Deepfake Video Detection Platform

Using Multimodal Learning

Software Design and Architecture Project Report

Group Members:

Hammad Zeb Khan (4347)
Aleena (4317)
Sara (4383)
Abeer (4225)

Supervisor:

Ma'am Faryal Ishfaq

February 7, 2026

Abstract

The rapid advancement of video manipulation technologies has resulted in a significant increase in deepfake content, posing serious challenges to digital media authenticity, security, and public trust. Existing deepfake detection techniques that rely on a single modality often fail to detect complex manipulations and lack robustness across diverse datasets. To address these limitations, this research proposes a Deepfake Video Detection Platform using multimodal learning. The system analyzes complementary visual, temporal, and audio features to identify subtle inconsistencies in manipulated videos. By leveraging multiple data modalities, the proposed approach enhances generalization and detection reliability. This report details the structured machine learning pipeline, including data preprocessing, multimodal feature extraction, model training, and performance evaluation.

Contents

1 Introduction

The rapid evolution of artificial intelligence and deep learning techniques has enabled the creation of highly realistic synthetic media, commonly known as deepfakes. These manipulated videos often combine altered facial expressions, synchronized speech, and realistic motion patterns, making them increasingly difficult to distinguish from authentic content. While such technologies have legitimate applications, their misuse poses serious threats to digital trust and information security. This research focuses on the design and development of a Deepfake Video Detection Platform using Multimodal Learning.

1.1 Problem Analysis

The primary problem addressed in this research is the reliable and scalable detection of deepfake videos. Existing detection methods mainly rely on a single modality, such as visual or audio cues, which limits their effectiveness against advanced deepfake techniques. Modern deepfakes manipulate visual, temporal, and audio components simultaneously, making unimodal systems ineffective. As a result, these systems struggle to generalize across different datasets and real-world scenarios.

1.2 Requirements

Based on the problem analysis, the system requirements are defined as follows:

Functional Requirements:

- The system shall accept video input files (MP4, AVI) for analysis.
- The system shall extract visual, temporal, and audio features from the input videos.
- The system shall apply multimodal machine learning models to classify videos as real or manipulated.
- The system shall provide detection results along with confidence scores (0-100%).

Non-Functional Requirements:

- **Accuracy:** The system shall ensure high detection accuracy ($> 85\%$) across different manipulation techniques.
- **Scalability:** The system shall be scalable to process large volumes of video data via a microservices architecture.
- **Efficiency:** The system shall maintain reasonable computational efficiency (inference < 20 seconds).

- **Ethics:** The system shall adhere to ethical guidelines, including responsible data usage and bias mitigation.

1.3 Software Design and Architecture

The proposed deepfake detection platform follows a modular machine learning pipeline architecture to ensure flexibility and scalability. A layered architectural pattern is used to separate data handling, model logic, and presentation layers.

Figure 1: System Architecture Diagram showing the Multimodal Pipeline.

- **Input Layer (UI):** A web-based interface (built with Streamlit) that handles user authentication and video file uploads.
- **Preprocessing Module:** Handles frame extraction using OpenCV and audio extraction using Librosa.
- **Inference Engine:** Uses Deep Learning models (Vision Transformers) for feature extraction.
- **Fusion Layer:** Combines visual and audio scores for the final verdict.

2 Literature Review

We conducted a structured literature review to understand current methodologies and identify gaps.

2.1 Visual Detection Methods

MesoNet (Afchar et al.): This work focuses on detecting facial tampering in compressed videos using a compact CNN. However, it relies solely on mesoscopic visual properties and ignores the audio channel, making it vulnerable to deepfakes with manipulated audio. This highlights the need for our multimodal approach.

FaceForensics++ (Rossler et al.): This paper introduced a large-scale dataset and benchmarked detectors like XceptionNet. The key limitation identified is that performance drops significantly on compressed videos (e.g., social media shared content). Our project aims to address this robustness gap.

2.2 Multimodal & Temporal Methods

Lip-Forensics (Haliassos et al.): This approach targets high-level semantic irregularities in mouth movement. While effective, it is a visual-only approach that infers speech but does not process the actual audio file. Our research validates that lip-sync inconsistency is key but requires actual audio signal processing.

Deepfake Detection Using Multimodal AI (Joshi et al.): This research addresses the failure of unimodal systems by using a Fusion Model that combines visual features (CNN) and audio features (Spectrograms). This serves as the foundational blueprint for our proposed architecture, though we improve upon the fusion mechanism.

3 Implementation

3.1 Technology Stack

We implemented the solution using **Python 3.9** due to its rich AI ecosystem. The core libraries include:

- **Computer Vision:** OpenCV (‘cv2‘) for frame extraction and face detection.
- **Deep Learning:** Hugging Face ‘transformers‘ library using a pre-trained Vision Transformer (ViT).
- **Audio Processing:** Librosa for spectral feature extraction.
- **Interface:** ‘Streamlit‘ for rapid application development.

3.2 The Inference Pipeline

The core logic follows a “Divide and Conquer” strategy:

1. **Frame Sampling:** The system extracts 1 frame every second using OpenCV to reduce computational load.
2. **Visual Analysis:** Each frame is passed to the ViT model (‘dima806/deepfake_vs_real‘) to detect pixel-level artifacts.
3. **Audio Analysis:** The system extracts the audio track and generates a spectrogram to visualize consistency.
4. **Fusion:** The scores are combined using a weighted average formula to determine the final probability.

4 Expected Results

Based on our implementation and testing on sample datasets, we observe the following:

Figure 2: User Interface displaying the Deepfake Analysis Dashboard.

- **Detection Accuracy:** The system correctly identifies high-quality deepfakes with an estimated accuracy of 85-90%.
- **Real-World Robustness:** The application successfully processes user-uploaded videos in standard formats (.mp4, .avi) with an average processing time of 10-15 seconds.
- **User Feedback:** The "Red Alert" UI provides clear, actionable feedback to the user when a threat is detected.

Figure 3: Result Screen showing a positive Deepfake detection (Red Alert).

5 Conclusion and Future Work

This project successfully designed and implemented a Multimodal Deepfake Detection Platform. By integrating visual artifact detection with audio analysis, we addressed the key limitations of unimodal systems. The system demonstrates that Transfer Learning is a viable strategy for rapid forensic tool development.

5.1 Future Work

- **Automated Takedown Integration:** Future versions will integrate with social media APIs to automatically report detected content.
- **Real-Time Browser Extension:** Development of a browser plugin to flag content in real-time.
- **Blockchain Evidence:** Storing detection logs on a blockchain to create an immutable record of authenticity.