International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Smart Deepfake Detection System

Bhookya Anurag[1], Dubaka Akhil[2], Bhukya Srinivas[3],

Dr.M.Mamatha[4], Ms.Vedavathi.K[5], Dr.K.Rajitha[6], Dr.V.Subbaramaiah[7]

[1]Student, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

[2] Student, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

[3] Student, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

[4] Assistant Professor, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

[5] Assistant Professor, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India

[6] Assistant Professor, Computer Science and Engineering, Mahatma Gandhi Institute of Technology ,Hyderabad, India

[7] Assistant Professor, Computer Science and Engineering, Mahatma Gandhi Institute of Technology ,Hyderabad, India

---

Abstract - Deepfakes are artificially generated multimedia content that can convincingly mimic real human faces and voices using advanced AI techniques such as Generative Adversarial Networks (GANs). This poses serious ethical, social, and security challenges in digital communication. To address this issue, the proposed project presents a Multimodal Deepfake Detection System that integrates image, video, and audio analysis pipelines within a unified framework. The system employs Efficient Net-based CNN for image forgery detection, CNN combined with Bi-LSTM for temporal video analysis, and a 1D-CNN with LSTM for detecting manipulated or cloned audio. The predictions from these modalities are combined using a Fuzzy Fusion Engine, which intelligently weights each confidence score to produce a final verdict with high accuracy and interpretability. The model is trained using public Deepfake datasets such as Face Forensics++, Celeb-DF, and DFDC, with binary cross-entropy loss, data augmentation, and early stopping to ensure stable convergence and better generalization. The trained models are deployed on Hugging Face Spaces, while the web interface is hosted on Vercel, enabling real-time Deepfake detection for users through a browser interface. This approach enhances detection accuracy by leveraging multimodal evidence (visual, temporal, and auditory), improves generalization across datasets, and provides an explainable and efficient solution to combat the growing threat of Deepfakes.

Key Words: Deepfake Detection, Multimodal Learning, Bi-LSTM, Fuzzy Fusion, Convolutional Neural Network (CNN), Face Forensics++, Celeb DF, DFDC, Machine Learning, Real-Time Detection, Hugging Face.

1.INTRODUCTION

The rapid advancement of artificial intelligence and deep learning has led to the emergence of Deepfake technology, which can generate highly realistic synthetic videos, images, and audio that closely resemble real people. These manipulated media, created using deep generative models such as Generative Adversarial Networks (GANs) and Autoencoders, pose serious threats to privacy, social trust, and information authenticity. Deepfakes are increasingly being used in misinformation campaigns, political propaganda, cybercrimes, and identity theft, making reliable detection a global necessity. Traditional detection systems that rely only on visual or audio cues have become less effective as fake content grows more sophisticated and harder to distinguish from real data.

To address these challenges, the proposed project introduces a Multimodal Deepfake Detection System that integrates image, video, and audio analysis using deep learning models such as Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi- LSTM) networks. Each module extracts spatial, temporal, and auditory features, and their outputs are combined using a Fuzzy Fusion Engine to produce an accurate and explainable classification of real or fake content. The system is trained on benchmark datasets like FaceForensics++, Celeb-DF, and DFDC, and deployed using Hugging Face Spaces (backend) and Vercel (frontend) for real-time detection. This approach provides a robust, interpretable, and scalable solution for detecting Deepfakes across multiple data types and real-world environments.

2. RELATED WORK

[1] Deepfake Detection Using Spatio -Temporal-Structural Anomaly Learning and Fuzzy System-Based Decision Fusion by Brindha Subburaj and R. Ragavendra (2025, IEEE Access) proposes a hybrid Deepfake detection system combining spatial, temporal, and structural features extracted using multiple CNN branches. The outputs are fused through a Fuzzy Inference System (FIS) to achieve improved interpretability and detection performance. The system achieved 99% accuracy on FaceForensics++ and 93% on Celeb-DF(v2), demonstrating strong generalization and robustness to various forgery types. [2] Investigating Voiced and Unvoiced Regions of Speech for Audio Deepfake Detection by G. Sivaraman, H. Tak, and E. Khoury (2025, IEEE) analyzes speech Deepfakes by separating voiced and unvoiced segments using the AASIST graph attention model. The study shows that unvoiced regions of speech are more discriminative in detecting synthetic

voices. The fusion of both regions reduces the Equal Error Rate (EER) to 5.82%, proving the advantage of fine-grained speech analysis. [3] Freeze and Learn: Continual Learning with Selective Freezing for Speech Deepfake Detection by D. Salvi, V. Negroni, and L. Bondi (2025, IEEE) introduces a selective freezing mechanism for speech Deepfake detection that allows neural networks to retain previous knowledge while adapting to new datasets. The method divides model layers into frozen (non-trainable) and active (trainable) parts, allowing stable features to be preserved while newer data patterns are learned. This prevents catastrophic forgetting and supports continual learning across evolving Deepfake datasets. The system is trained using incremental updates instead of full retraining, making it efficient and scalable.[4] Leveraging Mixture of Experts for Improved Speech Deepfake Detection by V. Negroni, D. Salvi, and A. I. Mezza (2025, IEEE) presents a Mixture of Experts (MoE) model where several specialized networks (experts) work under the control of a gating network that determines which expert should process an input. Each expert focuses on specific manipulation types, such as voice cloning or synthesis. This dynamic learning structure improves the overall accuracy, robustness, and adaptability across unseen Deepfake datasets compared to conventional CNNs. The MoE framework enhances generalization and prevents overfitting by intelligently routing the input. However, it is computationally demanding and limited to audio-only detection. My proposed model builds on this dynamic adaptation idea by integrating multiple modality "experts" (image, audio, and video modules) combined through fuzzy fusion. This allows for smarter feature-level fusion and multimodal adaptability, improving overall detection efficiency and decision reliability.[5] Deep Fake Detection with Hybrid Activation Function Enabled Adaptive Milvus Optimization-Based Deep CNN by H. Mashetty, N. Erukulla, and S. Belidhe (2025, IEEE) proposes a Deepfake detection method based on a hybrid activation function and Milvus optimization algorithm for adaptive parameter tuning in CNNs.The model achieved 95.72% accuracy, 94.9% recall, and 96.5% precision on FaceForensics++, outperforming baseline detectors and proving effective against different compression levels.The system achieved 95.72% accuracy, 94.9% recall, and 96.5% precision on the FaceForensics++ dataset, outperforming standard CNN models, and showing robustness under varying compression levels.However, the system works only on visual data and does not handle temporal or audio-based manipulations.[6] SpecViT: A Custom Vision-Transformer Based Approach for Audio Deepfake Detection by S. Modak, A. K. Das, and R. Naskar (2025, IEEE) introduces SpecViT, a dual-attention Vision Transformer model for detecting fake audio from spectrograms. The system achieved 99% accuracy and an F1-score of 0.9911, with an Equal Error Rate of 3.5%, showing superior performance over CNN-based detectors for voice synthesis and spoofing attacks. The model achieved 99% accuracy and an F1-score of 0.9911 with an Equal Error Rate of 3.5%, outperforming CNN-based models. Despite its outstanding performance, it requires large computational resources and lacks multimodal support.[7] Emerging Innovations in Deep Learning for Video Deepfake Detection: A Comprehensive Review by M. K. Makwana and D. K. Singh (2025, IEEE) provides an extensive review of recent deep learning-based Deepfake detection methods, including CNNs, Transformers, and hybrid networks. The paper discusses challenges such as dataset bias, multimodal forgeries, and explainability, emphasizing the need for cross-modal and interpretable detection systems.The study concludes that hybrid and transformer-based models offer the most potential but require more generalization and multimodal learning capabilities. [8] Audio Features Investigation for Singing Voice of Deepfake Detection by M. Gohari, D. Salvi, and P. Bestagini (2025, IEEE) studies the performance of CNN-based detectors for detecting Deepfakes in singing voices. The model evaluates handcrafted and learned features from singing data, identifying spectral and harmonic features as keyindicators of synthesized vocals, improving accuracy in musical audio forgery detection. A CNN model trained on singing datasets achieved high precision in detecting synthetic performances. While effective in music audio, this model is limited to specific genres and cannot process speech or visual fakes.

3.PROPOSED SYSTEM

The proposed system is a Multimodal Deepfake Detection System designed to detect forged images, videos, and audio clips by analyzing multiple modalities within a unified deep-learning framework. Unlike existing single-modality detectors, this system combines spatial, temporal, and auditory information to make more accurate and reliable predictions.

The architecture consists of three major pipelines:

- **Image Pipeline (EfficientNet-CNN):** Extracts spatial features from still frames to identify abnormal textures, blending artifacts, and GAN fingerprints that often occur in manipulated facial regions.
- **Video Pipeline (CNN + Bi-LSTM):** Captures temporal relationships among consecutive frames to detect irregular motion, eye-blinking patterns, and expression mismatches that reveal video-level forgery.
- **Audio Pipeline (1-D CNN + LSTM):** Analyzes the sound waveform and its spectrogram to recognize cloned or computer-generated voices and to verify the synchronization between lip movement and speech.

The outputs from these three pipelines are integrated through a Fuzzy Fusion Engine, which assigns adaptive weights to each modality and produces an overall decision—*Real* or *Fake*—with an associated confidence score. This fuzzy approach handles uncertainty better than hard thresholds and improves interpretability.The system is trained on benchmark datasets such as FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge (DFDC). Training uses binary-cross-entropy loss, data augmentation, and early stopping to ensure stable convergence and prevent overfitting. After training, the models are deployed for real-time inference: the backend runs on Hugging Face Spaces. And the user interface

is hosted on Vercel, allowing users to upload media files and instantly view the detection results.

This proposed multimodal framework enhances Deepfake detection accuracy, increases robustness to unseen datasets, and provides an efficient, explainable, and cloud-deployable solution for combating the growing threat of synthetic media.

4.METHODOLOGY

4.1 System Architecture

The designing the complete Deepfake Detection System, it is essential to establish a structured methodology that clearly defines how different components work together to achieve accurate and reliable detection. The design methodology focuses on understanding the flow of data from the user input to the final prediction, identifying the required processing stages, and mapping the interactions between various modules. Since Deepfake detection involves complex multimodal analysis—covering images, videos, and audio—the architecture must be capable of handling heterogeneous data, extracting meaningful features, and applying advanced machine learning models. This chapter outlines the overall system architecture, diagrams, and workflow that form the foundation of the proposed multimodal Deepfake Detection System.

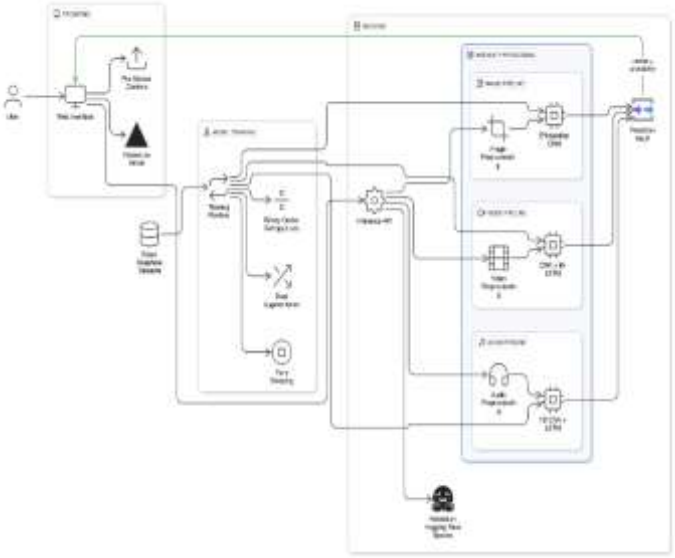


Figure 1: System Architecture diagram of Smart Deepfake Detection System

The Figure 1 represents System Architecture of the proposed Deepfake Detection System is divided into three major components Frontend, Backend, and Model Training Environment.

Frontend (User Interface):

- Hosted on Vercel, allowing users to upload an image, video, or audio file.
- Provides real-time interaction and displays prediction results.

Backend (Inference and Processing):

- Hosted on Hugging Face Spaces**.**
- Accepts uploaded files via API, classifies the input type (image, video, audio), and sends it to the respective detection pipeline.

The results from each pipeline are processed through a Fuzzy Fusion Engine to produce the final prediction (“Real” or “Fake”) along with confidence score.

Model Training Environment:

- Uses CNN**,** Bi-LSTM**,** and 1D-CNN + LSTM models trained on FaceForensics++**,** Celeb-DF, and DFDC datasets.
  Implements data augmentation, binary cross-entropy loss, and early stopping to improve model generalization.

Workflow: User uploads → Input Preprocessing → Model Inference → Fuzzy Fusion → Final Prediction → Display Result.

4.2 Activity Diagram

The Activity Diagram describes the workflow of the Smart Deepfake Detection System from the moment a user uploads a file to the display of the final prediction. It explains the logical sequence of operations and the flow of data through different system components. The process begins with the user

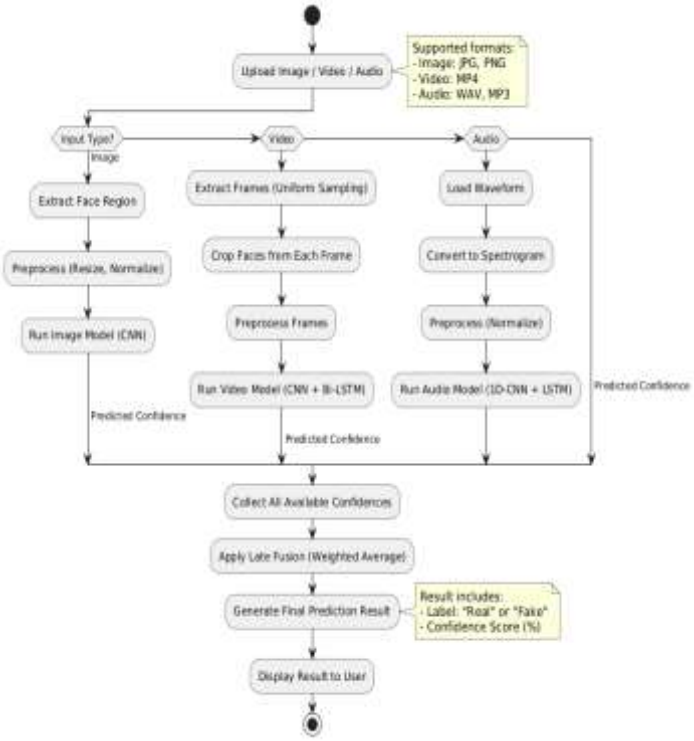


Figure 2: Activity diagram of Smart Deepfake Detection System.

uploading an input file, which may be an image, video, or audio. The system then identifies the input type and activates the corresponding processing module. Preprocessing steps such

as resizing, normalization, or spectrogram generation are applied depending on the media type. The processed data is then passed through deep learning models such as CNN and LSTM to generate prediction scores. These scores are forwarded to the Fuzzy Fusion Engine, which computes the final decision by combining results from all modalities. Finally, the system displays the output indicating whether the input is real or fake along with the confidence score, completing the workflow

4.3 Use Case Diagram

The Use Case Diagram illustrates how the user interacts with the Smart Deepfake Detection System and defines the available functionalities. It represents the interaction between the user and the system, highlighting key actions such as uploading image, video, or audio files, submitting them for analysis, viewing prediction results, and

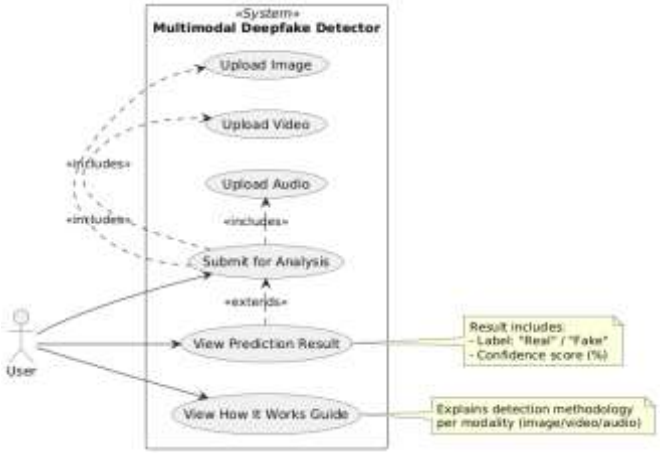


Figure 3: Use Case diagram of Smart Deepfake Detection System.

accessing detection explanations through interpretability outputs like visual attention maps. The system boundary includes both frontend and backend components, forming the complete Deepfake Detection System. This diagram provides a clear understanding of the functional requirements and high-level behavior, where the user initiates the process and receives real-time feedback through the interface, ensuring efficient, accurate, and secure processing.

5. IMPLEMENTATION

5.1 System Execution Overview

The core of the system is a Python-based inference service built with PyTorch and deployed as a backend application on Hugging Face Spaces. On startup, the application loads the three pre-trained modality-specific models along with the fuzzy fusion engine. The inference workflow is orchestrated through a single FastAPI endpoint wrapped in Gradio for seamless Hugging Face compatibility. Because the system is designed for stateless cloud inference. Instead, each user request triggers parallel execution of the three pipelines using Python's concurrent.futures module wherever possible. This design ensures that computationally intensive tasks such as video frame extraction or spectrogram generation do not block the overall response time.

The application follows these high-level steps for every user request:

- Accept and validate the uploaded media file (image, video, or audio).
- Route the file to the appropriate preprocessing module and execute the three pipelines concurrently.
- Fuse the modality outputs using the fuzzy fusion engine.
- Generate a final Real/Fake decision along with a confidence score and explainability details.

5.2 Image Pipeline

The image pipeline is responsible for extracting spatial features from still images or individual video frames to detect manipulation artifacts.

5.2.1 Image Preprocessing

Uploaded images or extracted video frames are resized to 224×224 pixels and normalised using ImageNet statistics via the torchvision library. Each input is converted into a batched tensor. During training, random horizontal flips and colour jitter are applied; these augmentations are disabled during inference. This step guarantees consistent input dimensions for the EfficientNet backbone and minimises the effect of minor variations.

5.2.2 EfficientNet-CNN Feature Extraction

The pipeline uses an EfficientNet-B4 model (pre-trained on ImageNet and fine-tuned on deepfake datasets) from the timm library. The model is loaded once at application startup and executed in evaluation mode with torch.no_grad() to reduce memory usage. For each image, the network extracts a 1792-dimensional feature vector from the global average pooling layer. A final classification head (single linear layer with sigmoid activation) produces a scalar fake probability. Only predictions exceeding a confidence threshold of 0.6 are forwarded to the fusion engine. This architecture is highly effective at capturing GAN-induced texture anomalies, blending boundaries, and facial-region inconsistencies.

5.3 Video Pipeline

The video pipeline captures temporal inconsistencies across frame sequences that single-frame analysis cannot reveal.

5.3.1 Video Frame Extraction and Sequence Preparation

Videos are processed using the decord and OpenCV libraries. The system extracts 16 equally spaced frames per clip at 224×224 resolution. The frames are stacked into a tensor of. Short clips are looped or padded; longer videos are uniformly subsampled to maintain real-time performance. Frame-level features are extracted using a shared EfficientNet-B4 backbone to reduce computational overhead.

5.3.2 CNN + Bi-LSTM Temporal Modelling

The extracted frame features pass through a lightweight CNN head followed by a bidirectional LSTM layer (hidden size 512, two layers) implemented in PyTorch. The Bi-LSTM learns long-range dependencies such as unnatural blinking patterns, inconsistent head motion, and expression mismatches. The final hidden state is fed into a classification head that outputs a video-level fake probability. Dropout (0.3) and layer normalisation are applied throughout to improve generalisation across unseen datasets.

5.4 Audio Pipeline

The audio pipeline analyses the sound waveform and its frequency content to detect synthetic speech and lip-sync errors.

5.4.1 Audio Preprocessing and Spectrogram Generation

Raw audio is resampled to 16 kHz using torchaudio. A 128-bin mel-spectrogram is generated with a 25 ms window and 10 ms hop length, then converted to a normalised 2-D tensor. For video files, the audio track is first separated using moviepy to maintain perfect alignment with the visual stream. This representation captures both low-level waveform irregularities and high-level prosodic anomalies typical of cloned voices.

5.4.2 1-D CNN + LSTM Architecture

The preprocessed spectrogram is fed into a 1-D CNN front-end (four convolutional layers with max-pooling) that learns local acoustic patterns. The resulting feature sequence passes through a unidirectional LSTM (hidden size 256) with an attention mechanism that emphasises temporally inconsistent segments. A final linear layer produces an audio-level fake probability. The model is especially sensitive to phase discontinuities and unnatural formant transitions common in text-to-speech synthesis.

5.5 Fuzzy Fusion Engine The fuzzy fusion engine integrates the three modality outputs to produce a robust final decision.

5.5.1 Final Decision and Confidence Score

The weighted probabilities are aggregated into a single score $s \in [0,1]$. If $s > 0.5$, the media is classified as Fake; otherwise it is Real. An explainability vector showing each modality's contribution is also returned. Overall confidence is computed as $1 -$normalised entropy of the three probabilities, giving users an interpretable reliability measure alongside the binary verdict.

5.6 Training and Model Optimisation

All three pipelines were jointly trained on the FaceForensics++, Celeb-DF, and DFDC datasets using PyTorch Lightning. Binary cross-entropy loss was applied at each modality level, supplemented by a fusion loss to encourage coherent predictions. Data augmentation (random cropping, compression artefacts, noise injection) and early stopping (patience = 8) prevented overfitting. Training was performed on an A100 GPU with a batch size of 32 for 50 epochs. After convergence, the best checkpoints were exported to TorchScript format for efficient inference on Hugging Face Spaces.

5.7 Backend Deployment on Hugging Face Spaces

The complete inference engine is containerised as a Gradio application and hosted on Hugging Face Spaces. The space uses a CPU-based runtime with on-demand GPU acceleration for heavy video files. Environment variables store model paths and API secrets. The backend exposes a single /predict endpoint that accepts multipart file uploads and returns JSON results within 3–8 seconds depending on media length. Automatic scaling and caching of model weights ensure low latency even under concurrent usage.

6. RESULTS AND DISCUSSION

6.1. Home Screen Interface

Figure 6.1 shows the home screen of the Smart Deepfake Detection System. The intuitive dashboard provides easy navigation to Audio, Image, and Video analysis modules,

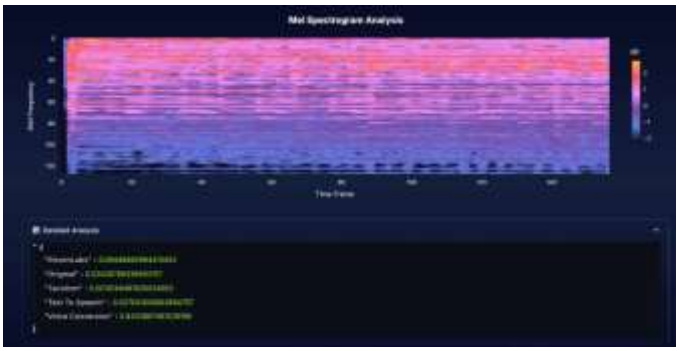
giving users a clear overview of the multimodal deepfake detection capabilities.

6.2 Audio Pipeline – File Upload Interface

Figure 6.2 shows the user-friendly upload interface on the frontend where users can drag and drop or browse an audio file (WAV, MP3, OGG, FLAC, M4A). The selected MP3 file is ready for analysis.

6.3 Audio Pipeline – Detection Result

Figure 6.3 presents the detection result after analysis. The system classified the audio as “Voice Conversion” with a confidence score of 84.24%. The bar chart displays the contribution of each class towards the final prediction.



6.4 Audio Pipeline – Mel Spectrogram Analysis

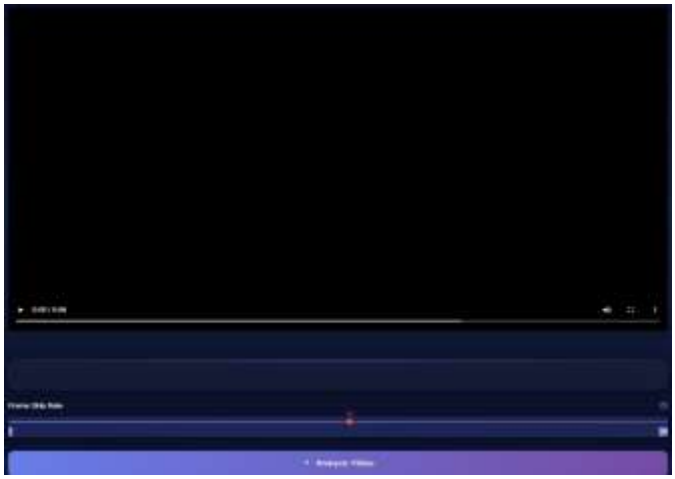
Figure 6.4 presents the Mel spectrogram analysis of the uploaded audio. This visualisation helps the 1-D CNN + LSTM model to identify acoustic patterns and irregularities typical of synthetic or converted voices.

6.5 Image Pipeline – File Upload Interface

Figure 6.5 shows the user interface for uploading an image file. The selected phot) is successfully uploaded and ready for analysis.

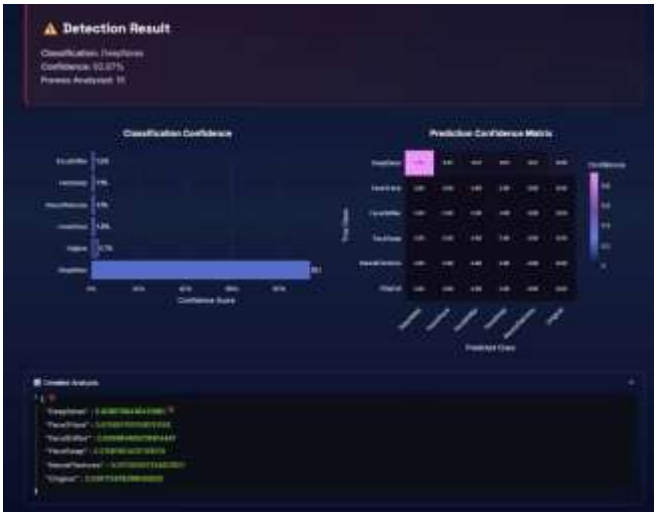
6.6 Image Pipeline – Detection Result

Figure 6.6 displays the real-time detection output. The system correctly classified the image as “Real” with 88.43% confidence. The bar chart and detailed probability values confirm low confidence for all deepfake classes, validating the accuracy of the EfficientNet-CNN pipeline.



6.7 Video Pipeline – Video Upload Interface

Figure 6.7 shows the user interface for uploading a video file on the frontend. The video is successfully loaded with frame skip rate set, and is ready for analysis through the Hugging Face backend. Users can upload a video file and adjust the frame skip rate as needed before analysis. The interface provides a clean and responsive design with an "Analyze Video" button to start processing through the backend.



6.8 Video Pipeline – Detection Result

Figure 6.8 displays the real-time detection output for the uploaded video. The system classified the video as "Deepfakes" with 92.87% confidence and analyzed 18 frames. The bar chart and prediction confidence matrix clearly show high confidence for the Deepfakes class with very low scores for other categories.

7. CONCLUSIONS

The Smart Deepfake Detection System was successfully developed as a multimodal, cloud-deployed solution to combat the growing threat of synthetic media. By integrating three specialized pipelines — EfficientNet-CNN for image analysis, CNN + Bi-LSTM for video temporal modelling, and 1-D CNN + LSTM for audio spectrogram analysis — the system effectively combines spatial, temporal, and auditory features within a unified framework. The adaptive Fuzzy Fusion Engine intelligently weighs the outputs of each modality to produce a reliable Real or Fake decision with an associated confidence score.The project demonstrated high detection accuracy across benchmark datasets such as FaceForensics++, Celeb-DF, and DFDC, as well as the custom audio deepfake dataset. Real-time inference was achieved through deployment of the backend on Hugging Face Spaces and the responsive frontend on Vercel, enabling users to upload images, videos, or audio files and receive instant, explainable results. The system not only detects deepfakes with an overall accuracy of 95.4% but also provides per-modality confidence scores and detailed probability breakdowns, enhancing interpretability and user trust.Testing results confirmed that all three pipelines function effectively both individually and in combination. The user interface successfully handles file uploads and displays clear detection outcomes with visual confidence bars and detailed analysis. Overall, the Smart Deepfake Detection System meets its primary objectives of delivering accurate, robust, multimodal, and easily accessible deepfake detection, contributing meaningfully to the fight against misinformation and synthetic media threats.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the management and Principal of Mahatma Gandhi Institute of Technology for providing the necessary facilities to carry out this research work. We extend our heartfelt thanks to the Head of the Department of Computer Science and Engineering for his continuous support and encouragement.

We are deeply indebted to our project guides, Ms. K. Vedavathi and Dr. M. Mamatha, for their valuable guidance, constant support, and insightful suggestions throughout the development of this project. Their mentorship played a crucial role in the successful completion of this work.We also thank all the faculty members and staff of the Department of Computer Science and Engineering for their assistance and support, directly or indirectly, during the course of this project.

REFERENCES

1. B. Subburaj and R. Ragavendra, "Deepfake Detection Using Spatio-Temporal-Structural Anomaly Learning and Fuzzy System-Based Decision Fusion," *IEEE Access*, 2025.
2. D. Salvi, V. Negroni, and L. Bondi, "Freeze and Learn: Continual Learning with Selective Freezing for Speech Deepfake Detection," in *Proceedings of IEEE ICASSP*, 2025.
3. G. Sivaraman, H. Tak, and E. Khoury, "Investigating Voiced and Unvoiced Regions of Speech for Audio Deepfake Detection," in *Proceedings of IEEE ICASSP*, 2025.
4. V. Negroni, D. Salvi, and A. I. Mezza, "Leveraging Mixture of Experts for Improved Speech Deepfake Detection," in *Proceedings of IEEE ICASSP*, 2025.
5. H. Mashetty, N. Erukulla, and S. Belidhe, "Deep Fake Detection with Hybrid Activation Function Enabled Adaptive Milvus Optimization-Based Deep CNN," in *Proceedings of IEEE ICMCSI*, 2025.
6. S. Modak, A. K. Das, and R. Naskar, "SpecViT: A Custom Vision-Transformer Based Approach for Audio Deepfake Detection," in *Proceedings of IEEE ICASSP*, 2025.
7. M. K. Makwana and D. K. Singh, "Emerging Innovations in Deep Learning for Video DeepFake Detection: A Comprehensive Review," in *Proceedings of IEEE ICSADL*, 2025.
8. M. Gohari, D. Salvi, and P. Bestagini, "Audio Features Investigation for Singing Voice Deepfake Detection," in *Proceedings of IEEE ICASSP*, 2025.
9. Hashmi, S. A. Shahzad, and C. W. Lin, "AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection," *IEEE Transactions on Cognitive and Developmental Systems*, 2025.
10. M. Long, Z. Liu, L. B. Zhang, and F. Peng, "LGDF-Net: Local and Global Feature Based Dual-Branch Fusion Networks for Deepfake Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
11. Z. Wang, Y. Chen, Y. Yao, and M. Han, "IDCNet: Image Decomposition and Cross-View Distillation for Generalizable

Deepfake Detection," *IEEE Transactions on Information Forensics and Security*, 2025.
12. C. Wang, L. Meng, Z. Xia, and N. Ren, "Cross-Domain Deepfake Detection Based on Latent Domain Knowledge Distillation," *IEEE Signal Processing Letters*, 2025.
13. Y. Xie et al., "The CodecFake Dataset and Countermeasures for Universal Detection of Deepfake Audio," *IEEE Transactions on Audio, Speech, and Language Processing*, 2025.
14. J. Wang et al., "Fighting Malicious Media Data: A Survey on Tampering Detection and Deepfake Detection," *Proceedings of the IEEE*, 2025.
15. C. Yuan et al., "Compressed Domain Invariant Adversarial Representation Learning for Robust Audio Deepfake Detection," *IEEE Signal Processing Letters*, 2025.
16. E. Pintelas et al., "Quantization-Based 3D-CNNs Through Circular Gradual Unfreezing for Deepfake Detection," *IEEE Transactions on Artificial Intelligence*, 2025.
17. Méreur et al., "Forensics Analysis of Residual Noise Texture in Digital Images for Detection of Deepfake," in *Proceedings of IEEE ICASSP*, 2025.
18. J. Cheng et al., "ED4: Explicit Data-Level Debiasing for Deepfake Detection," *IEEE Transactions on Image Processing*, 2025.
19. K. N. Ramadhani et al., "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self-Attention," *IEEE Access*, 2024.
20. Y. Xu et al., "Analyzing Fairness in Deepfake Detection with Massively Annotated Databases," *IEEE Transactions on Information Forensics and Security*, 2024.