A Novel Approach for detecting Deepfake Face using Machine Learning Algorithms

Manoj Kumar
CSE Department
G. L. Bajaj Institute of Technology and Management, Greater Noida,
Uttar Pradesh 201310, India
mksinghdeveloper@gmail.com

Praveen Kumar Rai
CSE Department
G. L. Bajaj Institute of Technology and Management, Greater Noida,
Uttar Pradesh 201310, India
praveen19jul@gmail.com

Pankaj Kumar
CSE Department
G. L. Bajaj Institute of Technology and Management, Greater Noida,
Uttar Pradesh 201310, India
unpankaj@gmail.com

2024 2nd International Conference on Disruptive Technologies (ICDT) | 979-8-3503-7105-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDT61202.2024.10489036

Abstract— **In today's digital age, the ability to identify, differentiate, and authenticate manipulated online content is essential. Being ability to discriminate between the real and the fake is crucial. Recent advances in technologies such as artificial intelligence, machine learning, and deep learning are playing a major role in the generation of deepfake media (images and videos). Very realistic deepfake images and videos can be produced by utilizing sophisticated deep learning models such as generative adversarial neural networks (GANs) and autoencoders, in conjunction with a sizable image collection pertaining to the subject matter. Deepfakes (DF) refer to artificially synthesized images or videos created using features such as face swapping and facial expression recombination. These face manipulation techniques have become extremely sophisticated. Deepfakes can be used to create child pornography, pornographic images of celebrities, revenge porn, fake news and harassment, spreading disinformation on social media platforms, financial fraud, election manipulation, and more. Therefore, there is a need to design and develop a robust framework to identify these deepfake images and videos.**

The purpose of this paper is to identify deepfakes from visual deepfake datasets and perform a comparative analysis of deepfake detection through machine learning algorithms.

Keywords—*Deepfake, Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, Generative Adversarial Neural Networks (GANs), Face Swapping*

I. Introduction

The advent of advanced technology, artificial intelligence, and deep learning methods has made it possible for very sophisticated false media, or "deepfakes," to proliferate. Deepfakes are artificially created media that use machine learning algorithms; they frequently contain movies and photos that have been altered [1]. Deepfakes represent a new subfield of artificial intelligence in which one person's facial features are algorithmically superimposed onto another person's representation. Specifically, various generative adversarial network (GAN)-based techniques have emerged to generate high-fidelity deepfake images. Like many rapidly evolving technologies, deepfakes raise innovative considerations regarding authenticity, ethics, and social impact, all of which merit a balanced and ongoing discussion from a variety of perspectives [2]. Deepfake involves the process that utilizes Generative Adversarial Networks (GANs) to produce fake images and videos. The word "deepfake" is derived from the fusion of "deep" and "fake" and originates from the utilization of artificial intelligence and deep learning methodologies (a form of machine learning involving numerous layers of processing) to create deceptive material.

In 2017, the term "deepfakes" emerged when Reddit administrators created a subreddit called "deepfakes" and started sharing videos using face-swapping technology in which celebrity appearances were superimposed onto preexisting pornographic content. While deepfake technology has valid applications in industries like the film industry and entertainment, there is a growing concern over its potential misuse for disseminating disinformation, defaming persons, or conducting blackmail. Although India does not have dedicated regulations expressly addressing deepfake technology, many existing laws can be utilized to impose legal repercussions on those who misuse deepfakes.

Deepfake technology can be used for fraudulent activity targets, namely to mislead the public by broadcasting false information and propaganda. For example, deepfake images or manipulated videos of politicians or celebrities saying something they didn't say are also called "misleading." Information that can help change public opinion. Deepfakes are generated by the use of two distinct AI deep learning algorithms. One algorithm generates very authentic replicas of authentic media, and the other detects whether a replica is counterfeit and, if so, how close it is to the original. while. The first algorithm generates a fabricated image that undergoes refinement through feedback from the second algorithm, resulting in a more authentic appearance. This method repeats itself until the second algorithm identifies zero instances of false positives[6]. Detecting deepfakes remains a difficult task as the technology for creating deepfakes is improving day by day. However, there is a growing number of users who are interested in generating fake content. This tool enables users to generate lifelike mock-ups without a significant financial commitment. Moreover, upon reviewing the community forums, one can observe a plethora of discussions and tests conducted by individuals who have developed mock-ups. An intriguing pattern emerges: an increased number of novices are engaging compared to past forums. Quickly learn from your mistakes [3].

As high-quality deepfake technology becomes more accessible, we should anticipate increased misunderstanding and debate around the production and dissemination of phony information. Several significant technology businesses have started campaigns to counteract deepfakes in an effort to stop the exploitation of this technology from jeopardizing information authenticity and privacy. These techniques primarily, but not exclusively, rely on deep learning and his form of artificial intelligence known as GAN. Communities and forums are the main driving force behind the growing scope of deepfakes and their open source creation software and applications.

A huge amount of work is being done on deepfakes in terms of improving applications, marketing, addressing negative impacts, and enforcing new laws. The process for developing and generating deepfakes is something we're going to study. Creating deepfakes may be done in several different methods. Utilizing Generative Adversarial Networks, often known as GANs. It can construct fake images and train themselves to recognize patterns using algorithms in Fig. 1

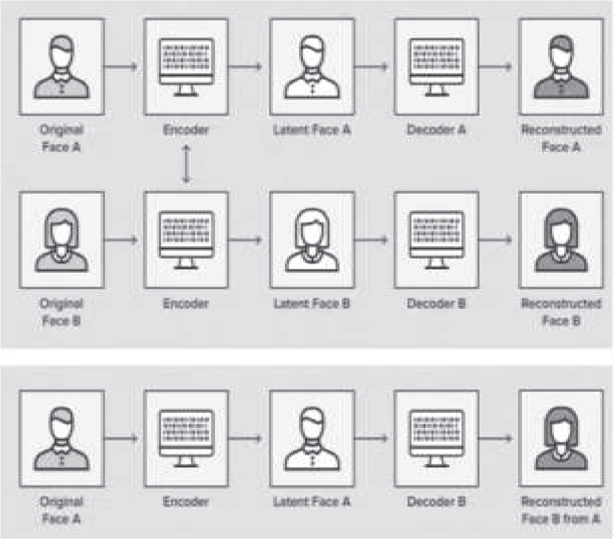

Fig 1. Working of Deepfake Generation

The rising popularity of deepfakes may be attributed to the high quality of the modified films and the user-friendly nature of the program, which caters to a diverse spectrum of users with varied levels of computer proficiency, including both experts and beginners. These apps mostly utilize deep learning techniques for their development. Deep learning is renowned for its capacity to depict intricate and multidimensional data. One particular sort of deep neural network that possesses this feature is a deep autoencoder, which is commonly employed for the dimensionality reduction and picture compression [5]. The initial effort to generate deepfakes involved the utilization of a software application called FakeApp, which was developed by a Reddit user. This application deployed a sophisticated architecture consisting of both an autoencoder and a decoder[7]. This approach involves using an autoencoder to extract latent characteristics from a face picture and then using a decoder to rebuild the facial image. Swapping faces between the source and final images requires two sets of encoder-decoders. Both network pairs have similar encoder configurations and are utilized to train a certain number of images. Deepfakes utilize advanced autoencoders to generate completely new images, surpassing the capabilities of traditional encoders.

Deepfake applications require the use of two autoencoders to swap images and actions between different images. Machine learning is essential for the development of deepfakes. Deepfakes use artificial neural networks for their operation. Creating a deepfake begins at that point. Developers start by feeding a large amount of genuine video material into a deep neural network to create the deepfake. This network is trained to accurately identify the intricate patterns and characteristics of an individual. Performing this task ensures that algorithm receives precise representation of person's appearance from different angles. [11].

II. Related works and studies

Early studies identified patterns of physiological behaviour, such as inconsistent head posture [3], unnatural eye blinking [4], and the relationship between facial expressions and head movements [5]. However, these artifacts were corrected in the second generation deepfake dataset, which limited the detection performance. Their study is based on both new problems with existing detection methods and possible future improvements. Recent research has also revealed deepfakes based on bio signals [8]. Detection methods based on deep neural networks (DNNs) have become popular.

For example, a two-stream CNN is used, Meso-4 focuses on the mesoscopic properties of the image, a VGG19-based capsule structure is used, ResNet is used to capture distortion artifacts, and ResNet is used to detect fake faces. Classic exceptions were used [9]. Because videos contain temporal features, some researchers combine CNNs and RNNs for classification. Although DNN-based methods have achieved some success due to their powerful feature extraction capabilities, they still have limitations for advanced deepfakes. To address this issue, learning-based methods are being further investigated. For example, Fake Spotter detects fake faces by monitoring the behaviour of neurons. Recently, researchers have combined useful modules and important features which detect face manipulations in videos, where the input is a sequence of video frames. Dunn et al. Attention mechanisms used to improve detection efficiency. Similarly, a vision transformer was used for detection. Gram-Net and Intel extract texture information from images to improve robustness. This study integrates attention processes and textural characteristics. Optimise the process of feature extraction to enhance achieving the preciseness of deepfake detection, rather than focusing on the development of sophisticated and complex neural networks. An additional approach that is based on eye-blinking detection was presented by the authors.

The authors think that eye-blinking is eliminated when Deepfakes are analysed. To improve generalization ability, Cozzino et al. propose an autoencoder-based learning integration. Wang et al. ResNet was trained using a multi-class ProGAN dataset and showed that generalization can be improved with proper pre-processing and post-processing [10]. Facial radiography examines the blended boundaries between the face and the background to detect changes in the face. HRNet is used as the framework. Resampling techniques used for deep generative models result in distortions in the frequency domain, which has led to the development of several detection methods based on spectrum analysis. However, detection based only on the frequency spectrum provides poor performance and generalization. Frequency-domain artifacts can be reduced by spectral regularization, focal frequency loss, or training with a spectral differentiator. Pseudo-polishers can generate surface reconstructions and reduce artifact patterns [6]. To effectively detect deepfakes, Deepfakes have a growing number of minor functional problems that need to be discovered. Keep in mind that you should not combine SI and CGS units, such as using amperes for current and Oersted for magnetic fields. Because the dimensions of the equation are not balanced, this phenomenon frequently results in misunderstanding. If you have to utilize mixed units, you must provide the units for each quantity in the equation [5].

III. PROPSED METHODLOGY

Even though there are several tools available for the creation of deepfakes, there are just a handful of tools that are available for the detection of deepfakes. For deepfake face detection, we require the significant amount of dataset of real and fake images. After getting the dataset we implement the proposed method which are shown in Fig 2.

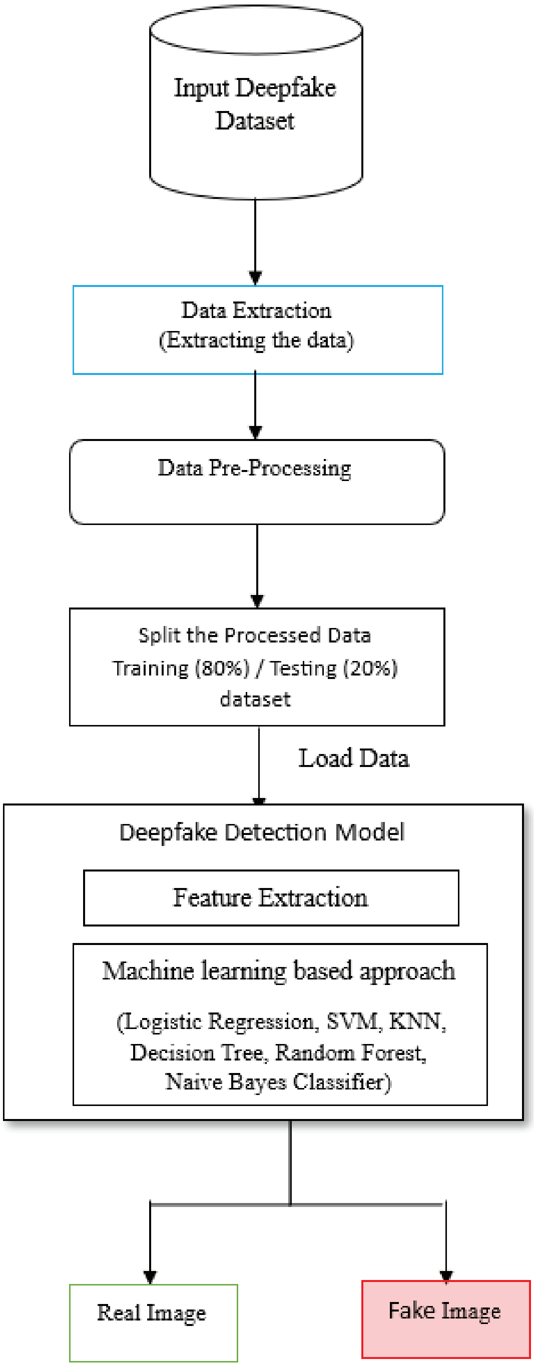
Input Deepfake Dataset
Data Extraction (Extracting the data)
Data Pre-Processing
Split the Processed Data Training (80%) / Testing (20%) dataset
Load Data
Deepfake Detection Model
Feature Extraction
Machine learning based approach
(Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Naive Bayes Classifier)
Real Image
Fake Image

Fig 2. Proposed model of Deepfake Detection

The proposed methodology explains the procedure, which involves the following phases:

- Input Deepfake dataset: - The data was gathered from several sources. Images were gathered from many sources to provide datasets for training our proposed models, including the CelebA dataset (10,000 photos) and a 100K fake dataset (10,000 images). We consolidated the collected data to promptly and precisely identify different forms of media.

- Extracting the data: - The process of collecting or assembling different types of data from multiple sources, some of which may be unstructured or unprocessed.

  Manual data extraction requires human participation in the tasks like reading, transferring, and replicating the data. Automated data extraction utilizes software tools like scripts, programs, or applications to carry out data extraction activities without human intervention.

- Data Pre-processing: - All images have been taken from different sources according to the information in the previous section. Therefore, it is crucial to standardize the preprocess of all images when providing them as input to the database during model training. Resizing of images, Data augmentation and cropping the face which perform all processed images in the collection to a uniform size. (varies depending on the model). When resizing an image, the total number of pixels is adjusted, while reconfiguration occurs when correcting lens distortion or spinning the image. It is crucial to adjust the overall pixel count by increasing or decreasing it. Zooming involves increasing the pixel count to provide greater information while viewing a picture. Data augmentation involves altering images in the input dataset by changing factors such as orientation, position, scale, and brightness. This strategy is utilised to enhance overall performance by making the model more generic. The primary objective of this approach is to train the model to recognise distinctive properties of each class and effectively differentiate across classes. Different methods can be used on the training picture dataset, such as rotation, flipping, zooming, and shearing. The augmentation techniques enhance the model's ability to generalise. After augmenting the data, we cropped the faces to improve the performance of Deepfake detection algorithms.

- Spilt the Processed data: - To perform a train-test split data. We generally separation the initial dataset into training and testing data. We train our model by utilizing a section of the original dataset known as the training dataset. Next, we assess its capacity to apply to a new or unknown dataset, referred to as the test set.Training and testing datasets are essential components of machine learning. During model training, the training dataset is utilized, while the test dataset is used for model evaluation. We utilise libraries such as scikit-learn in Python. Import the `train_test_split` function, provide the dataset, and define the test size as 20% with the train set being 80%. This function randomly splits the data into training and testing sets while maintaining the distribution of classes or outcomes.

- Deepfake detection Model: - In this section, the data loader is responsible for loading pre-processed photos that have cropped faces. Divides them into a training set and a testing set for further analysis. After being thoroughly examined by the deepfake detection algorithm, the process includes feature extraction and utilisation of other machine learning models. During feature extraction, the facial regions were isolated from the processed picture dataset. It greatly decreases the amount of parameters and calculations while yet achieving good accuracy.

We proposed extracting features and accurately recognising frame-level characteristics with little processing resources. Following the feature extraction. We utilise many types of machine learning models to distinguish between authentic and false images. These models utilise a supervised learning method. After implementing deepfake detection models, we conduct a comparison examination of several machine learning approaches. Each technique includes a validation step to ensure the precision of predictions and the advancement of categorization. The Results and Discussion section includes a comparative examination of how machine learning algorithms address the issue of false pictures using performance metrics including as accuracy, AUC, and confusion matrices.

IV. Result and discussion

This section outlines the assessment factors and the outcomes of the suggested resolution. The experiment's results are analysed in depth and presented utilising evaluation of performance confusion metrics. Six different supervised base machine learning algorithms were used in the pipeline, and their performance was measured using accuracy and AUC. However, these models use deep neural networks and are computationally expensive. This approach uses a simple machine learning algorithm to detect deepfakes in images using the frequency spectrum. Deepfakes have been observed to deviate from real images when viewed in the frequency domain. This deviation is then used and run through various machine learning classifiers to see how each algorithm performs. Table 1 shows the results of the pipeline using all 10,000 images.

Another crucial parameter to be used to consider when evaluating the proposed approach in this study is the confusion matrix. A confusion matrix is a square matrix with n rows and columns, where n is the number of classes in a classification problem. This statistic enables you to evaluate the algorithm's performance by contrasting its predictions with the real label values.

To determine accuracy, two classes are categorized as positive or negative. The four primary variables are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confusion matrix allows for determining various variables which are given below as:

True Positive (TP): - These are the classes that have provided accurate predictions, and the classes that have provided accurate predictions have been categorized as positive.

False Positive (FP): - These are the classes that have been incorrectly predicted, with these classes being identified as positive.

True Negative (TN): - These are the classes that have been accurately predicted, with those correctly predicted being classified as part of the negative category.

False Negative (FN): - These classes are considered to have been predicted inaccurately, while the classes that have been projected inaccurately are labelled as negative.

Accuracy: The accuracy is frequently utilized in task categorization to calculate the proportion of correctly predicted instances by the models under evaluation. The calculation involves dividing the number of accurately predicted issues by the total number of predictions made.

This formula is very useful to evaluate the accuracy of any model. It provides a fundamental understanding of a distinct categorization challenge.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision: This data is crucial for evaluating the accuracy of estimates by calculating the proportion of precise predictions to all projections made.

$$Precision = \frac{TP}{TP+FP}$$

Recall: This measurement indicates the calculation of the percentage of accurate positive predictions made by the equation and the overall number of true positive classes.

$$Recall = \frac{TP}{TP+FN}$$

When it comes to machine learning and statistics, a confusion matrix provides a helpful tabular representation to determine the performance of a classification model. The summary showcases the categorization results by providing the counts of true positive, true negative, false positive, and false negative predictions. Understanding the confusion matrix is crucial for assessing model performance, identifying misclassifications, and improving prediction accuracy.

When evaluating a classification model, a confusion matrix of size N x N is utilized, with N representing the total amount of target classes. It includes an evaluation of the real values and the machine-learning model's assumptions. Each technique has its own advantages and drawbacks, including the one we are proposing for our dataset. Figure 3 illustrates the confusion matrix of several machine learning methods. In the future, we will include other pattern recognition techniques in the different datasets.

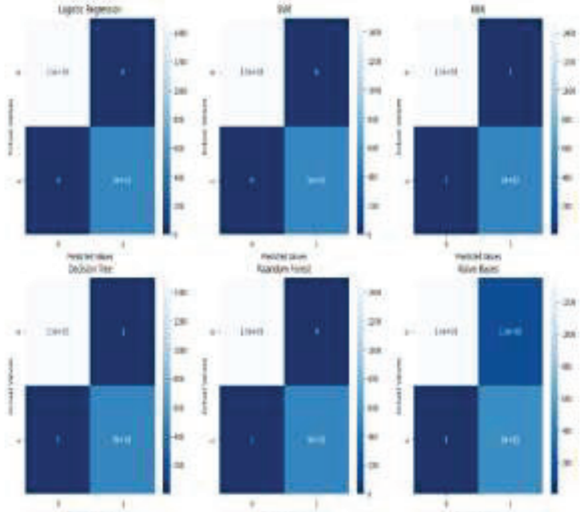

Fig 3: - Confusion matrix in the different ML models applied on the images test subset

We have determined the accuracy of the proposed model by analysing the confusion matrix on the testing data, resulting in an overall accuracy in Table 1. Deepfakes have been observed to deviate from real images when viewed in the frequency domain. This deviation is utilized and processed by different machine learning classifiers. Testing the accuracy of our model by applying an algorithm.

TABLE 1: PERFORMANCE OF DIFFERENT PROPOSED ML MODEL

Model	Accuracy	AUC
Logistic Regression	100.00%	1.00
Support Vector Machine (SVM)	99.70%	1.00
K nearest neighbors (K-NN)	99.85%	1.00
Decision Tree (DT)	99.80%	1.00
Random Forest	99.95%	1.00
Naive Bayes	92.35%	0.95

V. CONCLUSION

Deepfake detection is a crucial research topic that focuses on detecting manipulated images and videos that have been created with advanced artificial intelligence methods. Deep faking is a relatively recent technique that is being used extensively for spreading false information and hoaxes to the public. Even though not all deep fake contents are malicious, it is necessary to locate them since some of them represent a threat to the entire world.

This study's primary objective was to establish a reliable approach for recognizing deepfake images, and it was successful in doing so. Many researchers have been engaging in a lot of effort to identify deep fake elements by applying several different methods. Despite many obstacles in deepfake detection, including non-standardized datasets, advancements in technology, and adversarial attacks, researchers are consistently creating new models and methods to enhance detection efficiency. However, further improvements can be made to increase the usability and accuracy of deepfake models post-development. It is crucial to have a thorough understanding of the many strategies used for detecting deepfakes. Therefore, rather than developing a brand-new advanced model, Exploring the aim of a research study is to improve the knowledge of a dataset that is less well-known by employing fundamental machine learning approaches. We will do further research on various applications of feature generation and decision-making as well as ensemble detection that will improve performance.

REFERNCES

[1]. Aghasanli, A., Kangin, D., & Angelov, P. (2023). Interpretable-through-prototypes deepfake detection for diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 467-474).

[2]. Kumar, Pankaj, S. K. Singh, and Surya Deo Choudhary. "Reliability prediction analysis of aspect-oriented application using soft computing techniques." *Materials Today: Proceedings* 45 (2021): 2660-2665.

[3]. Wang, T., & Chow, K. P. (2023, June). Noise based deepfake detection via multi-head relative-interaction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 14548-14556).

[4]. Le, B. M., & Woo, S. S. (2023). Quality-agnostic deepfake detection with intra-model collaborative learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22378-22389).

[5]. A Mary, A Edison, A., & Edison, A. (2023, May). Deep fake Detection using deep learning techniques: A Literature Review. In 2023 International Conference on Control, Communication and Computing (ICCC) (pp. 1-6). IEEE.

[6]. Kumar, Pankaj, and S. K. Singh. "An emerging approach to intelligent techniques—soft computing and its application." *Internet of Things and Big Data Applications: Recent Advances and Challenges* (2020): 171-182.

[7]. FBI. Deepfakes and stolen pii utilized to apply for remote work positions. https://www.ic3.gov/Media/ Y2022/PSA220628, June2022.Accessed:2022-07-01.1

[8]. Lacerda, G. C., & Vasconcelos, R. C. D. S. (2022). A Machine Learning Approach for Deepfake Detection. arXiv preprint arXiv:2209.13792.

[9]. Kumar, Pankaj, et al. "Predictive analysis of novel coronavirus using machine learning model-a graph mining approach." *J. Math. Comput. Sci.* 11.3 (2021): 3647-3662.

[10]. Rana, M. S., Murali, B., & Sung, A. H. (2021, July). Deepfake Detection Using Machine Learning Algorithms. In 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 458-463). IEEE.

[11]. Web Link: https://www.fortinet.com/resources/cyberglossary/deepfake

[12]. Web Link: https://www.spiceworks.com/it-security/cyber-risk-management/articles/what-is- deepfake/

[13]. Kumar, Pankaj, Renuka Sharma, and S. K. Singh. "Predictive Analysis of Real-Time Strategy using Face book's Prophet Model on Covid-19 Dataset of India." *Journal of Pharmaceutical Research International* 33.51A (2021): 305-312.

[14]. Web Link: https://www.javatpoint.com/how-to-check-the-accuracy-of-your-machine-learning-model

[15]. Alanazi, F. (2022, December). Comparative Analysis of Deep Fake Detection Techniques. In 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 119-124). IEEE.

[16]. Deb, Nabamita, et al. "Suppressing the Spread of Fake News Over the Social Web of Things: An Influence Maximization-Based Supervised Approach." *IEEE Systems, Man, and Cybernetics Magazine* 9.4 (2023): 20-25.