JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Face Integrity: Comprehensive Facial AuthenticityAssurance

[1] Dr.Uttara Gogate, [2]Hariom Haushilaprasad Singh, [3]Ritik Ramchandra Prajapati, [4]Sahil Rajendra Patil

[1]Associate Professor, [2, 3, 4]Student
[1, 2, 3, 4]Department of Computer Engineering,
[1, 2, 3, 4]Shivajirao S. Jondhale College of Engineering, Dombivli, India

Abstract : Advancements in artificial intelligence and machine learning have made it easier to manipulate digital media, leading to the emergence of deepfake technology. Deepfakes refer to synthetic media, such as images and videos that have been altered or created using AI algorithms. These manipulated visual media can be used to spread misinformation, defame individuals, or deceive the public. This project aims to combat the negative implications of deepfakes by developing a website that can accurately detect and identify such manipulated content. The existing systems provides less accurately detect deepfakes. To overcome the limitations of existing system Convolutional Neural Networks (CNN) is used which provides high accuracy. In this project four types of deepfake i.e. FaceSwap, Face Attribute Manipulation, Face Reenactment and Lip-sync detection are detected successfully with high accuracy.

***IndexTerms* -** Deepfake generation and detection, Misinformation, Generative Adversarial Networks (GAN), Face Attribute Manipulation, Deepfake research, Face Reenactment.

I INTRODUCTION

An In an era where the authenticity of facial imagery is under constant threat from digital manipulations and deepfakes, the "Face Integrity" project emerges as a comprehensive solution. This innovative initiative seeks to safeguard the integrity of facial data through a multi-pronged approach, combining biometric authentication, image forensics, and deep learning. By addressing the limitations of existing work, it aims to provide a robust, adaptive system that can reliably verify the authenticity of facial images and videos. "Face Integrity" is driven by the imperative to combat deception, protect privacy, and ensure the credibility of facial data in an increasingly interconnected and digitally transformed world.

This project will detect the deepfakes to provide the security to the system by using the advance technique named CNNs (Convolutional Neural Networks). To implement the CNN in project it requires very much compatible environment. Deepfakes like FaceSwap, Face Reenactment, Facial Attributes Manipulation and Lip-sync detection.

A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is particularly well-suited for image recognition and processing tasks. It is made up of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

The convolutional layers are the key component of a CNN, where filters are applied to the input image to extract features such as edges, textures, and shapes. The output of the convolutional layers is then passed through pooling layers, which are used to down-sample the feature maps, reducing the spatial dimensions while retaining the most important information. The output of the pooling layers is then passed through one or more fully connected layers, which are used to make a prediction or classify the image.

II LITERATURE SURVEY

Table 2.1 Literature Survey

Sr. No.	Title	Methodology	Advantages	Limitations
1.	A Dataless FaceSwap Detection Approach Using Synthetic Images, Anubhav Jain, Nasir Memon, Julian Togelius. (2022)[1]	GAN (Generative Adversarial Networks)	Eliminates the need for real data Reduces biases created by facial image datasets	It may not work well for detecting deepfakes created using more advanced techniques.
2.	Masked Lip-Sync Prediction by Audio-Visual Contextual Exploitation in Transformers, Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, Hideki Koike. (2022)[2]	Convolutional-Transformer-Hybrid and Convolutional Refinement Network	Produces accurate lip-sync with photo-realistic quality. Incorporates audio-visual contextual information to improve lip-sync accuracy	Requires a large amount of training data to achieve optimal results. May not work well with non-standard or heavily accented speech.
3.	GGViT: Multistream Vision Transformer Network in Face2Face Facial Reenactment Detection, Haotian Wu, Peipei Wang, Xin Wang, Ji Xiang, Rui Gong. (2022)[3]	GAN (Generative Adversarial Networks)	The proposed GGViT architecture achieves state-of-the-art classification accuracy on the FF++ dataset.	The proposed model has only been tested on the FF++ dataset, and its performance on other datasets is unknown.
4.	FakeLocator – Robust Localization of GAN-Based Face Manipulations, Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, Geguang Pu. (2022)[4]	GAN (Generative Adversarial Networks)	Achieves high localization accuracy on manipulated facial images. It is robust against various real-world facial image degradations such as JPEG compression, low-resolution, noise, and blur.	It has limitations in terms of computational efficiency and scalability. The concept of method robustness and universality is too wide and cannot be guaranteed.
5.	Exposing Deepfake Videos by Detecting Face Warping Artifacts, Yuezun Li, Siwei Lyu. (2019)[6]	Proposes methods to identify distortions in facial features.	Novel approach to deepfake detection. Addresses challenges of early deepfake detection.	No information on computational complexity or detection time. Effectiveness in real-time scenarios not explicitly shown.

From the table 2.1, We can see that the current systems face various limitations in effectively detecting deepfakes. These include challenges such as inefficacy in detecting advanced deepfake creation techniques, the requirement of extensive training data for optimal performance, limited generalization beyond specific datasets, computational inefficiency and scalability issues, susceptibility to adversarial attacks, lack of discussion on user trust impact, and insufficient information on computational complexity, detection time, and real-time effectiveness. These limitations highlight the ongoing need for further advancements in deepfake detection technology to address these critical shortcomings.

Table 2.2 Research Gap

Performace Metric	GAN	LSTM Video Classification	ResNext Feature Extraction
Accuracy	Moderate	High	High
Robustness	Prone to mode collapse	Robust against variations	Robust against variations
Training Complexity	Complex	Moderate	Moderate
Interpretability	Limited Understanding	Moderate Understanding	Moderate Understanding
Computational Efficiency	Less Efficient	Efficient	Efficient
Scalability	Limited scalability	Moderate scalability	High scalability
Generalization	Limited generalization	High generalization	High generalization

From the table 2.2, GAN exhibits moderate accuracy, limited interpretability, and is prone to mode collapse, while LSTM and ResNext showcase high accuracy, robustness against variations, moderate training complexity, efficient computational efficiency, moderate interpretability, scalability, and high generalization capabilities, making them more efficient choices for deepfake detection.

After studying and analyzing the research papers, we found that each paper has several shortcomings which have been removed in our proposed system. These are:

- It may not work well for detecting deepfakes created using more advanced techniques.
- The created model may be prone to mode collapse.
- Requires a large amount of training data to achieve optimal results.
- The proposed model has only been tested on the FF++ dataset, and its performance on other datasets is unknown which give rise to biases or inaccuracies in detecting deepfakes across different scenarios and contexts.
- It has limitations in terms of computational efficiency and scalability. The concept of method robustness and universality is too wide and cannot be guaranteed.

III METHODOLOGY

- Data Collection and Preprocessing:
  - Gather diverse dataset with real and deepfake files (images/videos).
  - Implement preprocessing techniques tailored to deepfake detection (e.g., frame extraction, feature manipulation).
- Model Architecture Design:
  - Choose specialized architecture for analyzing spatial and temporal information, especially for video analysis.
  - Incorporate pretrained models or design custom architectures for detection.
- Model Compilation:
  - Define appropriate loss functions specialized for deepfake detection (e.g., binary cross-entropy, adversarial loss).
  - Select optimizer (e.g., stochastic gradient descent, Adam) for model training.
- Model Training:
  - Split dataset into training, validation, and testing sets with balanced real and deepfake samples.
  - Train model with different configurations, adjusting learning rates, batch sizes, and regularization techniques.
- Performance Evaluation:
  - Evaluate model using metrics like precision, recall, and F1-score for accuracy assessment.
  - Analyze model's behavior under various deepfake manipulations to ensure robustness.
- Deployment and Testing:
  - Deploy trained model in a secure environment for real-time deepfake detection.
  - Test model with diverse deepfake samples to verify effectiveness against latest manipulation techniques.

IV PROBLEM STATEMENT AND OBJECTIVES

6.1 Problem Statement

The project "Face Integrity- Comprehensive Facial Authenticity Assurance" aims to address the issue of deepfakes, which can potentially harm individuals, societies, and organizations by manipulating images and videos for malicious purposes. The project utilizes advanced techniques, particularly Convolutional Neural Networks (CNN), to detect various types of deepfakes and ensure the security and integrity of individuals' photos. By preventing the swapping of faces or any other form of image manipulation, the project aims to safeguard the reputation and well-being of individuals in both personal and professional contexts. The project recognizes the potential social, psychological, and reputational consequences that can arise from the misuse of manipulated images and videos and seeks to provide a comprehensive solution for facial authenticity assurance.

6.2 Objectives

- The "Face Integrity: Comprehensive Facial Authenticity Assurance" project aims to combat deepfakes, which can cause harm to individuals, societies, and organizations.
- The project employs advanced techniques, specifically Convolutional Neural Networks (CNN), to detect three primary types of deepfakes:
  - FaceSwap: To determine if an image is generated from the original person, protecting individuals and organizations from malicious actors.
  - Facial Attribute Manipulation: To debunk rumors and prevent damage to an individual's reputation by detecting facial attribute manipulations.
  - Face Reenactment: To prevent impersonation by identifying instances of face reenactment or complex masking.
- The project's focus on these four types of deepfakes aims to protect individuals and organizations from the negative impacts of deepfake manipulations.
- By implementing CNN, the project provides a comprehensive solution for facial authenticity assurance, ensuring the security and integrity of individuals' photos.

V PROPOSED SYSTEM

7.1 Introduction

The proposed Web Application aims to detect various forms of deepfakes, including FaceSwap, FaceReenactment, Facial

Attribute Manipulation, and Lip-sync. These deceptive technologies have significant adverse effects on both society and individuals, potentially leading to mental instability. By addressing these challenges, this project endeavors to mitigate the negative impacts of deepfakes and present the truth to society.

Through the development of this system, efforts are directed towards combating the harmful consequences of deepfake manipulation. By detecting and preventing the spread of falsified content, the project seeks to uphold the integrity of information and safeguard individuals from the detrimental effects of deceptive media. This initiative strives to contribute to a more truthful and reliable societal discourse by countering the proliferation of misleading deepfake technologies.

7.2 System Architecture

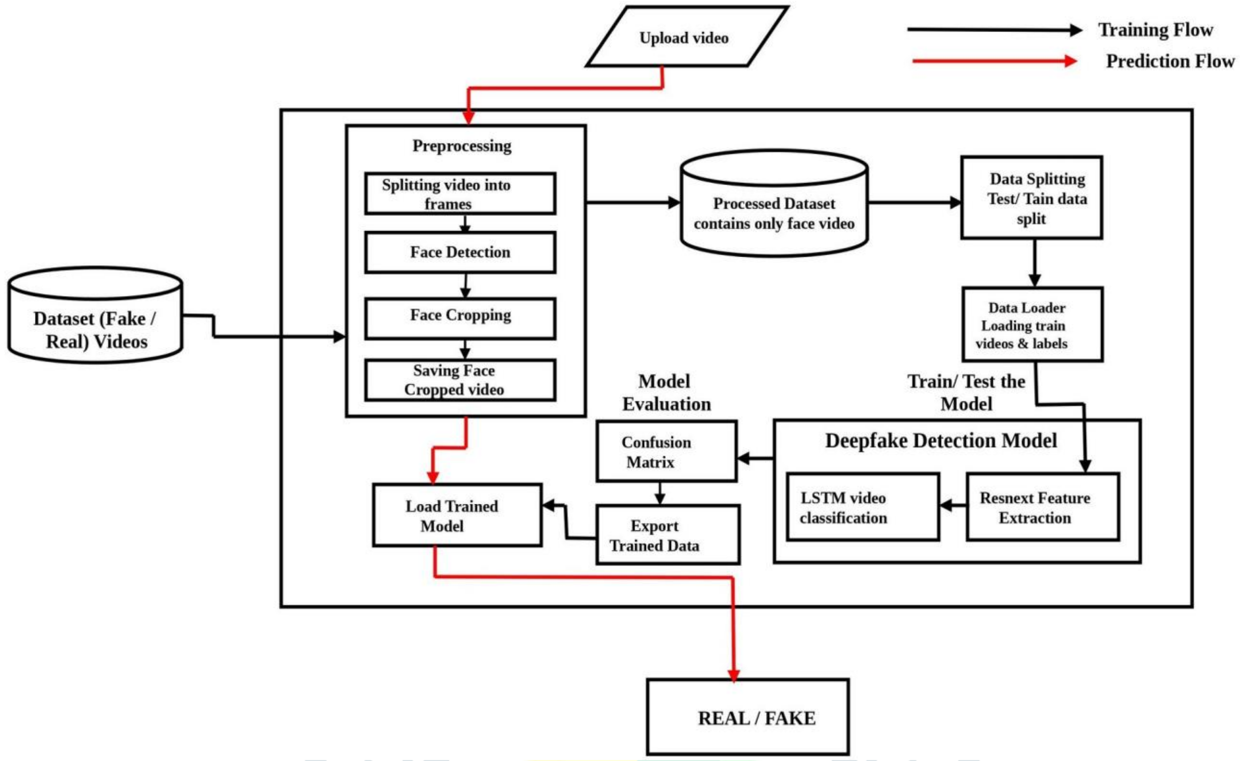

Upload video
Training Flow
Prediction Flow
Preprocessing
Splitting video into frames
Face Detection
Face Cropping
Saving Face Cropped video
Dataset (Fake / Real) Videos
Processed Dataset contains only face video
Data Splitting Test/ Tain data split
Data Loader Loading train videos & labels
Model Evaluation
Train/ Test the Model
Deepfake Detection Model
Confusion Matrix
LSTM video classification
Resnext Feature Extraction
Load Trained Model
Export Trained Data
REAL / FAKE

Fig. 7.1 System Architecture

From figure 7.1, we can seek the Architecture of Proposed System. The deepfake detection system architecture involves several key steps, starting with data-set exploration. In this stage, the face-cropped video is saved, a new face-cropped video is created, the face is cropped, and the face is detected. The video is then split into frames for further processing.

Pre-processing of the frames involves resizing them to a standard size, converting them to grayscale, and extracting the face region. This ensures that the model can effectively learn the features of the faces in the images.

The model architecture utilizes the ResNext-50 model, which has been shown to be effective in image recognition tasks. Two LSTM layers are included, each with a 2048 shape input vector and 2048 latent features. A dropout layer with a 0.4 chance of dropout and ReLU activation function is incorporated to prevent overfitting and improve generalization. The LSTM layers are stacked using sequential layers to form the final model architecture.

The training workflow involves training the model using specific parameters such as the number of epochs, batch size, and loss function. This allows the model to learn the features of deepfakes and pristine images, enabling it to accurately distinguish between the two.

The prediction workflow describes the process of predicting whether an image is a deepfake or pristine using the trained model. This is achieved by inputting the processed image into the model and obtaining the predicted label.

The project employs several tools and technologies, including the PyTorch framework for building and training the model, Google Cloud Platform for cloud services, Python3 and Python programming languages. The IDEs used in the project include Google, Jupyter Notebook, and Visual Studio Code.

In summary, the deepfake detection system architecture involves data-set exploration, pre-processing, model architecture with ResNext-50 and LSTM layers, training and prediction workflows, and the tools and technologies employed in the project. This comprehensive approach allows for accurate detection of deepfakes, contributing to the fight against misinformation and deception.

7.3 Pseudo Code

Step 1: Data-set Exploration:

- SAVING_THE_FACE_CROPPED_VIDEO

- CREATING_NEW_FACE_CROPPED_VIDEO
- CROP_FACE
- FACE_DETECTION
- SPLIT_VIDEO_INTO_FRAMES

Step 2: Pre-processing:

- Resize frames
- Convert frames to grayscale
- Extract face region

Step 3: Model Architecture:

- Load the ResNext-50 model
- Add a LSTM layer with 2048 shape input vector and 2048 latent features, along with a 0.4 chance of dropout and ReLU activation function
- Add another LSTM layer with the same parameters as the previous layer
- Add two sequential layers to stack the LSTM layers

Step 4: Training Workflow:

- Define the training parameters (number of epochs, batch size, loss function)
- Train the model with the cropped and pre-processed frames

Step 5: Prediction Workflow:

- Load the trained model
- Pre-process the input frames
- Pass the pre-processed frames through the model to predict whether they are deepfake or pristine

Step 6: End of Algorithm.

7.4 Model Architecture

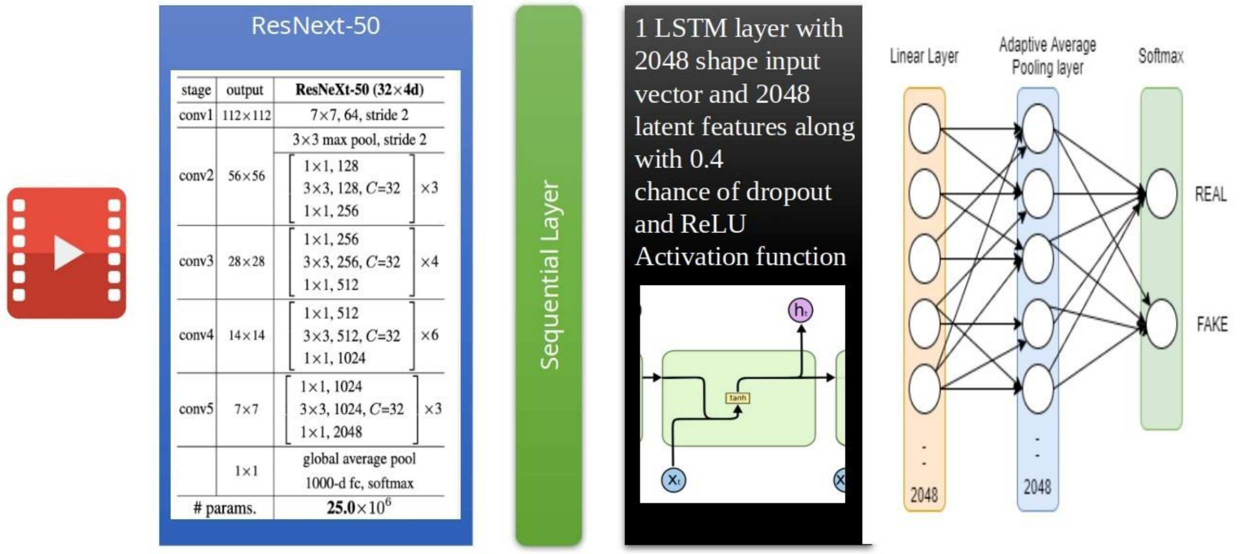
The model architecture shows the use of ResNext-50, LSTM layers, and sequential layers in the deep learning model.

The ResNext-50 model is a deep residual network that is designed to improve the accuracy of image classification tasks. It uses a residual block structure that allows for the training of deeper networks without encountering the vanishing gradient problem.

Two LSTM (Long Short-Term Memory) layers are used in the model architecture, each with a shape input vector of 2048 and 2048 latent features. LSTM layers are a type of recurrent neural network (RNN) that is well-suited for processing sequential data, such as time series or natural language text. The LSTM layers in this model are used to capture the temporal dependencies in the video frames.

The model also includes two sequential layers, which are used to stack the LSTM layers and create a deep learning model. The first sequential layer contains the first LSTM layer, and the second sequential layer contains the second LSTM layer. The model architecture also includes a dropout layer with a 0.4 chance of dropout and a ReLU activation function. The dropout layer is used to prevent overfitting by randomly setting a fraction of the input units to zero during training. The ReLU activation function is used to introduce non-linearity into the model.

In summary, the figure 7.2 explains the use of ResNext-50, LSTM layers, and sequential layers in the deep learning model for deepfake detection. The ResNext-50 model is used for image classification, while the LSTM layers are used to capture the temporal dependencies in the video frames. The dropout layer and ReLU activation function are used to prevent overfitting and introduce non-linearity into the model.



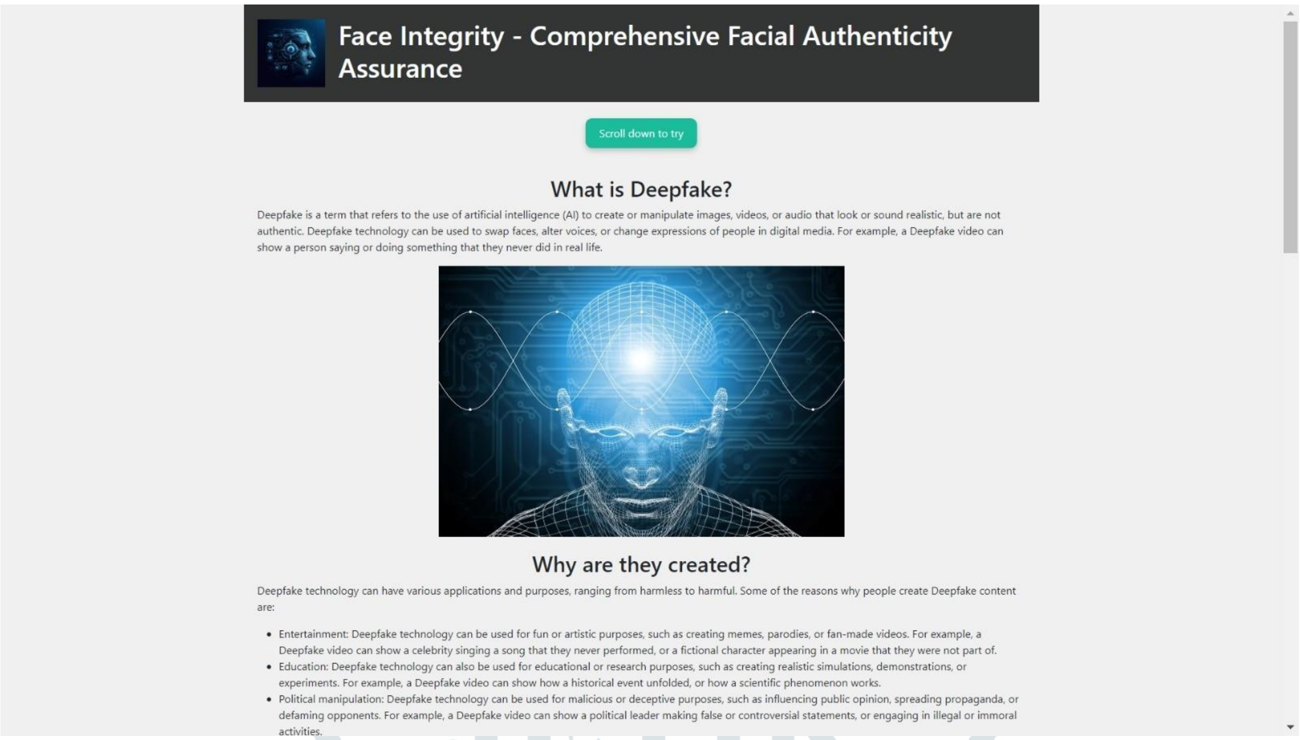
ResNext-50

stage	output	ResNeXt-50 (32×4d)
conv1	112×112	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 [1×1, 128 3×3, 128, C=32 1×1, 256] ×3
conv3	28×28	[1×1, 256 3×3, 256, C=32 1×1, 512] ×4
conv4	14×14	[1×1, 512 3×3, 512, C=32 1×1, 1024] ×6
conv5	7×7	[1×1, 1024 3×3, 1024, C=32 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax
# params.		**25.0**×10^6

1 LSTM layer with 2048 shape input vector and 2048 latent features along with 0.4 chance of dropout and ReLU Activation function

Fig. 7.2 Model Architecture

VI RESULT



Face Integrity - Comprehensive Facial Authenticity Assurance

What is Deepfake?

Deepfake is a term that refers to the use of artificial intelligence (AI) to create or manipulate images, videos, or audio that look or sound realistic, but are not authentic. Deepfake technology can be used to swap faces, alter voices, or change expressions of people in digital media. For example, a Deepfake video can show a person saying or doing something that they never did in real life.

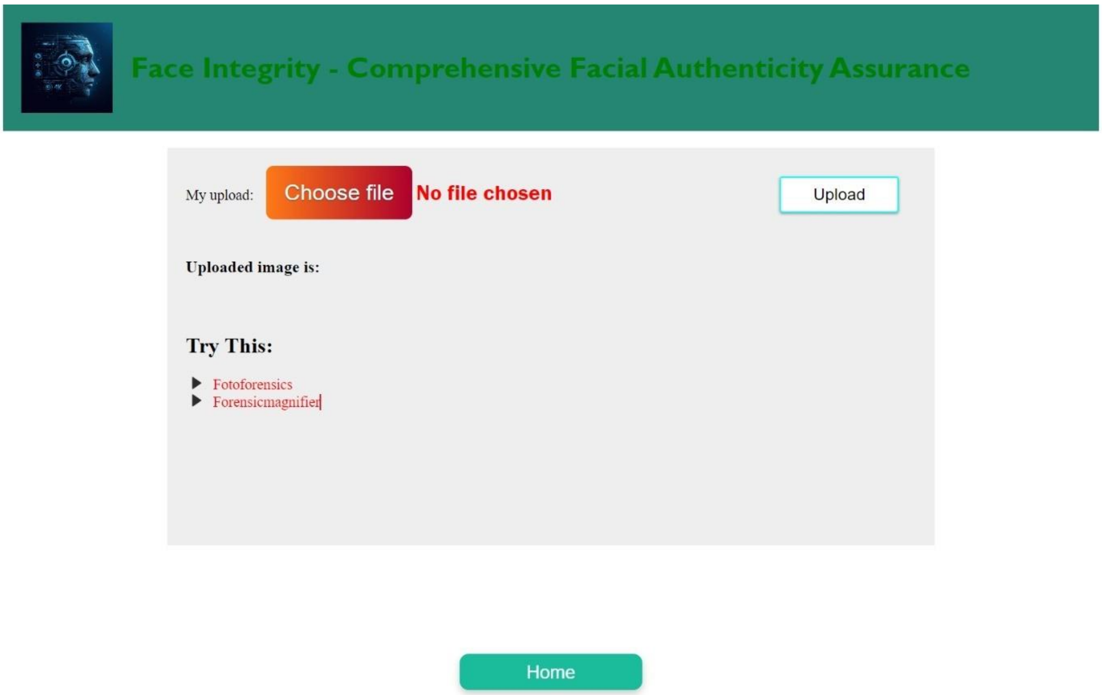
Why are they created?

Deepfake technology can have various applications and purposes, ranging from harmless to harmful. Some of the reasons why people create Deepfake content are:

- Entertainment: Deepfake technology can be used for fun or artistic purposes, such as creating memes, parodies, or fan-made videos. For example, a Deepfake video can show a celebrity singing a song that they never performed, or a fictional character appearing in a movie that they were not part of.
- Education: Deepfake technology can also be used for educational or research purposes, such as creating realistic simulations, demonstrations, or experiments. For example, a Deepfake video can show how a historical event unfolded, or how a scientific phenomenon works.
- Political manipulation: Deepfake technology can be used for malicious or deceptive purposes, such as influencing public opinion, spreading propaganda, or defaming opponents. For example, a Deepfake video can show a political leader making false or controversial statements, or engaging in illegal or immoral activities.

Fig. 8.1 Home Page

Fig 8.1 is the home page of our website which is a blog post providing information on what exactly are deepfakes and its related contents. It the consist of a button that takes us to the main page.



Face Integrity - Comprehensive Facial Authenticity Assurance

My upload: Choose file No file chosen

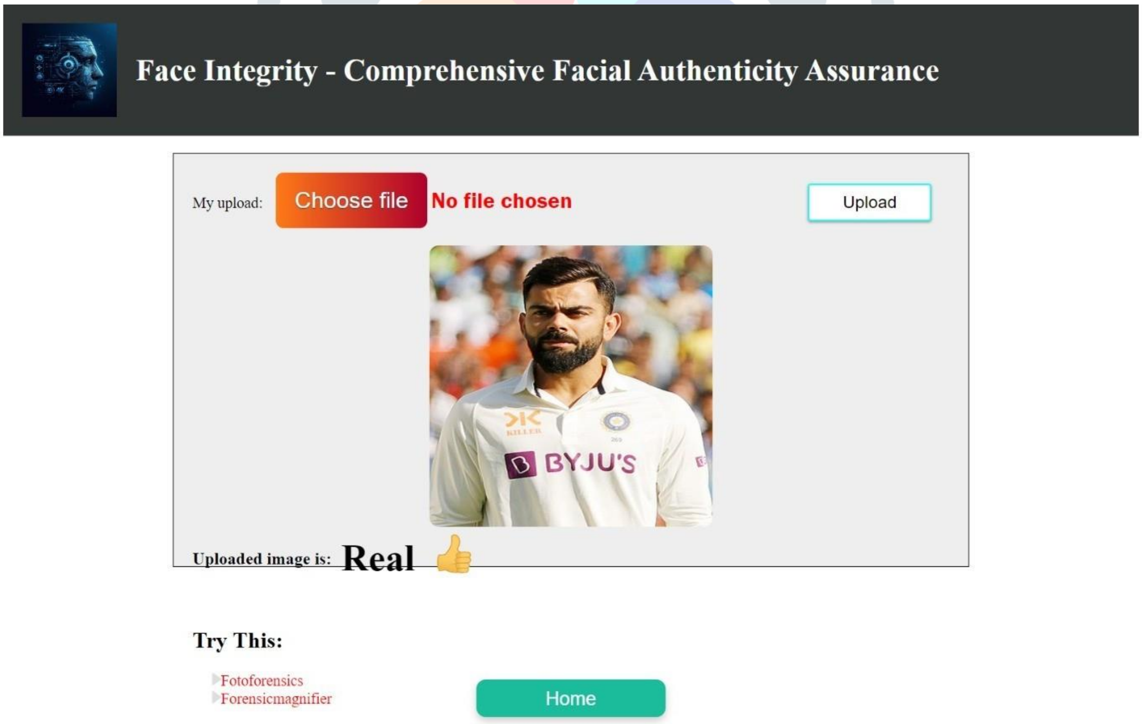
Uploaded image is:

Try This:

- Fotoforensics
- Forensicmagnifier

Fig. 8.2 Detection Page

Fig 8.2 is the main page of our website which helps user detect their image whether it is deepfake or not. It also consist of external links which would help them see which part of their image is deepfake.



Face Integrity - Comprehensive Facial Authenticity Assurance

My upload: Choose file No file chosen

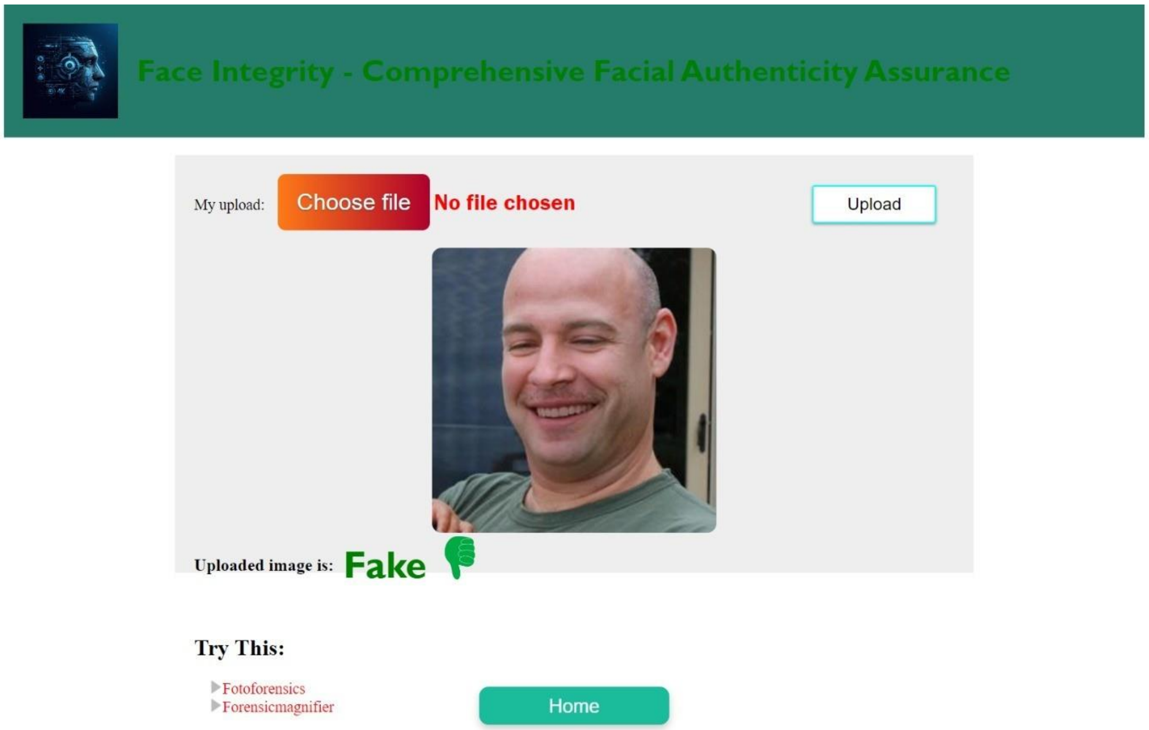
Uploaded image is: Real 👍

Try This:

- Fotoforensics
- Forensicmagnifier

Fig. 8.3 Real Image

Fig 8.3 depicts result when the image uploaded is real.



Face Integrity - Comprehensive Facial Authenticity Assurance

My upload: Choose file No file chosen

Uploaded image is: Fake

Try This:

Fig. 8.4 Fake Image

Fig 8.4 depicts result when the image uploaded is deepfaked.

VII CONCLUSION

The Face Integrity Web Application is a platform that aims to elevate the trustworthiness of facial image verification by conquering current constraints. With a core mission of combatting deceit, protecting individual privacy, and upholding the integrity of facial data within our digitally interconnected society, this project introduces a resilient and adaptive system. By harnessing the power of Convolutional Neural Networks (CNN), the project strives to deliver sophisticated functionalities for authenticating facial content. To meet the rigorous hardware demands, the implementation leverages a cloud-based infrastructure, specificall Kaggle, for model development, while Python acts as the primary programming language for seamless website integration. This innovative detection system not only offers heightened security measures but also holds the promise of widespread deployment in diverse organizational settings and governmental institutions to efficiently identify falsified content. The Face Integrity Web Application serves as a valuable resource for users seeking to uncover various types of visual deepfake content, such as Face Reenactment and Face Attribute Manipulation, providing a comprehensive solution in the realm of facial integrity verification.

References

[1] Jain, A., Memon, N., & Togelius, J. (2022, October). A dataless faceswap detection approach using synthetic images. In 2022 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-7). IEEE.

[2] Sun, Y., Zhou, H., Wang, K., Wu, Q., Hong, Z., Liu, J., ... & Hideki, K. (2022, November). Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In SIGGRAPH Asia 2022 Conference Papers (pp. 1-9).

[3] Wu, H., Wang, P., Wang, X., Xiang, J., & Gong, R. (2022, August). Ggvit: Multistream vision transformer network in face2face facial reenactment detection. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 2335-2341). IEEE.

[4] Huang, Y., Juefei-Xu, F., Guo, Q., Liu, Y., & Pu, G. (2022). Fakelocator: Robust localization of gan-based face manipulations. IEEE Transactions on Information Forensics and Security, 17, 2657-2672.

[5] Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.