Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00052-8

Original Article

Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation

Sanjana Shetty [1], Ketaki Sakhadeo [1], Tejashree Deore [1], Shehnaz Siddique [1]

1. *Computer Science and Technology, Usha Mittal Institute of Technology, Mumbai, IND*

Received: March 12, 2026 | Review began: March 19, 2026 | Review ended: March 31, 2026 | Published: April 07, 2026

Abstract

The rapid spread of deepfake content on social networks, video conferencing platforms, and voice communication systems means we need to find ways to detect it that are fast and work well. This paper presents a lightweight multimodal deepfake detection framework designed for real-time deployment under resource-constrained environments. The system integrates a hybrid architecture combining ResNet18-based convolutional neural networks for spatial feature extraction, EfficientNet for frame-level video analysis, and Wav2Vec2 for audio representation learning. We use these tools to get information from each type of media and then combine them to make a decision.

We tested our system using some datasets like FaceForensics++ Celeb-DF and ASVspoof 2019. Experimental results demonstrate an accuracy of 88.25% for image-based detection, 70.56% for video frame analysis, and 81.50% for audio classification under CPU-only deployment. The system achieves real-time performance with low latency and reduced computational overhead, making it suitable for practical applications.

Our approach provides an effective trade-off between detection accuracy and computational efficiency, enabling deployment in real-world scenarios such as social media content moderation, secure video conferencing, and voice phishing prevention.

Categories: Explainable AI, Computer Vision, Deep Learning

Keywords: audio forensics, deepfake detection, lightweight models, multimodal learning, real-time ai, video forensics

Introduction

The problem with deepfake technology is that it is making it hard for us to trust what we see and hear online [1]. Deepfakes are made using computer programs called generative adversarial networks and diffusion-based models [2,3]. These programs manipulate facial expressions, lip movements, and speech patterns [1,4]. Deepfake technology poses significant risks, including misinformation dissemination, political manipulation, financial fraud, and identity theft.

Existing deepfake detection methods usually look at either what we can see or what we can hear. They often use big computer models like XceptionNet or large Vision Transformers [5,6]. These models are good at figuring out what is real and what is not when everything is perfect [7]. They are not good for using in systems that need to work fast because they need a lot of computer power and memory [2]. Deepfake detection methods that use one way of checking, like looking at pictures or listening to sound, are not very good at finding fake things [6,8]. These methods focus on only one type of signal, making them easier to deceive by attackers who design manipulations specifically targeting that modality

[6]. Deepfake detection is not reliable when it only uses one kind of signal [9]. Furthermore, single-modality detection systems are highly vulnerable to adversarial attacks that exploit weaknesses unique to the selected modality, significantly reducing their effectiveness in real-world scenarios where deepfake generation techniques continue to evolve [8].

To address these challenges, this work introduces a multimodal deepfake detection system that combines visual frame analysis, temporal consistency, and audio cues [10-12]. The framework is built with practical deployment in mind, allowing it to run in real time on consumer-grade hardware while remaining robust against a wide range of deepfake techniques.

The key contributions of this paper are as follows:

- A unified framework that brings together image, video, and audio information for more reliable deepfake detection.
- An effective frame sampling and preprocessing approach that reduces unnecessary data while keeping important visual and temporal details.
- An attention-based fusion strategy that combines information from different modalities to improve overall detection performance.
- A detailed experimental evaluation that highlights how the system balances detection accuracy with computational efficiency in resource-constrained environments.

Materials And Methods

Threat model and security considerations

The proposed framework exists in a realistic adversarial environment, which means that the adversary aims to produce crafted multimedia content that may evade human evaluation as well as verification procedures. The adversary possesses publicly available deepfake generation tools that utilize generative adversarial networks, diffusion models, or voice cloning techniques [12].

The potential objectives of the attack could be identity impersonations, political misinformation, biometric spoofing, social engineering, financial fraud, among others [13]. In addition, the attacker could modify one or more modality types, including video facial synthesis, video temporal frame rate consistency, or generated speech waveform synthesis using neural text-to-speech systems [8].

The objective here is to detect cross-modal inconsistencies and forensic artifacts in real time without access to the original source media. This system is developed to be effective in an open environment where post-generation detection capabilities are ever-evolving.

The framework assumes no privileged access to watermarking and signatures, and therefore it is inclined towards detection-based mitigation techniques rather than prevention-based techniques.

Proposed methodology

System Architecture

The proposed system consists of independent modality-specific pipelines in image, video, and audio that operate in parallel and are integrated through a fusion mechanism. The results of each modality are then combined to make a strong decision. This means that each part can be made to work well on its own and used only when it is needed, depending on what the platform can handle. The system has four pipelines: (I) image-based detection, (II) video-based frame-level detection, (III) audio-based deepfake detection with explainability, and (IV) multi-modal feature fusion.

Let input sample be

$$X = \{I, V, A\}$$

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

where *I*, *V*, and *A* represent image input, video input, and audio waveform, respectively.

The final deepfake prediction is

$$\hat{y} = \Phi(F_I, F_V, F_A)$$

where F_I, F_V, and F_A denote feature representations extracted from image, video, and audio modalities, respectively, and Φ represents the fusion and classification function.

The proposed design is different from models that do everything. It focuses on being easy to understand using computer power wisely and being able to add things. Each part of the process makes its judgment about how sure it is and creates its own special features. Then, these are all combined by a part to make the final decision, about what something is. The proposed design emphasizes interpretability, computational efficiency, and modular extensibility.

Recent studies show that it is important to develop deepfake detection models that are not too heavy. These models need to operate on devices such as smartphones and other small devices. By simplifying the models, they can run faster and use less memory, which is desirable. This makes them suitable for real-world applications [5,13]. The overall system architecture is shown in Figure *1*.

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

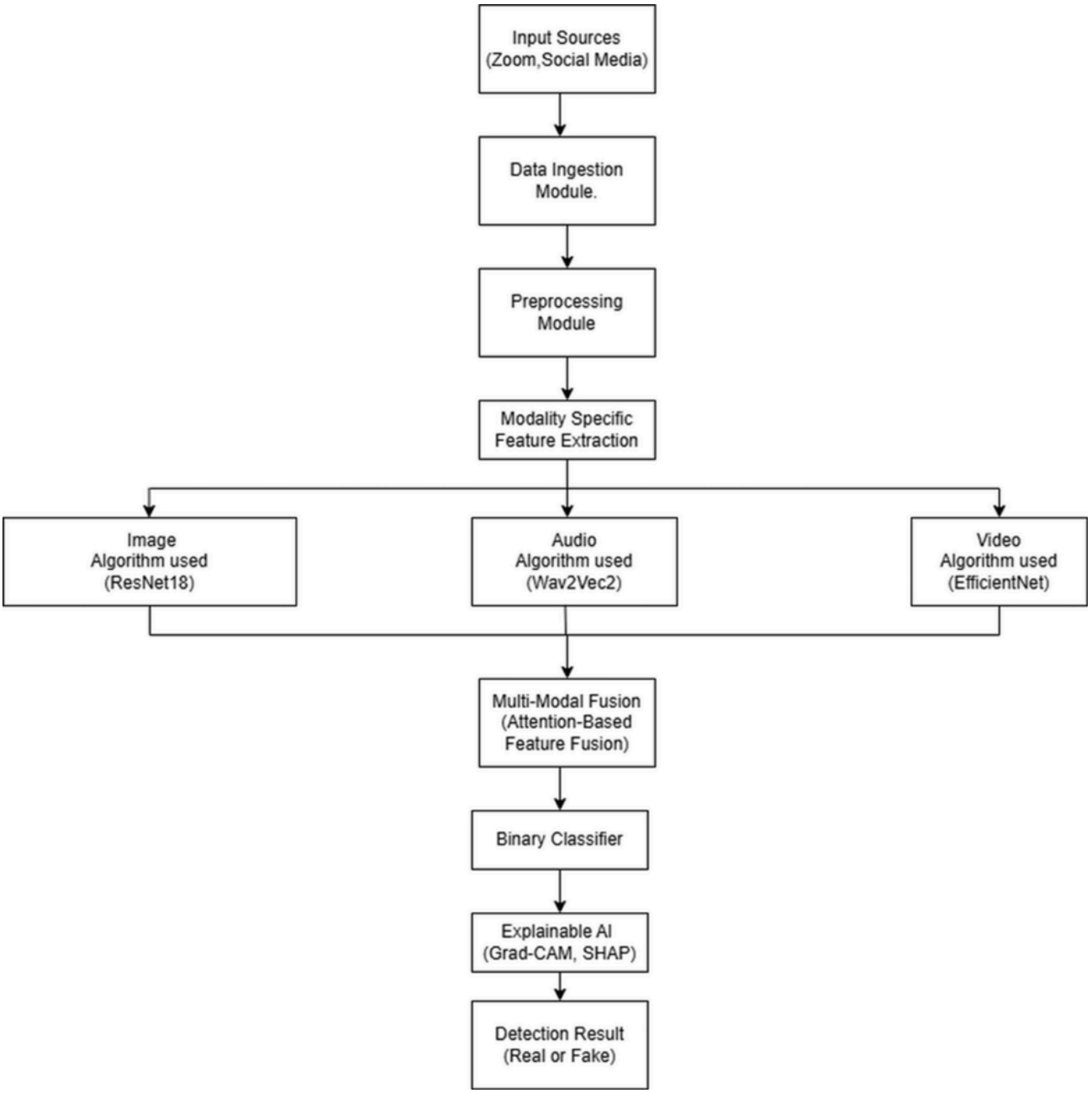

FIGURE 1: Proposed multi-modal deepfake detection architecture

Grad-CAM, Gradient-weighted Class Activation Mapping; ResNet, Residual Network; SHAP, Shapley Additive Explanations

Image-Based Deepfake Detection Using ResNet

For static image analysis, a Residual Network (ResNet) architecture is employed due to its proven effectiveness in visual forensics and its ability to mitigate vanishing gradient issues in deep networks [3]. Residual connections enable the network to learn subtle manipulation artifacts such as abnormal texture patterns, color inconsistencies, and facial boundary distortions commonly introduced during image-based deepfake generation [7].

Input images are resized and normalized before being passed through the ResNet backbone. Feature maps extracted from the final convolutional layers are globally averaged and fed into a lightweight, fully connected classifier. Global average pooling effectively captures spatial manipulation cues such as texture irregularities and boundary distortions while reducing model complexity [5].

Image feature extraction:

$$F_I = \mathbf{GAP}(\mathbf{ResNet}(I))$$

where GAP denotes global average pooling, and ResNet represents the residual network used for feature extraction.

Final image classification:

$$y_I = \mathbf{Softmax}(W_I F_I + b_I)$$

where W_I and b_I are learnable parameters of the classifier, and Softmax converts the output into class probabilities.

Although ResNet demonstrates strong generalization across datasets, visual-based detectors may suffer performance degradation under heavy compression, noise, or aggressive post-processing that suppresses forensic artifacts [2,7].

Video Deepfake Detection Using EfficientNet-Based Frame Analysis

Video-based detection is performed using a frame-level analysis strategy built upon an EfficientNet backbone, which provides improved accuracy with fewer parameters through compound scaling of network width, depth, and resolution [4].

To reduce computational overhead, a uniform frame sampling strategy extracts representative frames from videos instead of processing every frame. Frame-level analysis significantly reduces processing time while preserving temporal diversity necessary for identifying manipulation artifacts [5,13].

Each sampled frame is independently evaluated by the EfficientNet model, producing frame-level predictions. Statistical aggregation techniques such as mean and variance pooling are used to derive video-level predictions. Variance-based analysis helps identify temporal inconsistencies caused by deepfake generation processes [5].

Frame sampling:

$$\{f_1, f_2, \ldots, f_k\} = \mathbf{Sample}(V)$$

Frame feature extraction:

$$F_{v_i} = \mathbf{EfficientNet}(f_i)$$

Video aggregation: $F_V = (1/k)\Sigma_{i=1}^{k}(F_{v_i} + \sigma(F_{v_i}))$

where F_{v_i} represents features extracted from the i^{th} frame, and σ denotes variance capturing temporal inconsistencies across frames.

Frame-based video analysis is widely adopted for real-time deepfake detection systems due to its ability to balance computational efficiency and temporal artifact detection [13]. However, high-quality deepfakes often maintain strong temporal coherence, which reduces detectable inconsistencies and presents ongoing challenges for detection frameworks [7].

Audio-Based Deepfake Detection Using Wav2Vec2

Audio deepfake detection in the proposed framework is done using the Wav2Vec2 model. This model teaches itself to understand speech. It does not use ways of taking out features like Mel Frequency Cepstral Coefficients. Instead, Wav2Vec2 learns what speech sounds like from the audio sounds. This helps it find problems in fake speech that are hard

to hear. Audio deepfake detection is better because of this. The Wav2Vec2 model is effective at finding these problems, in audio deepfakes.

The Wav2Vec2 model uses a kind of learning called contrastive self-supervised learning. This helps the Wav2Vec2 model understand speech better by looking at how sounds change over time in signals. The Wav2Vec2 model can find mistakes in how sounds go the way someone speaks and who is speaking. These mistakes often happen when computers try to make voices sound real or change one voice to sound like another. Other studies have shown that models like Wav2Vec2 are good at figuring out if a voice is fake or not. They do a better job than other methods that use special audio features made by people.

The proposed system processes audio inputs by applying normalization and segmentation, followed by feature extraction using the Wav2Vec2 encoder. This is done by making sure the volume is the same and the speed is right. The Wav2Vec2 encoder helps us get a sense of what is in the audio. We use these findings to figure out if the audio is real or not. The audio signal is checked to see if it is authentic.

Wav2Vec2 is good for finding deepfakes because it looks at the speech in two ways. It checks for problems in the sound that happen quickly. It also checks for problems that happen over a longer period of time. This means the system can find speech that might not be found using other methods. Also, Wav2Vec2 models that have been trained already work well with different types of recordings and environments. Wav2Vec2 is very useful for deepfake detection.

Wav2Vec2 has some things about it but it also makes things more complicated when it comes to computing. This is compared to the way of extracting audio features by hand. To make sure Wav2Vec2 can still be used in time, the framework uses special ways to make inference faster and simpler classification layers that do not use a lot of resources. Wav2Vec2 needs these methods to work properly.

$$F_A = \mathbf{Classifier}(\mathbf{Wav2Vec2}(A))$$

where *A* represents the raw audio waveform, and Wav2Vec2 is a pretrained speech representation model used to extract audio features.

Audio classification

$$y_A = \mathbf{Softmax}(W_A F_A + b_A)$$

where W_A and b_A are trainable parameters, and Softmax outputs class probabilities.

Multi-Modal Fusion and Classification

The fusion module integrates visual, temporal, and audio features using an attention-based aggregation strategy. Multi-modal fusion enables cross-modal correlation analysis, which improves detection robustness by identifying inconsistencies between audio and visual signals [10].

Feature concatenation:

$$F_{concat} = [F_I; F_V; F_A]$$

where [;] denotes feature concatenation across modalities.

Attention fusion:

$$F = \mathbf{Softmax}\left(\frac{Q(K_V + K_A)^T}{\sqrt{d}}\right)V$$

where Q, K_V, and K_A represent query and key matrices, and *d* is the scaling factor.

Attention mechanisms dynamically assign importance weights to modality-specific features, allowing the system to prioritize the most reliable modality under varying conditions. The fused feature representation is passed to a binary classifier trained using cross-entropy loss for final deepfake classification [5].

Binary classification using cross entropy:

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

where *y* is the ground truth label, and $\hat{y}$ is the predicted probability.

Multimodal integration significantly improves detection performance by leveraging complementary information from different data modalities, which is better than using one kind of data. The reason is that we can get information from each type of data that the others do not have [10].

Explainable AI (XAI) Integration

Explainability plays an important role in security applications. It is essential to understand the decisions that the model makes such as digital forensics and misinformation analysis [9,13].

For image and video pipelines, Gradient-weighted Class Activation Mapping (Grad-CAM) is used to generate saliency heatmaps highlighting manipulated facial regions. Grad-CAM enables visualization of spatial features that influence classification decisions, allowing analysts to verify whether the model focuses on meaningful forensic artifacts [10].

For fusion-level interpretation, Shapley Additive Explanations (SHAP)-based feature attribution methods are employed to evaluate modality contributions toward classification decisions. SHAP provides quantitative explanations by estimating feature importance across modalities, enhancing transparency in multimodal decision-making [10].

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where α_k^c represents the importance weight of feature map *k*, and A_{ij}^k is the activation map

Recent studies show that making systems explain themselves in ways helps people check the work of computers. This makes people trust computers more. It also helps fix problems with the systems and makes sure they follow the rules. Systems like these are very important, for things that need to be explained like model debugging and regulatory compliance [9,11].

However, incorporating XAI introduces additional computational overhead, requiring careful optimization to maintain real-time performance constraints [14].

Methodology

Figure *2* illustrates the complete training and inference workflow of the proposed system, including preprocessing, feature extraction, fusion, explainability, and evaluation. The framework processes three modalities: image, video, and audio. Each modality undergoes independent preprocessing and feature extraction. The outputs from preprocessing are directly fed into their respective feature extraction models, and the extracted features are fused using an attention-based mechanism for final classification.

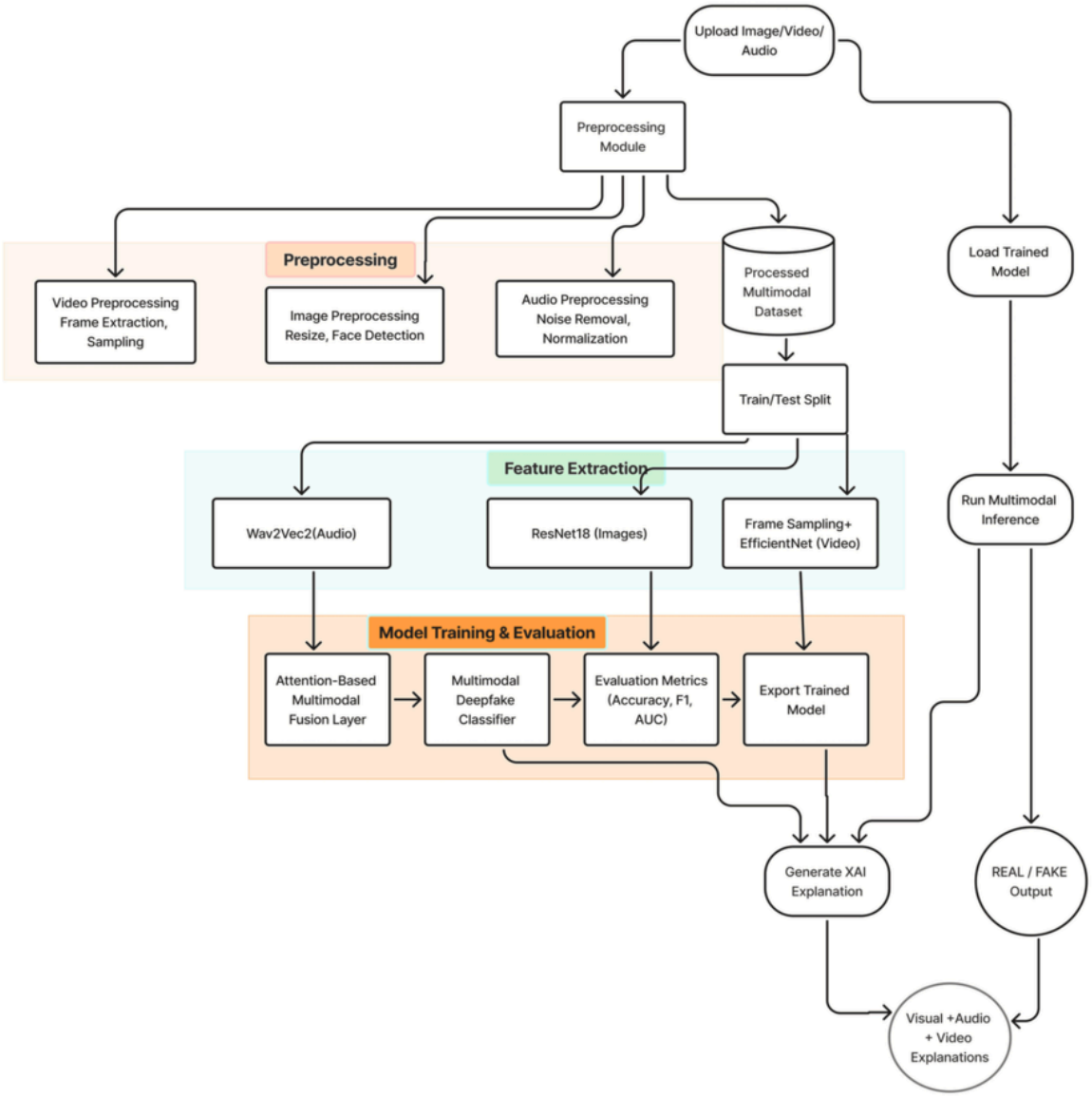

FIGURE 2: Training and inference workflow of the proposed deepfake detection framework

AUC, Area under the Curve; XAI, Explainable Artificial Intelligence

Datasets and Data Preparation

We evaluate the proposed multi-modal deepfake detection system using a set of publicly available datasets including real and manipulated images, videos, and audio samples [2,8,12]. These datasets were selected as they represent real-world scenarios that can be found on social media websites and communications systems online [1,7]. Subject identities, poses, facial expressions, lighting conditions, video durations, audio quality, and compression artifacts are largely uncontrolled.

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

For the video mode, our dataset includes authentic and deepfake videos with a few hundred to thousands of frames from short clips up to longer ones. This large variability has potential implications for real-time inference, especially in terms of the computationally expensive aspect thereof. This is achieved through the splitting of the dataset into respective sections for training, validation, and testing, with fixed ratios to ensure a high level of consistency and fairness is maintained throughout the evaluation phase. The subject identities are handled very carefully in an attempt to avoid overlapping data to reduce chances of overestimation [2].

The images are acquired directly from images or video data, aiming to extract representative frames from video data, essentially to maintain a high degree of consistency between the image and video modes of data capture. The audio data is acquired from video soundtrack data as well as speech data, including both natural human voices and artificially produced or manipulated sounds.

A description of the datasets used for each modality is offered here. At this stage, information about the number of samples in the dataset and the input type and resolution, as well as the duration in the audio modality, is gathered. This provides clarity and transparency regarding the experiment's reproducibility. FaceForensics++ and Celeb-DF are mainly used for image and video-based deepfake detection, while ASVspoof is utilized for audio-based spoofing and deepfake speech analysis. Table *1* presents the dataset distribution used in the proposed framework.

Modality	Dataset	Train	Validation	Test	Input Details
Image	FaceForensics++, Celeb-DF	1,200	399	400	RGB face images (224×224)
Video (Frame-based)	FaceForensics++, Celeb-DF	16,447 frames	3,736 frames	—	Sampled video frames
Video (Clip-level)	FaceForensics++	64	22	—	Short video clips
Audio	ASVspoof 2019	1,100	400	—	Speech waveforms (16 kHz)

TABLE 1: Dataset distribution and input details for each modality

Preprocessing Strategy

At this stage, efficient preprocessing is very critical in striking a balance between detection accuracy and real-time performance.

For datasets involving images as inputs, the first preprocessing step is to normalize and resize all images to a specified size and resolution. Facial region detection is performed using a simple face detection mechanism. The use of facial regions improves the overall detection and ensures that the ResNet-based image detector identifies facial artifacts related to manipulation and ignores other content.

The cost of processing each frame in a video is computationally expensive and is expected to be highly redundant due to the high similarity between two consecutive frames. Therefore, the uniform frame sampling approach is introduced [6]; i.e., a fixed number of frames are uniformly sampled over the duration of the video. This method reduces the computational cost and provides diversity at the same time. Then, the resized and normalized frames are fed into the EfficientNet-based video detector.

For multimodal inputs, the audio and video streams are extracted from the same source media, ensuring inherent temporal correspondence. Each sampled video frame is associated with a corresponding segment of the audio signal based on its timestamp. Since strict frame-level synchronization is not always required, the proposed framework adopts a feature-level and decision-level fusion strategy. Features from image, video, and audio modalities are extracted independently and later combined through an attention-based fusion mechanism. This design allows the system to handle minor temporal misalignments and missing modalities, making it robust for real-world scenarios. The user interface is shown in Figure *3*.

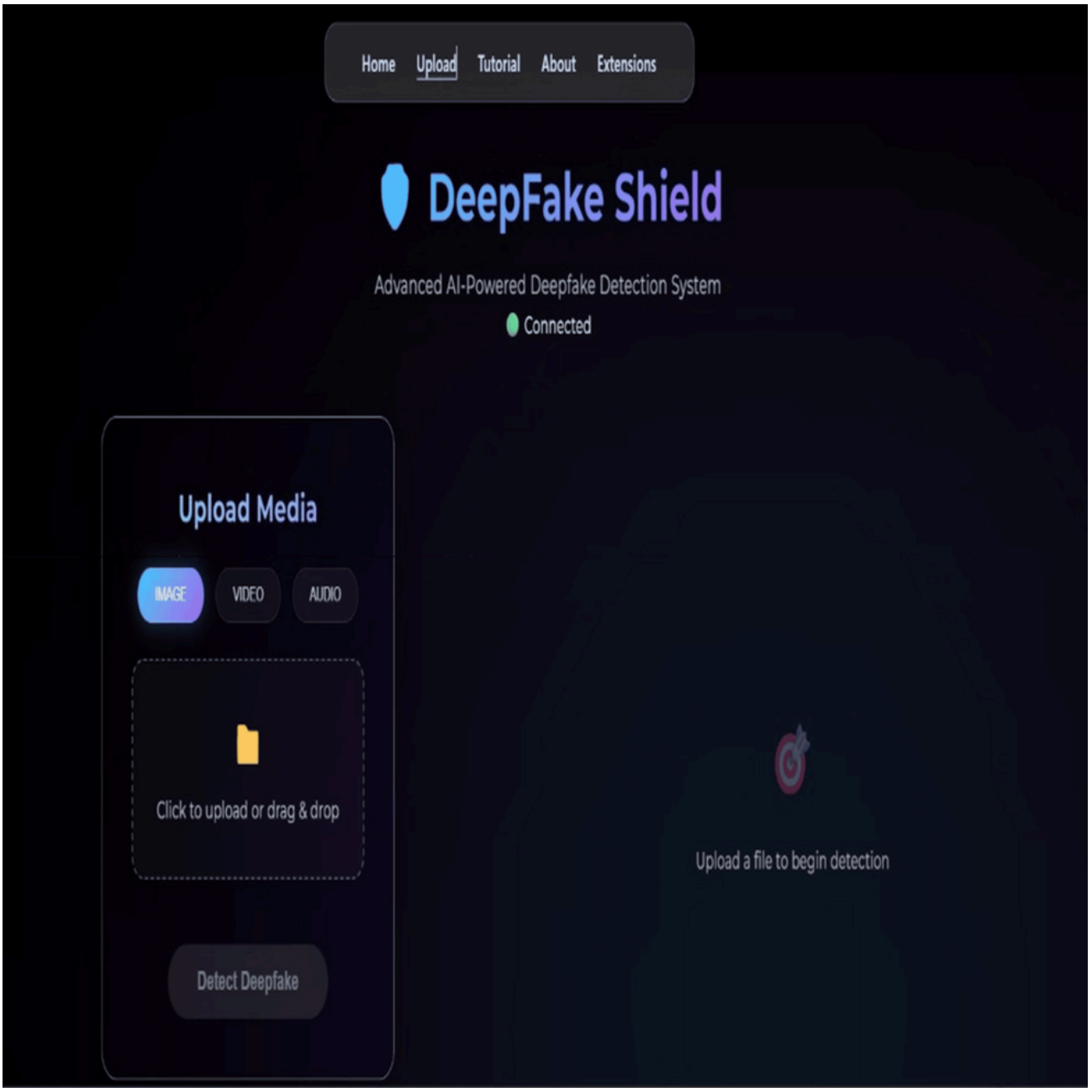

FIGURE 3: Multi-modal media upload interface of the proposed system

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

The audio preprocessing steps include removing the silent periods, using simple noise reduction techniques, and segmenting the audio signal into fixed-size windows. Then, from the windowed audio signal, learned feature representations extracted directly using the Wav2Vec2 encoder and spectral flux are extracted [9,13]. The static and dynamic characteristics present in these features are highly efficient in detecting the subtle frequency - irregularities introduced due to speech synthesis and voice conversion. Such frequency-based inconsistencies and artifacts are commonly observed in synthetic media and have been explored in prior deepfake detection methods [15].

Model Training Configuration

Each modality-specific model is trained independently to allow flexible optimization and easier deployment in real-world scenarios.

The image-based ResNet model is trained using labeled real and fake images with a binary cross-entropy loss function [3]. To improve robustness against visual variations, data augmentation techniques such as random horizontal flipping, brightness adjustment, and contrast variation are applied during training.

The video-based EfficientNet model is trained using sampled frames extracted from videos [4]. Frame-level labels are inherited from the corresponding video labels during training. During evaluation, individual frame predictions are aggregated using statistical techniques such as mean probability estimation and majority voting to produce a final video-level prediction. This aggregation strategy helps stabilize predictions for longer videos.

The audio classifier is trained on extracted spectral features using a lightweight neural architecture designed for fast inference. Since audio data is particularly sensitive to noise, compression, and recording conditions, regularization methods including dropout and weight decay are employed to improve generalization.

All models are trained using the Adam optimizer with carefully selected learning rates to balance convergence speed and stability. Training is intentionally performed on consumer-grade hardware to reflect realistic deployment conditions rather than ideal laboratory settings. Table *2* summarizes the training configuration and validation accuracy.

Modality	Backbone	Epochs	Optimizer	Device	Best Validation Accuracy (%)
Image	ResNet	10	Adam	CPU	83.46
Video (Frame-based)	EfficientNet	12	Adam	CPU	70.56
Audio	Wav2Vec2 + Classifier	10	Adam	CPU	81.50

TABLE 2: Training configuration and validation accuracy of modality-specific models

Explainability Integration During Evaluation

In the evaluation step, there are several XAI techniques implemented to ensure the transparency and trustworthiness of the produced outcome.

For image and video object detection models, saliency maps are generated using gradient-based attribution methods such as Integrated Gradients and Grad-CAM to identify the locations on the image responsible for determining the object class [10,11].

Additionally, for the case of videos, the XAI method is applied both for individual frames and overall videos, with the focus being the identification of the relevant features over the time domain. Grad-CAM visualization is presented in Figure *4*.

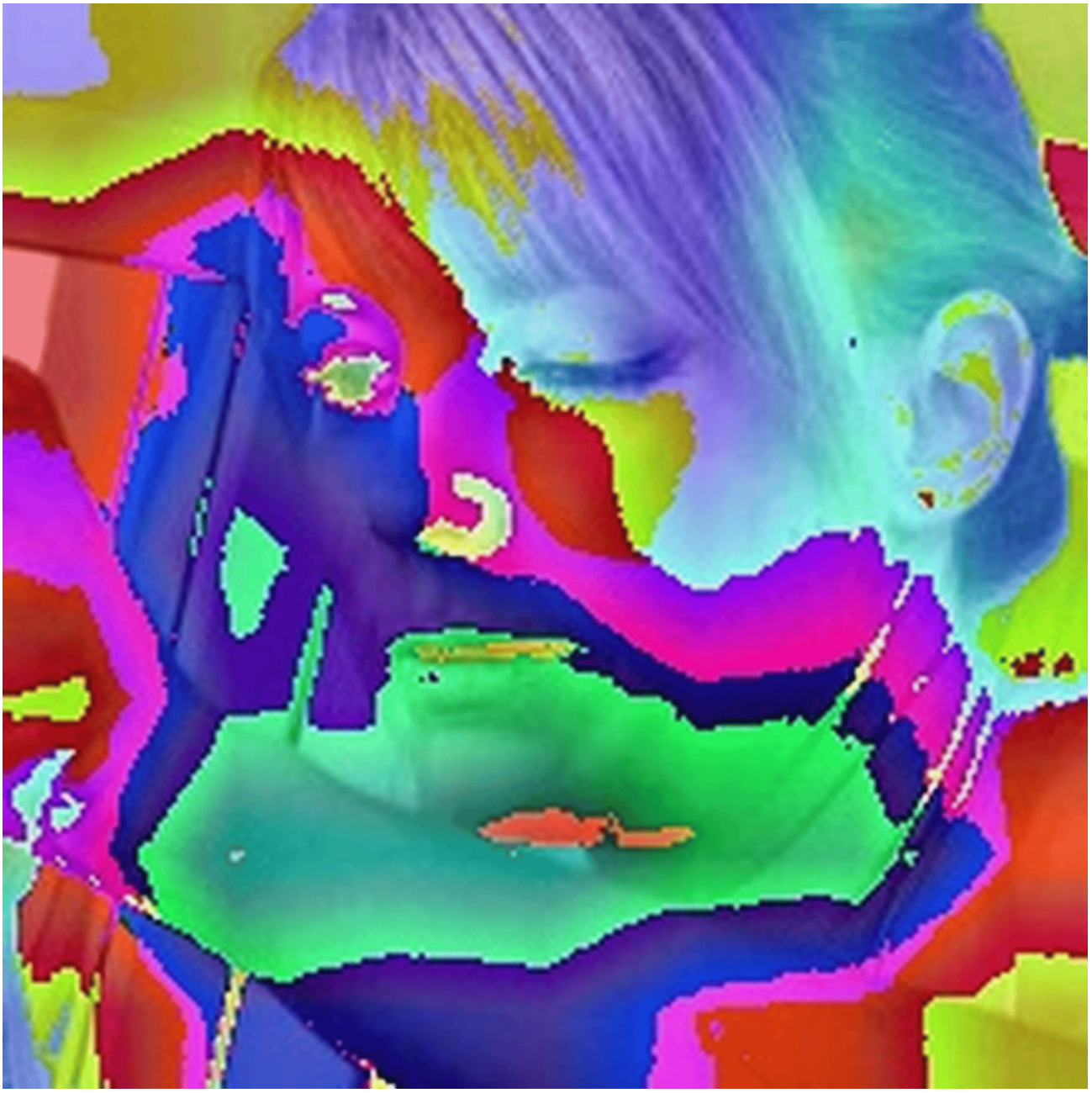

FIGURE 4: Grad-CAM visualization highlighting manipulated facial regions

Grad-CAM, Gradient-weighted Class Activation Mapping

In the audio sample-based models, the XAI method is applied to identify the relevant frequency bands with the largest influence on the outcome. The spectrogram representation is shown in Figure *5*.

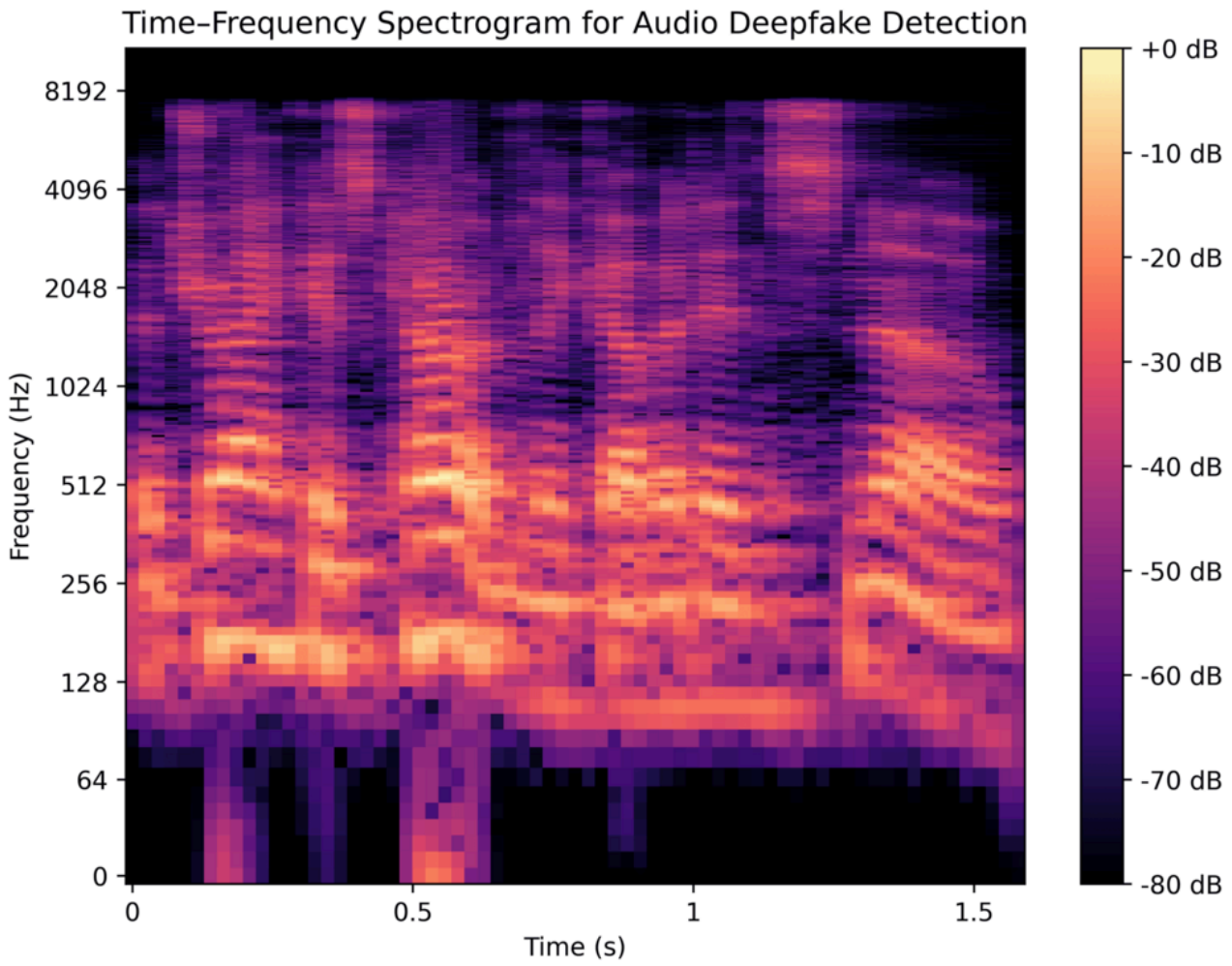

FIGURE 5: Time-frequency spectrogram for audio deepfake detection

The outcome from the XAI method is also required to ensure, during the evaluation step, that the models are utilizing relevant forensic features rather than spurious correlations, based on the confidence factor incorporated. Furthermore, the outcome from the XAI method is also required in the fusion step, facilitating the determination of the outcome based on the individual modality used.

Decision Fusion Strategy

The system adopts a decision-level fusion approach to combine predictions from image, video, and audio modalities. Each modality produces a confidence score indicating the likelihood of manipulation. These scores are combined using a weighted confidence fusion scheme, where modality weights are assigned to the sources based on their reliability and availability.

This design choice allows the system to remain functional even if some modalities are missing or not working well. As a result, the system works well in real life situations like videos, with no sound bad audio or pictures that are not complete such as silent videos, low-quality audio, or incomplete visual data.

Evaluation Metrics

The system's performance is measured using standard binary classification metrics such as accuracy, precision, recall, and F1-score, which are widely used in deepfake detection and media forensics research [1,16]. Moreover, these metrics are calculated individually for real classes and fake classes.

This is significant, as in deepfake detection, the error rate, i.e., the number of fake data classified as real, is generally far higher than the number of real data classified as fake. For the systematic evaluation of the proposed system in terms of the effectiveness of multimodal integration, the system is compared with various single-modality systems.

For a clear idea regarding the contribution of single modalities toward the final prediction, ablation tests are carried out on individual modalities. Along with accuracy, the system's computational efficiency is also measured.

Implementation Environment

The framework works purely in Python. Moreover, for building the model, PyTorch is used, whereas for pipelines, OpenCV and audio processing libraries are used. Experiments have been conducted on a standard laptop-grade computer that lacks access to any high-end GPU, proving the viability of the approach. The real-time pipeline is shown in Figure *6*.

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

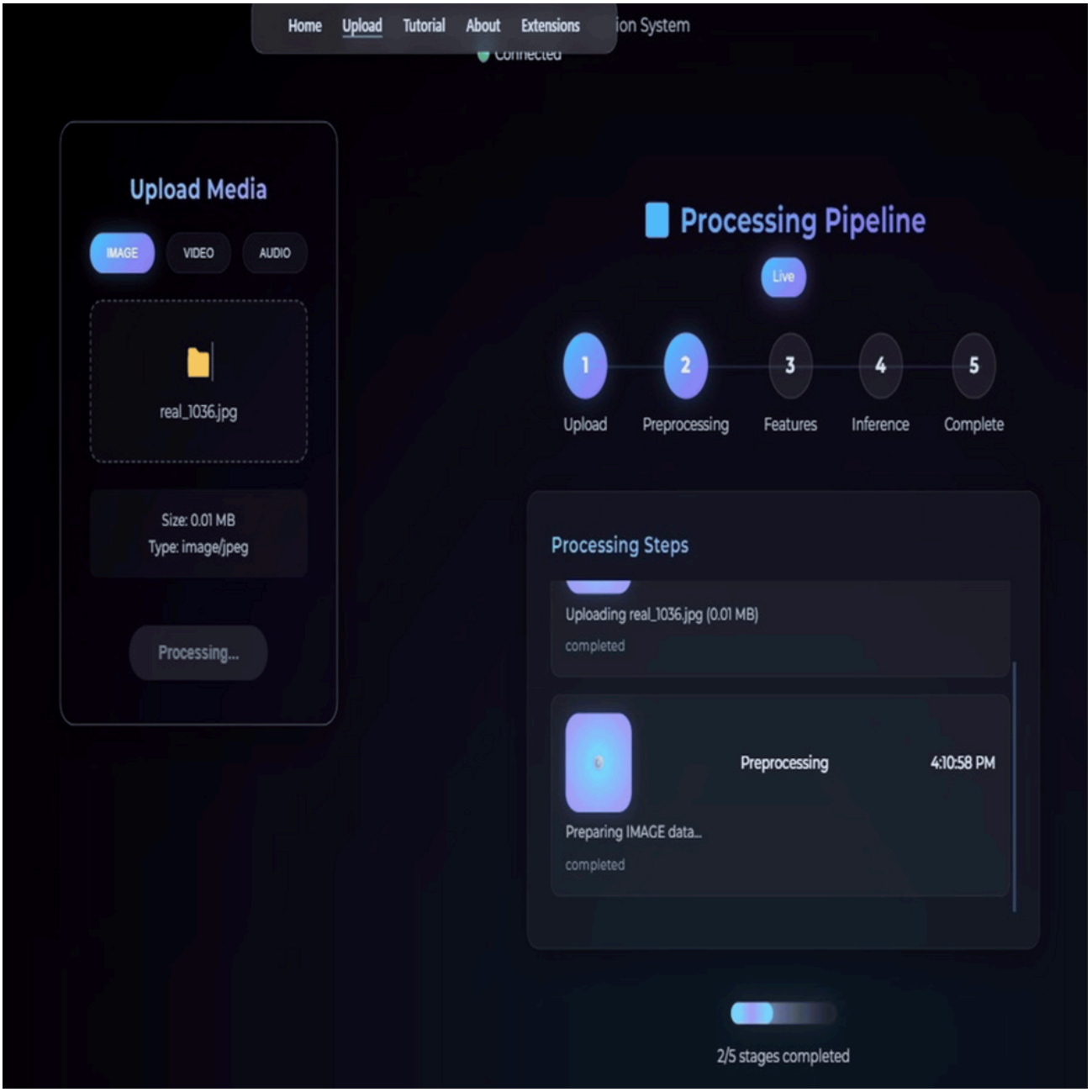

FIGURE 6: Real-time processing pipeline showing preprocessing, feature extraction, and inference stages

In order to ensure reproducibility, the experiment is conducted under controlled conditions, fixing the random seeds during execution. In the experiment, detailed logs are maintained to track the statistics associated with the sampling rate, model prediction, confidence, inference latency, and explainability results.

Experimental Setup

All experiments were conducted on a standard consumer-grade laptop equipped with an Intel i5 processor, 8GB RAM, and no dedicated GPU. The models were implemented using Python with PyTorch as the deep learning framework. OpenCV was used for video processing, and Librosa was used for audio preprocessing.

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

The dataset was divided into training, validation, and testing sets using an 80:10:10 split. Image inputs were resized to 224×224 resolution. Audio samples were processed at a sampling rate of 16 kHz.

Training was performed using the Adam optimizer with a learning rate of 0.001. Each model was trained for 10-12 epochs depending on convergence. The batch size was set to 16 to ensure efficient processing under limited computational resources.

Evaluation was carried out using standard metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

Dataset and Code Availability

The datasets used in this study include FaceForensics++, Celeb-DF, and ASVspoof 2019. These datasets are publicly available and widely used in deepfake detection research.

The implementation code and processed dataset samples are available at https://drive.google.com/drive/folders/1U3Uf1UXTlCAYfJZfBtObMPhsPJBxVzxC?usp=sharing

This ensures reproducibility and transparency of the proposed framework.

Results And Discussion

Quantitative results

The effectiveness of the proposed lightweight multi-modal framework is evaluated under realistic computational constraints for different image, video, and audio modalities. Table *3* shows the performance comparison across modalities.

Modality	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Image	88.25	0.8768	0.8900	0.8834	0.9442
Video (Frame-based)	70.56	0.69	0.65	0.67	0.71
Audio	81.50	0.82	0.81	0.81	0.86

TABLE 3: Performance comparison of the proposed system across different modalities

ROC-AUC, Area Under the Receiver Operating Characteristic Curve

For video-based detection, two evaluation strategies were considered: frame-level classification and clip-level aggregation. The frame-based EfficientNet model achieved a validation accuracy of 70.56%, demonstrating the ability to detect spatial artifacts in sampled frames under computational constraints.

However, when predictions were aggregated at the clip level using statistical fusion, the overall clip-level accuracy decreased to 52.33%, with 64.53% correct classification for real videos and 40.34% for fake videos. This performance gap highlights the challenges of maintaining temporal robustness under lightweight deployment settings without advanced temporal modeling architectures.

While they are lower compared to heavyweight deep learning models reported in prior literature, they are more indicative of performance in the absence of high-end GPU support or dataset-specific fine-tuning [5,6].

The accuracy for the image-based detection using a lightweight version of ResNet is 88.25%. Sample prediction output is shown in Figure *7*. The precision and recall are balanced for both fake and real classes. Good results have been achieved for spatial face artifacts, but low accuracy is observed for highly compressed low-resolution images. Table *4* presents the confusion matrix for image-based detection.

	Predicted Real	Predicted Fake
Actual Real	175	25
Actual Fake	22	178

TABLE 4: Confusion matrix for image-based deepfake detection

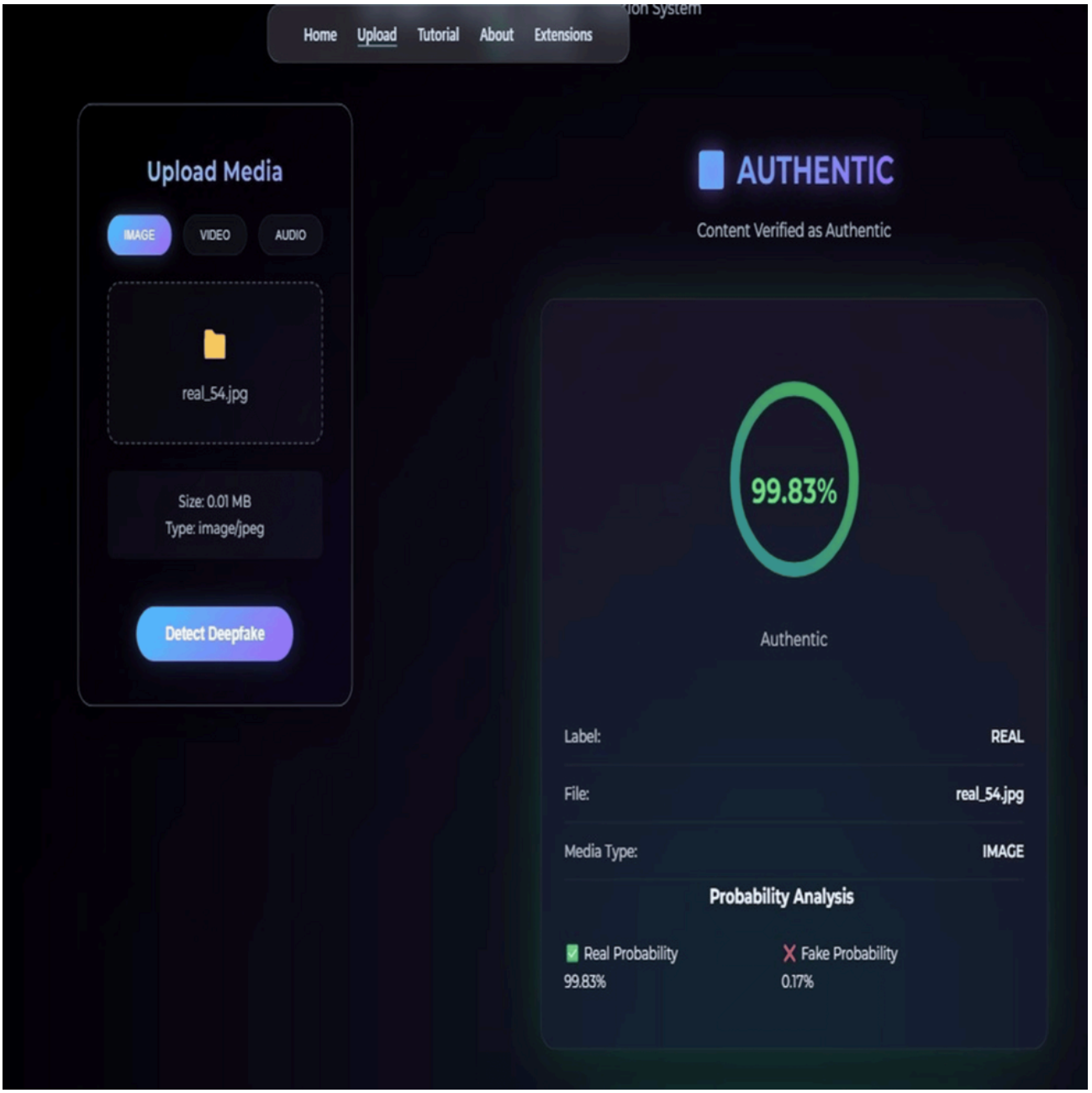

FIGURE 7: Example output for an authentic input with predicted class and confidence score

For the case of the audio-based detection, the classifier has an accuracy of 81.6% when the speech is clean. However, this accuracy reduces for cases involving noisy or compressed audio, where inconsistencies in the spectra are less distinct.

The above findings show that video-based detection is still the most challenging under the assumption of limited resources, while the image- and audio-based modalities provide potent cues for multimodal deepfake detection [12].

Qualitative analysis

Qualitative analysis of misclassified samples shows that false positives occur very frequently in real videos and images containing extreme facial expressions, motion blur, or poor illumination. On the other hand, false negatives are more common for high-quality synthetic videos and images with consistent blending and lip synchronization from a visual

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8

point of view. Sample prediction output is shown in Figure *7*.

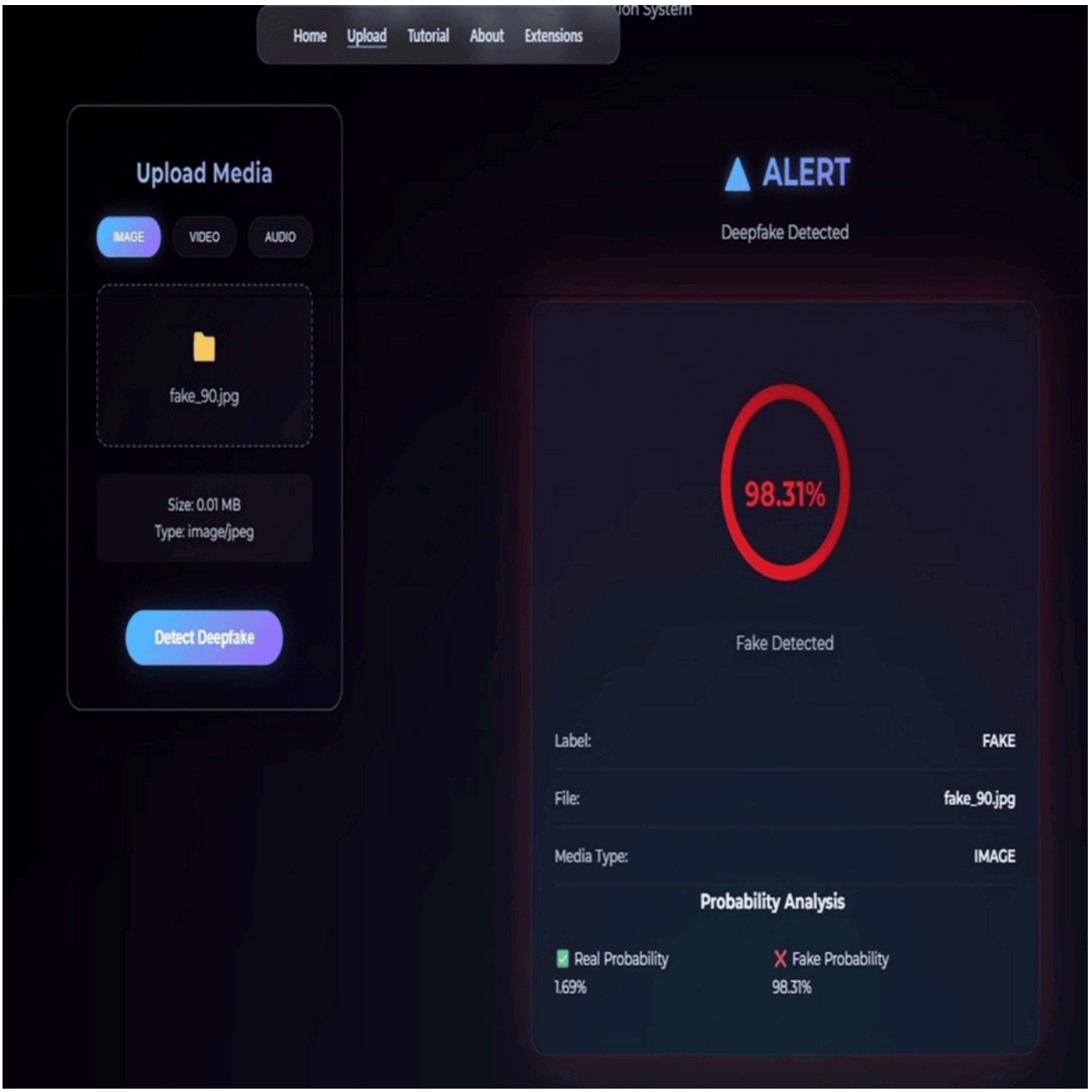

FIGURE 8: Example deepfake detection output with fake-class probability analysis

Indeed, for the audio samples, misclassifications are found mainly in the presence of background noise or when overlapped speech is present. These findings confirm that no single modality can be used alone and provide the justification for integrating image, video, and audio cues in the proposed framework.

Computational efficiency analysis

The lightweight nature of the design greatly reduces inference time and memory consumption for all modes of input [4,6]. Frame sampling is utilized to bound the computation required for video, while compact feature extraction is utilized for fast processing of images and audio. This makes the deepfake detection framework suitable for use in situations for

near real-time execution without needing specialized hardware and computational power.

As a result, the system works well on standard laptop computers, providing a real-world balance between detection accuracy, speed and real-time feasibility.

Ethical and privacy implications

Deepfake detection systems must carefully balance security enforcement of people while protecting their privacy. While the proposed framework analyzes multimedia inputs to see if they are real, it does not perform identity tracking or store biometric identifiers.

When we use this system in environments where many people are being monitored we have to follow the rules about protecting people information and be transparent about what we are doing. The integration of XAI techniques helps us understand how the system works making it more trustworthy and helps us use it in a way especially in situations where security is very important.

Conclusions

This paper presented a lightweight multi-modal deepfake detection framework designed to achieve real-time performance in resource-constrained environments. By integrating image, video, and audio analysis with XAI mechanisms, the system provides a practical defense against emerging synthetic media threats. The proposed framework addresses several key challenges associated with heavyweight deepfake detection approaches. Hence, experimental results affirm that even if detection accuracy reduces for lightweight models, there are significant advantages to consider, specifically in terms of increased computational efficiency, scalability, and feasibility for deployment within consumer devices. The framework contributes toward scalable, real-time protection of digital identity systems, online communication platforms, and cybersecurity infrastructures. These investigations, therefore, emphasize that understanding the design trade-offs for detection systems is essential for effective deepfake detection, as opposed to simply focusing on benchmarking accuracy.

The future work will concentrate on the improvement of the robustness of the fake detection method to noise via the adoption of adaptive sampling of frames, the attention-based fusion mechanism, as well as the development of a noise-resilient audio artifact modeling mechanism. In addition, the system is anticipated to be expanded to incorporate other security applications such as voice phishing detection as well as spoofing detection, which could be done via the incorporation of cryptographic verification systems.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Sanjana Shetty, Tejashree Deore, Ketaki Sakhadeo, Shehnaz Siddique

Acquisition, analysis, or interpretation of data: Sanjana Shetty, Tejashree Deore, Ketaki Sakhadeo, Shehnaz Siddique

Drafting of the manuscript: Sanjana Shetty, Tejashree Deore, Ketaki Sakhadeo, Shehnaz Siddique

Critical review of the manuscript for important intellectual content: Sanjana Shetty, Tejashree Deore, Ketaki Sakhadeo, Shehnaz Siddique

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors

have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request. All data generated or analyzed during this study are included in this published article and/or its appendices.

Acknowledgements

The authors would like to thank Shehnaz Siddique for her guidance, support, and valuable insights during the development of this research. Her feedback helped improve the conceptualization and clarity of the study.

References

1. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion. 2020, 64:131-148. 10.1016/j.inffus.2020.06.014
2. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M: FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019, 1-11. 10.1109/ICCV.2019.00009
3. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, 770-778. 10.1109/CVPR.2016.90
4. Tan M, Le Q: EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning (ICML). 2019, 97:6105-6114.
5. Boháček M, Farid H: Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. Proceedings of the National Academy of Sciences. 2022, 119:e2216035119. 10.1073/pnas.2216035119
6. Zhang T: Deepfake generation and detection, a survey. Multimedia Tools and Applications. 2022, 81:6259-6276. 10.1007/s11042-021-11733-y
7. ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. (2019). Accessed: March 18, 2026: https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf.
8. Shen J, Pang R, Weiss RJ, et al.: Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018, 4779-4783. 10.1109/ICASSP.2018.8461368
9. Sundararajan M, Taly A, Yan Q: Axiomatic attribution for deep networks. ICML'17: Proceedings of the 34th International Conference on Machine Learning. 2017, 70:3319-3328.
10. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision. 2017, 618-626. 10.1109/ICCV.2017.74
11. Li Y, Yang X, Sun P, Qi H, Lyu S: Celeb-DF: A large-scale challenging dataset for deepfake forensics. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020, 3204-3213.
12. Torkzadehmahani R, Kairouz P, Paten B: DP-CGAN: Differentially private synthetic data and label generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019, 98-104. 10.1109/CVPRW.2019.00018
13. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. (2019). Accessed: March 18, 2026: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.
14. Li H, Li B, Tan S, Huang J: Identification of deep network generated images using disparities in color components. Signal Processing. 2020, 174:107616. 10.1016/j.sigpro.2020.107616
15. Li Y, Lyu S: Exposing deepfake videos by detecting face warping artifacts. arXiv. 2019, 10.48550/arXiv.1811.00656

16. Verdoliva L: Media forensics and DeepFakes: An overview. IEEE Journal of Selected Topics in Signal Processing. 2020, 14:910-932. 10.1109/JSTSP.2020.3002101

How to cite this article:
Shetty S, Sakhadeo K, Deore T, et al. (April 07, 2026) Real-Time Generalized Deepfake Detection via Multi-Modal Fusion and Explainable Artificial Intelligence for Cross-Platform Validation. Cureus J Comput Sci 3 : es44389-026-00052-8. DOI https://doi.org/10.7759/s44389-026-00052-8