Deep Sight: Unveiling Digital Deception

Deepa N.[1], Edwin Joel P.[2], Cletus Sylphia P.[3], Ahalya G.[4]

Artificial Intelligence and Data Science, PSNA College of Engineering and Technology, Anna University, Dindigul, India.

E-mail: [1]deepanatrayan@psnacet.edu.in, [2]edwin810joel@gmail.com, [3]cletusjact@gmail.com, [4]ahalyagnanamurugan3@gmail.com

Abstract

Deepfakes are rapidly increasing and generating severe problems with fake news, fraud and criminal behavior. The issue with AI-generated media has evolved into more accurate and recognizing fake data. The proposed "Deep Sight: Unveiling Digital Deception" is an AI/ML-driven system detects deepfake in various kinds of media including images, videos and audio. It compares differences in facial features, voice structures and action behavior to recognize fake data. This system separates each component (video, audio, images and contents) into real and fake data to improve transparency. It provides subtitle to show the sections of the text to be edited by AI. A built-in monitoring tool allows user to report possible data directly transferred to cybersecurity authorities for immediate action. The "Deep Sight" improves digital responsibility using advanced detection techniques and addressing AI-driven fraud. It enables people and organizations to protect media quality by avoiding modification. This concept proposes to create a safe global network by solving the increasing risk of deepfake technologies.

Keywords: Deepfakes, AI Detection, Media Forensics, Cybersecurity, Digital Integrity.

1. Introduction

An advanced artificial intelligence (AI) leads to increase in number of deepfakes include videos, images and audio recordings. These AI-generated fake data can accurately replicate the human features, voices and emotions proving it difficult for automated systems and people to differentiate between real and fake data. As a result, deepfakes are widely used to transmit fake data, create identity theft, influence public opinion and support a variety of cybercrimes. These fake data affects digital security, decreases public trust in online media and also has a possible chance to damage national security and social stability. This developed

DOI: https://doi.org/10.36548/rrrj.2025.2.008

challenge solved by using smart, advanced detecting systems to maintain the developed processes. This system creates an advanced deepfake detection based on AI and machine learning to handle this issue.

The proposed deep sight method is designed to evaluate different media types by identifying differences in face, vocal features and movement patterns usually unpredictable by the human eye or ear. This system uses visual signals and detailed contents to identify data as real or modified to detect and highlight AI-generated data. Furthermore, this work has a monitoring tool to allow users to identify fake data and send it to the relevant cybersecurity authorities for evaluation. It also improves social media responsibility by using advanced deepfake detecting algorithms. This solution provides organizations, people and government authorities with reliable methods for detecting and responding to AI-driven fraud. Advanced deepfake technologies help to secure and trust digital environment. The advanced artificial intelligence (AI) leads to a rise in the amount of deepfakes include videos, images and audio recordings.

2. Related Work

Deepfake detection has gained momentum as synthetic media becomes increasingly sophisticated. Rossler et al. [1] proposed FaceForensics++ to highlight the dataset that enables training and evaluation of facial forgery detection models. Their dataset played a crucial role in advancing spatial-based detection systems. Dolhansky et al. [2] introduced the DeepFake Detection Challenge (DFDC) dataset to support the development of robust AI models capable of identifying manipulated content. This large-scale dataset focuses on real-world scenarios, improving model generalization. Korshunov and Marcel [3] examined vulnerabilities in biometric systems highlighting the deepfakes that can compromise face recognition. Their work reinforced the importance of integrating anti-spoofing measures within detection frameworks. Ahmed et al. [4] provided a comprehensive survey of deepfake detection techniques, categorizing them based on the underlying architecture and data types. Their taxonomy serves as a valuable reference for system developers. Li et al. [5] introduced a method that detects deepfakes by analyzing abnormal eye blinking behavior is often missing in synthesized videos. Similarly, Güera and Delp [6] applied Recurrent Neural Networks (RNNs) to capture temporal inconsistencies across video frames. Tolosana et al. [7] provide a comprehensive survey of face manipulation and DeepFake detection techniques. The study reviews major face synthesis approaches, including identity swapping and facial reenactment, and examines state-of-the-art detection methods based on deep learning. It also discusses the highlighted datasets, evaluation protocols and key challenges in developing robust and generalizable fake detection systems. Kumar, Rai, and Kumar [8] propose a novel machine learning–based approach for deepfake face detection. Their method extracts discriminative facial features and applies multiple classifiers to identify manipulated images. Experimental results demonstrate improved detection accuracy compared to existing techniques, highlighting the effectiveness of traditional machine learning algorithms for deepfake identification. Yang, Li, and Lyu [9] introduce a deepfake detection method based on inconsistencies in estimated 3D head poses. By analyzing geometric relationships between facial landmarks, their approach effectively reveals manipulation artifacts, demonstrating that physical and spatial inconsistencies can expose synthesized or altered facial videos. Mirsky and Lee [10] present a comprehensive survey on deepfake generation and detection techniques, discussing creation pipelines, detection methods, datasets, threats and future challenges serving as a foundational reference for deepfake research. Verdoliva [11] provides an overview of media forensics with emphasis on deepfakes, reviewing manipulation techniques, forensic detection strategies and open challenges highlighting the evolving arms race between content generation and forensic analysis. Sabir et al. [12] proposes a recurrent convolutional architecture for detecting face manipulations in videos, effectively modelling both spatial and temporal inconsistencies and demonstrating improved performance over frame-based deepfake detection approaches. Korshunov and Marcel [13] analyze the impact of deepfakes on face recognition systems evaluating vulnerabilities and proposing detection strategies showing that synthetic media poses a significant threat to biometric security systems. Agarwal et al. [14] address the risks of deepfakes targeting public figures, particularly world leaders, and propose detection and protection strategies based on facial artifacts and identity inconsistencies to mitigate misinformation and security threats. Table 1 represents the overall summary of the key deepfake dataset.

Table 1. Summary of Key Deepfake Datasets

S.No	Author(s)	Year	Contribution	Technique Used
1	Rossler et al. [1]	2019	Created FaceForensics++ dataset	Facial forgery detection
2	Dolhansky et al. [2]	2020	Introduced DFDC dataset	Real-world video manipulation
3	Korshunov et al. [3]	2019	Analyzed deepfake threats to biometric systems	Anti-spoofing techniques

4	Ahmed et al. [4]	2021	Surveyed deepfake detection models	CNN-based categorization
5	Li et al. [5]	2018	Detected deepfakes via eye-blinking analysis	Temporal behavior detection

3. Proposed Work

The proposed deepfake detection system identifies modified data like videos edited using AI-based algorithms. It uses multiple layers for preprocessing, feature extraction, classification and evaluation to achieve high accuracy in identifying between real and fake data.

This figure 1 shows a deepfake detection system analyzes videos to determine they are real or fake. The system uses a structured pipeline that includes preprocessing, model training, assessment and prediction. This approach begins with a dataset includes both real and fake media. During preprocessing, videos are separated into frames and facial detection is used to identify and crop human faces. These compressed images are stored as a processed file contains facial video data. The processed data is separated into training and testing sets. A data processor protects the trained images and their labels are appropriately loaded. The basic concept of this system is based on a deepfake detection model uses Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for video categorization. CNNs record spatial data from individual frames but LSTM networks examine temporal connections between frames allow the model to accurately identify deepfake patterns.

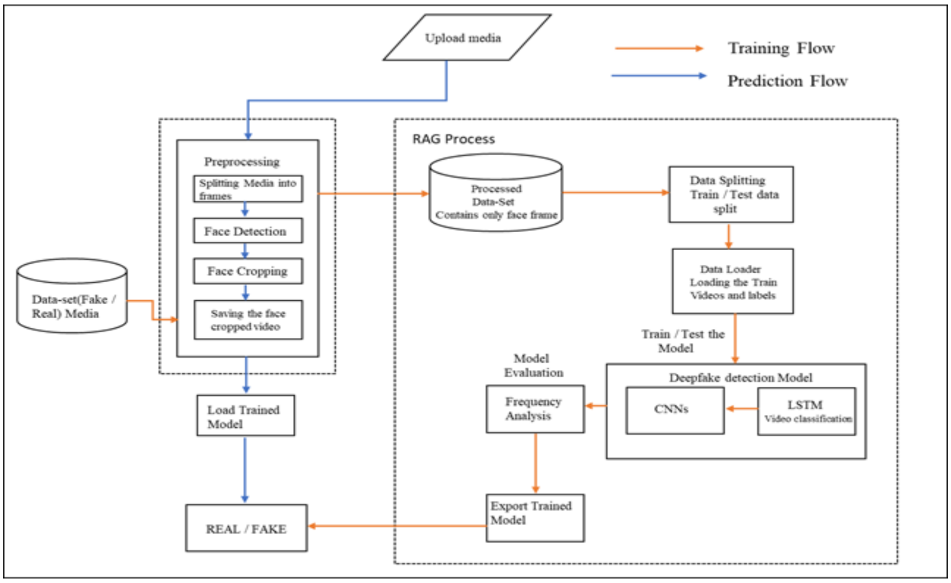

Upload media
Training Flow
Prediction Flow
Preprocessing
Splitting Media into frames
Face Detection
Face Cropping
Saving the face cropped video
Data-set(Fake / Real) Media
Load Trained Model
REAL / FAKE
RAG Process
Processed Data-Set Contains only face frame
Data Splitting Train / Test data split
Data Loader Loading the Train Videos and labels
Train / Test the Model
Deepfake detection Model
CNNs
LSTM Video classification
Model Evaluation
Frequency Analysis
Export Trained Model

Figure 1. System Architecture

The model is evaluated using frequency analysis to verify its accuracy and robustness for training. Once verified, it is used for prediction to identify uploaded media files as real or fake. The system's flow represented with blue arrows for prediction and orange for training follows a structured method to allow accurate deepfake detection and effective generalization to new inputs. The figure 2 explains the working of hidden layers.

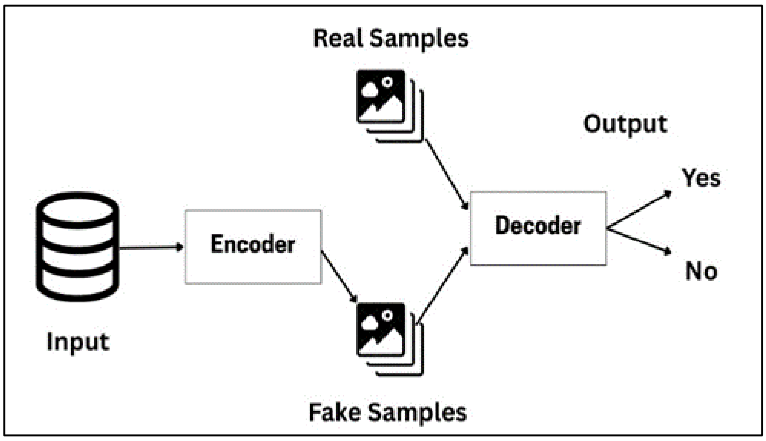


Figure 2. Working of Hidden Layers

3.1 Algorithms/Classifiers and Datasets Used

The proposed system uses Convolutional Neural Networks (CNNs) to extract spatial data including face texture, edge differences and pixel-level variations from images and video frames. A Stacked Autoencoder with Gated Long Short-Term Memory (SAE-GLSTM) used to explain temporal dependencies across successive frames effectively capturing mobility variations and temporal inconsistencies are normal in deepfake videos. Additionally, a Graph CNN (GCNN) is used to evaluate connecting data between face points, improving adaptability to slight changes. The collected characteristics consist of face feature geometry, texture descriptors, temporal motion vectors and frequency-domain data. A completely connected Softmax layer is used to make binary decisions (real vs. false). The experiments are performed on freely available datasets including FaceForensics++ and the DeepFake Detection Challenge (DFDC) dataset ensuring accurate evaluation and consistency. Algorithmic system worked along with a formal mathematical model of the proposed deepfake detection system. The overall algorithm consists of four main stages: Preprocessing, feature extraction, temporal modelling and classification.

Given an input video V={f_1,f_2,…,f_T }, individual frames are extracted and facial regions are detected and normalized. Spatial feature extraction is performed using a

Convolutional Neural Network (CNN) where each convolutional layer computes feature maps as

$$X\_l=\sigma(W\_l*X\_(l-1)+b\_l) \qquad (1)$$

W_l and b_l denoting the weights and biases
* representing convolution
σ the ReLU activation function.

The retrieved spatial characteristics are compressed by a Stacked Autoencoder (SAE) develops low-dimensional model with reducing reconstructing loss. These encoded variables are sent to a Gated Long Short-Term Memory (GLSTM) network to represent temporal connections between images. The GLSTM unit is computationally defined by input, forget and output gates allow the model to maintain the long-term temporal variations found in deepfake videos. Finally, the output sequence representation is entered into a fully connected Softmax layer for binary classification (real or fake). The model is trained with the Binary Cross-Entropy loss function and optimized with the Adam optimizer. This work includes pseudocode, mathematical formulas and an accurate algorithm for flow diagram to increase clarity, accuracy, and technical rigor.

4. Methodology

4.1 The Utilization of CNNs

CNNs for image analysis recognized for visual signals proposed to manipulating individual frame bases. It detects small details including texture issues, damaged features, edge and facial defects. These variations are common in deepfake images. CNNs can highlight and demonstrate visual differences by applying convolution and pooling layers directly to raw pixel inputs allow the model to identify between real and modified video data.

4.2 The Utilization of SAEs

The primary purpose of SAEs is to generate an accurate and detailed representation of each feature identified in the training data. The samples are encoded into an inactive space reduces noise by maintaining important data. The results of this model includes these types of features are more effective with regard to processing and generalization than models developed from raw pixel data, making deepfake categorization much more accurate.

4.3 The Utilization of GLSTMs

GLSTMs was developed to study the visual data changes continuously by analyzing video frame patterns. They can identify temporal patterns such as real facial expression movements and effortless motion will help to maintain frame-to-frame consistency. The GLSTM model maintains essential temporal data and may detect fake transitions or unpredictable movements are common features of deepfake manipulations due to the gates used in GLSTMs.

4.4 Description of Dataset

The features of the dataset have been improved including an overview of its size and the distribution of samples between real and fake videos. The process of collecting a frame from a video has also been provided and a discussion to use inputs into the models to maintain accuracy.

4.5 Experimental Setup and Evaluation

More details on different areas of the study (e.g., simulation settings, model architecture and training setup) have been provided to improve the ability to replicate the work. The results section has been enhanced to explicitly interpret the model to detect Deepfake videos and each piece of the model contributed to the overall detection.

A. Binary Classification Output

For the final fully connected layer with Softmax:

$$\hat{y} - \mathrm{Softmax}(Wz + b) \tag{2}$$

For binary classification (Real =0, Fake =1), prediction is:

$$\hat{y} = \begin{cases} 1 & \text{if } P(\text{ fake } \mid x) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where τ is the decision threshold (typically 0.5).

B. Loss Function (Binary Cross-Entropy)

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{4}$$

This validates the model was optimized during training.

C. Confusion Matrix-Based Metrics

Let:

- TP = True Positives (Fake correctly detected) = 177
- TN = True Negatives (Real correctly detected) = 172
- FP = False Positives (Real → Fake) = 28
- FN = False Negatives (Fake → Real) = 23

Now define:

Accuracy

$$\text{Accuracy } = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

Precision

$$\text{Precision } = \frac{TP}{TP+FP} \tag{6}$$

Recall

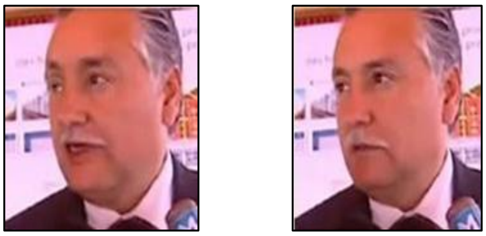
$$\text{Recall } = \frac{TP}{TP+FN} \tag{7}$$

F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

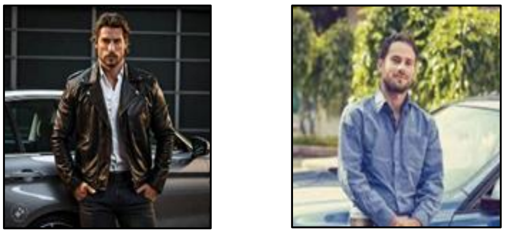
5. Results and Discussion

The system handles a dataset includes real and fake images using preprocessing techniques such as data augmentation and grayscale conversion to optimize feature extraction during the training phase. These images are processed using Convolutional Neural Networks (CNNs) and an encoder-decoder model to differentiate between real and fake data by analyzing and recreating visual characteristics shown in Fig 3–Fig 5. In testing phase, the system evaluates media using technologies such as OpenCV and NumPy with an 80-20 ratio between training and testing data. It uses SAE-GLSTM and a dual Graph CNN to detect temporal and spatial irregularities to categorize data by confidence scores. This method has an accuracy of 86% for deepfake detection and 72% for fake image identification. A simple online application allows user to upload media and verify its validity, but a cybersecurity feedback form allows user to report inaccurate data. The device detects modifications such as face swapping and verifies deepfake content using advanced AI algorithms.



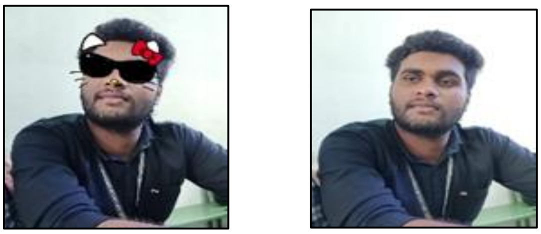
Fake Image Real Image

Figure 3. Real-Time Image



Fake Image Real Image

Figure 4. AI Generated



Fake Image Real Image

Figure 5. Filter Image

5.1 The Testing and Validation Processes

The performance of the proposed work dataset is evaluated using traditional classification measures. This model achieves overall accuracy of 86% and shows better differences between real and fake data. Precision represents the accuracy of fake data detection by reducing false positives and recall (sensitivity) measures the ability of the system to accurately identify true deepfakes by reducing false negatives. The F1-score calculated the balanced mean of accuracy and recall to provide an accurate evaluation of the model's performance. These metrics confirm the robustness and generalization capability of the proposed deepfake detection system shown in Table 2.

Table 2. Performance Metrics

Metric	Value (%)
Accuracy	86.0
Precision	84.2
Recall	88.5
F1-Score	86.3

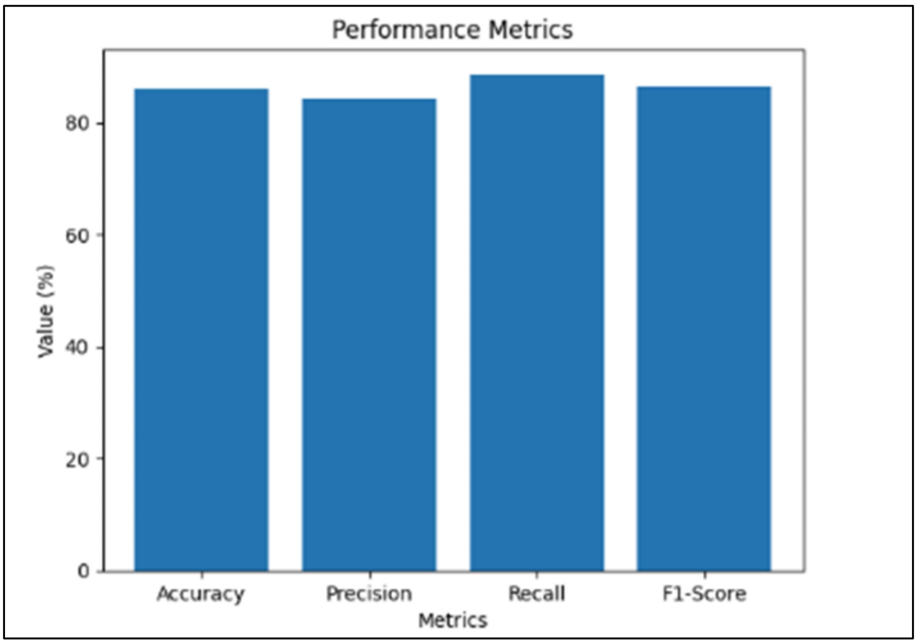


Figure 6. Performance Evaluation Metrics

The confusion matrix shows the classification performance of the proposed work using the Kaggle test dataset [15]. There are 172 accurately identified as real data and 28 inaccurately identified as fake data. Similarly, 177 fake samples were detected accurately but 23 data incorrectly identified as real. The results show a possible detection with a low false negative rate illustrating in fig 6 shows the value of the proposed model identified in modified media.

The 23 fake images mistakenly identified as real data represents the accurate deepfakes with few visual and temporal issues. These examples were created using advanced GAN structures maintain facial texture accuracy, realistic lighting and continuous temporal transformations. Additionally, limited clips of video or static facial expressions reduced temporal inconsistency signals reducing the GLSTM's capacity to identify strange moving patterns. This demonstrates the challenge of identifying next-generation deepfakes are identical to real-world data distributions.

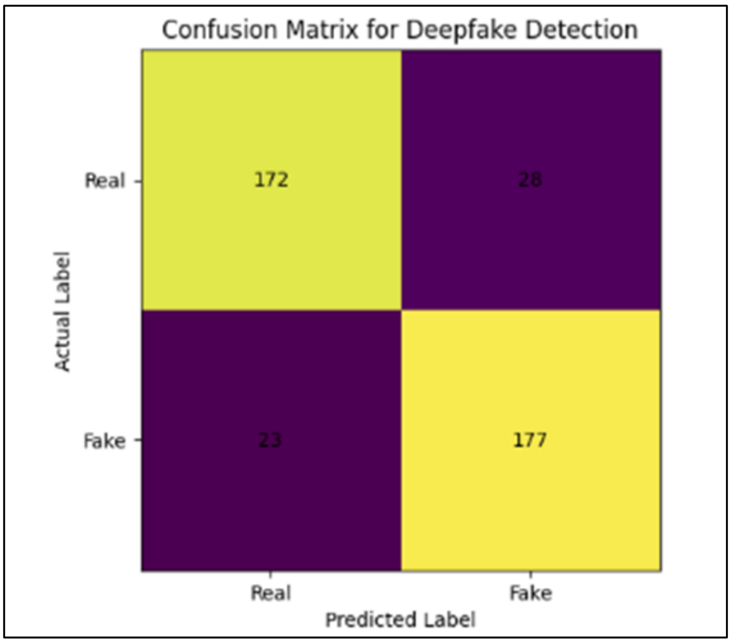


Figure 7. Confusion Matrix for Deepfake Detection on the Test Dataset

5.2 Feature Analysis of Genuine vs. Fake Images

Many real images have better balance between their colors and contrast symmetry in the direction of their facial features point, smooth texture over all parts of the face and even lighting across the whole face and between the eyes. Therefore, images have stable patterns of frequency and consistent level of sensor noise. Fake images (including manipulated images) show the subtle signals help to determine if they were created with the help of GANs, such as: checkerboard patterns (from GANs), blurry edges of facial features, irregular patterns around the eyes, shadows thrown in odd directions across the face and/or unusual high frequencies in noise levels on digital camera sensors. It is important to understand these visual inconsistencies provide the opportunity to use them to identify whether an image is a deepfake or not. In the proposed system, detecting deepfakes are provided by using Convolutional Neural Networks (CNNs) extract detailed spatial features and by the use of SAE-GLSTM models provide an analysis of temporal inconsistencies, enabling CNNs to classify deepfakes.

6. Conclusion

The proposed work combines artificial intelligence and machine learning to handle the increasing risk of deepfake technology. It helps to protect the digital environment by differentiating between real and AI-generated data accurately. This platform is user-friendly design allow users to identify and support interaction with cybersecurity experts to efficiently prevent fake data. Further improvements will focus on real-time identification for online content, better audio deepfake recognition and support in multiple languages for connecting with viewers around the world. A mobile application was also developed to make deepfake detection more simple and affordable. This system improves deepfake method by increasing safety and trust in digital world using monitoring tools and training the model contiuously. The advanced artificial intelligence and improved detection methods are used to provide accurate methods for digital fraud.

References

[1] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images." In Proceedings of the IEEE/CVF international conference on computer vision, (2019): 1-11.

[2] Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. "The deepfake detection challenge (dfdc) dataset." arXiv preprint arXiv:2006.07397 (2020).

[3] Korshunov, Pavel, and Sébastien Marcel. "Vulnerability assessment and detection of deepfake videos." In 2019 International Conference on Biometrics (ICB), IEEE, (2019): 1-6.

[4] Ahmed, Naveed Ur Rehman, Afzal Badshah, Hanan Adeel, Ayesha Tajammul, Ali Daud, and Tariq Alsahfi. "Visual deepfake detection: Review of techniques, tools, limitations, and future prospects." IEEE Access 13 (2024): 1923-1961.

[5] Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai created fake videos by detecting eye blinking." In 2018 IEEE International workshop on information forensics and security (WIFS), Ieee, (2018): 1-7.

[6] Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, (2018): 1-6.

[7] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion, 64, 131-148.

[8] Kumar, Manoj, Praveen Kumar Rai, and Pankaj Kumar. "A novel approach for detecting deepfake face using machine learning algorithms." In 2024 2nd International Conference on Disruptive Technologies (ICDT), IEEE, (2024): 1588-1592.

[9] Yang, Xin, Yuezun Li, and Siwei Lyu. "Exposing deep fakes using inconsistent head poses." In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, (2019): 8261-8265.

[10] Mirsky, Yisroel, and Wenke Lee. "The creation and detection of deepfakes: A survey." ACM computing surveys (CSUR) 54, no. 1 (2021): 1-41.

[11] Verdoliva, Luisa. "Media forensics and deepfakes: an overview." IEEE journal of selected topics in signal processing 14, no. 5 (2020): 910-932.

[12] Sabir, Ekraam, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. "Recurrent convolutional strategies for face manipulation detection in videos." Interfaces (GUI) 3, no. 1 (2019): 80-87.

[13] Korshunov, Pavel, and Sébastien Marcel. "Deepfakes: a new threat to face recognition? assessment and detection." arXiv preprint arXiv:1812.08685 (2018).

[14] Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting world leaders against deep fakes." In CVPR workshops, vol. 1, no. 38. 2019.

[15] https://www.kaggle.com/datasets