**National University of Computer & Emerging Sciences**

**Karachi Campus**



**AI PROJECT**

**Deepfake Detection**

**Group Members**

**Muhammad Qasim (21K-4901)**

**Sooraj Kumar (21K-4884)**

**Introduction**

The rapid evolution of artificial intelligence has led to significant advancements in digital content creation, particularly through the development of "deepfakes." These synthetic media, created by superimposing one person’s likeness over another in videos, present both opportunities and challenges. While they have promising applications in entertainment and education, deepfakes also pose serious risks by enabling the creation of convincingly fraudulent media that can spread misinformation and violate personal privacy. Consequently, there is a pressing need for effective tools to detect and differentiate between genuine and manipulated content.

**Background or Literature Review**

Deepfakes leverage advanced neural networks, especially Generative Adversarial Networks (GANs), to produce photorealistic videos. The technology's potential misuse in malicious ways has escalated ethical and legal concerns, highlighting the necessity for robust detection methods. Traditional techniques, which primarily focused on detecting physical inconsistencies in videos, are increasingly inadequate due to the improving quality of deepfake generation.

**Propose Solution**

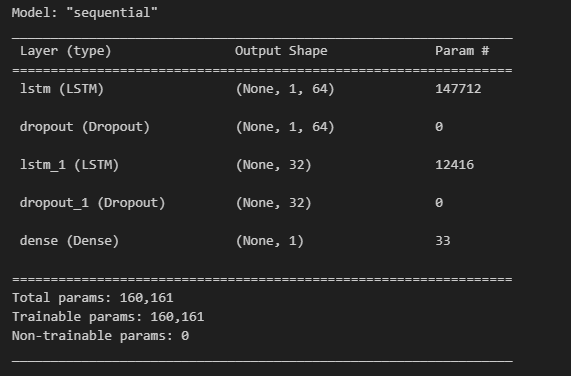
Our project introduces a deepfake detection system that integrates OpenCV for frame extraction, FaceNet for feature extraction, and LSTM networks for analyzing temporal inconsistencies in videos. This combination allows for a detailed examination of both static and dynamic aspects of video content, enhancing the ability to identify manipulated media effectively. By employing this advanced approach, our system aims to contribute to digital media forensics, ensuring the integrity of digital content in the face of evolving AI capabilities. This introduction outlines the methodologies employed, the challenges addressed, and the potential impact of our efforts on media authenticity and security.

**Methodology**

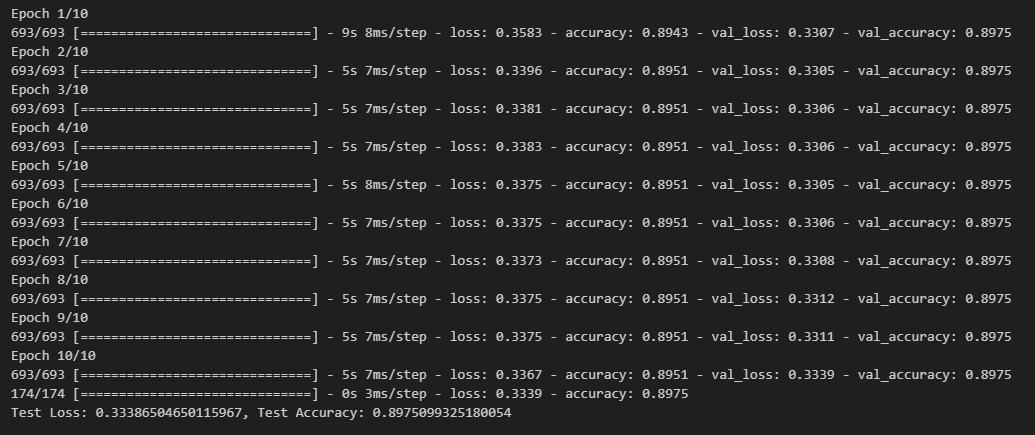
Deepfake Detection System employs a comprehensive approach combining frame extraction, feature extraction, and classification to effectively identify manipulated media. This system utilizes a combination of OpenCV for frame extraction, FaceNet for generating high-quality facial embeddings, and a Long Short-Term Memory (LSTM) network for classification, implemented using the TensorFlow framework. Below is a detailed description of the methodology employed in the system:

* **Input Data for the Algorithm:** The input data consists of video files potentially containing deepfake content. These videos are processed to extract individual frames, which serve as the basis for subsequent analysis.
* **Frame Extraction using OpenCV:** Using OpenCV, a powerful image and video processing library, each video is broken down into its constituent frames. This step is crucial as it isolates each moment of the video for individual inspection, ensuring comprehensive analysis across the entire video sequence.
* **Feature Extraction using FaceNet:**Once frames are extracted, they are input into FaceNet, a deep convolutional neural network renowned for its ability to generate fixed-size feature vectors from face images. These vectors, known as Mel-frequency cepstral coefficients (MFCCs) in audio processing, capture critical facial characteristics essential for distinguishing real faces from artificially generated ones. In this system, the embeddings produced by FaceNet serve a similar purpose by encapsulating key aspects of each face in the video frames.
* **Specific Output:** The specific output of the algorithm is the classification of each video as either containing deepfake content or not. This binary classification is performed by analyzing the sequence of extracted feature vectors to detect anomalies indicative of manipulation.
* **Relevant Factors for Predicting Output:**
* **Mel-frequency cepstral coefficients (MFCCs):** The extracted feature vectors are processed using an LSTM network, which is particularly adept at handling sequences and temporal data. The LSTM analyzes the embeddings over time to identify inconsistencies or patterns that are characteristic of deepfake videos.
* **Model Architecture and Training:** The LSTM network is implemented within the TensorFlow framework, featuring multiple layers including dense layers with ReLU activation functions, and a softmax output layer for binary classification. The architecture and training regimen of the LSTM are critical for its ability to accurately detect deepfakes.
* **Regularization Techniques:** During training, various regularization techniques are applied, such as dropout and hyperparameter tuning. These techniques are essential for preventing overfitting and ensuring that the model generalizes well to new, unseen videos.

**Model Architecture**



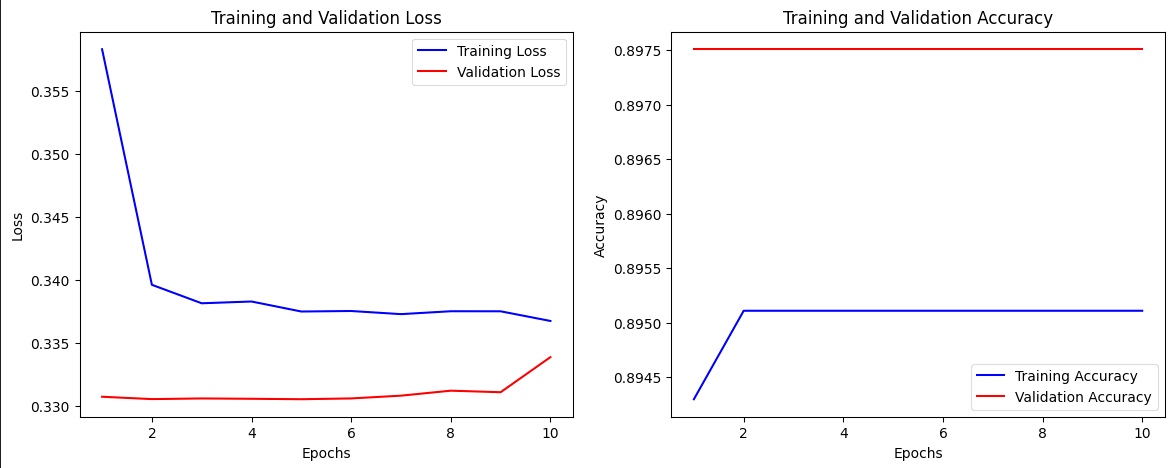
**Training Epochs**



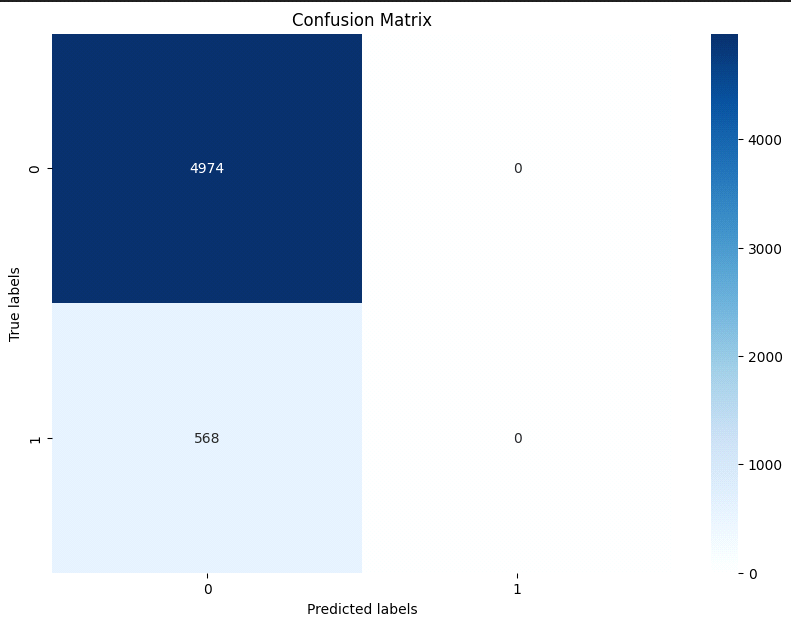
**Analysis**

Our analysis of the training logs for the model, which utilizes FaceNet for feature extraction in distinguishing between real and fake faces, indicates early convergence with minimal variation in loss and accuracy metrics. Specifically, training accuracy hovers around 89.51%, and validation accuracy stabilizes at 89.75%, suggesting the model generalizes well but does not significantly improve with further training. Similarly, the loss decreases slightly from 0.3583 to 0.3367, reflecting modest learning without substantial progress across epochs.The consistently small range in both accuracy and loss values points to potential underfitting, implying that the model may be too simplistic or the extracted features are not adequately capturing the complexities needed to differentiate more effectively between the classes. To address these issues, we propose increasing data variability through augmentation, enhancing the model's complexity, and refining FaceNet’s feature extraction capabilities. These steps aim to expand the model's learning capacity and improve its ability to discern real from fake facial features more distinctly.

**Accuracy**



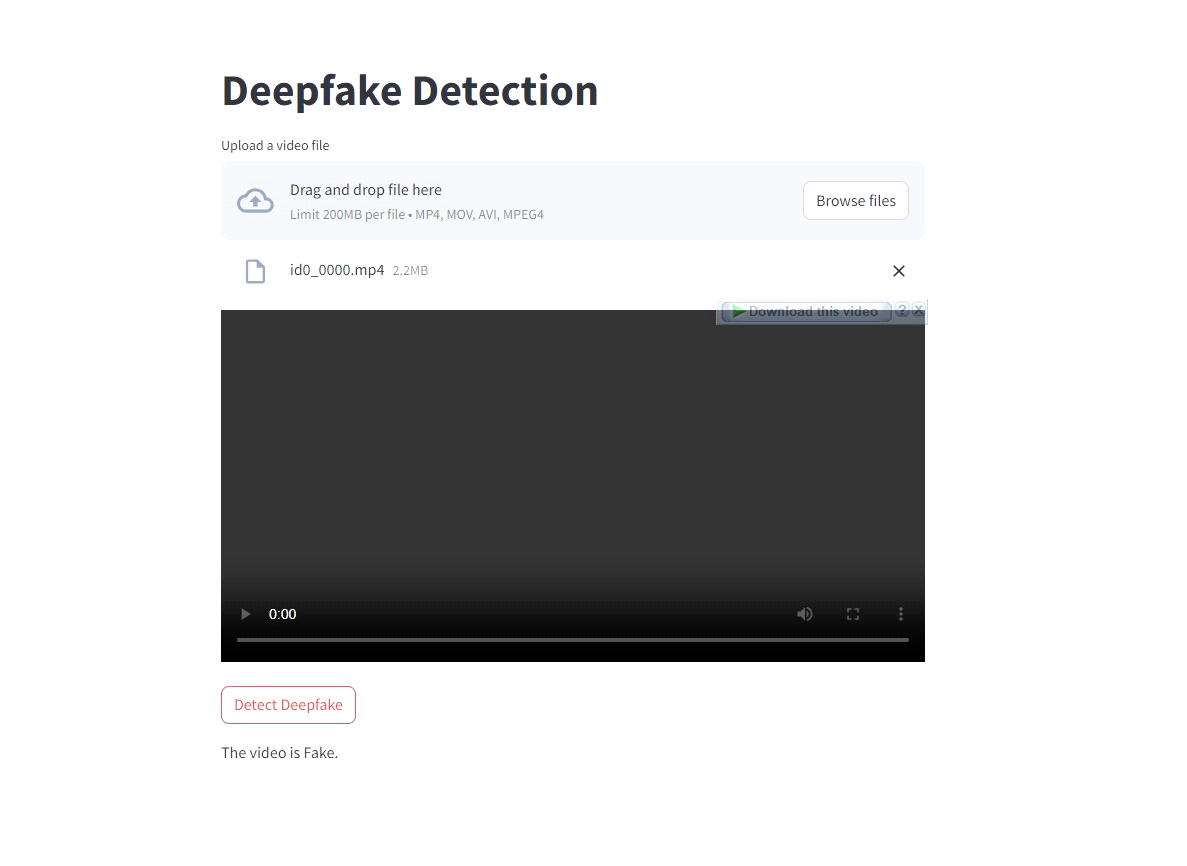
**Confusion Matrix**



**Precision And Recall**



**Output**



**Comclusion**

Our project aimed to detect deepfakes using OpenCV, FaceNet, and LSTM networks; however, we did not fully achieve our goal, as our system struggled to keep pace with the evolving sophistication of deepfake technology. Recognizing these limitations, we propose a shift towards using Generative Adversarial Networks (GANs) for future improvements. GANs, integral to creating deepfakes, can also be pivotal in detecting them by training a discriminator to identify subtle discrepancies characteristic of manipulated content. This approach promises to enhance our detection capabilities by employing the same advanced technologies used in deepfake generation, potentially offering a more effective solution in the ongoing effort to combat digital manipulation.