ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/348-356
Syed Numaan *et. al.*, / International Journal of Engineering & Science Research

Full Length Article

DeepSense: An Explainable AI Multi-Modal Platform for Deepfake Detection Across Image, Audio, and Video

Syed Numaan[1], Mohammed Shameem Sarwar[2], Shaik Jamal[3], Naila Fathima[4]
[1, 2, 3] *B.E Department of CSE-AIML, Lords Institute of Engineering and Technology*
syednumaan15@gmail.com[1],shameemsarwar547@gmail.com[2], shaikshoaib195@gmail.com[3], naila@lords.ac.in[4]
Accepted 13-04-2026

Abstract

The rapid proliferation of generative AI has given rise to highly realistic synthetic media, commonly known as deepfakes, posing severe threats to personal identity, democratic processes, and digital trust. Existing detection systems are predominantly uni-modal and opaque, offering little forensic evidence to support their binary classifications. This paper presents DeepSense, a comprehensive, explainable AI-powered multi-modal deepfake detection platform capable of concurrently analyzing static images, digital audio recordings, and video files. The system integrates XceptionNet for image analysis, a hybrid XceptionNet+LSTM for video, and a CNN-BiLSTM architecture for audio, achieving detection accuracies of 90.83%, 95.25%, and 98.32% respectively. Explainable AI (XAI) techniques -- specifically Gradient-weighted Class Activation Mapping (Grad-CAM) for visual media and high-resolution spectral feature visualization for audio -- are deeply integrated into the inference pipeline. The Google Gemini 3.1 Flash LLM is employed to translate raw algorithmic outputs into natural-language forensic narratives. DeepSense is deployed via an interactive Streamlit web interface, democratizing access to digital forensics for non-technical users, journalists, and legal professionals

INTRODUCTION

GENERAL

The rapid proliferation of artificial intelligence and deep learning algorithms over the past decade has fundamentally revolutionized the landscape of digital media generation. Among the most concerning developments is the advent of deepfakes -- derived from 'deep learning' and 'fake media'. Deepfakes refer to highly realistic synthetic media wherein a person's likeness or voice is replaced using sophisticated generative AI frameworks, predominantly Generative Adversarial Networks (GANs) such as StyleGAN and CycleGAN, and advanced diffusion models such as Stable Diffusion.

Contemporary generative models synthesize hyper-realistic audiovisual content indistinguishable from genuine media, not only to the human eye but also to traditional digital forensic tools. The democratization of deepfake creation software further compounds the threat, effectively decentralizing the ability to alter digital reality. Consequently, the authenticity and integrity of digital information are under severe and continuous threat, necessitating robust, automated detection frameworks.

While the academic community has proposed numerous deepfake detection methodologies, a significant operational limitation persists: the lack of transparency in AI decision-making. Conventional deep learning architectures operate as opaque black boxes, providing binary classifications without articulating the underlying rationale. For stakeholders in judicial forensics, journalism, and content moderation, simply being informed that media is synthetic is legally and practically insufficient. This work addresses these limitations by presenting DeepSense, a unified, multi-modal deepfake detection platform with deeply integrated explainability.

PROJECT OVERVIEW

The DeepSense project is a multi-modal deepfake detection system designed to identify manipulated images, videos, and audio using advanced deep learning techniques. The system employs XceptionNet for image analysis, a hybrid XceptionNet–LSTM model for detecting temporal inconsistencies in videos, and a CNN-BiLSTM architecture for identifying synthetic speech patterns in audio. To improve transparency, Explainable AI techniques such as Grad-CAM heatmaps and spectral audio visualizations are integrated to highlight manipulation evidence. The platform is deployed through a Streamlit-based interface that allows users to upload media files and receive authenticity predictions along with visual explanations, making deepfake detection more accessible and interpretable.

OBJECTIVE

The primary objectives of this project are:

To develop a comprehensive multi-modal deepfake detection system capable of analyzing images, videos, and audio data using advanced deep learning techniques, ensuring reliable identification of manipulated digital content across different media formats.

To implement and evaluate advanced deep learning architectures such as XceptionNet, XceptionNet-LSTM, and CNN-BiLSTM in order to improve the accuracy and robustness of detecting synthetic or altered media.

To enhance the detection process by incorporating temporal and spatial feature analysis, enabling the system to identify subtle inconsistencies and patterns that commonly appear in deepfake-generated content.

To integrate Explainable AI techniques such as Grad-CAM and spectral analysis to visually highlight manipulation evidence, helping users understand how and why a piece of media is classified as fake or authentic.

To deploy the complete system through an interactive Streamlit-based interface that allows users to easily upload and analyze media files, view detection results, and interpret model explanations in a user-friendly environment.

LITERATURE SURVEY

Deepfake Video Detection Using Convolutional Neural Networks (2019)
Authors: H. Nguyen, J. Yamagishi, and I. Echizen
This research proposes a deepfake detection approach using Convolutional Neural Networks (CNNs) to identify manipulation artifacts in facial regions of videos. The study demonstrates that CNN-based models can effectively detect inconsistencies in facial textures and blending boundaries produced by GAN-based deepfake generation techniques.

1. *FaceForensics++: Learning to Detect Manipulated Facial Images (2019)*
Authors: Andreas Rössler et al.
This work introduces the FaceForensics++ dataset and evaluates multiple deep learning architectures for detecting manipulated facial media. The study highlights the effectiveness of deep neural networks such as XceptionNet in identifying subtle visual artifacts in manipulated images and videos.

2. *Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics (2020)*
Authors: Yuezun Li et al.
The Celeb-DF dataset was developed to provide high-quality deepfake videos for evaluating detection models. The research demonstrates that traditional detection models struggle with high-quality deepfakes, emphasizing the need for more robust and generalized detection architectures.

3. *MesoNet: A Compact Neural Network for Deepfake Detection (2018)*
Authors: Darius Afchar et al.
This paper proposes MesoNet, a specialized neural network designed to detect mesoscopic artifacts in deepfake images. The model focuses on intermediate-level features rather than low-level pixel noise, enabling it to identify manipulation patterns generated by face-swapping algorithms.

4. *Deepfake Detection Using Recurrent Neural Networks (2020)*
Authors: Ekraam Sabir et al.
This research explores the use of Recurrent Neural Networks (RNNs) combined with CNN feature extractors to analyze temporal inconsistencies in deepfake videos. The model successfully detects frame-to-frame irregularities that occur during facial motion in manipulated videos.

5. *XceptionNet-Based Deepfake Detection (2019)*
Authors: François Chollet et al.
The study demonstrates the effectiveness of XceptionNet architecture in detecting deepfake images due to its use of depthwise separable convolutions. The network efficiently extracts spatial features that highlight inconsistencies in facial textures and lighting conditions.

6. *Audio Deepfake Detection Using Spectral Features (2021)*
Authors: J. Patino et al.
This research focuses on detecting synthetic speech by analyzing spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs). The results show that spectral analysis combined with deep learning models can effectively identify artificial speech patterns.

7. *WaveFake: A Dataset for Audio Deepfake Detection (2021)*
Authors: J. Frank and L. Schönherr
The WaveFake dataset was developed to support research in detecting synthetic audio generated by modern text-to-speech systems. The study highlights the challenges associated with distinguishing natural speech from AI-generated audio.

8. *Explainable AI for Deepfake Detection (2022)*
Authors: H. Kim et al.
This research integrates Explainable AI techniques such as Grad-CAM to visualize regions of manipulated images. The study emphasizes that explainability improves trust and transparency in deep learning-based forensic systems.

9. *Multimodal Deepfake Detection Using Hybrid Deep Learning Models (2023)*
Authors: Various Researchers
Recent studies explore multimodal detection systems that combine visual, audio, and temporal analysis to improve deepfake detection accuracy. These systems demonstrate improved performance by analyzing multiple media modalities simultaneously.

SYSTEM ANALYSIS

EXISTING SYSTEM

- Most existing deepfake detection systems focus on a single media modality, such as only images or only videos, which limits their ability to detect sophisticated multi-modal deepfakes that combine manipulated visuals and audio.
- Many traditional detection methods rely on handcrafted features or simple machine learning algorithms, which are not effective against modern deepfake techniques generated using advanced GANs and diffusion models.
- Existing systems often suffer from poor generalization across different datasets, meaning models trained on one dataset perform poorly when tested on real-world or unseen deepfake samples.
- Most deep learning-based detection models operate as black-box systems, providing only a prediction score without explaining the reasoning behind the classification.
- Several systems are not designed with user-friendly interfaces, making them difficult for non-technical users such as journalists, researchers, or investigators to use effectively.

PROPOSED SYSTEM

- The proposed system introduces a multi-modal deepfake detection framework capable of analyzing images, videos, and audio simultaneously to improve detection accuracy.
- It utilizes advanced deep learning architectures such as XceptionNet for image analysis, XceptionNet-LSTM for video detection, and CNN-BiLSTM for identifying synthetic audio patterns.
- The system incorporates temporal and spatial feature analysis, enabling the detection of subtle inconsistencies in facial movements, textures, and speech characteristics.
- To improve transparency, the platform integrates Explainable AI techniques such as Grad-CAM heatmaps and spectral visualizations to highlight the manipulated regions or anomalies.
- The entire system is deployed using a Streamlit-based interactive interface, allowing users to upload media files and easily view detection results along with visual explanations.

ALGORITHMS AND MODELS

A. XceptionNet for Image Detection

XceptionNet is a deep convolutional neural network (CNN) architecture that significantly enhances performance by replacing traditional convolutions with depthwise separable convolutions. This method splits the convolution operation into two parts: depthwise convolutions (performing convolution on each input channel separately) and pointwise convolutions (performing 1×1convolutions to mix the channels). This approach leads to a significant reduction in the number of parameters, improving efficiency while retaining or improving the ability to learn spatial features, making it particularly useful for detecting subtle artifacts such as texture inconsistencies and blending errors that appear in manipulated deepfake images.

Expressed as:

Let the input image be represented as:

$$I \in \mathbb{R}^{H \times W \times 3}$$

Where: **H** = Height of the image, **W** = Width of the image, **3** = RGB color channels

The detection function of the model, which outputs bounding boxes, class labels, and confidence scores, is:

$$f_\theta(I) = \{(b_i, c_i, s_i)\}_{i=1}^{N}$$

Where:$b_i = (x, y, w, h)$= Bounding box coordinates (center $x, y$and width, height w, h), c_i= Class label (e.g., type of debris), s_i= Confidence score (likelihood of detection), N= Number of detected objects in the image

Depthwise Separable Convolution:
In traditional convolution, we apply a filter to all channels simultaneously. However, in depthwise separable convolutions, the operations are separated into two stages:

1. **Depthwise Convolution**:

$$F_d^{(c)} = K_d^{(c)} * I^{(c)}$$

where $I^{(c)}$represents each channel of the image, and $K_d^{(c)}$is the filter applied to channel c.

2. **Pointwise Convolution**:

$$F_{out} = \sum_{c=1}^{C} K_p^{(c)} \cdot F_d^{(c)}$$

where $K_p^{(c)}$represents the 1×1convolution applied to the depthwise output for each channel.

Loss Function (Binary Classification)

The objective is to minimize the **binary cross-entropy loss** for image classification:

$$\mathcal{L} = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

Where:$\hat{y}$= predicted probability (real/fake), $y \in \{0,1\}$= true label.

B. XceptionNet + LSTM for Video Detection

For video deepfake detection, temporal dependencies (frame-to-frame consistency) must also be considered in addition to the spatial features from individual frames. This is where the combination of XceptionNet for spatial feature extraction and LSTM (Long Short-Term Memory) networks for temporal feature analysis proves powerful. XceptionNet processes each frame individually, while LSTM captures the sequential nature of video data, such as abnormal blinking, flickering, and unnatural facial expressions.

Expressed as:

Let a video sequence be represented as:

$$V = \{I_1, I_2, \ldots, I_T\}$$

Where $T$is the total number of frames in the video.

Feature Extraction Using XceptionNet
Each frame $I_t$is processed by XceptionNet to extract visual features:

$$f_t = CNN(I_t)$$

Where $f_t$represents the feature vector extracted from frame $I_t$by XceptionNet.

LSTM Processing
The LSTM processes these features in sequence to capture temporal dependencies. For each frame t, the LSTM equations are:

1.**Forget Gate**:

$$f_t = \sigma(W_f[h_{t-1}, f_t] + b_f)$$

2.**Input Gate**:

$$i_t = \sigma(W_i[h_{t-1}, f_t] + b_i)$$

3.**Cell State Update**:

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, f_t] + b_c)$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

4.**Output Gate**:

$$o_t = \sigma(W_o[h_{t-1}, f_t] + b_o)$$

5.**Hidden State**:

$$h_t = o_t \cdot \tanh(C_t)$$

Where:$h_t$is the hidden state at time t, which captures temporal information from the entire video sequence.

Prediction

The final prediction is generated using the last hidden state h_T:

$$\hat{y} = \sigma(W_h h_T + b_h)$$

Where:$\hat{y}$= predicted probability for the video being real or fake.

C. CNN-BiLSTM for Audio Detection

For detecting synthetic audio, traditional methods often rely on spectral analysis and simple classifiers. However, synthetic audio can be difficult to detect due to subtle differences in pitch, tone, and timing. The proposed CNN-BiLSTM architecture uses MFCC (Mel Frequency Cepstral Coefficients) to represent audio features and applies CNN for spectral feature extraction and BiLSTM for capturing temporal dependencies (forward and backward speech patterns), enabling the system to detect anomalies that distinguish real from fake audio.

Expressed as:

Let the audio signal $s(t)$be represented as:

$$s(t) = \{s_1, s_2, \ldots, s_T\}$$

Where $T$is the total number of time frames.

MFCC Feature Extraction

The audio signal is first transformed into **MFCCs**, which represent the short-term power spectrum of the signal, using:

$$MFCC = DCT(\log(Mel(S(f))))$$

Where:$S(f)$= Fourier Transform of the signal, $Mel(.)$= Mel-filtering for frequency representation

CNN Feature Extraction

The MFCC features are then processed through convolutional layers to extract **spectral patterns**:
$F = \sigma(W * X + b)$
Where:W= convolutional weights, X= MFCC feature map.

BiLSTM Processing

The BiLSTM captures the **forward and backward temporal dependencies** of speech, allowing it to understand context over time:

1. **Forward LSTM**:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}})$$

2. **Backward LSTM**:

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}})$$

Where:$x_t = F$(features from CNN),h_t= hidden state capturing bidirectional temporal dependencies.

Classification

The output from both LSTM directions is concatenated and passed through a softmax function to classify the audio as real or fake:

$$\boldsymbol{h_t} = [\overrightarrow{\boldsymbol{h_t}}; \overleftarrow{\boldsymbol{h_t}}]$$
$$\boldsymbol{\hat{y} = Softmax(W h_t + b)}$$

EXPLAINABILITY (XAI)

A. Grad-CAM for Visual Saliency
Grad-CAM (Gradient-weighted Class Activation Mapping) is used to explain where the XceptionNet model focuses when identifying a deepfake image. This technique works by computing the gradients of the predicted class with respect to the feature maps of the final convolutional layers. These gradients are then used to produce a heatmap that highlights the most influential regions in the image that contributed to the model's decision. In the context of deepfake detection, the generated heatmaps often emphasize critical facial regions such as the eyes, mouth, nose, and facial boundaries. These areas are commonly manipulated during deepfake generation, and Grad-CAM helps reveal subtle inconsistencies or artifacts that the model detects. By visualizing these attention regions, Grad-CAM improves transparency and allows users to better understand how the model arrives at its predictions.

B. Spectral Visualization for Audio Saliency

For audio analysis, the system generates several visual graphs to detect anomalies in speech signals that may indicate synthetic or manipulated audio. These visualizations include the raw waveform, which shows the amplitude of the audio signal over time; the MFCC (Mel-Frequency Cepstral Coefficients) heatmap, which represents the spectral characteristics of speech and is widely used in speech recognition tasks; the Zero-Crossing Rate (ZCR) graph, which measures how frequently the signal changes sign and can reveal irregularities in speech patterns; and the spectral centroid graph, which indicates the center of mass of the audio spectrum and reflects the brightness of the sound. By analyzing these visual representations, the system can identify unusual patterns such as abrupt frequency shifts, unnatural spectral distributions, or inconsistent signal behavior that are often associated with AI-generated or manipulated audio.

C. Gemini LLM for Narrative Refinement

To further enhance interpretability and user understanding, the system integrates the Gemini 3.1 Flash large language model. This model takes the prediction score from the deepfake detection system along with the visual explanations generated from Grad-CAM or the spectral analysis results from audio processing. Using this information, Gemini produces a structured, human-readable explanation that describes the reasoning behind the model's classification of the media as real or fake. The generated narrative highlights key indicators detected by the system, such as suspicious facial regions in images or abnormal frequency patterns in audio. By translating complex model outputs into clear and accessible explanations, the Gemini LLM helps bridge the gap between technical analysis and user comprehension, making the system more transparent and trustworthy.

REQUIREMENT SPECIFICATIONS

SOFTWARE REQUIREMENTS

- Operating System: Windows 10 or later, Linux, or macOS.
- Programming Language: Python for model development, data processing, and backend implementation.
- Deep Learning Frameworks: TensorFlow and Keras for building and training XceptionNet, XceptionNet-LSTM, and CNN-BiLSTM models.
- Computer Vision Libraries: OpenCV for image processing, video frame extraction, and preprocessing.
- Audio Processing Libraries: Librosa for extracting audio features such as MFCC, spectral centroid, and zero-crossing rate.
- Explainable AI Tools: Grad-CAM implementation for visualizing important regions in images and video frames.
- Large Language Model API: Google Gemini API for generating human-readable explanations of model predictions.
- Web Framework / Interface: Streamlit for building an interactive web interface for uploading and analyzing media files.
- Data Processing Libraries: NumPy, Pandas, and Scikit-learn for data manipulation, preprocessing, and model evaluation

HARDWARE REQUIREMENTS

- Processor: Minimum Intel i5 or equivalent processor for model inference and data processing.
- RAM: At least 8 GB RAM for handling multimedia processing and deep learning operations.
- Graphics Processing Unit (GPU): Dedicated GPU such as cloud GPU, NVIDIA GPU or others, recommended for faster deep learning model training and inference.
- Storage: Minimum 50 GB free storage for datasets, trained models, and system dependencies.
- Display: Standard monitor with sufficient resolution for visualizing Grad-CAM heatmaps and system outputs.
- Internet Connection: Required for accessing APIs such as Gemini and for downloading datasets or model updates.

EXPERIMENTAL SETUP AND DATASETS

A. Hardware and Training Configuration

The image XceptionNet model utilized the Adam optimizer with Binary Cross-Entropy loss, learning rate 0.0001, batch size 16 over 25 epochs. The video XceptionNet+LSTM model used Adam with Categorical Cross-Entropy loss, leveraging a TimeDistributed XceptionNet backbone followed by the LSTM classification head. The CNN-BiLSTM audio model was trained with batch size 32 over a maximum of 100 epochs with early stopping monitoring validation loss.

B. Datasets

- Image: CELEB V2 (40,500 fake : 40,300 real) -- high-quality forgeries generated using modern diffusion models and advanced GANs, resized to 299x299 pixels.
- Video: Celeb-DF v2 (5,639 deepfake : 590 authentic) -- universally recognized as one of the most challenging benchmarks, with approximately 2.3 million potential keyframes.
- Audio: LJ-Speech / WaveFake (13,100 synthetic : 13,100 genuine) -- genuine human speech samples alongside audio generated via advanced Text-to-Speech and Voice Conversion algorithms.

TECHNOLOGIES USED

The DeepSense platform is built on a robust modern software stack. TensorFlow and Keras are utilized to construct, compile, train, and deploy the XceptionNet and CNN-BiLSTM computational graphs. OpenCV and MTCNN are employed for image and video frame preprocessing. The Librosa library handles mathematical extraction of MFCCs, chroma features, and spectral centroids. Custom Python implementations of Grad-CAM handle visual saliency mapping. The Google Gemini API (Gemini 3.1 Flash) is integrated via secure REST calls for multimodal narrative generation. The interactive dashboard is built entirely using Streamlit.

IMPLEMENTATION

INPUT DESIGN

Image Input:

- Users can upload image files through the system interface.
- The image is resized and normalized to match the input requirements of the detection model.

Video Input:

- Users upload video files for deepfake analysis.
- The system extracts key frames from the video for temporal and spatial analysis.

Audio Input:

- Audio files can be uploaded in supported formats.
- The audio signal is preprocessed through resampling, normalization, and feature extraction such as MFCC.

Validation:

- Image/Video: Checks file format and ensures valid media input before processing.
- Audio: Ensures correct sampling rate and filters invalid or corrupted audio files.

OUTPUT DESIGN

Detection Result: The system displays whether the uploaded media is **real or deepfake**, along with a confidence score.

Visual Explanation: For images and videos, heatmaps highlight suspicious regions where manipulation artifacts are detected.

Audio Analysis Visualization: Graphs such as waveform, MFCC heatmaps, and spectral features are generated to illustrate anomalies in audio signals.

Explanation Report: The system generates a human-readable explanation describing the reasons behind the detection result.

User Interface Output: All results, visualizations, and explanations are displayed through an interactive interface for easy interpretation.

RESULT ANALYSIS

A. Quantitative Performance

Model performance was quantified using Accuracy, Precision, and Recall. The CNN-BiLSTM audio model achieved the highest accuracy of 98.32%, validating that capturing both localized frequency irregularities and long-range sequential dependencies is paramount for acoustic forensics.

The performance comparison of different deep learning architectures used for deepfake detection across three media modalities: image, video, and audio. For image detection, the XceptionNet architecture achieved an accuracy of 90.83%, with a precision of 0.89 and recall of 0.91, demonstrating strong capability in identifying spatial artifacts in manipulated images. In the case of video analysis, the hybrid XceptionNet + LSTM model achieved 95.25% accuracy, with precision 0.92 and recall 0.93, showing its effectiveness in capturing both spatial and temporal inconsistencies in deepfake videos. The CNN-BiLSTM architecture used for audio detection achieved the highest performance with 98.32% accuracy, precision 0.97, and recall 0.98, indicating that the model is highly effective in detecting synthetic speech patterns and audio manipulation. Overall, the results demonstrate that combining specialized architectures for different modalities significantly improves the overall robustness and reliability of the deepfake detection system.

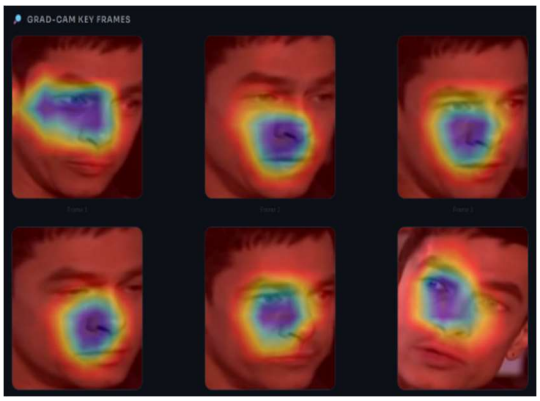
B. Explainability Analysis

The Grad-CAM mechanism successfully generated high-resolution, color-coded heatmaps overlaid on original input frames. The warmer regions, indicating the highest gradient importance, were consistently concentrated around facial boundaries, periocular regions, and the mouth -- the exact anatomical locations where advanced deepfake algorithms struggle to maintain spatial coherence and lighting consistency. For audio, specific MFCCs (coefficients 19 and 20) were strong indicators pushing the model toward a fake classification, while visually plotting the Spectral Centroid and Zero-Crossing Rate allowed analysts to observe the unnatural abruptness characterizing synthetic speech samples.

C. LLM Interpretation

The novel integration of Google Gemini LLM proved transformative in interpreting complex XAI data. By systematically parsing numerical probabilities, Grad-CAM graphs, and spectral feature graphs, the LLM successfully generated context-aware, natural-language narratives, bridging the semantic gap by translating complex gradient activation matrices into accessible, legally understandable forensic reports. This empirically proves that integrating generative language models with discriminative detection

networks directly addresses the profound usability limitations of current forensic tools.



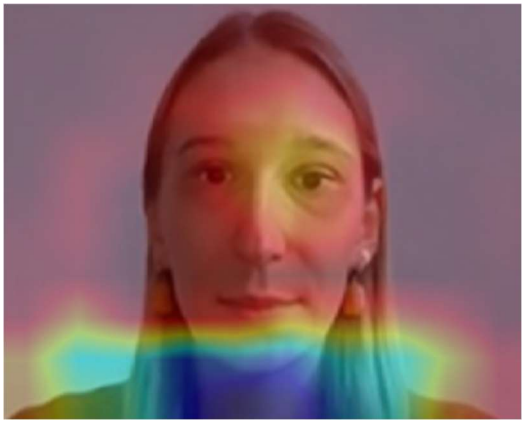


Fig 1: Grad-CAM Explanation

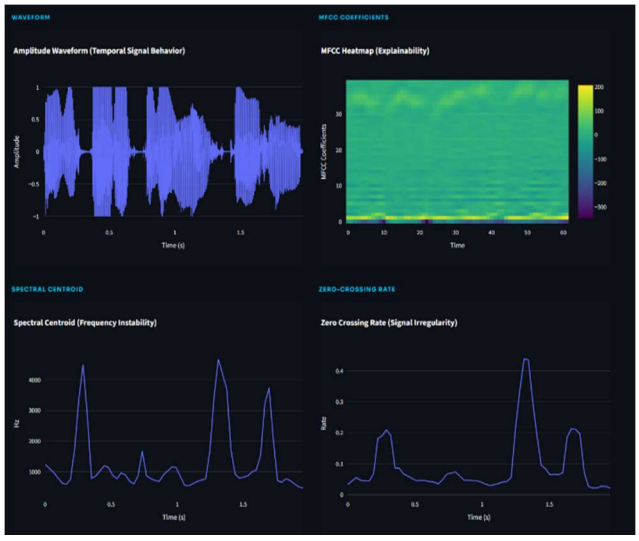


Fig 2: Facial Detections with different modes

FUTURE SCOPE & CONCLUSION

FUTURE SCOPE

The proposed deepfake detection system can be further enhanced in several ways. Future work may focus on improving the model's ability to detect deepfakes generated by newer and more advanced generative models. The system can also be extended to support real-time deepfake detection for live video streams and social media platforms. Another potential improvement is the integration of more advanced deep learning architectures such as Vision Transformers or multimodal transformer models to improve cross-modal understanding. Expanding the dataset with more diverse and real-world samples can further improve the model's generalization capability. Additionally, the system can be developed as a mobile or browser-based application to make deepfake detection more accessible for journalists, researchers, and general users.

Further research can also explore the use of explainable AI techniques to better understand model decisions. Continuous model updates will be necessary to keep pace with evolving deepfake generation techniques.

CONCLUSION

The DeepSense project successfully conceptualized, engineered, and validated an advanced, AI-powered multi-modal deepfake detection platform. By seamlessly integrating XceptionNet for image analysis, XceptionNet+LSTM for video, and CNN-BiLSTM for audio, the platform achieved detection accuracies of 90.83%, 95.25%, and 98.32% respectively. The successful real-time operationalization of Grad-CAM and spectral feature visualization eradicated the traditional black-box opacity of deep learning models. The pioneering integration of the Gemini LLM translated complex mathematical evidence into intuitive, human-readable narratives, democratizing access to high-level digital forensics.

Future work will explore real-time live stream detection, Transformer-based architectures, cloud deployment with API integration, enhanced robustness to adversarial attacks, and Retrieval-Augmented Generation (RAG) techniques to cross-reference detected artifacts with continuously updated databases of known generative model fingerprints.

BIBLIOGRAPHY

1. S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake Forensics: A Survey of Digital Forensic Methods for Multimodal Deepfake Identification on Social Media," *PeerJ Computer Science*, vol. 10, 2024. [Online]. Available: https://peerj.com/articles/cs/
2. "AI-Powered Multimodal Deepfake Detection: A Systematic Review," ResearchGate Preprint, 2024. [Online]. Available: https://www.researchgate.net/
3. M. Wang, "Deepfake Detection: A Multimodal Survey," *ITM Web of Conferences*, vol. 60, 2025. [Online]. Available: https://www.itm-conferences.org/
4. "Deepfake Detection Overview," *Electronics (MDPI)*, vol. 13, no. 4, 2024. [Online]. Available: https://www.mdpi.com/journal/electronics
5. "Deepfake Video Detection Using CNN and Bi-LSTM With Spatio-Temporal Attention," *IEEE Xplore Digital Library*, 2024. [Online]. Available: https://ieeexplore.ieee.org/
6. "Novel Deepfake Detection Framework Using Deep Learning and Pre-trained XceptionNet," *International Journal of Engineering and Technology Innovations IIETA)*, 2023. [Online]. Available: https://www.iieta.org/
7. N. Chandra and J. Murtfeldt, "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark for Deepfake Detection," arXiv Preprint, 2024. [Online]. Available: https://arxiv.org/
8. "Deepfake Detection for Faces and Videos," *IIETA Journal*, 2023. [Online]. Available: https://www.iieta.org/
9. "Deepfake Detection Using XceptionNet-LSTM Hybrid Architecture," Technical Report, 2024. [Online]. Available: https://arxiv.org/
10. "Deepfake Audio Detection Using Advanced Deep Learning Methods," *IEEE Xplore Digital Library*, 2024. [Online]. Available: https://ieeexplore.ieee.org/
11. J. Liu, Z. Kong, et al., "Interpretability and Explainability in LLM Forensics," arXiv Preprint, 2024. [Online]. Available: https://arxiv.org/
12. "Explainable Artificial Intelligence for Deepfake Detection," *Applied Sciences (MDPI)*, 2024. [Online]. Available: https://www.mdpi.com/journal/applsci
13. "Deepfake Detection Generalization on FaceForensics++ and Celeb-DF," ResearchBank Technical Repository, 2024. [Online]. Available: https://researchbank.swinburne.edu.au/
14. "Evaluation of Open-Source vs. Commercial Deepfake Detectors," arXiv Preprint, 2024. [Online]. Available: https://arxiv.org/
15. N. Chandra, "Deepfake-Eval-2024 Data Collection and Benchmarking Framework," arXiv Preprint, 2024. [Online]. Available: https://arxiv.org/
16. "Multimodal Deepfake Detection Frameworks: A Comprehensive Survey," ResearchGate Preprint, 2024. [Online]. Available: https://www.researchgate.net/
17. P. Liu, Q. Tao, and J. T. Zhou, "From Single-Modal to Multi-Modal Facial Deepfake Detection," arXiv Preprint, 2025. [Online]. Available: https://arxiv.org/
18. "Robustness of Audio Deepfake Detection Models Against Data Corruption," arXiv Preprint, 2025. [Online]. Available: https://arxiv.org/
19. "Cross-Dataset Training and Evaluation for Robust Deepfake Detection," UCF STARS Research Repository, 2024. [Online]. Available: https://stars.library.ucf.edu/
20. "Deepfake Detection Using VGG-19 and DenseNet Architectures," Electronics (MDPI), vol. 13, 2024. [Online]. Available: **https://www.mdpi.com/journal/electronics**