object. Specifically, it considers the prediction of the Euler angles of an individual's head. The Euler angles consist of 3 values: yaw, pitch, and roll. These 3 values describe the rotation of an object in a 3D area. By accurately predicting these 3 values, we can understand the direction an individual's head is facing. The 3D head pose relates to the rotation and translation of the world coordinates to the camera coordinates. In particular, [U, V, W] T be the world coordinates of one facial landmark, [X, Y, Z] T be its camera coordinates, and (x, y) T be its image coordinates. The change between the world and the camera coordinate systems can be figured out.

In 3D head pose assessment, we need to tackle the reverse problem of estimating s, R, and ~t utilizing the 2D image coordinates and 3D world coordinates of the same set of facial landmarks got from a standard model, e.g, a 3D average face model, expecting we know the camera boundary. In particular, for a group of n facial landmark points, this can be defined as an improvement issue that can be settled efficiently using the Levenberg-Marquardt algorithm. The assessed R is the camera pose which is the rotation of the camera concerning the world coordinate, and the head pose is acquired by reversing it as RT (as R is an orthonormal matrix). Having a computer helps us to understand the direction an individual's head is facing and this provides several helpful applications. For example, it will be accustomed to plot a 3D object to join the direction of the head almost like those seen in TikTok, Snapchat, and Instagram filters. Additionally, it may be utilized in self-driving cars to trace whether or not a driver is that specialized in driving.

The alignment error of head pose and eye blinking data is passed to the SVM classifier as a feature Vector which detects whether the given video is deepfake or real. These features are flattened into a vector, which is standardized by subtracting its mean to predict the desired result of deepfake or real.

IV. EXPERIMENT

A. Dataset

UADFV[33] dataset contains a total of 98 videos. From YouTube videos 49 are real videos and 49 videos are fake ones generated by FakeAPP.Celeb-DF(v2) [32] dataset contains real and synthesized videos. The Celeb-DF (v2) dataset greatly extends from Celeb-DF (v1), which contains 795 DeepFake videos.Celeb-DF includes 590 original videos collected from YouTube with subjects of different ages, ethnic groups, and genders and 5639 corresponding DeepFake videos.

B. Eye Blinking Evaluation Metrics

The number of eye blinks in the video is found by incorporating the temporal relationship between consecutive frames which memorizes the previous state to produce output on the basis of sequence learning. To evaluate the videos using the number of eye blinks we will be using the following metrics and as shown in Table 5.1 If the Eye Blinking Rate is 4-6 blinks, then it is a real video If the Eye Blinking Rate is 0-2 blinks, then it is a fake video.

File Category	Average video Length	Eye blinking rate
Real videos	10-11 sec	4-6 blinks
Fake videos	10-11 sec	0-2 blinks

Table 5.1 Eye Blinking Evaluation metrics

C. Headpose Evaluation Metrics

The headpose is estimated and cosine differences between the whole face and central region are found and taken as alignment error. To evaluate the videos using the alignment error we will be using the following metrics and as shown in Table 5.2 If the alignment error value e <0.02 then it is a real video If the alignment error value 0.02-0.08 then it is a fake video.

File Category	Average video Length	Alignment error
Real videos	10-11 sec	<0.02
Fake videos	10-11 sec	0.02-0.08

Table 5.2 Headpose Evaluation Metrics

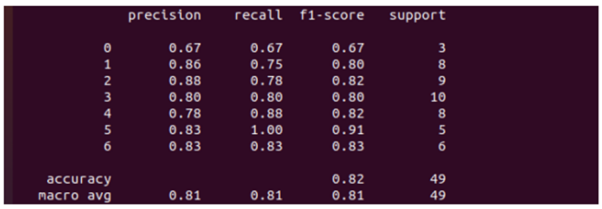
D. Overall Evaluation Metrics

Alignment error of head pose and eye blinking data are passed to SVM classifier as feature Vector which detects whether the given video is deep fake or real. The output metrics are shown in Table 5.3

Eye blink rate	Alignment error (AE)	Probability	Classification
4-6	AE<0.02	0.5<=x<=1	Real
0-2	0.02<=AE<=0.08	0<=x<=0.5	Pristine

Table 5.3 Overall Evaluation metrics

E. Performance Observation



	precision	recall	f1-score	support
0	0.67	0.67	0.67	3
1	0.86	0.75	0.80	8
2	0.88	0.78	0.82	9
3	0.80	0.80	0.80	10
4	0.78	0.88	0.82	8
5	0.83	1.00	0.91	5
6	0.83	0.83	0.83	6
accuracy			0.82	49
macro avg	0.81	0.81	0.81	49

Fig 4 - Performance Observation

It is observed that the accuracy of our Deepfake classifier comes as 0.82 and the macro average is obtained as 0.81.

V. CONCLUSION AND FUTURE WORK

In this work, we express a new method to expose fake face videos which are generated with neural networks.This method is based on the detection of eye blinking which is a spontaneous and unconscious human reflex function along with head pose estimation in the videos, which is a physiological signal that is not well presented in the synthesized fake videos that can be used as an approach to detect the Deepfakes. This work can be improved through several measures because cyber-security attacks and defense evolve continuously. Firstly, we will explore other deep neural network architectures for more effective methods to detect closed eyes and roll, yaw, and the pitch of head pose instead of pickle file, and also explore other types of physiological signals to detect fake videos. The proposed system was implemented for stored videos. As future work, the proposed system can be scaled up to support live video or streaming video from an IP camera for a live feed.

REFERENCES

[1] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in NIPS, 2017, pp. 700–708.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672– 2680.

[3] E. L. Denton, S. Chintala, R. Fergus et al., “Deep generative image models using a laplacian pyramid of adversarial networks,” in Advances in neural information processing systems, 2015, pp. 1486–1494.

[4] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[5] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” arXiv preprint arXiv:1701.07875, 2017.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint, 2017.

[7] K. Jalui, A. Jagtap, S. Sharma, G. Mary, R. Fernandes and M. Kolhekar, "Synthetic Content Detection in Deepfake Video using Deep Learning," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 01-05, doi: 10.1109/GCAT55367.2022.9972081.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818-2826, doi:10.1109/CVPR.2016.308

[9] D. Yadav and S. Salmani, "Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 852-857, doi:10.1109/ICCS45141.2019.9065881

[10] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, N. D. Bokde, J. H. Yoon and Z. W. Geem, "An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture," in IEEE Access, vol. 8, pp. 101293-101308, 2020, doi: 10.1109/ACCESS.2020.2998330

[11] Ji, Q., Li, X., Qu, Z. and Dai, C., 2019. Research on urine sediment images recognition based on deep learning. IEEE Access, 7, pp.166711-166720.

[12] Kang, R., Liang, Y., Lian, C. and Mao, Y., 2018. CNN-based automatic urinary particles recognition. arXiv preprint arXiv:1803.02699.

[13] Sun, Q., Yang, S., Sun, C. and Yang, W., 2018, May. An automatic method for red blood cells detection in urine sediment micrograph. In 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC) (pp. 241-245). IEEE.

[14] Aglibot, K.P., Angeles, J.A., Gecana, J.F., Germano, A.B., Macalindong, J.A. and Tolentino, R.E., 2022, November. Urine Crystal Classification Using Convolutional Neural Networks. In 2022 International Visualization, Informatics and Technology Conference (IVIT) (pp. 245-250). IEEE.

[15] F. Chollet, "Xception: Deep Learning with Depth Wise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Honolulu, HI, USA, 2017, pp. 1800-1807.
doi: 10.1109/CVPR.2017.195

[16] Y. Li, M. -C. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1-7. doi: 10.1109/WIFS.2018.8630787

[17] A. Tewari et al., "High-Fidelity Monocular Face Reconstruction Based on an Unsupervised Model-Based Face Autoencoder," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 357-370, 1 Feb. 2020. doi: 10.1109/TPAMI.2018.2876842

[18] M. -Y. Liu, X. Huang, J. Yu, T. -C. Wang and A. Mallya, "Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications," in Proceedings of the IEEE, vol. 109, no. 5, pp. 839-862, May 2021. doi: 10.1109/JPROC.2021.3049196

[19] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6. doi: 10.1109/AVSS.2018.8639163

[20] H. Rajaguru and V. K. Bojan, “Performance Analysis of Em, Svd and Svm Classifiers in Classification of Carcinogenic Regions of Medical Images,” International Journal of Imaging Systems and Technology, vol. 24, no. 1, pp. 16-22, 2014.

[21] S. C. Tan, C. P. Lim, R. F. Harrison, and R. L. Kennedy, “A New RBFNDDA-KNN Network and Its Application to Medical Pattern Classification,” in Soft Computing in Industrial Applications. Springer, 2014, pp. 71-82.

[22] A. Chintha et al., "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 1024-1037, Aug. 2020. doi: 10.1109/JSTSP.2020.2999185

[23] L. Guarnera, O. Giudice and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 2841-2850.
doi: 10.1109/CVPRW50498.2020.00341

[24] H. R. Hasan and K. Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts," in IEEE Access, vol. 7, pp. 41596-41606, 2019.
doi: 10.1109/ACCESS.2019.2905689