IMPROVING DEEP FAKE DETECTION USING DEEP LEARNING

Manvi Tandon
B. Tech CSE
Bennett University
Greater Noida, India
E21CSEU0415@bennett.edu.in

N Bala Yashaswini
B. Tech CSE
Bennett University
Greater Noida, India
E21CSEU0432@bennett.edu.in

Naman Agarwal
B. Tech CSE
Bennett University
Greater Noida, India
E21CSEU0952@bennett.edu.in

Abstract— Deep fake creation has evolved by leaps and bounds in recent years and can be used to spread disinformation around the world, which will soon pose a serious threat. Deepfakes are synthesized audio and video content created using artificial intelligence algorithms. It is common practice to use videos as evidence in legal disputes and criminal courts. Especially as deep fakes become more difficult to create, this is expected to become a difficult task. The project "Improving Deepfake Detection Using Deep Learning" aims to address the escalating threat of deepfake technology by developing an advanced detection system leveraging deep learning techniques. Utilizing Python, Pandas, NumPy, and Matplotlib for data preprocessing, manipulation, and visualization, alongside TensorFlow/PyTorch for model implementation, the project creates a comprehensive platform for deep fake detection. Through the integration of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), including Bicyclic and Deep variants, the system achieves heightened accuracy in distinguishing authentic from manipulated multimedia content. Custom CNN architectures are designed and optimized for deepfake detection, with GANs augmenting feature extraction and refining discernibility. Comparative analyses of GAN types provide insights into their effectiveness, ensuring transparency and accountability. The project's interdisciplinary approach, combining artificial intelligence, machine learning, and data science methodologies, culminates in a robust solution to combat the proliferation of deepfake content, safeguarding the integrity of digital media and fostering trust in online information sources.

Keywords—Deepfake detection, Real and Fake Images, Deep learning, Convolutional Neural Network (CNN), Generative Adversarial Network (GANs), Bicyclic GAN, Deep GAN.

I. INTRODUCTION

Amidst a period marked by swift progressions in artificial intelligence and digital media, the advent of deepfake technology poses unparalleled obstacles to the genuineness and reliability of multimedia content. The

boundaries between reality and fiction can be blurred by deepfakes, which are produced using advanced deep-learning algorithms to create remarkably realistic images, videos, and audio recordings. Significant risks are associated with the widespread distribution of deepfake content on online platforms. These risks include the spread of false information, a decline in public confidence in the objectivity of the media, and possible dangers to both national security and individual privacy. The goal of the project "Improving Deepfake Detection Using Deep Learning" is to use cutting-edge deep learning techniques to create a reliable detection system in response to these difficulties. Using Python, Pandas, NumPy, matplotlib, and TensorFlow/PyTorch, the project aims to develop a cutting-edge platform that can precisely detect and lessen the effects of deep fake content. The project aims to improve detection accuracy and transparency by integrating custom CNN architectures and advanced GAN models, such as the Bicyclic and Deep variants. This will ultimately protect the integrity of digital media and promote trust in online information sources.

II. LITERATURE REVIEW

Deepfake is the term for artificial intelligence-generated, realistic-looking, but phony images, sounds, and videos. Deepfake is now easier to create and more realistic thanks to recent advancements in the field. Deepfake has been a significant threat to national security, democracy, society, and our privacy, which calls for deep fake detection methods to combat potential threats[13]. The proposed framework in the research paper initially performs an Error Level Analysis of the image to determine if the image has been modified. That image is then fed into convolutional neural networks for deep feature extraction. The resulting feature vectors are then classified by support vector machines and K-nearest neighbors, performing hyperparameter optimization. The suggested approach had the highest accuracy, coming in at 89.5% through residual network and K-nearest neighbor. The dataset was created specifically for use in the investigation and advancement of facial recognition and verification systems, especially those intended to identify altered or phony photos. Every image in the dataset has been classified as either real or fake. The dataset also contains additional details about the images, including the subject's age, gender, and ethnicity as well as the method of image manipulation used for the fakes. The results obtained from ResNet's confusion matrix and ML classifiers achieved the highest accuracy of 89.5% via KNN. They tested various hyper-parameters for both classifiers before concluding.

The proposed method achieved 89.5% accuracy in correlation distance measurement through KNN and a total of 881 neighbors. SVM achieved an accuracy of 88.6% in a Gaussian kernel with a scale of 2.3. In addition, the feature vector obtained from GoogLeNet achieved the highest accuracy of 81% in Chebyshev distance measurement with a total of 154 neighbors by KNN. The SVM classified the feature vector with 80.9% accuracy as a Gaussian kernel with a scale of 0.41 kernel. SVM and KNN classified the SqueezeNet feature vector as 69.4% and 68.8%, respectively. The classifiers were evaluated with different parameters[1].

In another research, the primary goal of the research was to identify deep fake media using an effective framework. This study proposes a novel deepfake predictor (DFP) approach based on a hybrid VGG16 and convolutional neural network architecture. The deep fake dataset, which includes both real and fake faces, is used to train neural network techniques. The transfer learning techniques used in the comparison include Xception, NAS-Net, Mobile Net, and VGG16. The proposed DFP approach achieved 95% precision and 94% accuracy in deepfake detection. Our novel DFP approach outperformed transfer learning and other cutting-edge studies. Our novel research approach assists cybersecurity professionals in preventing deepfake-related cybercrimes by accurately detecting deepfake content and protecting deepfake victims from blackmail [5]. This paper evaluates humans' ability to detect image deepfakes of human faces (uncurated output from the StyleGAN2 algorithm as trained on the FFHQ dataset) from a pool of non-deepfake images (random selection of images from the FFHQ dataset), as well as the effectiveness of some simple interventions designed to improve detection accuracy. Using an online survey, participants (N = 280) were randomly assigned to one of four groups: a control group and three assistance interventions. Each participant was shown a series of 20 images chosen at random from a pool of 50 deepfake human faces and 50 real human faces [8].

Participants were asked whether each image was AI-generated, their level of confidence, and the reasoning behind their responses. Overall, detection accuracy was only slightly above chance, and none of the interventions significantly improved it. Equally concerning was the high level of confidence in participants' answers, which was unrelated to accuracy. When the results are analyzed per image, it is clear that participants found certain images easy to label correctly and others difficult, but they all reported high levels of confidence. Thus, while participant accuracy was 62% overall, it varied quite evenly between 85 and 30% across images, with one in every five images having an

accuracy of less than 50% [8]. To detect deepfake videos, a new project was introduced: You Only Look Once Convolution Recurrent Neural Networks (YOLO-CRNNs). The YOLO-Face detector detects face regions in each frame of the video, and a fine-tuned EfficientNet-B5 extracts their spatial features. These features are fed as a batch of input sequences into a Bidirectional Long Short-Term Memory (Bi-LSTM), which extracts temporal features. This shows the new scheme is then tested on a new large-scale dataset called CelebDF-FaceForencics++ (c23), which is a combination of two popular datasets: FaceForencies++ (c23) and Celeb-DF [11]. It has an Area Under the Receiver Operating Characteristic Curve (AUROC) score of 89.35%, accuracy of 89.38%, recall of 83.15%, precision of 85.55%, and F1-measure of 84.33% for the pasting data method.

In the "Deep Fakes Detection Techniques Using Deep Learning: A Survey" paper, we present a comprehensive review of deep fake creation and detection technologies that employ deep learning approaches. In addition, we provide a thorough analysis of various technologies and their applications in deepfake detection. Our study will benefit researchers in this field because it will cover the most recent cutting-edge methods for detecting deepfakes in social media content. Furthermore, the detailed description of the most recent methods and datasets used in this domain will aid in comparison with previous works [2]. Another research paper introduces a new deepfake detection method called YOLO-CNN-XGBoost. The YOLO face detector extracts the face area from video frames, and the InceptionResNetV2 CNN extracts features from these faces. These features are fed into XGBoost, which functions as a recognizer at the top level of the CNN network. On the CelebDF-FaceForencics++ (c23) merged dataset, the proposed method achieves 85.39% sensitivity, 85.39% recall, 87.36% precision, 93.73% accuracy, 93.53% specificity, and 86.36% F1-measure. The experimental study confirms the superiority of the proposed method over state-of-the-art methods [4]. In another research, we look at the deepfake detection technologies Xception and MobileNet as two approaches for classification tasks that automatically detect deepfake videos. We use FaceForensics++ training and evaluation datasets, which consist of four datasets generated using four different and well-known deepfake technologies [7]. The proposed technique in the "DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection" paper combines several DL-based state-of-the-art classification models to produce an improved composite classifier. Our experiments show that DeepfakeStack outperforms other classifiers in detecting Deepfake, with an accuracy of 99.65% and an AUROC of 1.0. As a result, our method provides a solid foundation for developing a real-time deepfake detector [10]. The other research proposes a temporal-aware pipeline for automatically detecting deepfake videos. We use a convolutional neural network (CNN) to extract frame-level features. These features are then used to train a recurrent neural network (RNN), which learns to determine whether a video has been manipulated or not [3]. Different research uses a structured case study to address the detection of such attacks. We specifically evaluate eight different machine learning algorithms in distinguishing between tampered and untampered images, including three traditional machine learning methods (Support Vector Machine, Random Forest, Decision Tree) and five deep learning models (DenseNet121, DenseNet201, ResNet50, ResNet101, and VGG19). The five deep-learning models are used for feature extraction, followed by fine-tuning each pre-trained model. The results of this study show near-perfect accuracy in detecting instances of tumor injections and removals [6].

In another significant research paper, we use an augmented real and fake face-detection dataset to compare the most common, cutting-edge face-detection classifiers, including Custom CNN, VGG19, and DenseNet-121. Data augmentation is used to improve performance while reducing computational resources. When compared to the other models analyzed, our preliminary results show that VGG19 has the best performance and highest accuracy (95%) [9]. Analysis has revealed that the suggested model's performance is excellent and consistent. The loss of discriminator is minimized compared to generator loss with successive iterations. Its fake detection strengthens with higher iterations. Adversarial training without mode collapse and convergence showed good predictive performance. It is also analyzed that good accuracy can be achieved with fewer images under controlled conditions by optimizing factors like a sufficient number of epoch cycles, normalized batch size of images, noise value, and effective model layers [12]. The images are re-scaled and fed to the D-CNN model in the paper for Deepfake Image Detection and a binary-cross entropy and Adam optimizer are utilized to improve the learning rate of the D-CNN model. We have considered seven different datasets from the reconstruction challenge with 5000 deepfake images and 10000 real images. The proposed model yields an accuracy of 98.33% in AttGAN, [Facial Attribute Editing by Only Changing What You Want (AttGAN)] 99.33% in GDWCT,[Group-wise deep whitening-and-coloring transformation (GDWCT)] 95.33% in StyleGAN, 94.67% in StyleGAN2, and 99.17% in StarGAN [A GAN capable of learning mappings among multiple

domains (StarGAN)] real and deep fake images, that indicates its viability in experimental setups [15]. We trained eight different CNN models in a comparative study of using convolutional neural networks (CNN) to identify genuine and deeply fake images. Three of these models were trained with the DenseNet architecture (DenseNet121, DenseNet169, and DenseNet201); two were trained in the VGGNet architecture (VGG16, VGG19); one was with ResNet50 architecture, one with VGGFace architecture, and one with custom CNN architecture. We also implemented a custom model that includes methods such as termination and completion to help determine if other models are serving their purpose. The results were ranked according to five evaluation metrics: precision, accuracy, recall, F1 score, and subregion. ROC (receiver operating characteristic) curve. Of all the models, VGGFace performed best, with 99% accuracy. Besides, we obtained 97% from the ResNet50, 96% from the DenseNet201, 95% from the DenseNet169, 94% from the VGG19, 92% from the VGG16, 97% from the DenseNet121 model, and 90% from the custom model [13]. One of the strategies finalized in another research paper identifies input by looking at the facial zones and their encompassing pixels by parting the video into outlines separating the highlights with a ResNext-v1 CNN and utilizing the MTCNN to catch the transient irregularities between frames presented by GANs during the remaking of the pixels [14].

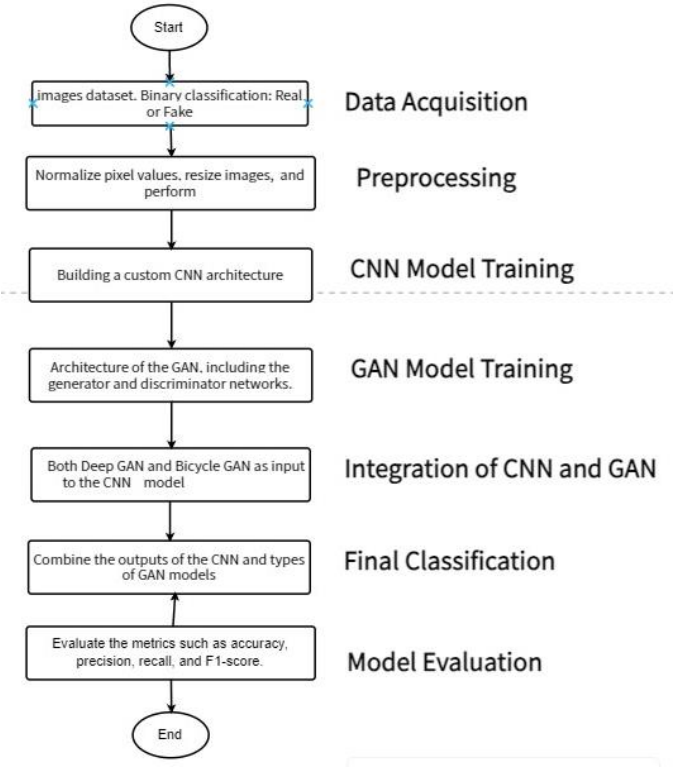
DATASET DESCRIPTION

We meticulously curated two primary datasets to underpin our project's deep fake detection endeavors. The first dataset, labeled **faces__224**, comprises a collection of images sourced from the esteemed **Kaggle deepfake-detection challenge**. Each image within this dataset boasts a resolution of 224 x 224 pixels, meticulously chosen to facilitate comprehensive analysis and model training. The high-resolution nature of these images ensures that intricate details crucial for accurate detection are preserved and leveraged effectively.

Complementing the visual data, our second dataset, aptly named **metadata.csv,** serves as a valuable repository of supplementary information. This metadata provides essential context, aiding in the categorization and understanding of the images contained within the faces__224 dataset. By incorporating this additional layer of information, we enhance the richness and depth of our dataset, enabling more nuanced and informed analyses of deepfake content. Through the synergistic utilization of these meticulously curated datasets, we empower our deep learning models to discern patterns, anomalies, and distinguishing features crucial for robust and reliable deep fake detection.

Figure 1: Dataset

III. METHODOLOGY



The methodology employed in our project for improving deepfake detection involved a systematic approach encompassing various stages. Initially, we meticulously trained and tested our dataset to discern between real and fake images. Subsequently, we embarked on the development of a custom Convolutional Neural Network (CNN) architecture tailored specifically for deep fake detection. This involved preprocessing and normalizing the dataset to optimize model training efficacy, as well as visualizing training images to glean insights into dataset characteristics and ensure data quality.

Moving forward, our methodology capitalized on the observation of pixel-level details in images to ascertain their authenticity. We then seamlessly integrated Generative Adversarial Networks (GANs) into our custom CNN framework to augment detection accuracy. This integration entailed engineering a dual-module system comprising a class generator and discriminator, leveraging ReLU, tanh, leaky ReLU, and sigmoid activation functions to refine feature extraction and enhance the discernibility of manipulated content.

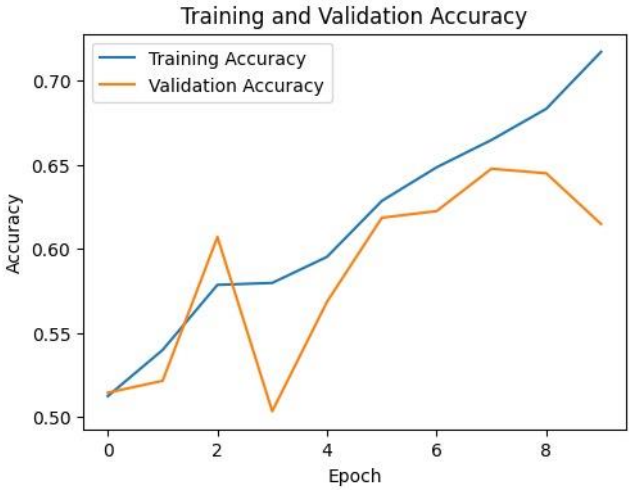


Figure 2: Training vs Validation Accuracy Graph

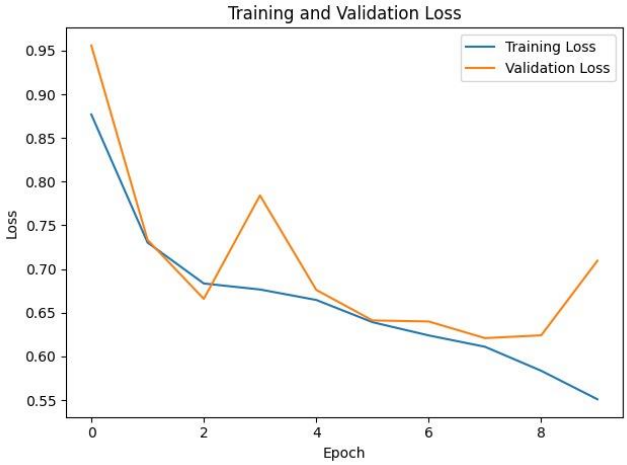


Figure 3: Training vs Validation Loss Graph

Central to our methodology was the comparative analysis of GAN types, namely Bicyclic GAN and Deep GAN. With this analysis, we assessed how well each GAN architecture performed on a feature-by-feature basis in differentiating between phony and real images. Our methodology guarantees transparency and accountability to stakeholders by offering insights into the advantages and disadvantages of these GAN variants, thereby progressing the field of deepfake detection.

IV. RESULT

The architecture used in the presented code can be classified as a Convolutional Neural Network (CNN) architecture. CNNs are particularly suitable for image classification tasks because they can automatically learn hierarchical representations of features from raw pixel values.
It follows a typical CNN structure with alternating convolutional and convolutional layers followed by fully connected dense layers, which makes it suitable for image classification tasks \as shown in the code. The architecture of the models consists of several convolutional layers that are controlled by dense layers. Here are the following:

1. Convolution layers:

The model starts with a convolution layer (Conv2D) with 64 filters and a kernel size of 7x7. The activation function used here is ReLU ("relu"). Two more convolutional layers (Conv2D) follow, each with 128 filters and a kernel size of 3x3. ReLU also serves as the activation function for these layers.

2. Pooling Layers:

Two max-pooling layers (MaxPooling2D) are used to reduce the spatial dimensions of the feature maps.

3. Batch Normalization:
Batch normalization layers (BatchNormalization) are added after the convolutional layers to normalize the activations and stabilize training.

4. Dense Layers:
After flattening the feature maps, the model includes three dense layers (Dense) with ReLU activation functions. The first dense layer has 128 units, followed by batch normalization and dropout layers (Dropout) with a dropout rate of 0.5. The second dense layer has 64 units, followed by batch normalization and dropout layers with a dropout rate of 0.5. The final dense layer has 1 unit with a sigmoid activation function, which is used for binary classification.

5. Activation Functions:
ReLU activation function ("relu") is used for convolutional and dense layers, except for the final dense layer where sigmoid activation ("sigmoid") is used for binary classification.

6. Model Type:
The model architecture follows a common pattern for image classification tasks, consisting of convolutional layers followed by pooling layers to extract features from input images. This architecture is commonly used for tasks like object detection and image classification.

The confusion matrix provides a summary of the model's performance by showing the counts of true positive, true negative, false positive, and false negative predictions. It helps evaluate the effectiveness of the classification algorithm by revealing the model's ability to correctly classify instances belonging to each class and identify misclassifications.

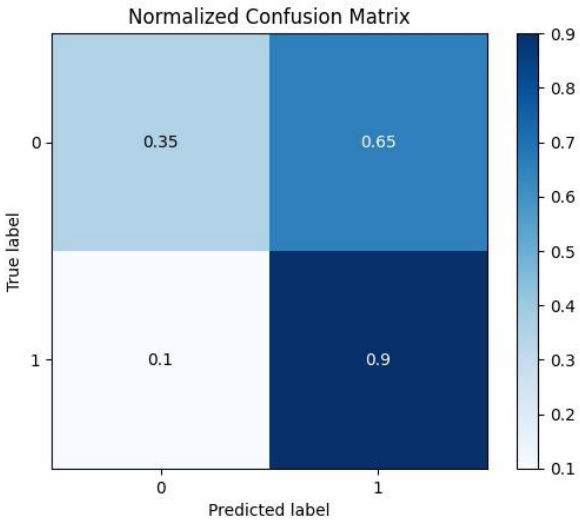


Figure 4: Confusion Matrix

In summary, the model includes convolutional layers for feature extraction, followed by dense layers for classification, with ReLU activation functions used throughout the network and a sigmoid activation function for binary classification at the output layer. The provided code outlines a comprehensive approach to deepfake detection using a custom CNN model along with different types of Generative Adversarial Networks (GANs), including BicycleGAN, Deep Convolutional GAN (DCGAN), and Simple GAN. The custom CNN model is designed for image classification tasks, leveraging convolutional and pooling layers followed by dense layers for feature extraction and classification.

BicycleGAN, DCGAN, and Simple GAN classes are also defined, providing a framework for implementing and training these GAN architectures for generating realistic fake images. The models are trained on a dataset containing both real and fake images, with the training process tailored to each GAN architecture. This approach enables a versatile approach to deep fake detection by combining the strengths of CNNs for classification with the ability of GANs to generate and detect fake images.

V. CONCLUSION

To sum up, our project is a major advancement in the ongoing fight against deepfake technology. Using the methodical construction and fusion of bespoke CNN topologies and Generative Adversarial Networks (GANs), we have established a resilient and efficient mechanism to identify and thwart the dissemination of manipulated multimedia content. Through the

utilization of sophisticated deep learning methodologies and comprehensive GAN-type comparisons, we have acquired a significant understanding of the advantages and disadvantages of different deepfake detection strategies.

Our method, which focuses on careful data preprocessing and pixel-level observation, has produced encouraging results in accurately classifying real from fake images. Our unique CNN architecture and GANs' seamless integration have improved our detection capabilities even more, enabling us to better distinguish manipulated content and refine feature extraction. We're still dedicated to developing the field of deepfake detection and protecting the integrity of digital media even as we continue to hone and improve our strategy. Our goal is to create a more secure and reliable digital environment for everyone by encouraging accountability and openness and providing stakeholders with the resources and information they need to stop the spread of deepfake content.

VI. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to all individuals and Bennett University who have contributed to the completion of this research paper.
First and foremost, we extend our sincere appreciation to the Mentors, scientists, and experts whose groundbreaking
work laid the foundation for this study. Their insights and contributions have been invaluable in shaping our research direction and methodology.
We extend our gratitude to team members for their dedication, expertise, and teamwork throughout the course of this project. Their tireless efforts and commitment to excellence have been essential in achieving our research goals. We would also like to acknowledge the assistance and support provided by School of applied science and engineering Bennett university, CSE Department, our Dean Dr. Abhay Bansal and mentor Dr. Rohit Kumar Kaliyar for their technical guidance, access to resources, and logistical assistance.
This research paper is a testament to the collective efforts and collaboration of all involved, and we are grateful for the opportunity to contribute to the advancement of knowledge and innovation in our field.

REFERENCES

[1] Rafique, R., Gantassi, R., Amin, R. et al. “Deep fake detection and classification using error-level analysis and deep learning.” Sci Rep 13, 7422 (2023). https://doi.org/10.1038/s41598-023-34629-3

[2] 2.Almars, Abdulqader. (2021).“ Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications.” 09. 20-35. 10.4236/jcc.2021.95003.

[3] 3.Guera, David and Edward J. Delp. “Deepfake Video Detection Using Recurrent Neural Networks.” 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS) (2018): 1-6.

[4] 4. Ismail, Aya, Marwa Elpeltagy, Mervat S. Zaki, and Kamal Eldahshan. 2021. "A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost" Sensors 21, no. 16: 5413. https://doi.org/10.3390/s21165413

[5] 5. Raza, Ali, Kashif Munir, and Mubarak Almutairi. 2022. "A Novel Deep Learning Approach for Deepfake Image Detection" Applied Sciences 12, no. 19: 9820. https://doi.org/10.3390/app12199820

[6] 6. Siddharth Solaiyappan, Yuxin Wen,” Machine learning based medical image deepfake detection: A comparative study”, Machine Learning with Applications, Volume 8, 2022, 100298, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2022.100298.

[7] 7.D. Pan, L. Sun, R. Wang, X. Zhang and R. O. Sinnott, "Deepfake Detection through Deep Learning," 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Leicester, UK, 2020, pp. 134-143, doi: 10.1109/BDCAT50828.2020.00001.

[8] 8. Sergi D Bray, Shane D Johnson, Bennett Kleinberg, Testing the human ability to detect ‘deep fake’ images of human faces, Journal of Cybersecurity, Volume 9, Issue 1, 2023, tyad011, https://doi.org/10.1093/cybsec/tyad011

[9] 9. Taeb, Maryam, and Hongmei Chi. 2022. "Comparison of Deepfake Detection Techniques through Deep Learning" Journal of Cybersecurity and Privacy 2, no. 1: 89-106. https://doi.org/10.3390/jcp2010007

[10] 10. M. S. Rana and A. H. Sung, "DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection," 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), New York, NY, USA, 2020, pp. 70-75, doi: 10.1109/CSCloud-EdgeCom49738.2020.00021.

[11] 11. Ismail A, Elpeltagy M, Zaki M, ElDahshan KA. Deepfake video detection: YOLO-Face convolution recurrent approach. PeerJ Comput Sci. 2021 Sep 21;7:e730. doi: 10.7717/peerj-cs.730. PMID: 34712799; PMCID: PMC8507472.

[12] 12. Preeti, Manoj Kumar, Hitesh Kumar Sharma, “A GAN-Based Model of Deepfake Detection in Social Media”, Procedia Computer Science Volume 218, 2023, Pages 2153-2162, https://doi.org/10.1016/j.procs.2023.01.191

[13] 13. Hasin Shahed Shad, Md. Mashfiq Rizvee, Nishat Tasnim Roza, S. M. Ahsanul Hoq, Mohammad Monirujjaman Khan, Arjun Singh, Atef Zaguia, Sami Bourouis, "[Retracted] Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network", Computational Intelligence and Neuroscience, vol. 2021, Article ID 3111676, 18 pages, 2021. https://doi.org/10.1155/2021/3111676

[14] 14. Sharma, H.K., Khan, S.S., Choudhury, T., Khurana, M. (2023). CNN-Based Model for Deepfake Video and Image Identification Using GAN. In: Reddy, K.A., Devi, B.R., George, B., Raju, K.S., Sellathurai, M. (eds) Proceedings of Fourth International Conference on Computer and Communication Technologies. Lecture Notes in

Networks and Systems, vol 606. Springer, Singapore. https://doi.org/10.1007/978-981-19-8563-8_47

[15] 15. Y. Patel et al., "An Improved Dense CNN Architecture for Deepfake Image Detection," in IEEE Access, vol. 11, pp. 22081-22095, 2023, doi: 10.1109/ACCESS.2023.3251417.