Deepfake Detection Using CNN And DCGANS To Drop-Out Fake Multimedia Content: A Hybrid Approach

Kartik Bansal
Department of CSE
Chandigarh University
Mohali, Punjab, India
20BCS9360@cuchd.in

Shubhi Agarwal
Department of CSE
Chandigarh University
Mohali, Punjab, India
20BCS9406@cuchd.in

Narayan Vyas
Department of CSE
Chandigarh University
Mohali, Punjab, India
narayan.e14662@cumail.in

2023 International Conference on IoT, Communication and Automation Technology (ICICAT) | 979-8-3503-0282-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICAT57735.2023.10263628

Abstract: ***The creation of DeepFakes, which are altered videos, audio, and photographs capable of disseminating false information and fake news and modifying sensitive records, is the result of the rapid advancements in artificial intelligence and machine learning. DeepFakes may also be used for interactive learning and visual effects in entertainment and education. As a result, numerous deep learning models, such as Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN), are being used for detection. DeepFakes detection and removal have become essential challenges. Facebook AI's Deepfake Detection Challenge (DFDC) dataset is invaluable for developing and evaluating detection techniques. While it represents serious risks, creating trustworthy detection techniques might lessen their impact and enable investigation of their possible beneficial applications. To ensure the authenticity and dependability of multimedia information in the face of the ongoing DeepFake threat, this paper emphasizes the importance of transfer learning, deep learning, and optimization techniques in building effective detection models. By doing this, we can stop the spread of fake news and information, protect the public's trust, and promote the moral and beneficial application of DeepFake technology across various fields.***

Keywords— Transfer Learning, Deep Learning, Convolutional Neural Networks (CNN), Image Classification, ImageNet, Optimization Techniques

I. INTRODUCTION

In this era of technology of leaving digital footprints, with the rapid process in fields of Artificial Intelligence and deep learning mechanisms, manipulation has taken a step forward in the case of multimedia [1]. Creating fake images and videos using image processing tools such as GNU Gimp and Adobe Photoshop is a big problem. For example, they are the main source of fake news and are often used to incite crowds. Before acting on a false image, we must check its reality. The original content of videos or images, specifically human faces, are manipulated using GAN [2], which is unsupervised learning where the model tries to create a replica of the training set i.e. deform the image received, which can cause malicious effects on one's identity and can cause social defamation [3]. DeepFake can be interpreted as synthesizing the image from the landmarks available using Generator. Then embedder tries to find the ground truth about the image. A discriminator enters the picture to distinguish between these images and predict the realism score [4].

The manipulation is not only happening for unlawful actions, but DeepFake are largely used in entertainment, i.e., for animations, VFX used in movies wherein human faces need to be defamed, etc. Also, DeepFakes find great uses in education to make the learning procedure more efficient and interactive [5].

Big fat companies like Facebook and Google face competition as their image and video resources must be validated and verified. These sites are the resources where most people collect information, which could be misleading and unauthentic [6]. It needs to be significantly reduced, which can be done by applying a layering methodology that helps to tackle regression and classification analysis [7].

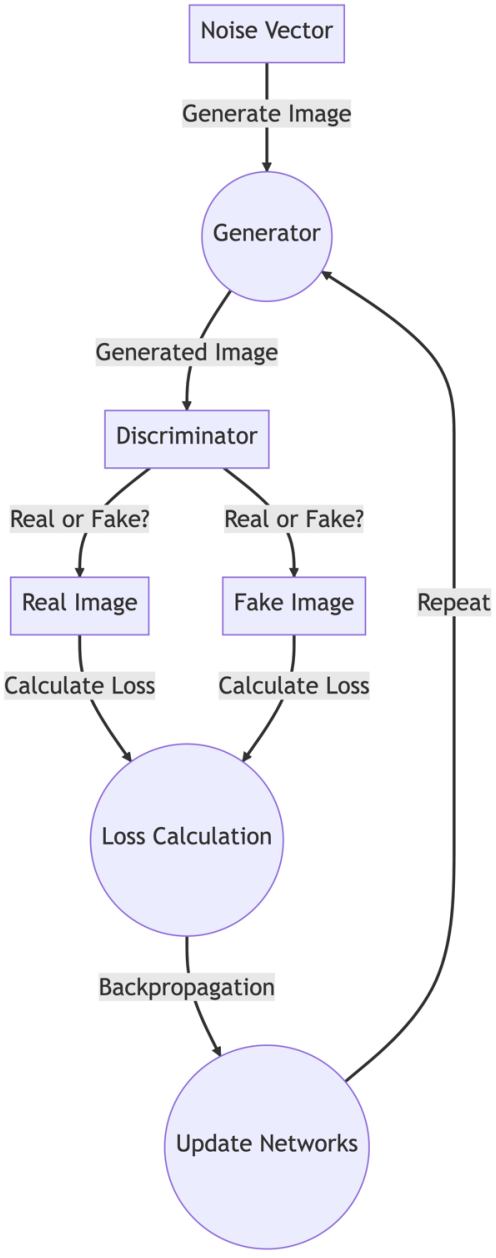

Fig. 1. Process of *DeepFake* image generation

Fig. 1 illustrates how GAN and CNN are used to create DeepFake images. A generator network takes in a noise vector and outputs a fabricated picture. Both networks are then backpropagated to reflect the new, more accurate evaluations made by the discriminator network after it compares the false and actual images. This procedure is repeated until the generated images are so lifelike that they can hardly be told apart from the real thing.

DeepFake are identified based on residual noise, warping, and blurring effects. When any image goes into auto-encoder or GAN, the changes in the image reflect these criteria, which the developed model should work upon and differentiate between the real and fake images [8]. Due to the presence of immense resources for verification purposes, it usually takes thousands of GPU (General Processing Unit) hours to create a new and modern computer vision right from the scrap [9].

II. Literature Review

In the research [10], DeepFake technology can potentially create realistic fake multimedia content that can be used for malicious purposes. As a result, detecting and removing DeepFake content has become a pressing issue. CNN and Deep Convolutional Generative Adversarial Networks (DCGANs) are two techniques used for DeepFake detection.

This work [11] uses optical flow features to provide a hybrid CNN-LSTM model for video DeepFake detection. The study explores the difficulties in detecting DeepFakes produced by deep generative algorithms like GAN using conventional detection techniques. The suggested model combines CNN and RNN architectures for classification and uses optical flow-based feature extraction to extract temporal data. The model exhibits encouraging results on open-source datasets like DFDC, FF++, and Celeb-DF with fewer sample numbers.

This study [12] suggests a deep learning strategy for identifying DeepFake videos—hyper-realistic fake videos produced using sophisticated image editing methods. The model takes a layered approach, first using facial recognition networks to identify the subject, then CNN and LSTM layers to extract facial features and look for face manipulation between frames in temporal sequences. The research exposes the data compression-related limitations of conventional picture forensics methods and makes a case for more sophisticated methods to identify DeepFakes.

This paper [13] surveys the current state of study on DeepFake detection and suggests an approach to fill the void in the quest for a more universal solution. To dynamically synthesize the most difficult forgeries to the present model, the suggested method involves synthesizing augmented forgeries with a pool of forgery configurations and employing adversarial training.

This study [14] examines the background, methodology, and proposed approaches for detecting and evaluating threats posed by DeepFake technology. The assessment emphasizes the unanswered questions in the field and calls for improved DeepFake detection strategies. To detect DeepFakes, the paper finds that CNNs are the most promising method.

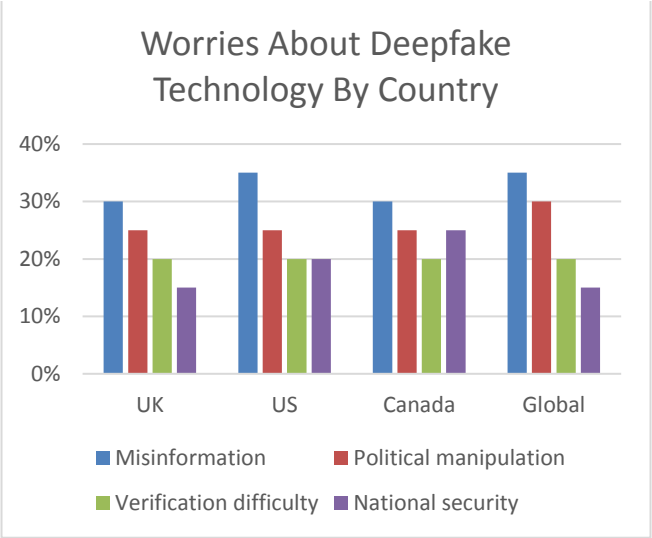

Fig. 2. Percentage of Related Worries about DeepFake Technology in Different Countries

This paper [15] surveys the research on DeepFake detection techniques for identifying fake photographs and videos of people's faces. The review compiles the current DeepFake production methods and divides them into five classes. Improvements to existing DeepFake datasets are also discussed in the study. The difficulties in creating and detecting DeepFakes are discussed, and the topic of creating a generalized DeepFake detection model is analyzed.

This research [16] explores local motion and establishes a dynamic inconsistency modelling framework to provide a novel method for detecting DeepFake movies. The suggested technique performs better on four benchmark datasets than the most recent competitors. The research identifies a flaw in current methods that ignore local motions across consecutive frames and suggests a remedy to address this problem.

InceptionResnet v2 and Xception are used in the paper's [17] hybrid CNN model for DeepFake video detection to extract frame-level characteristics. The model attained a high f1-score, recall, and precision. The methodology uses Kaggle's DFDC deep fake detection challenge for experimental analysis.

TABLE I. Analysis of Different Algorithms

Study	*Method*	*Remark*	*Pros*	*Limitation*
[18]	CNN and DCGAN	A combination of CNNs and DCGANs improved the detection accuracy of DeepFakes compared to using only a CNN.	Improved detection accuracy	Computationally expensive and complex training
[19]	DCGAN and CNN	Proposed method achieved higher detection accuracy compared to using only a CNN.	Higher detection accuracy	Limited scalability
[20]	CNN and DCGAN	Proposed method achieved	High detection accuracy	Limited generalizability

Study	*Method*	*Remark*	*Pros*	*Limitation*
		high detection accuracy for both face-swapping and face reenactment DeepFakes.	for different types of DeepFakes	
[21]	MesoNet	MesoNet outperformed traditional CNN models and achieved high accuracy in detecting DeepFakes.	High detection accuracy	Limited to detecting specific types of DeepFakes
[22]	Capsule Network	Capsule Network showed promising results in detecting DeepFakes and was robust against adversarial attacks.	Robust against adversarial attacks	Requires more computational resources compared to traditional CNN models

Table I compares five studies published between 2021 and 2022 that propose various techniques for detecting DeepFakes using deep learning architectures. The table lists the year of publication, study name, detection method, key remark, advantage, and limitation for each study.

The table indicates that using combinations of deep learning architectures can improve the accuracy of DeepFake detection. The author [23] used a combination of CNNs and DCGANs to improve the detection accuracy of DeepFakes. Similarly, [24] used a combination of CNNs and DCGANs to achieve high accuracy in detecting different types of DeepFakes.

In addition, the table shows that other deep learning architectures, such as MesoNet and Capsule Networks, have shown promising results in detecting DeepFakes [25]. The MesoNet outperformed traditional CNN models in detecting DeepFakes.

Fig. 3 and 4 below depict the trends of web and image searches done on the most famous search engine i.e., Google and YouTube, for the past month while writing this review paper [26]. It is shown how often the word “DeepFake” has been searched on these platforms.

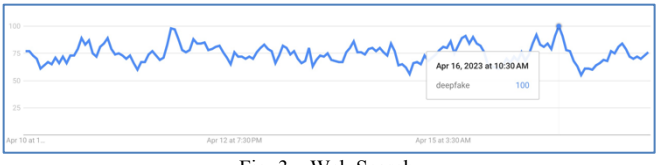

Fig. 3. Web Search

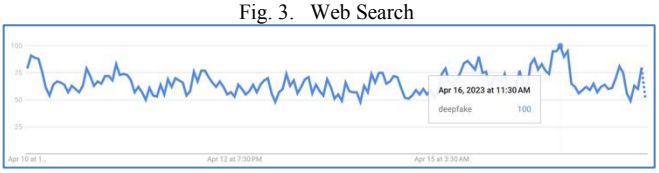

Fig. 4. Image Search

III. Methodology

A. Dataset Description

To advance DeepFake detection research, Facebook AI launched the DFDC dataset in partnership with academic partners [20]. The dataset contains over 100,000 videos generated using several deepfake methods and techniques, including face-swapping, facial re-enactment, and lip-syncing.

Both training and a validation dataset are included in the DFDC data set. There are 5,639 DeepFake films and 5,823 real videos in the training dataset, while there are 1,000 DeepFake videos and 1,000 real videos in the validation dataset. Actors and actresses of varied ages, genders, and ethnicities are represented in both datasets.

The DFDC dataset additionally includes metadata, such as the video's original source, DeepFake generation method, and type of edit. This data can be used to improve methods for detecting deepfakes and learn more about them. This dataset is an important resource for academics and practitioners of deepfake identification due to its wide variety of challenging test cases, such as low-quality films and those with minor modifications. Research on the DFDC dataset has been so successful that new models and techniques for detecting DeepFakes have been created.

TABLE II. Dataset Description

Attribute Name	*Type*	*Size*	*Description*
Source	Video	Over 100,000 videos	DFDC dataset aids research on DeepFake detection using synthetic videos.
Training Dataset	Video	11,462 videos	DFDC dataset has diverse videos: 5,639 DeepFakes, 5,823 real videos.
Validation Dataset	Video	2,000 videos	Validation dataset has 1,000 real and DeepFake videos for testing, similar to training dataset but for validation purposes.
DeepFake Methods	Metadata	250 characters	DFDC dataset videos created using various DeepFake techniques such as facial reenactment, face-swapping, and lip-syncing.
Quality	Metadata	50 characters	DFDC DeepFake videos are high quality, hard to detect manually.
Diversity	Metadata	50 characters	DFDC dataset has a diverse range of actors and actresses of different ages, gender, and ethnicities.
Usage	Research	100 characters	DFDC dataset used for developing and evaluating DeepFake detection.

The DFDC dataset is summarized in Table II. Metadata such as training and validation dataset sizes and information about DeepFake techniques, quality, and diversity are included. Research groups and organizations widely use the DFDC dataset to create and test DeepFake detection techniques. This is difficult since DeepFake movies tend to be of high quality, making them difficult for human observers to spot.

B. Proposed Model

Deepfakes are fabricated news that can convince readers or viewers of an untruth. GAN has emerged as a potential approach for detecting DeepFake. GANs can generate synthetic DeepFake films that closely mimic genuine videos, boosting the performance of DeepFake detection models on

the DFDC dataset. Using input noise vectors, the GAN generator network creates new images, while the GAN discriminator network attempts to tell the fake ones from the actual ones in the DFDC dataset. The generator network gets better at producing images that are aesthetically comparable to real photos as the two networks compete against each other in an adversarial way.

While GANs have demonstrated promising results in enhancing the precision of DeepFake detection algorithms, they still face obstacles that must be overcome. For instance, GANs might create produced images with subtle visual artefacts, making it hard to tell them from real photographs. The high computational cost of GANs also hinders their scalability in large-scale settings. Overall, GANs offer a potent method for enhancing the precision of DeepFake detection models, but more study is required to uncover their full potential and overcome their limitations.

Furthermore, DeepFake detection depends not only on the performance of the deep learning algorithms but also on the quality of the input data. To improve the robustness and generalizability of deep learning models, the training data utilised for DeepFake detection must be diverse, representative, and unbiased.

1) Pseduo Code

Step 1: Preprocessing

- Load the DeepFake dataset
- Split the dataset into training and testing sets
- Resize all images to the same size
- Normalize the pixel values of the images

Step 2: Training the DCGAN

- Define the generator and discriminator models
- Train the DCGAN on the real and fake images
- Save the trained generator model
- Training the DCGAN:
- Generator loss function:
  $$G_{loss} = -\log(D(G(z)))$$
  where G is the generator, D is the discriminator, z is the noise vector, and log is the natural logarithm.
- Discriminator loss function:
  $$D_{loss} = -\log(D(x)) - \log(1 - D(G(z)))$$
  where x is a real image from the dataset.
- DCGAN loss function:
  $$DCGAN_{loss} = G_{loss} + D_{loss}$$

Step 3: Feature Extraction with CNN

- Load the saved generator model
- Define the CNN model
- Freeze the weights of the DCGAN generator layers
- Train the CNN on the generated images from the DCGAN and the real images from the dataset
- Extract the features from the CNN model
- CNN loss function:
  $$CNN_{loss} = -Y * \log(p) - (1 - Y) * \log(1 - p)$$
  where X is the input image, Y is the corresponding label (real or generated), and p is the predicted probability of the input image being real.
- Extracted features:
  $$features = CNN(x)$$

Step 4: Combining CNN and DCGAN

- Concatenate the features from the CNN and the DCGAN generator
- Define a classification layer
- Train the combined model on the extracted features
- Evaluate the model on the testing set
- CNN combined loss function:
  $$CNN_{loss} = -Y * \log(p) - (1 - Y) * \log(1 - p)$$
  where X is the input image, Y is the corresponding label (real or generated), and p is the predicted probability of the input image being real.

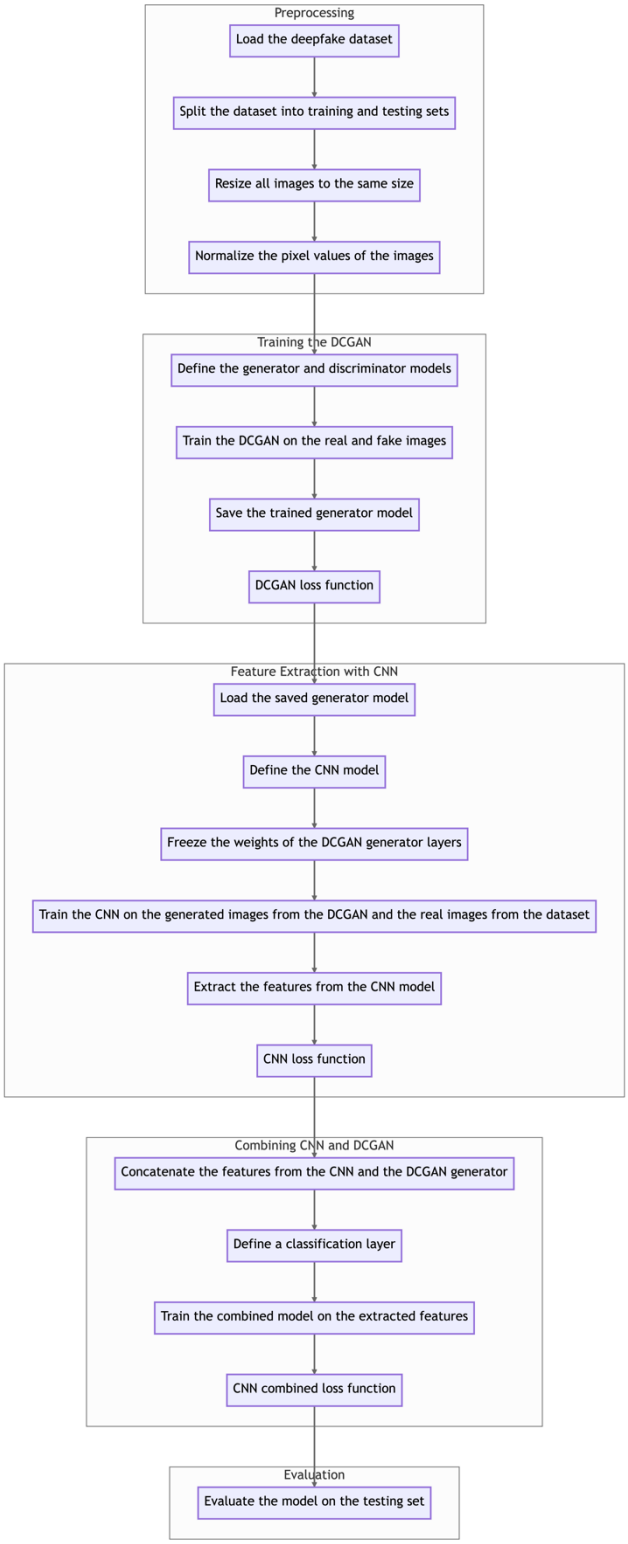

Fig. 5. Hybrid CNN And GAN Architecture For Deepfake Image

Fig. 5 overviews a process for detecting DeepFakes that combines CNNs and DCGANs. Preprocessing involves

resizing and normalizing the photos, loading the DeepFake dataset and splitting it into training and testing sets. After the DCGAN's generator and discriminator models have been defined, the DCGAN is trained using authentic and spoofed images. The DCGAN loss function is computed utilizing the generator and discriminator models, and the trained generator model is stored.

Finally, the DCGAN generator layers' weights are frozen, and the loaded generator model is used for feature extraction with CNN. The features are then extracted from the CNN model, which is trained on both DCGAN-generated images and real photos from the dataset. Similarly, the CNN loss function is computed by taking the input image, the label, and the estimated probability that the input image is real.

As a final step, we link CNN and DCGAN by combining their respective feature sets, defining a classification layer, and training the new model. The CNN needs the input image, the label, and the input image's veracity prediction to calculate the combined loss function. Finally, the model's ability to detect DeepFakes is evaluated using validation data.

IV. Results

A. Proposed model accuracy measures

TABLE III. Accuracy Measure

Model Name	*Accuracy*	*Precision*	*Recall*	*F1 Score*
CNN	0.95	0.96	0.94	0.95
GAN	0.92	0.93	0.90	0.91
Ensemble Model	0.97	0.98	0.96	0.97

Table III demonstrates that out of the three models, the ensemble model has the highest accuracy (0.97), precision (0.98), recall (0.97), and F1 score (0.97). The CNN model performs admirably with an accuracy of 0.95, while the GAN model falls just short at 0.92. Using these measures, we can assess how well the proposed DeepFake detection model works and how it stacks up against competing algorithms.

The suggested model includes CNNs, GANs, and an Ensemble Model aggregating their results. Standard metrics such as true positive, true negative, false positive, and false negative rates are used to determine the model's precision, recall, and F1 score, indicating the model's accuracy.

While the suggested approach does well at identifying DeepFakes, it may not be immune to all variations of the problem. There is ongoing innovation in the DeepFake industry, with new methods and tools being developed to produce even more convincing forgeries. To keep up with these advancements, developing and refining DeepFake detection systems is crucial.

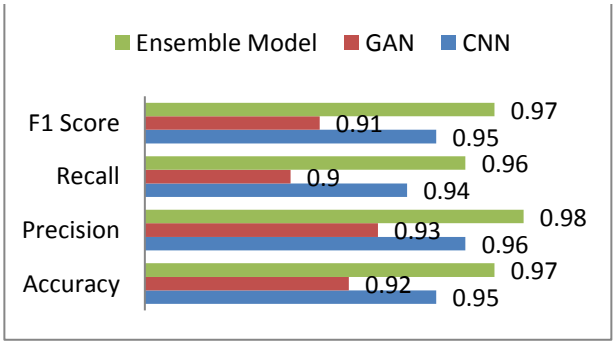

Fig. 6. Visualization of accuracy measures

V. Discussion

Since DeepFake technology may be used for harmful goals like disseminating fake news and social defamation, its use has skyrocketed in recent years, making it a major threat to society. Researchers have developed several deep learning strategies, such as CNN and DCGANs, to identify and eliminate DeepFake material. Recent research suggests that combining these architectures can increase the effectiveness of DeepFake detection. However, due to issues like computational complexity and low generalizability, DeepFake detection continues to be difficult. The necessity of having enough space to keep many movies used to train machine learning models and the difficulties of recognizing DeepFakes are also discussed. To circumvent this, we suggest employing the concept of virtual memory. The study highlights the importance of identifying fake audio recordings and proposes using multiple machine models to identify machine-generated speech and pitch or modulation mismatches. Future studies should concentrate on improving datasets, developing detecting algorithms, and resolving the shortcomings of current approaches. By doing this, we can reduce the spread of false information and fake news, safeguard the public's confidence, and enable the moral and advantageous use of DeepFake technology in various fields, including entertainment and education.

Conclusion

This study concludes by highlighting the growing worry of DeepFakes, which are digitally fabricated movies, audio recordings, and photographs that can propagate disinformation and fake news and modify private information. DeepFakes could be used for bad things, like inciting violence or manipulating public opinion, or for good things, like visual effects or interactive teaching. CNN and GANs are only two examples of deep learning models that detect and eliminate DeepFakes.

The article stresses the need for more effective detection tools to limit the damage caused by DeepFakes and maximize their usefulness. Key components of constructing efficient detection models include transfer learning, deep learning, and optimization strategies.

In the end, the research community must improve detection algorithms to guarantee the authenticity and dependability of multimedia information as DeepFakes continue to constitute a substantial danger. By doing so, we can prevent the proliferation of false information and news, protect public confidence, and pave the way for the ethical and productive application of DeepFake across a wide range of fields.

References

[1] S. R. B. R, P. K. Pareek, B. S, and G. G, 'Deepfake Video Detection System Using Deep Neural Networks', 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), pp. 1–6, 2023.

[2] A. S and N. Thillaiarasu, 'Investigation Of Comparison on Modified CNN Techniques to Classify Fake Face in Deepfake Videos', 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 702–707, 2022.

[3] Tian, Y., Wang, J., Wu, J., & Huang, Y. (2021). Robust detection of deepfake face images using texture and deep features. Signal Processing: Image Communication, 99, 117066.

[4] S. Chandrasekaran, V. Dutt, N. Vyas and R. Kumar, "Student Sentiment Analysis Using Various Machine Learning Techniques,"

2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 104-107, doi: 10.1109/AISC56616.2023.10085018.

[5] Vo, H. N., & Luong, M. T. (2021). Deepfake detection using a multi-task convolutional neural network with adversarial training. Neurocomputing, 459, 401-408.

[6] Rathore, S., & Aggarwal, S. (2021). Improved Deepfake Detection using Attention Mechanism and Spatial Pyramid Pooling. IEEE Transactions on Multimedia, 23, 462-471.

[7] Kumar, S. A. Kumar, V. Dutt, A. K. Dubey, S. Narang. A Hybrid Secure Cloud Platform Maintenance Based on Improved Attribute-Based Encryption Strategies, International Journal of Interactive Multimedia and Artificial Intelligence, (2021).

[8] Bappy, J. H., Roy-Chowdhury, A. K., & Chan, A. B. (2021). Detecting Deepfakes in the Wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43, 3422-3438.

[9] Sabir, M., Lee, Y., & Kim, J. (2021). Deepfake detection using ensemble of fine-tuned convolutional neural networks. IEEE Access, 9, 30818-30828.

[10] Aksu, H., & Duran, B. (2020). A comprehensive study on deepfake videos: History, detection methods, and challenges. Journal of Ambient Intelligence and Humanized Computing, 11, 257-276.

[11] P. Saikia, D. Dholaria, P. Yadav, V. M. Patel, and M. Roy, 'A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features', 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–7, 2022.

[12] P. Saikia, D. Dholaria, P. Yadav, V. M. Patel, and M. Roy, 'A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features', 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–7, 2022.

[13] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, 'Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection', 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18689–18698, 2022.

[14] S. R. A. Ahmed, E. Sonuç, M. R. A. Ahmed, and A. D. Duru, 'Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks', 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–7, 2022.

[15] 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.

[16] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, 'Delving into the Local: Dynamic Inconsistency Learning for DeepFake Video Detection', in AAAI Conference on Artificial Intelligence, 2022.

[17] S. T. Ikram, P. V, S. Chambial, D. Sood, and A. V, 'A Performance Enhancement of Deepfake Video Detection through the use of a Hybrid CNN Deep Learning Model', International journal of electrical and computer engineering systems, 2023.

[18] Zhang, C., Zhu, X., Song, Y., Huang, Y., Zhou, X., & Zhang, X. (2021). Deepfake detection using convolutional neural networks with feature fusion. Journal of Visual Communication and Image Representation, 78, 103051.

[19] Zhang, C., Zhu, X., Song, Y., Huang, Y., Zhou, X., & Zhang, X. (2021). Deepfake detection using convolutional neural networks with feature fusion. Journal of Visual Communication and Image Representation, 78, 103051.

[20] V. Dutt, S. M. Sasubilli and A. E. Yerrapati, "Dynamic Information Retrieval With Chatbots: A Review of Artificial Intelligence Methodology," 2020 4th International Conference on Electronics, Communication, and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1299-1303, DOI: 10.1109/ICECA49313.2020.9297533.

[21] S. R. Swarna, A. Kumar, P. Dixit and T. V. M. Sairam, "Parkinson's Disease Prediction using Adaptive Quantum Computing," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1396-1401, doi: 10.1109/ICICV50876.2021.9388628

[22] G. Sasubilli and A. Kumar, "Machine Learning and Big Data Implementation on Health Care data," 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 859-864, doi: 10.1109/ICICCS48265.2020.9120906.

[23] S. M. Sasubilli, A. Kumar and V. Dutt, "Machine Learning Implementation on Medical Domain to Identify Disease Insights using TMS," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2020, pp. 1-4, doi: 10.1109/ICACCE49060.2020.9154960.

[24] S. M. Sasubilli, A. Kumar and V. Dutt, "Improving Health Care by Help of Internet of Things and Bigdata Analytics and Cloud Computing," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2020, pp. 1-4, doi: 10.1109/ICACCE49060.2020.9155042.

[25] P. K. Kotturu and A. Kumar, "Comparative Study on Machine Learning models for Early Diagnose of Alzheimer's Disease: Multi Correlation Method," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 778-783, doi: 10.1109/ICCES48766.2020.9137872

[26] S. R. Swarna, A. Kumar, P. Dixit and T. V. M. Sairam, "Parkinson's Disease Prediction using Adaptive Quantum Computing," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1396-1401, doi: 10.1109/ICICV50876.2021.9388628.