Deep Fake Detection System

Gurukiran D P[1], Joywin Monteiro[2], Karan Anjan[3] , Kushal C[4] , Nivyashree R[5]

[1]Department Of Computer Science and Engineering, Malnad College Of Engineering, Hassan
[2] Department Of Computer Science and Engineering, Malnad College Of Engineering, Hassan
[3] Department Of Computer Science and Engineering, Malnad College Of Engineering, Hassan
[4] Department Of Computer Science and Engineering, Malnad College Of Engineering, Hassan
[5] Department Of Computer Science and Engineering, Malnad College Of Engineering, Hassan

---***---

Abstract - Deep learning methods are used by the Deep Fake Detection System to recognize "deepfakes," or distorted media content. Deepfakes are artificial media produced by sophisticated artificial intelligence algorithms that threaten the credibility of media. The goal of our project is to create a reliable system that can discriminate between authentic and modified content in order to stop the spread of false information and protect media integrity. Our goal is to improve deepfake detection efficiency and accuracy by conducting a thorough evaluation of deep learning-based detection techniques. Our technology aims to offer real-time detection capabilities by utilizing sophisticated neural networks and machine learning techniques. This will aid in the continuous endeavors to tackle the widespread occurrence of deepfakes in digital media.

Key Words: Deep Fake Detection, Deep Learning, Media Integrity, Misinformation, Neural Networks

1.INTRODUCTION

A method for synthesizing human images based on neural network tools, auto encoders, etc. is called deep fake. These tools super impose target images onto source videos using a deep learning techniques and create a realistic looking deep fake video. Deepfakes are seen as the main threat posed by AI in the world of ever-expanding social media networks. Several scenarios involving the deployment of realistic face swapping deepfakes to incite political distress, fake terrorism events, produce revenge porn, and blackmail people are readily imagined. Some of the examples are Brad Pitt, Angelina Jolie videos. Knowing how to distinguish between immaculate and deepfake footage becomes crucial. AI is being used against AI by humanity. Deepfakes are produced with the aid of programmes such as Face App and Face Swap, which use auto encoders or pretrained neural networks like GAN. In our approach, the sequential temporal analysis of the video frames is processed by an artificial neural network based on long short-term memory (LSTM), and the frame-level features are extracted using a pre-trained CNN. In order to categories a video as either real or deepfake, an artificial recurrent neural network based on long short-term memory is trained using the frame-level data extracted by a convolution neural network. We trained our model using a huge quantity of balanced and combination of several available datasets, such as Deepfake detection challenge, Face Forensic++, and Celeb-DF, in order to mimic the real-time scenarios and improve the model's performance on real-time data. In addition, we have created an interface where users may upload videos to make the system suitable for usage by clients. After the video is processed by the model, the user will receive an output that includes the model's confidence level and a categorization of the video as either legitimate or deepfake.

2. CREATION OF DEEPFAKE

It is crucial to know about the deepfake's production process in order to identify deepfake videos. The majority of tools, such as autoencoders and GANs, require a source image and a target video as input. These tools deconstruct the video into individual frames, identify faces in the footage, then swap out the source and target faces in each frame. Next, several pre-trained models are used to integrate the replaced frames. By eliminating the remnants of the deepfake production model, these models also improve the quality of the video. which lead to the production of a deepfake that has an appearance that is real. The same method has also been applied by us to identify deepfakes. The deepfakes produced with pretrained neural network models are so accurate that it is nearly hard to tell the difference with one's own eyes. In the real world, however, the technologies used to create deepfakes leave certain artefacts or traces in the video that are invisible to the human eye. This research aims to detect these minute details and noticeable artefacts of these videos and categories them as authentic or deepfake.

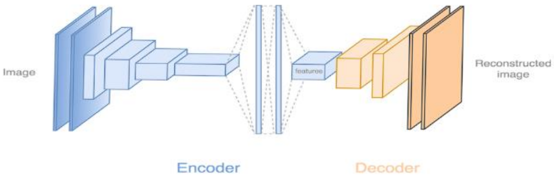

Fig -1: Creation of deepfake

Fig -2: Detection Process

3. METHODOLOGY

1. **Upload video:**
   A video file is uploaded to the system in this first stage.

2. **Preprocessing:**
   Preprocessing is applied to the submitted video in order to get it ready for facial recognition.
   a. A frame sequence of images is what makes up a video. By breaking up a video into its individual images, or frames, it becomes simpler to examine each frame separately to check for the presence of faces.

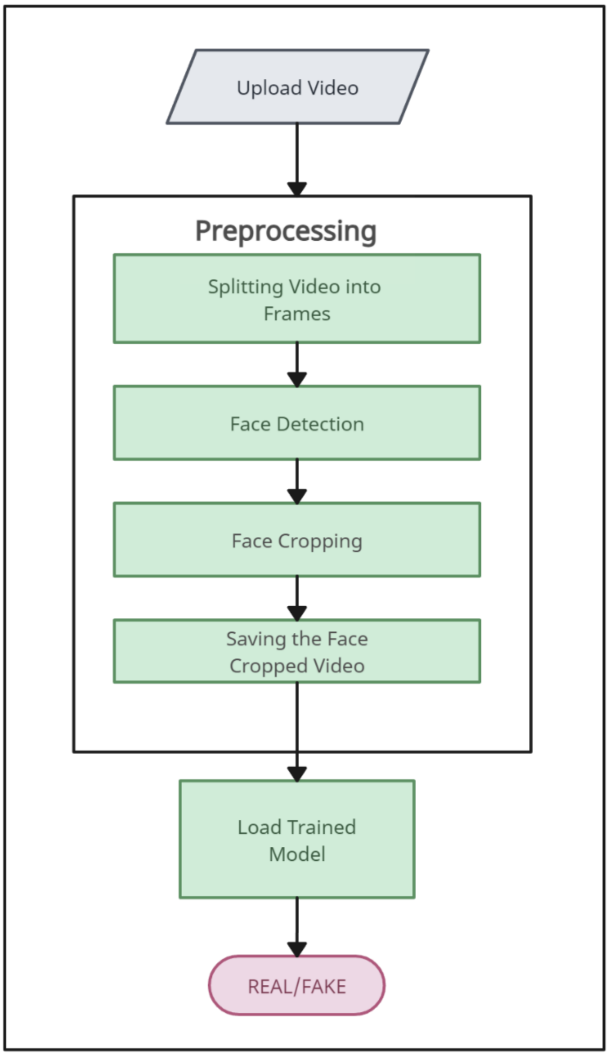

Fig -1: Data Flow Diagram

   b. To determine whether faces are present and where they are located in each frame, a scan is performed.
   c. When a face is identified in a frame, the algorithm removes or reduces the face's region from the frame while removing unnecessary background information.
   d. Next, a new video file containing the cropped faces is saved.
   e. The system loads a facial identification model that has been trained on a sizable dataset of images of faces and the identities that go with them.

3. **Real/Fake:**
   To distinguish between real and artificial faces, the system compares the faces in the preprocessed video with the faces in the trained model.

4. SYSTEM ARCHITECTURE

Input Video: Putting the video footage into the computer system is the first step.

Pre-processing: After that, the video is changed into a form that the manipulation detection system can examine. The video might need to be resized or changed to a new color space for this.

Frame Extraction: Next, the video is divided into its component frames. The video is composed of individual images, called frames.

Encoder: In this case, an encoder processes every frame. One kind of neural network used for information compression is an encoder. In this instance, the video frame is compressed into the lower-dimensional representation using the encoder.

Face Detection, Cropping, and Alignment: Next, any faces within the frame are identified by the system. When a face is identified, it is cut from the frame and placed in a particular alignment. This is carried out due to the fact that the algorithm used for manipulation detection frequently detects alterations in faces more successfully than in other areas of the image.

CNN (Convolutional Neural Network): After the facial image has been cropped and aligned, it is run into a convolutional neural network (CNN). A particular kind of neural network called a CNN is made especially for image identification. In order to ascertain whether or not a picture has been altered, characteristics are extracted from it using a CNN.

RNN: A recurrent neural network (RNN) is used to process the CNN's output after it has been processed. One kind of neural network system that can handle sequential data is an RNN. In this instance, the features that were taken out of the video frames are processed using an RNN.

Output: Next, the RNN generates a percentage score indicating how likely it is that the video is a fake. A higher score suggests a higher probability of video falsification.

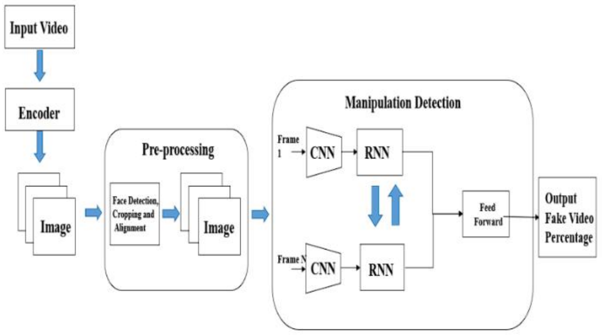

Fig -1: System Architecture

5. CONCLUSIONS

In order to differentiate between real and deepfake films, we presented a neural network-driven method that assigns confidence ratings to the model's predictions. Our method shows that it is capable of accurately analyzing 10 frames per second-, or one-second's worth of video data. We extract facial features from each frame using a pre-trained Multi-Task Cascaded Convolutional Neural Network (MTCNN), and we use an LSTM (Long Short-Term Memory) network for temporal sequence analysis to identify changes between consecutive frames (t and t-1). Our model performs well on video sequences that have frame intervals of ten, twenty, forty, sixty, eighty, and Hundred frames, guaranteeing strong detection performance in a range of situations. Apart from demonstrating a strong performance over multiple frame intervals, our model demonstrates adaptability in practical scenarios by furnishing dependable deepfake video identification together with comprehensible confidence ratings. Utilizing the pre-trained MTCNN for effective extraction of facial features and the LSTM for processing temporal sequences, our method demonstrates flexibility and scalability across various video lengths and levels of complexity. Furthermore, by offering a reliable and accurate mechanism for detecting modified information, our approach supports continued efforts to stop the spread of synthetic media. As deep fake technologies continue to evolve, our model has the potential to improve the security and integrity of digital media platforms through ongoing refinement and integration with current frameworks for video analysis.

REFERENCES

1. Prof. Vilas Jarali Anjali Mahantesh Mudakavi Lalitha Virupakshi Mudakavi. Prateek Kataraki, Shreya Mahesh Desai. Deep Fake Detection Using Deep Learning. 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), December,2023.
2. Omar Bilal Ashfaq Ahmed Saima Waseem1 Adel Hafeezallah Syed Abdul Rahman, Syed Abu Bakar Zaid and Saba Baloch. Multi-attention-based approach for deepfake face and expression swap detection and localization. 2022.
3. Mohamad Nur Nobi Beddhu Murali and Andrew H.Sung. Deepfake detection: A systematic literature review. 2022.
4. Xin Yang Pu Sun Honggang Yuezun and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics,, 2021.
5. Luisa Verdoliva Davide Cozzolino and Christian Riess. Faceforensics++: Learning to detect manipulated facial images. 2019.