AN EFFECTIVE GENERATIVE ADVERSARIAL NETWORK FOR DEEP FAKE DETECTION IN SOCIAL MEDIA PLATFORM

Kolli Sai Chandra, Department of Computer Science Engineering, Indian Institute of Industry Interaction Education and Research, Chennai, Tamil Nadu 600066
Mohammad Irfan Alam, Department of Computer Science Engineering, Indian Institute of Industry Interaction Education and Research, Chennai, Tamil Nadu 600066
Anaparthi Akash, Department of Computer Science Engineering, Indian Institute of Industry Interaction Education and Research, Chennai, Tamil Nadu 600066
Vandana Kiran Kumar, Department of Computer Science Engineering, Indian Institute of Industry Interaction Education and Research, Chennai, Tamil Nadu 600066

ABSTRACT

Deep fake technology has become increasingly prevalent on social media platforms, leading to a rise in the dissemination of manipulated or false information. Detecting deepfake content on social media poses a unique challenge due to the realism and sophistication of the technology. Hence, the developed framework aims to put forth a new generative model for identifying the existence of deep fake on the social media images. The proposed technique consists of multiple steps: data acquisition, pre-processing, and classification. Initially, the input social media images are obtained and it is pre-processed using normalized gamma-based contrast-limited improved histogram equalization (NG-CLIHE) to remove noises from the images. Preprocessed images are then fed into the Progressive Wasserstein generative adversarial network (GAN) for the detection of deep fakes accurately. The developed framework is processed on platform and a freely accessible deepfake-real image dataset is utilized. In the simulation part, the accuracy of 98.92%, precision of 98.76%, recall of 98.02% and f-measure of 98.9% are obtained by the developed method.

Keywords:
Online Social Media Platform, Generative Adversarial Network, Deep Fake, Forgery Detection, Contrast Enhancement, Normalized Gamma-Based Contrast-Limited Improved Histogram Equalization

I. Introduction

Deepfake technology has become a growing concern on social media platforms, as it allows users to create realistic and often misleading or fake videos or other forms of media. With the advancement of artificial intelligence and machine learning algorithms, deepfake technology has become more accessible to the general public, posing a threat to the authenticity and trustworthiness of online content [1]. As a result, researchers and tech companies have been working on developing tools and methods to detect and prevent the spread of deepfake content on social media. Detecting deepfake content on social media poses a unique challenge due to the realism and sophistication of the technology. Deepfake images can be created by swapping faces, altering voices, or manipulating the context of the original footage to create a false narrative. As a result, traditional methods of content moderation and fact-checking may not be sufficient to detect deepfakes, as they can easily deceive even trained human eyes [2]. In recent years, researchers have been developing automated tools and algorithms to detect deepfake content on social media platforms. These tools often rely on machine learning algorithms that are trained on a large dataset of deepfake and authentic videos to learn the patterns and characteristics of deepfake content. By analyzing factors such as facial movements, lighting, and audio discrepancies, these algorithms can identify inconsistencies that may indicate the presence of a deepfake [3]. One common approach to deepfake detection is the use of neural networks, a type of artificial intelligence that processes information similar to the way a human brain does. Neural networks can be trained to detect deepfakes by analyzing patterns and discrepancies in the video or audio data. These algorithms are constantly being improved and updated to keep up with the evolving technology of deepfake creation. Another approach to deepfake detection is the use of blockchain technology, which can help verify the authenticity of media content by creating a secure and transparent record of its origin and editing history. By storing metadata and timestamps on a blockchain, social media platforms can track the provenance of content and identify any alterations that may indicate the presence of a deepfake. In addition to automated detection tools, tech companies like Facebook and Twitter have also implemented policies and measures to combat the spread of deepfake content on their platforms. These measures include partnering with fact-checking organizations, implementing content moderation algorithms, and providing resources for users to report suspicious or misleading content [4]. Despite these efforts, the battle against deepfakes on social media is ongoing and requires collaboration between tech companies, researchers, and policymakers to develop effective strategies for detecting and preventing the spread of manipulated content. As deepfake technology continues to advance, it is crucial for the online community to remain vigilant and sceptical of the content they encounter, and for social media platforms to continue investing in tools and technologies to combat the spread of deceptive content [5]. In this research paper, we will explore the current state of deep fake detection technology, examining the methods and algorithms that have been developed to distinguish between real and fake videos. We will also discuss the ethical implications of deep fake technology and the challenges that researchers face in developing effective detection methods. By gaining a deeper understanding of the capabilities and limitations of deep fake detection technology, we can better protect ourselves from the potentially harmful effects of this emerging technology.

Motivation: Deep fake technology has become increasingly prevalent on social media platforms, leading to a rise in the dissemination of manipulated or false information. As a result, there is a pressing need for effective detection techniques to combat the spread of deep fakes and protect the integrity of online content. One of the primary motivations for detecting deep fakes on social media is the potential for these manipulated videos to deceive and mislead the public. Deep fakes can be used to create fake news, spread disinformation, and manipulate public opinion, leading to serious consequences for individuals, organizations, and society as a whole. By detecting and removing deep fakes from social media platforms, we can minimize the impact of these malicious activities and preserve the trust and credibility of online content. Furthermore, deep fake detection is essential for ensuring the safety and security of social media users. Deep fakes have the potential to violate the privacy and rights of individuals by creating fake videos or images of them without their consent. These manipulated media can be used for cyberbullying, harassment, or even extortion, posing a significant threat to the safety and well-being of social media users. By implementing robust detection mechanisms, social media platforms can prevent the spread of deep fakes and protect their users from potential harm. Additionally, deep fake detection is crucial for maintaining the authenticity and reliability of online content. As deep fake technology continues to advance, it becomes increasingly difficult for users to discern between genuine and manipulated media. This can lead to a loss of trust in social media platforms and a decline in the credibility of online information. By actively detecting and removing deep fakes, social media platforms can preserve the integrity of their content and ensure that users are exposed to accurate and truthful information. Motivated by these concerns, the developed study put forth new generative model for identifying the existence of deep fake on the social media images.

The main contributions of the proposed framework are deliberated as follows:

- To propose an innovative deep learning (DL) model (GAN) for detecting the existence and non-existence of fake attacks using extensive feature learning.
- To preprocess the raw neuroimages using normalized gamma-based contrast-limited improved histogram equalization (NG-CLIHE) thereby enhancing the image quality.
- To present a new generative adversarial network (GAN) to detect the deep fakes on social media platforms.

- To validate the performance of the proposed framework with the conventional technique by evaluating various assessment measures like accuracy, precision, recall and f-measure.

The upcoming sections are organized as follows: Section 2 outlays the related work, Section 3 deliberates over the suggested approaches, Section 4 presents the results and discussion, and Section 5 represents the conclusion of the proposed framework.

II. Related Works

Paten et al., [6] defined the DL model for detecting deep fakes on online social media platform. Here, a novel and improved deep-CNN (D-CNN) architecture was introduced for deepfake detection with reasonable accuracy and high generalizability. Images from multiple sources were captured to train the model, enhancing overall generalizability capabilities. The images were re-scaled and fed to the D-CNN model. A binary-cross entropy and Adam optimizer were utilized to improve the learning rate of the D-CNN model. Seven different datasets from the reconstruction challenge were considered, consisting of 5000 deepfake images and 10000 real images. On analysing the resultant part, accuracy, recall, precision, and F-measure were computed and compared with other techniques. However, the deep CNNs may also fail to accurately detect more sophisticated deep fakes that have been specifically designed to evade detection by AI systems. This could lead to a false sense of security for users and platforms, allowing malicious actors to continue spreading disinformation and fake content. Hamid et al. [7] introduced the improved DL model for identifying deep fake on social media. This study introduced a computer vision model based on Convolutional Neural Networks (CNN) for fake image detection. A comparative analysis of 6 popular traditional machine learning models and 6 different CNN architectures was conducted to select the best possible model for further experimentation. The proposed model, based on ResNet50, was employed with powerful preprocessing techniques resulting in a perfect fake image detector with a total accuracy of 0.99. This model showed an improvement of around 18% in performance compared to other models. Nevertheless, the CNNs may also be susceptible to adversarial attacks, where malicious actors intentionally manipulate the content to evade detection. This can undermine the effectiveness of using CNNs to detect deep fakes on social media platforms, as these attacks can be used to trick the model into misclassifying the content as genuine. Lee et al. [8] established the detection of GAN induced face images and manual face manipulations using DL model. Here, a Handcrafted Facial Manipulation (HFM) image dataset and soft computing NN models (Shallow- FakeFaceNets) with an efficient facial manipulation detection pipeline were introduced. The neural network classifier model, Shallow-FakeFaceNet (SFFN), demonstrated the ability to focus on manipulated facial landmarks to detect fake images. The detection pipeline solely relied on detecting fake facial images based on RGB information, without leveraging any metadata, which could be easily manipulated. The results showed that the method achieved the best performance of 72.5% in Area Under the Receiver Operating Characteristic (AUROC), with a 3.9% increase in F1-score and 2.9% increase in AUROC for detecting handcrafted fake facial images. It also achieved a 93.99% accuracy in detecting small GAN-generated fake images, with a 1.98% increase in F1-score and 10.44% increase in AUROC compared to other models. Nevertheless, the deepfake technology is constantly evolving and becoming more sophisticated, there is a risk that SFFN may quickly become outdated and ineffective at detecting the latest deepfake techniques.

Khalil et al. [9] put forth combined DL model for detecting deep fake on video and images. Here, a deepfake detection approach, iCaps-Dfake, was emphasized that addressed the other low generalization problem. Two feature extraction methods were combined, including texture-based Local Binary Patterns (LBP) and CNN based modified High-Resolution Network (HRNet), along with the application of capsule neural networks (Caps Nets) implementing a concurrent routing technique. Experiments were conducted on large benchmark datasets to evaluate the performance of the proposed model. Several performance metrics were applied and experimental results were analysed. The proposed model was primarily trained and tested on the Deep Fake Detection Challenge-Preview (DFDC-P) dataset and then tested on Celeb-DF to examine its generalization capability. Experiments achieved an AUC score improvement of 20.25% compared to other schemes. Nevertheless, integrating capsule nets into existing systems may require significant resources and expertise, which could pose challenges for smaller social media platforms with limited technical capabilities.

Wu et al. [10] determined an effective image forgery detection scheme on shared online social media networks. Here, a robust training scheme was proposed and a thorough analysis of the noise introduced by OSNs was conducted, and it was decoupled into two parts, predictable noise and unseen noise, which were modelled separately. The former simulated the noise introduced by the disclosed (known) operations of OSNs, while the latter was designed to not only complement the previous one but also take into account the defects of the detector itself. The modelled noise was then incorporated into a robust training framework, significantly improving the robustness of the image forgery detector. Extensive experimental results were presented to validate the superiority of the proposed scheme compared to several state-of-the-art competitors. Finally, to promote the future development of image forgery detection, a public forgeries dataset was built based on four existing datasets and three of the most popular OSNs. Nevertheless, this method may not be able to adapt quickly to new types of deep fake videos that emerge on social media platforms.

Problem statement

Traditional studies on detecting deep fakes on social media have encountered several challenges and limitations. These studies primarily focus on developing algorithms and techniques to identify and verify the authenticity of multimedia content shared across various social media platforms. However, these studies often struggle to keep pace with the rapidly evolving technology and sophistication of deep fake creation. Additionally, existing detection methods may lack scalability and efficiency, making it difficult to detect deep fakes in real-time. Moreover, there is a lack of standardized benchmarks and datasets for evaluating the performance of different detection techniques, hindering the development of reliable and robust detection systems. Hence, this study proposes a innovative generative models for detecting fake images effectively.

III. Proposed Methodology

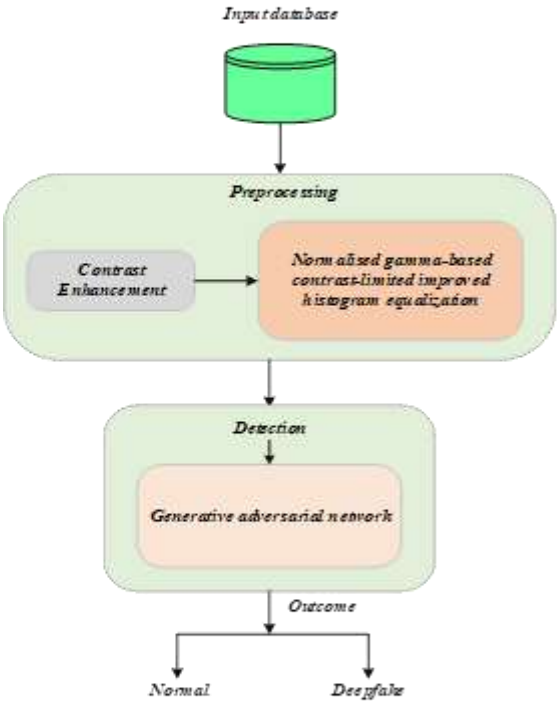
Input database
Preprocessing
Contrast Enhancement
Normalised gamma-based contrast-limited improved histogram equalization
Detection
Generative adversarial network
Outcome
Normal
Deepfake

Figure 1: Workflow of the developed method

The proposed technique consists of multiple steps: data acquisition, pre-processing, and classification. Initially, the input social media images are obtained and it is pre-processed using normalized gamma-based contrast-limited improved histogram equalization (NG-CLIHE) to remove noises from the images. Preprocessed images are then fed into the Progressive Wasserstein generative adversarial network (GAN) for the detection of deep fakes accurately. Figure 1 deliberates the workflow of the proposed approach

3.1 Preprocessing stage

The acquiesced images from the database consist of high noises that cannot be directly fed into the proposed network model. Nowadays, the histogram equalization (HE) technique plays an essential role in removing unwanted blurriness and enhancing the contrast of the image. However, the useful structural components are eliminated during the equalization process. To manifest this issue, a normalized gamma-based contrast-limited improved histogram equalization (NG-CLIHE) technique is introduced that preserves useful information and improves the illumination level of the neuroimages. The gamma modification process is a grey-level alteration function implemented on images to improve the image contrast level. This can be achieved by a power-law (PL) function that uses the spatial property for adjusting the image illumination level and it is formulated in equation (1).

$$X = zY^n \tag{1}$$

Here, X indicates the gamma-modified image, Y represents the actual image and it is in the range of 0 to 1, z represents the constant positive parameter, and n represents the constant positive parameter indicating gamma values. The parameter z is eliminated due to increased loss of essential details and vulnerable illumination in the image. Hence, the optimized PL is utilized and it can be formulated in equation (2),

$$P_l = Y^n \tag{2}$$

Here, K indicates the input image, and $norm$ indicates the normalized image, $\min$ and $\max$ operators are used to reduce and increase the pixel intensities. Here, an innovative normalized gamma function is used to minimize the illumination and improve the contrast of the image. It can be mathematically formulated in equation (3),

$$norm_{gamma} = \frac{\left[V - \min\left(V\right)\right]}{\left[\max\left(V\right) - \min\left(V\right)\right]} \tag{3}$$

Finally, the normalized gamma function is combined with CLIHE to enhance the contrast and brightness of the image. After applying CLIHE, the updated pixel value y in quadrant 1 with the region (a,b) is mathematically expressed in equation (4),

$$y_{new} = \frac{c}{c+d}\left(\frac{j}{i+j} f_{a-1,b-1}\left(y_{old}\right) + \frac{i}{i+j} f_{a,b-1}\left(y_{old}\right)\right) + \frac{d}{c+d}\left(\frac{j}{i+j} f_{a,b-1}\left(y_{old}\right) + \frac{i}{i+j} f_{a,b-1}\left(y_{old}\right)\right) \tag{4}$$

Here, c, d, i and j manipulates the pixel distances, $f(.)$ indicates the cumulative distribution function. The updated pixel values for quadrants 2, 3, and 4 with the region (a,b) can be evaluated using the above-mentioned procedure. It is to be noted that quadrants 1 and 3 are the same for the inner pixel region whereas quadrants 2 or 4 are different, then the updated pixel value y in quadrant 2 with the region (a,b) can be mathematically formulated in equation (5),

$$y_{new} = \frac{d}{c+d} f_{a,b-1}\left(y_{old}\right) + \frac{c}{d+c} f_{a,b}\left(y_{old}\right) \tag{5}$$

However, quadrant 1 is completely different from other quadrants, and the updated pixel value y in quadrant 1 with the region (a,b) is mathematically expressed in equation (6),

$$y_{new} = f_{a,b}(y_{old}) \qquad (6)$$

Here the corner pixels are mapped in the same way and the updated values of pixels are recovered in the updated array size that is similar to the actual image to obtain the enhanced image.

3.2 Deep fake detection using GAN technique

The pre-processed images are then fed into a Generative Adversarial Network (GAN) technique to classify deep fake images on the social media platform. The developed architecture is composed of three major stages: generator, weighted anatomically delicate integrated loss function, and discriminator. The advanced generator is comprised of multiple convolutional (Conv) layers that learn the abnormal condition by training large amounts of data. It contains six different phases that learn the relevant features for the classification process. In every stage, Conv layers are introduced that take model input. For extracting the features, LSTM layers are utilized and it is extended up to six times. A stacked network is introduced to enhance the model parameters and balance the classification performance. Moreover, intra-phase iterative unfold residual blocks (RB) are utilized to minimize the network parameters. The detailed analysis of the developed architecture is conquered below:

Generator CNN:

In this stage, six phases are concatenated comprised of ReLU activation function, ConvLSTM layers, and RBs. The single RB is iteratively extended five times thereby minimizing network parameters. The strides are 1 with its kernel size of three and a total of 32 filters are considered in ReLU for both input and output. In addition to this, 32 filters are considered in Conv layers present in LSTM and RB layers. As a result, the final Conv layers accept the outcome of RBs with 32 filters and output the generated outcome with a solitary filter.

Discriminator CNN:

The outcome of the generator is given as input to the discriminator blocks. In the discriminator block, 8 layers are present that provide adversarial outcomes based on the generated outcome. The initial six layers consist of Conv layers and the final two contain the fully connected (FC) layers. For all the layers, the kernel size is set to three, and Conv kernels are 64 for the initial two layers. For the next two layers, the Conv kernels are folded as 128 and the fifth-sixth layer has a total of 256 Conv kernels. Alternately, stride 1 is set for all Conv layers, and the final layer outputs the decision variable for multiclass classification. In the discriminator block, the Leaky ReLU function is utilized as the activation function. Figure 2 illustrates the architecture of the GAN technique.

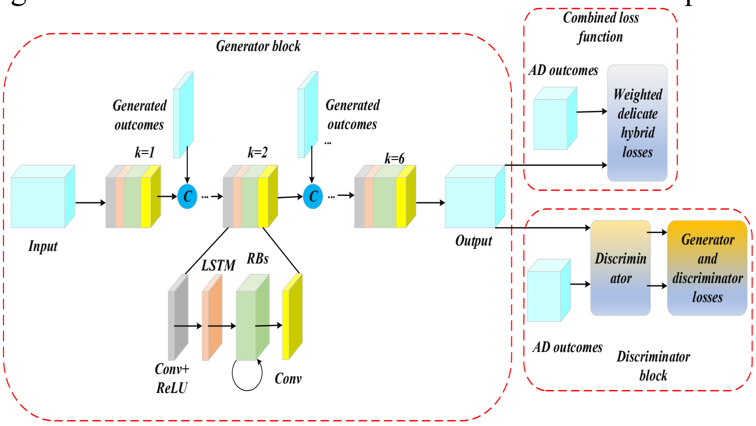

Figure 2: Architecture of GAN technique

Weighted Anatomically Delicate Integrated Loss Function (WAD_LF):

For enhancing the accuracy performance, a WAD_LF is intended for processing the generator phase. In this framework, the normal and the outcome of the generator are fed as input to the WAD_LF, and the error obtained is backpropagated to optimize the model parameters. The LF is the complex in training the network model that affects the robustness and accuracy of the developed model. During the actual adversarial process, the generator plans to set classified outcomes more similar to the actual outcome. However, the discriminator fails to differentiate the generated outcome from the actual outcome due to ineffective learning of complicated brain patterns. To tackle this issue, WAD_LF is introduced which considers effective LF to enhance the accuracy performance. The detailed analysis of the GAN technique is deliberated as follows:

L2 Loss: It is considered the default loss function that is caused due to data regularity, differentiability, and convexity. It can be mathematically determined in equation (7),

$$L_2 = \frac{1}{D}\left\|G(x) - x\right\|_2^2 \tag{7}$$

Here, D indicates the dimension of the data, and $G(x)$ represents the data generated during the training process. However, L2 LF inclines to be wedged at local minima and makes the outcome inaccurate.

L1 Loss: It is more operative than L2 LF in minimizing training error and the minimum value of L1-LF is lower than that of L2 LF. It can be expressed mathematically in equation (8),

$$L_2 = \frac{1}{D}\left|G(x) - x\right|_2^2 \tag{8}$$

The total LF for enhancing the PW_GAN model is formulated in equation (9),

$$L_{total} = L_1 + L_2 \tag{9}$$

Finally, from the SoftMax layer normal and deep fake images are accurately classified.

VI. Results and Discussion

The proposed framework is processed and experimented with via the Python platform. The simulation process is carried out via the openly accessible deepfake-real image dataset [11] that consist of more than 190,000 real and fake images. Each image present in this database are of 256×256 size collected under different environmental condition. Figure 3 depicts the Sample images taken from deepfake-real image dataset.

Figure 3: Sample images from deepfake-real image dataset

4.1 Assessment measures

Performance indicators like Accuracy, recall, precision, F-score, and computation time are computed to better understand the proposed approach.

Accuracy: It determines the model's overall recognition accuracy, accounting for both TP and TN. It is calculated as using equation (11),

$$Accuracy = \frac{w + x}{w + x + y + z} \tag{11}$$

Recall: Recall measures the positive outcomes that are accurately predicted model has successfully captured. The calculation is performed using equation (12),

$$\mathrm{Re}\,call = \frac{x}{x + y} \tag{12}$$

Precision: Precision is the proportion of the suggested model's positive recognition that are correct out of all the positive forecasts. It is computed using equation (13),

$$\Pr ecision = \frac{x}{x+z} \qquad (13)$$

Here, w, x, y, and z indicates the true positive (TN), true positive, (TP), False negative (FN), and false positive (FP) respectively.

F-measure: The F-measure displays the consonant mean of recall and precision. It is computed using equation (14) as follows,

$$F1 - score = 2 * \left(\frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \right) \qquad (14)$$

4.2 Performance analysis of developed method over existing studies

In this section, the performance achieved by the proposed technique is scrutinized via the table illustration. Various existing approaches are compared with the proposed method by computing various measures like accuracy, precision, recall, F-score, and time consumption. A brief analysis of the achieved performances is depicted below:

Table 1: Comparative analysis of developed scheme over conventional scheme

Methods	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Proposed	**98.92**	**98.76**	**98.02**	**98.9**
Deep-CNN	97	97.05	96.9	96.88
Meso-Net	57	60.5	71.32	71.2
Meso-Inception	50.73	66.39	69.8	70.67
CapsNet	86	84.4	96.9	91.82

Table 1 depicts the Comparative analysis of developed scheme over conventional scheme. From the experimental outcome, it is clear that the developed model showed effective performance compared to other deep-fake models.

IV. Conclusion

The proposed model introduced and investigated GAN model for identifying forgery images from the genuine image using social media images. The proposed technique undergone multiple steps: data acquisition, pre-processing, and classification. The input social media images are obtained and it is pre-processed using normalized gamma-based contrast-limited improved histogram equalization (NG-CLIHE) and provided better contrast images effectively. Preprocessed images are then fed into the generative adversarial network (GAN) for the detection of deep fakes and proved better accuracy performance than other existing studies. The developed framework is processed on platform and a freely accessible deepfake-real image dataset is utilized. In the simulation part, the accuracy of 98.92%, precision of 98.76%, recall of 98.02% and f-measure of 98.9% are obtained by the developed method.

References

[1] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," IEEE Access, vol. 10, pp. 18757-18775, 2022.

[2] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "GAN-generated faces detection: A survey and new perspectives," arXiv preprint arXiv:2202.07145, 2022

[3] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, et al., "Deepfakes detection with automatic face weighting," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 668-669, 2020.

[4] S. R. Ahmed, E. Sonuç, M. R. Ahmed, and A. D. Duru, "Analysis survey on deepfake detection and recognition with convolutional neural networks," in 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1-7, IEEE, 2022.

[5] S. Tyagi and D. Yadav, "A detailed analysis of image and video forgery detection techniques," The Visual Computer, vol. 39, no. 3, pp. 813-833, 2023.

[6] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I. E. Davidson, and T. F. Mazibuko, "An improved dense CNN architecture for deepfake image detection," IEEE Access, vol. 11, pp. 22081-22095, 2023..

[7] Y. Hamid, S. Elyassami, Y. Gulzar, V. R. Balasaraswathi, T. Habuza, and S. Wani, "An improvised CNN model for fake image detection," International Journal of Information Technology, vol. 15, no. 1, pp. 5-15, 2023.

[8] S. Lee, S. Tariq, Y. Shin, and S. S. Woo, "Detecting handcrafted facial image manipulations and GAN-generated facial images using Shallow-FakeFaceNet," Applied Soft Computing, vol. 105, p. 107256, 2021

[9] K. S. Samir, S. M. Youssef, and S. N. Saleh, "iCaps-Dfake: An integrated capsule-based model for deepfake image and video detection," Future Internet, vol. 13, no. 4, p. 93, 2021.

[10] H. Wu, J. Zhou, J. Tian, and J. Liu, "Robust image forgery detection over online social network shared images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 13440-13449.

[11] https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images