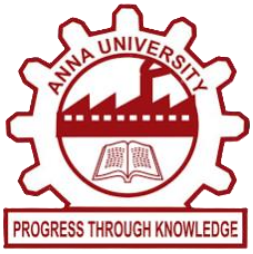


MADRAS INSTITUTE OF TECHNOLOGY

ANNA UNIVERSITY

CHENNAI – 600044

AZ5613

SOCIALLY RELEVANT PROJECT LABORATORY

IDENTIFYING SYNTHETIC VIDEOS

A DEEPFAKE DETECTION APPROACH

PROJECT

Submitted by

RHITHIK S – 2022510009
DIVYA BHARATHI S – 2022510063
KARTHIKEYAN U - 2022510311

B.Tech (6/8)

DEPARTMENT OF INFORMATION TECHNOLOGY

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

January/April 2025

Abstract:

Deepfake technology has rapidly evolved, enabling the creation of highly realistic synthetic videos that pose significant challenges to digital authenticity and security. This project presents a deepfake detection system that leverages two advanced deep learning models: MobileNet and Multiscale Vision Transformer (MViT). MobileNet, known for its lightweight structure, offers real-time detection but exhibits lower accuracy. In contrast, MViT, a transformer-based model, provides superior accuracy but requires higher computational resources. By utilizing the Celeb-DF v2 dataset, this study compares both models' performance in terms of accuracy, precision, recall, F1-score, and AUC. The results demonstrate that MViT outperforms MobileNet, making it more suitable for applications requiring high detection accuracy.

How Our Project is Socially Relevant

Deepfake videos have become increasingly prevalent in recent years, leading to a significant rise in online misinformation and digital identity theft. These manipulated videos are used maliciously to spread false information, defame individuals, and cause public unrest. Detecting deepfakes is crucial for preserving social integrity, combating cybercrime, and protecting individuals' digital identities. Our project addresses this social issue by developing a robust system that accurately identifies synthetic videos, helping authorities and digital platforms ensure the authenticity of media content.

Problem Statement

The primary challenge is to accurately identify synthetic videos generated using AI-based deepfake techniques. This project aims to develop a robust detection system that distinguishes between real and synthetic videos, comparing the performance of MobileNet and Multiscale Vision Transformer (MViT) models.

Objectives

- Develop an accurate and reliable system to detect synthetic videos (deepfakes).
- Compare the performance of MobileNet and MViT in terms of accuracy, precision, recall, F1-score, and AUC.
- Evaluate the efficiency of each model in real-time application scenarios.

Scope of the Project

- **Inclusions:**
  - Deepfake detection from video data (real and synthetic)
  - Comparison of MobileNet and MViT models
- **Exclusions:**
  - Audio deepfake detection
  - Real-time integration with social media platforms

Dataset Information:

The project utilizes the **Celeb-DF v2 dataset**, a widely used benchmark for deepfake detection research. This dataset includes:

- **Total Videos:** Approximately 5,639 videos
- **Real Videos:** 890 genuine videos from celebrities
- **Fake Videos:** 4,749 synthetic videos generated using advanced deepfake techniques
- **Video Quality:** High resolution and realistic facial movements, making detection challenging
- **Dataset Features:**
  - Realistic lip synchronization
  - High-quality face swapping
  - Minimal artifacts compared to earlier datasets
- **Data Format:** MP4 video files categorized into real and synthetic folders
- **Preprocessing Steps:**
  - Frame extraction from videos
  - Normalization of frame size (224x224 pixels)
  - Data augmentation techniques (flipping, rotation) to increase diversity
- **Data Split:**
  - Training: 70%
  - Validation: 15%
  - Testing: 15%

Literature Review / Existing Systems

Deepfake detection has become a vital research area due to the increasing prevalence of manipulated videos. Traditional methods, such as CNN-based frame analysis, have been used to detect inconsistencies in facial features, frame transitions, and pixel-level anomalies. However, as deepfake generation techniques have advanced, these traditional methods often fail to detect high-quality synthetic videos effectively.

To address these challenges, modern deep learning approaches have been introduced, significantly improving detection accuracy. Among these approaches, MobileNet and Multiscale Vision Transformer (MViT) stand out:

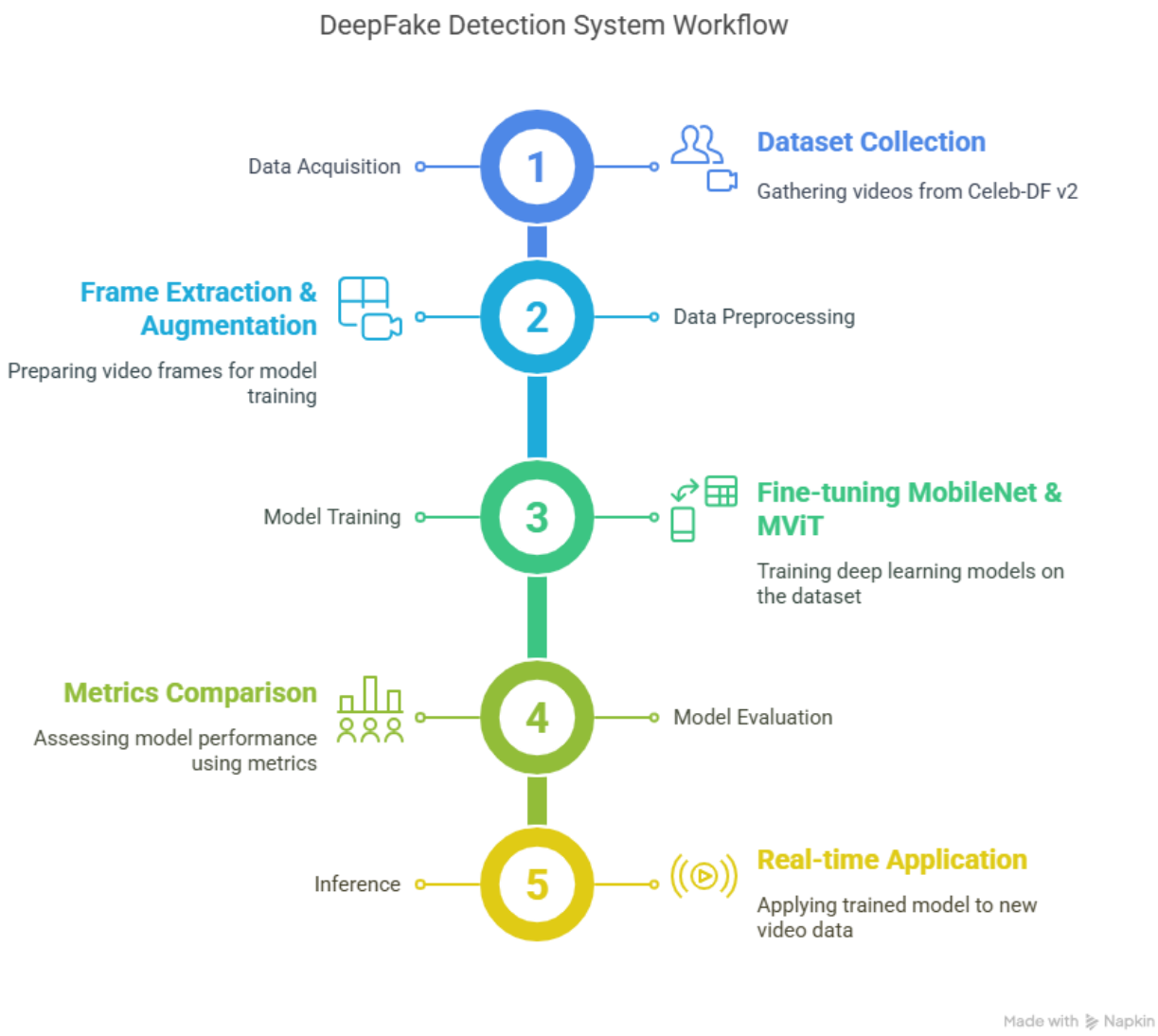
- **MobileNet:** A lightweight CNN architecture that balances speed and accuracy, making it ideal for real-time applications. However, its lower accuracy makes it less suitable for detecting complex and subtle deepfakes.
- **MViT:** A transformer-based model designed to capture fine-grained details in synthetic videos. It offers high accuracy but is computationally intensive, posing challenges for real-time implementation.

Given these contrasting characteristics, it is essential to compare MobileNet and MViT to determine the most practical and effective model for deepfake detection, particularly in applications where both accuracy and real-time performance are critical.

Proposed System Overview

The project employs two deep learning models: MobileNet and Multiscale Vision Transformer (MViT). MobileNet is known for its lightweight structure, making it suitable for real-time applications, but it has lower accuracy compared to MViT. On the other hand, MViT, being a transformer-based model, offers higher accuracy but requires more computational resources. The system leverages the Celeb-DF v2 dataset to train and evaluate both models, comparing their performance in terms of accuracy, precision, recall, F1-score, and AUC. Through this comparative study, the project aims to determine the best approach for detecting synthetic videos in various application contexts. The proposed system integrates MobileNet and MViT for detecting deepfake videos. MobileNet serves as a lightweight solution, while MViT enhances detection accuracy. The system uses the Celeb-DF v2 dataset, containing both real and fake videos, and applies data augmentation to improve robustness.

System Architecture / Block Diagram



The system architecture includes:

1. Data Acquisition: Collecting videos from the Celeb-DF v2 dataset.
2. Data Preprocessing: Frame extraction, normalization, and augmentation.
3. Model Training: Fine-tuning MobileNet and MViT on the dataset.
4. Model Evaluation: Comparing model metrics.
5. Inference: Applying the trained model to new video data.

Modules Description

- **Data Collection:** Gathering real and synthetic video data.
- **Data Preprocessing:** Frame extraction and augmentation.
- **Model Training:** MobileNet and MViT model training.
- **Model Evaluation:** Accuracy, precision, recall, F1-score, and AUC comparison.
- **Inference Module:** Applying trained models to detect deepfakes.

Technology Stack

- **Programming Language:** Python
- **Libraries:** PyTorch, OpenCV, TorchVision, Scikit-learn
- **Framework:** Streamlit for app deployment
- **Hardware:** GPU (NVIDIA Tesla V100), 32 GB RAM, 1 TB SSD

Implementation

- **MobileNet:** Efficient for real-time applications, but less accurate.
- **MViT:** High accuracy, better for detecting complex manipulations.
- **Data Pipeline:** Video data -> Frame extraction -> Model input
- **Comparison:** MobileNet (Accuracy: 73.04%, AUC: 0.6049) vs. MViT (Accuracy: 77.88%, AUC: 0.7931)

Methodology:

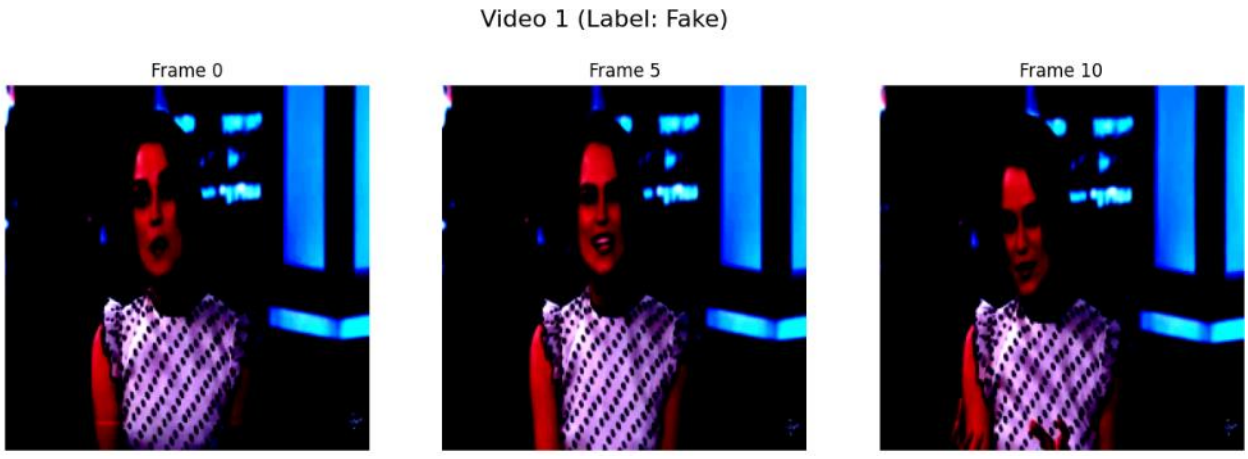
The methodology for deepfake detection in this project follows a structured pipeline as follows:

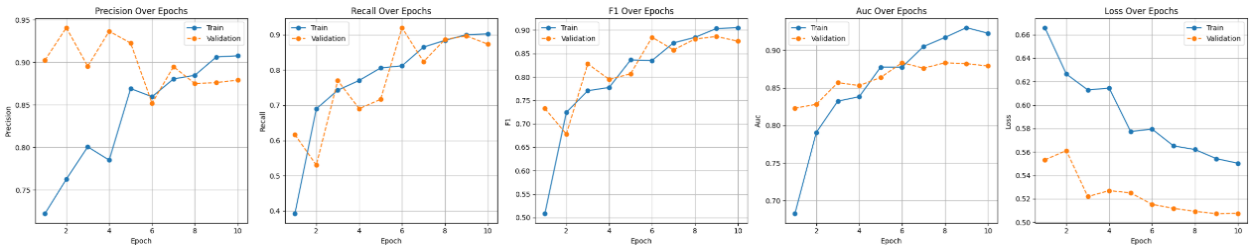
1. **Data Acquisition:**
    - Collect real and synthetic videos from the Celeb-DF v2 dataset.
    - Store videos in organized folders for easy access.
2. **Data Preprocessing:**
    - Extract frames from each video using OpenCV.
    - Normalize frames to ensure consistent size (224x224 pixels).
    - Apply data augmentation techniques to enhance model robustness.
    - Label videos as **real (0)** or **fake (1)** for supervised learning.
3. **Model Selection and Training:**
    - Use two models for comparison:
        - **MobileNet:** Lightweight CNN suitable for real-time detection.

        - **Multiscale Vision Transformer (MViT):** Advanced transformer-based model offering higher accuracy.
    - Fine-tune both models on the preprocessed dataset.
    - Use **Binary Cross-Entropy Loss** as the loss function.
    - Apply **AdamW optimizer** with a learning rate of 0.0001.
4. **Model Evaluation:**
    - Evaluate models on the validation set after each training epoch.
    - Track metrics such as **Accuracy**, **Precision**, **Recall**, **F1-score**, and **AUC**.
    - Visualize the training and validation performance using loss curves and ROC plots.
5. **Comparison and Analysis:**
    - Compare the performance of MobileNet and MViT on the test set.
    - Analyze model accuracy and processing time.
    - Generate confusion matrices and ROC curves to illustrate model efficacy.
6. **Inference:**
    - Load the best-performing model (MViT) for final testing.
    - Detect synthetic videos in new, unseen data.
    - Output classification results with confidence scores.

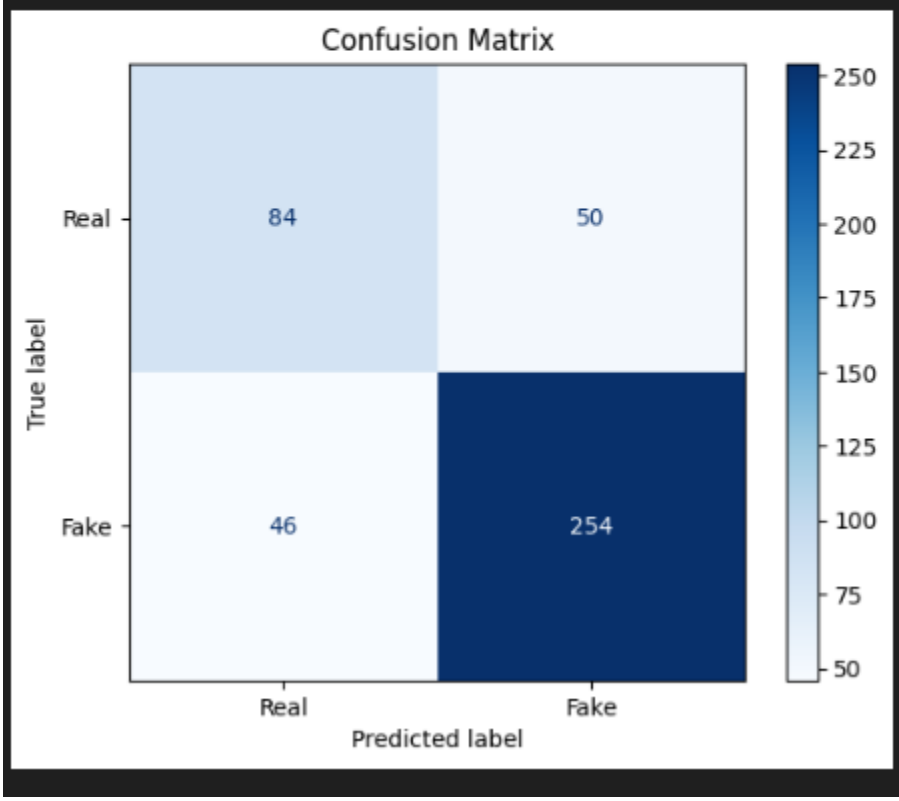
OUTPUT:

Multiscale Vision Transformer:

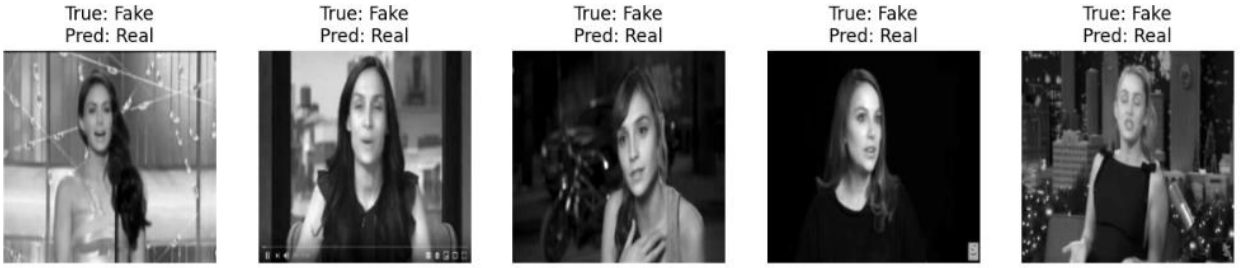




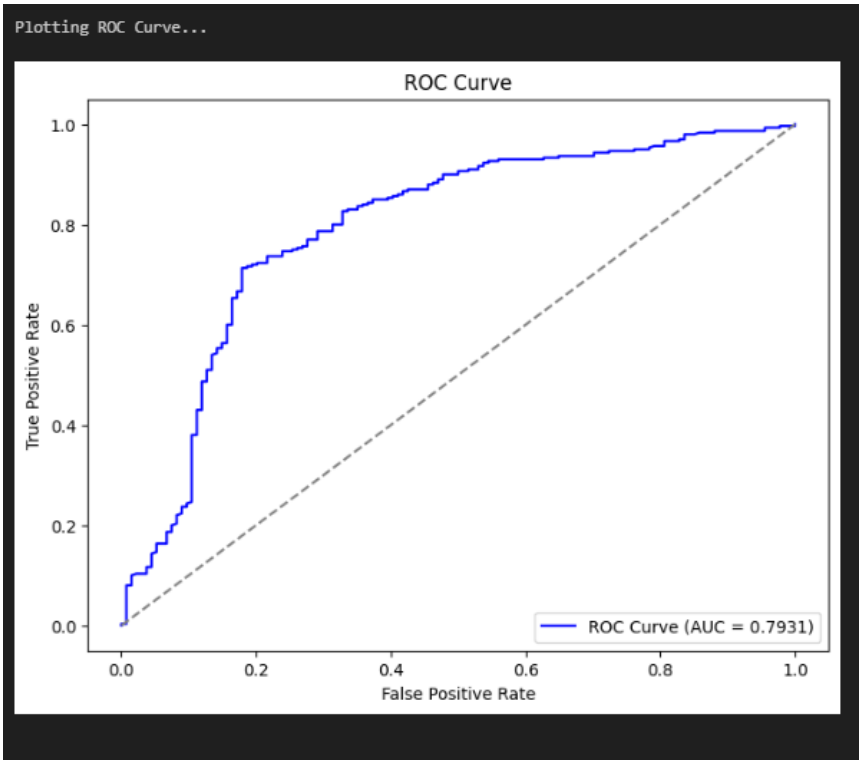

```
Evaluating Model...

Metrics:
Precision: 0.8355
Recall: 0.8467
F1: 0.8411
Auc: 0.7931
Accuracy: 0.7788

Plotting Confusion Matrix...
```



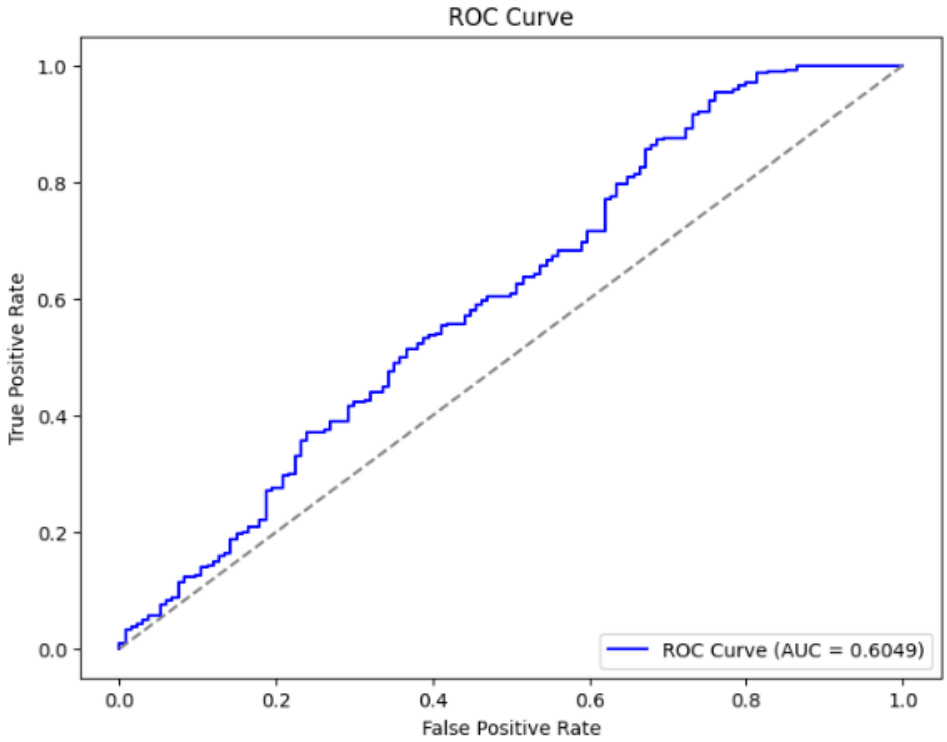
Misclassified Examples



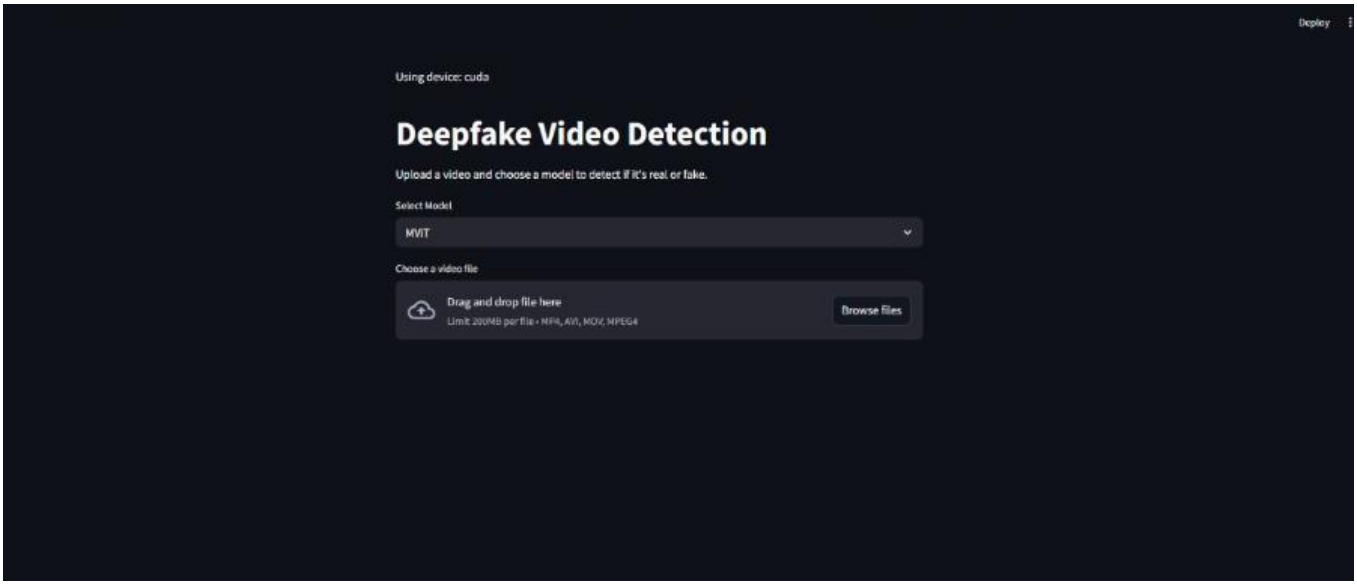
Correctly Classified Examples

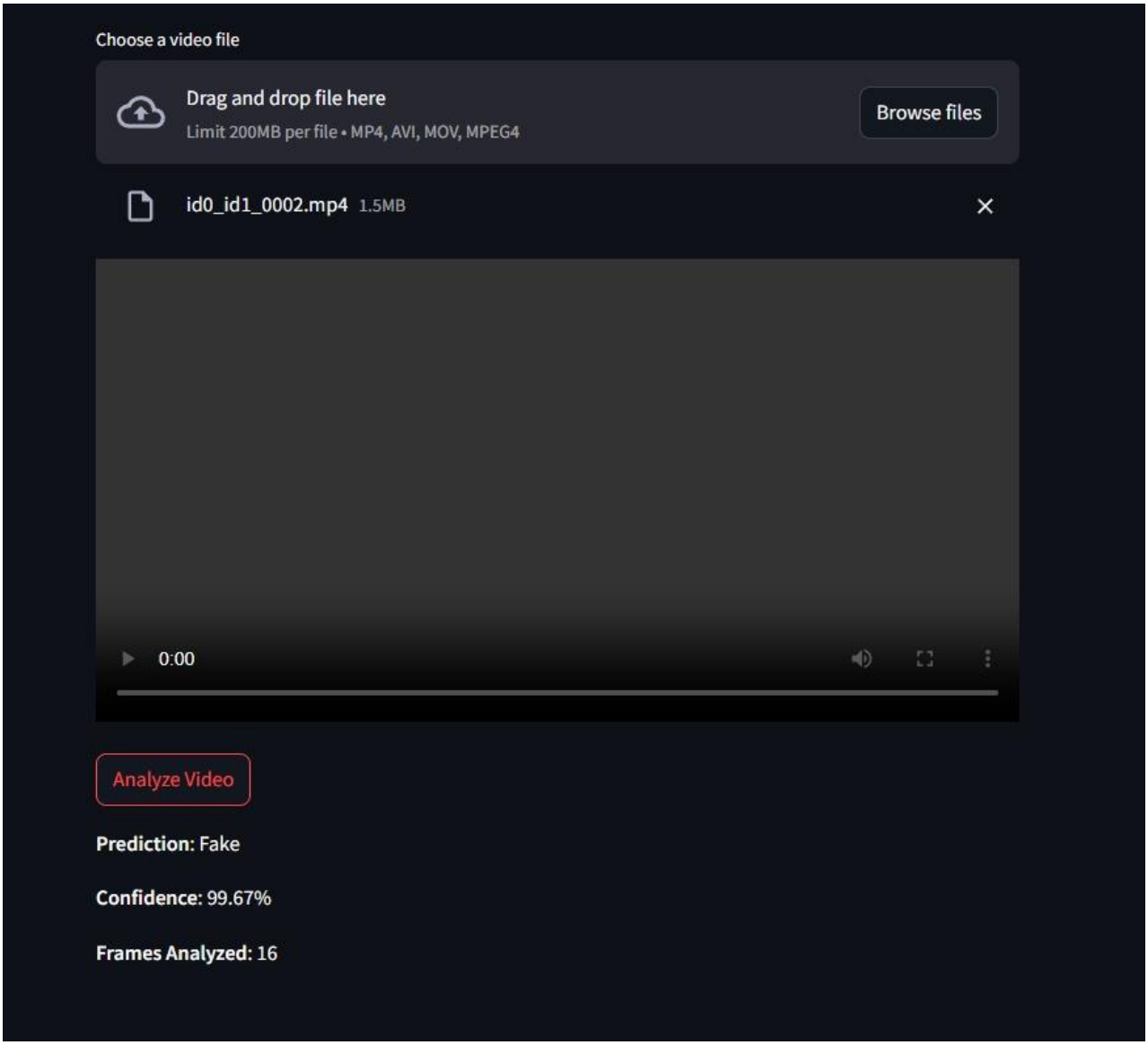


MobileNet Approach for Deepfake Video Detection:



Project Demo





Testing

- **Unit Testing:** Validating frame extraction and augmentation.
- **Integration Testing:** Ensuring model and data pipeline compatibility.
- **Performance Testing:** Evaluating model accuracy and processing speed.

Results and Analysis

- **MobileNet Results:**
  - Accuracy: 73.04%
  - AUC: 0.6049
- **MViT Results:**
  - Accuracy: 77.88%
  - AUC: 0.7931
- **Analysis:** MViT shows better performance due to its transformer architecture, especially in capturing subtle manipulations.

Limitations & Future Enhancements

- **Limitations:**
  - MobileNet's lower accuracy
  - High computational demand of MViT
- **Future Enhancements:**
  - Integrating ensemble models
  - Exploring real-time deployment using optimized versions

Conclusion

The project successfully demonstrates the effectiveness of MViT over MobileNet in deepfake detection. While MobileNet is computationally efficient, it falls short in accuracy, making MViT a preferable choice for critical applications.