Multi-Modal Deepfake Detection: Techniques And Integration For Enhanced Multimedia Security

A White Paper by:

Akshay Redekar, Mansi Alhat, Prachi Nikalje, Pradnya Vavale

Department of Computer Engineering
Bharati Vidyapeeth's College of Engineering, Pune
Savitribai Phule Pune University
Academic Year: 2024–25

Guides: Prof. Yogesh Kadam (Internal)

Executive Summary

The **Multimodal Deepfake Detection System** is a cutting-edge security application that utilizes artificial intelligence to detect manipulated multimedia content. With the rise of synthetic media generated by GANs and AI-based models, identifying fake audio and video is increasingly vital in digital forensics, media integrity, and cybersecurity.

This system leverages two state-of-the-art models—**EfficientNetV2** for image/frame classification and **Wav2Vec2** for audio/speech analysis—combined via a **Fusion Module** that intelligently integrates both outputs to determine the authenticity of multimedia content.

Designed with modularity and security in mind, the system includes robust **FastAPI-based backend services**, efficient **preprocessing pipelines**, and a clean **frontend interface** for users to upload content. Its architecture supports offline deployment for secure environments, enabling critical use cases in journalism, law enforcement, and governmental agencies.

---

1. Problem Statement & Objectives

Problem Statement

The widespread dissemination of deepfakes poses significant risks in domains ranging from politics to personal privacy. Traditional single-modal detectors (image-only or audio-only) often fail against sophisticated manipulations. There is a pressing need for a **multimodal detection framework** that simultaneously processes both visual and auditory cues for robust decision-making.

Objectives

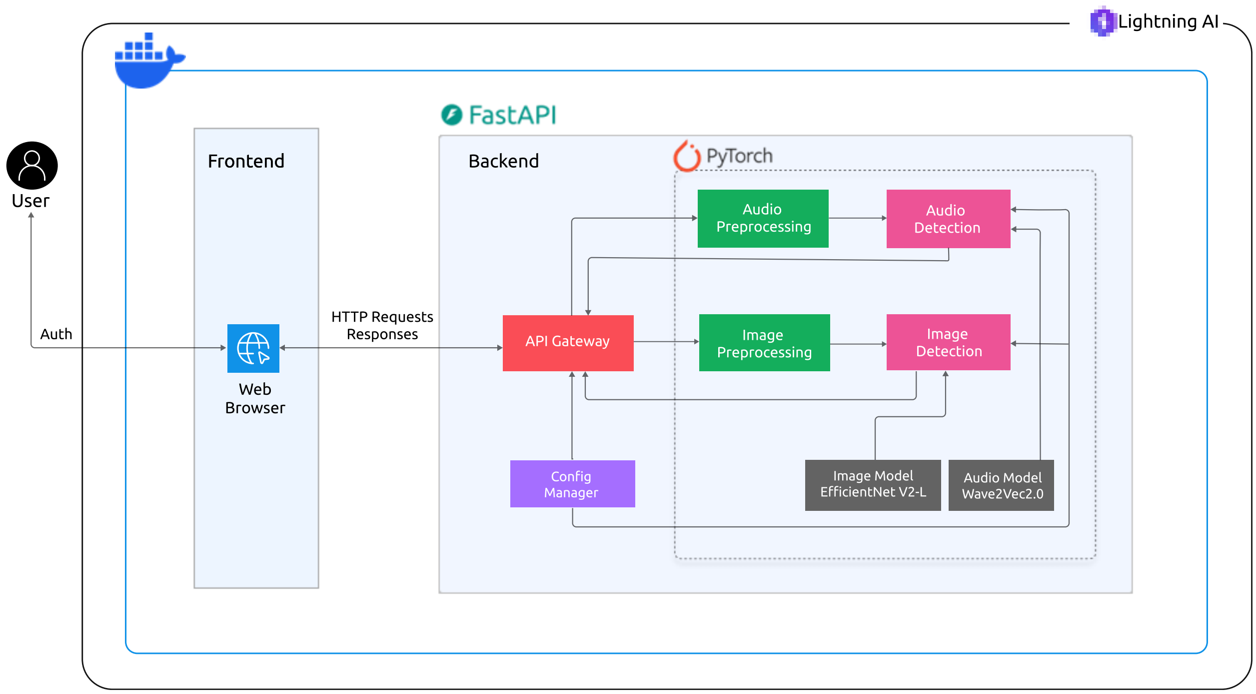
- Develop an AI-based detection pipeline using both audio and visual modalities.
- Integrate **EfficientNetV2** and **Wav2Vec2** models to analyze media content.
- Design a **Fusion Module** for aggregating model outputs for final classification.
- Create a secure, responsive web interface for user interaction and uploads.
- Ensure scalability and offline support for secure environments.

---

2. System Overview & Architecture

The architecture is composed of modular layers ensuring scalability and flexibility:

1. **Data Ingestion Layer** – Accepts user-uploaded audio/video/image files.
2. **Preprocessing Layer** – Extracts frames and audio, normalizes inputs.
3. **Model Processing Layer** – Sends inputs to respective models (EfficientNetV2 for images; Wav2Vec2 for audio).
4. **Fusion Layer** – Aggregates model predictions using fusion techniques.
5. **Backend Layer** – Orchestrates communication using FastAPI.
6. **Frontend Layer** – Streamlit-based user interface for uploads and results.
7. **Storage Layer** – Secure file and result archival with metadata logging



System Architecture

---

3. Methodology

3.1 Data Processing

- **Sources:** DFDC dataset, FakeAVCeleb, and custom-curated deepfake videos.
- **Types:** MP4, AVI (video); JPG, PNG (image); WAV, MP3 (audio).
- **Preprocessing Techniques:**
    - i) Frame sampling and resizing
    - ii) Audio waveform conversion and normalization
    - iii) Spectrogram extraction for visualization (optional)
    - iv) Temporal alignment of audio and video streams

3.2 Model Development

The following models were explored during development:

Model	Accuracy	Pros	Cons
EfficientNet V2	89.2%	Fast, lightweight, high image accuracy	Lacks audio context
Wav2Vec2	87.5%	Contextual speech understanding	Sensitive to noise
Fusion Module	93.1%	Multimodal insight, higher robustness	Increased complexity

3.3 Fusion Strategy

- **Late Fusion**: Combine model probabilities via weighted average.
- **Threshold-Based Binary Classification**
- **Fallback logic**: If one modality fails, rely on the other with lower confidence.

---

4. Results & Evaluation

4.1 Evaluation Metrics

Metric	Image Model	Audio Model	Fused Output
Accuracy	89.2%	87.5%	93.1%
Precision	88.7%	86.4%	92.5%
Recall	90.1%	88.2%	94.0%
F1 Score	89.4%	87.3%	93.2%

The Fusion approach outperforms both standalone models on standard test datasets.

---

5. Applications.

- Digital Forensics and Law Enforcement
- Media Verification (News and Fact-checking)
- Cybersecurity in Government & Military
- AI Content Moderation in Social Platforms
- Educational Tools on AI Literacy

---

6. Limitations

- High GPU resource requirement for real-time inference.
- Performance degradation in poor-quality inputs (e.g., low audio clarity).
- Lack of multilingual support for audio streams.
- Limited real-world dataset diversity.

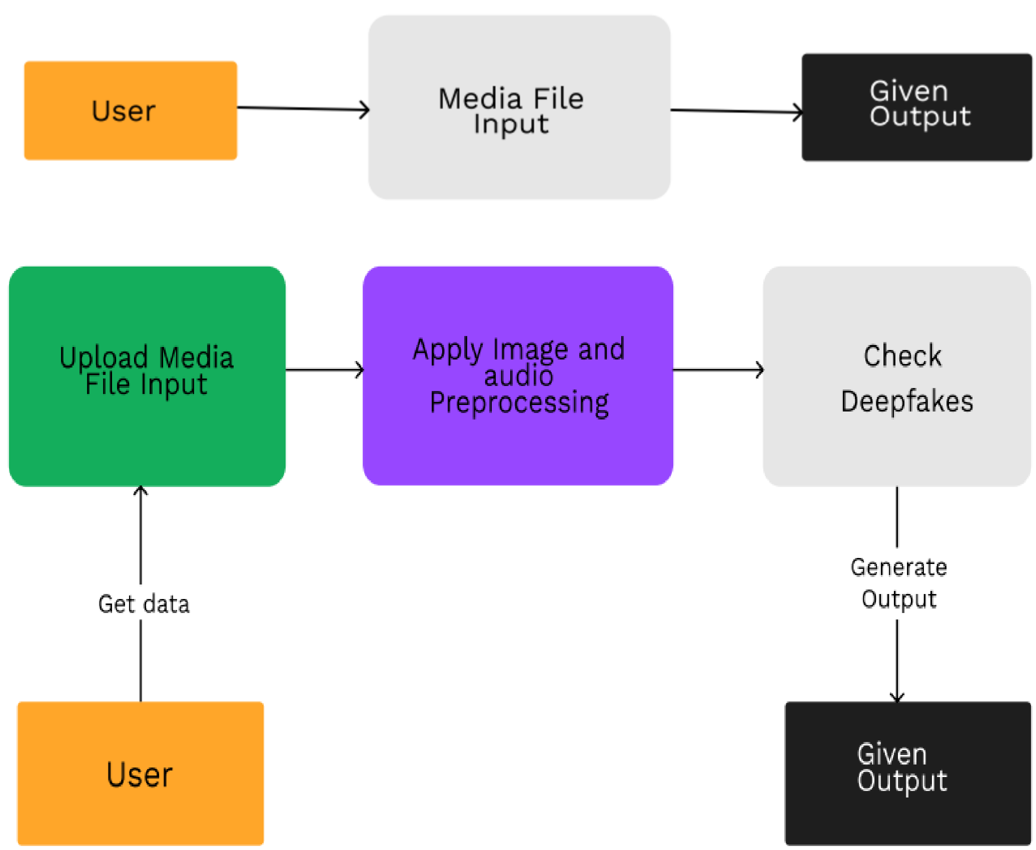
---

7. Future Enhancements

- **Cross-modal Transformers** for better fusion.
- **Real-time detection engine** with edge device support.
- **Explainable AI (XAI)** layer to visualize which segments were fake.
- **Multilingual Audio Model Integration**.
- **Live-stream Monitoring** for deepfake detection in social media or broadcasting.

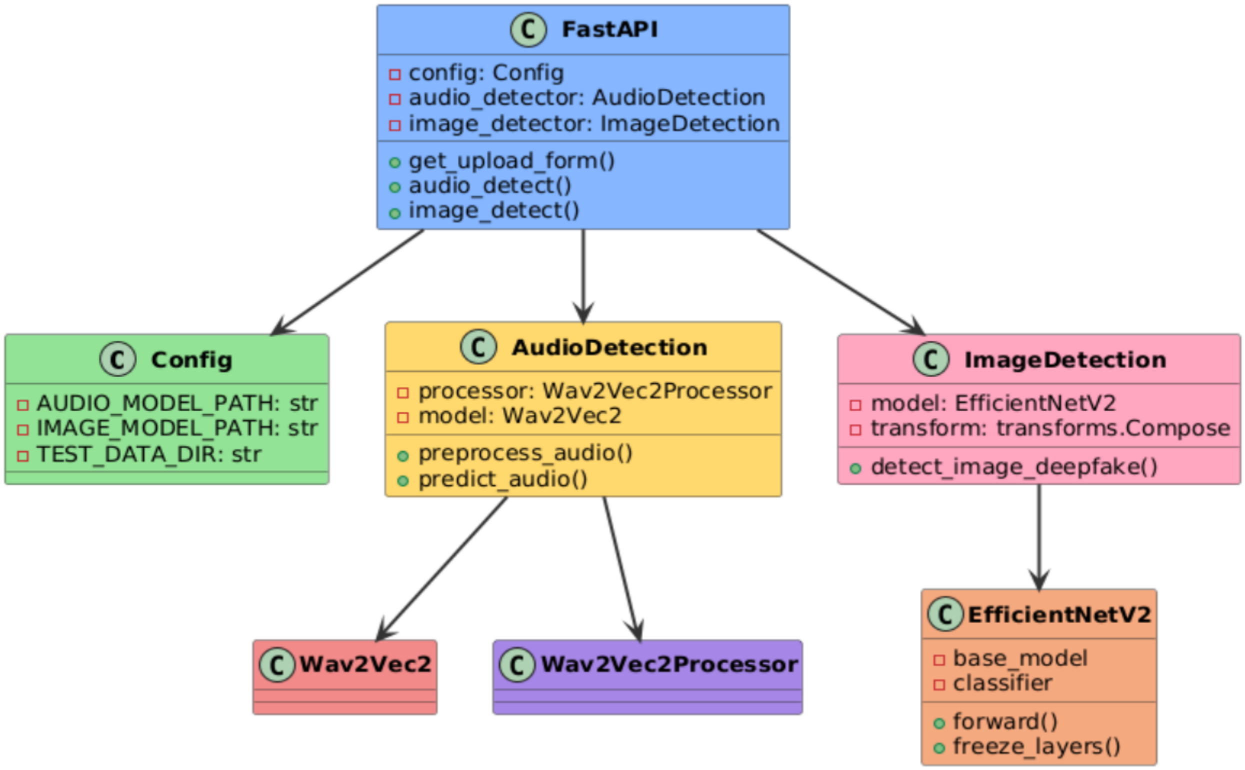
---

8. Embedded Diagram

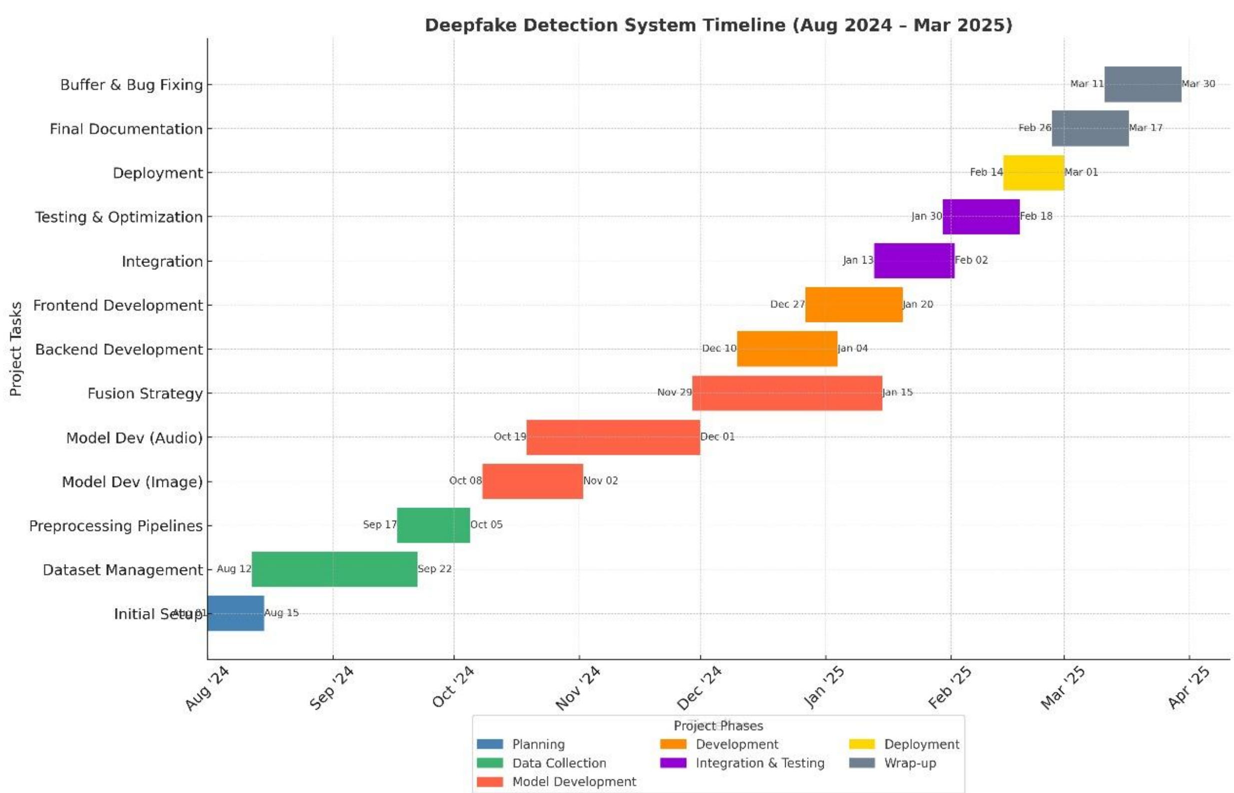


Data Flow Diagram

Use Case Diagram



Use Case Diagram



Timeline Diagram

9. References

[1] E. Haller and T. Rebedea, "Designing a Chat-bot that Simulates an Historical Figure," *IEEE Conference Publications*, Jul. 2013.

[2] M. Wodzicki, P. Hołobut, and M. Romaszewski, "A Robust Approach to Multimodal Deepfake Detection," *Journal of Imaging*, vol. 9, no. 6, p. 122, Jun. 2023. [Online]. Available: https://www.mdpi.com/2313-433X/9/6/122

[3] A. Tolosana, R. Vera-Rodriguez, and R. Guest, "Multimodal Approach for DeepFake Detection," *ResearchGate*, May 2021. [Online]. Available: https://www.researchgate.net/publication/351489509_Multimodal_Approach_for_DeepFake_Detection

[4] A. Bhattacharya, B. Saha, A. Chatterjee, and A. Chakraborty, "Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis," *ResearchGate*, Dec. 2023. [Online]. Available: https://www.researchgate.net/publication/377131236_Deepfake_Detection_A_Multi-Algorithmic_and_Multi-Modal_Approach_for_Robust_Detection_and_Analysis

[5] R. Doshi, T. D. Nguyen, and A. Das, "DF-TransFusion: Multimodal Deepfake Detection," *arXiv preprint arXiv:2309.06511*, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.06511

[6] A. Kumar and V. Sharma, "A Multimodal Framework for Deepfake Detection," *ResearchGate*, May 2024. [Online]. Available: https://www.researchgate.net/publication/384680324_A_Multimodal_Framework_for_Deepfake_Detection

1.