IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510155

Unmasking Deepfakes: Audio-Visual Deepfake Detection Using AI

Akhila P. V[1], Sahana J[2], Sagarika Sujil[3], and Sreehitha P. C[4]

Department of Computer Science and Engineering (Internet of Things & Cyber Security Including Blockchain Technology), K S Institute of Technology, Bengaluru, India[1,2,3,4]

Abstract: The increasing realism of synthetic and manipulated media has made the authentication of digital video and audio a significant multimedia-forensics problem. Visual-only detectors may miss audio manipulation, while audio-only detectors cannot inspect facial artifacts or temporal inconsistencies. This paper presents a source-grounded description of "Unmasking Deepfakes: Audio-Visual Detection Using AI," a multimodal detection system documented in the project's Phase-2 presentation. The visual pathway samples video frames, detects facial regions using YOLOv8, extracts spatial representations using XceptionNet, and models frame-sequence dynamics using a bidirectional long short-term memory (BiLSTM) network. In parallel, the audio pathway extracts the soundtrack, converts it into a Mel-spectrogram representation, and applies a convolutional neural network (CNN) for audio classification. The documented design combines modality-specific information through an attention layer and late-fusion mechanism to produce a Real/Fake prediction and confidence score. A Gradio interface supports video upload and result presentation, while Grad-CAM is identified for visual explanation. The project presentation records the principal processing modules as implemented but states that module/system testing and final performance evaluation remain in progress. Accordingly, this paper contributes a formalized system description, mathematical formulation, inference procedure, and reproducible evaluation plan; it does not claim unverified performance. Dataset identity, split protocol, training configuration, fusion implementation, and measured results must be supplied from experiment logs before quantitative conclusions or deployment claims can be made.

Index Terms*:* Deepfake detection; audiovisual forensics; multimodal learning; YOLOv8; XceptionNet; BiLSTM; Mel-spectrogram; late fusion.

I. INTRODUCTION

Deepfakes are synthetic or manipulated media generated with machine-learning methods that can alter facial identity, expression, lip motion, or speech. Their increasing realism creates risks for misinformation, impersonation, fraud, harassment, and the evidentiary reliability of digital media. Human inspection is not a dependable verification method because manipulation traces may be subtle, localized, or concealed by compression and post-processing.

A video with an audio track contains complementary evidence. Spatial facial patterns can reveal texture anomalies, blending boundaries, and unnatural local details. Temporal analysis can identify irregularities across successive frames. Audio analysis can detect synthetic speech characteristics and, where alignment is modelled, inconsistencies between spoken audio and visible mouth motion. A detector that processes both modalities can therefore examine a broader set of cues than a single-modality system. However, multimodal processing also introduces synchronization, missing-modality, and computational challenges.

The Phase-2 project presentation describes a pipeline combining YOLOv8 face detection, XceptionNet spatial feature extraction, BiLSTM temporal modelling, Mel-spectrogram-based CNN audio analysis, attention-assisted late fusion, binary prediction, confidence estimation, Grad-CAM visualization, and a Gradio interface [1]. This paper formalizes that documented design. It does not treat the presence of two branches as proof of superior accuracy; that claim requires controlled comparison on the same held-out data.

A. *Research Gap*

Recent work explores spatial-temporal video features, self-supervised representations, and multimodal local-global fusion [2]–[5]. Yet results reported across different datasets and protocols cannot be compared directly. Generalization to unseen generators, robustness to compression, computational cost, explainability, and evaluation leakage remain important concerns [3]. The supplied project materials do not specify a dataset, split protocol, training configuration, or final test metrics. The immediate research gap addressed here is therefore an implementation-focused, clearly specified audiovisual pipeline whose contribution can be evaluated through modality ablations, leakage-resistant testing, and cross-domain validation.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510155

B. *Contributions*

The paper makes four bounded contributions: (1) it documents a dual-branch architecture consistent with the Phase-2 project presentation; (2) it gives a mathematical description of frame-level visual processing, temporal aggregation, Mel-spectrogram audio representation, and late fusion; (3) it specifies an inference procedure and an evaluation protocol intended to prevent identity/source leakage; and (4) it records the boundary between completed implementation modules and experimental claims that remain unverified.

II. RELATED WORK AND LITERATURE REVIEW

Spatial CNN-based detectors learn local facial texture and manipulation traces. Temporal approaches extend frame-level classification by modelling changes over time, since a single frame may not expose inconsistencies that become visible across a sequence. Zafar et al. describe a hybrid framework combining an enhanced EfficientNet-B0 with temporal convolutional processing, illustrating the use of spatial and temporal evidence in video detection [2]. Its backbone and temporal module differ from the XceptionNet–BiLSTM design documented in the present project.

Khormali and Yuan propose a self-supervised graph-transformer framework in which facial patches are represented through learned features and graph relationships. Their method combines a vision-transformer feature extractor, graph convolution, transformer discrimination, and relevancy maps [3]. This work is relevant to generalization and interpretability, although it is a visual-image framework rather than the same audiovisual pipeline.

Javed et al. investigate multimodal local-global feature integration using CNNs, vision transformers, and diffusion-based preprocessing. They report evaluation on FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF [4]. Their reported scores belong to their own experimental configuration and are not transferable to the present project. Tian et al. propose unsupervised detection by explicitly discovering intra-modal and cross-modal inconsistencies between video segments [5]. Their work motivates the importance of cross-modal evidence but uses a different learning formulation.

Kaur et al. survey deepfake-video detection challenges, including dataset imbalance, limited labelled data, computational requirements, reliability, generalization, and insufficient real-world testing [6]. These concerns directly inform the evaluation plan in this paper. In contrast to methods that introduce additional transformer, graph, or diffusion components, the present project uses a modular CNN/recurrent architecture with separate audio and visual pathways. Its novelty should be described as an implementation-specific integration and evaluation target, not as a claim that late fusion or audiovisual detection is itself new.

TABLE I LITERATURE COMPARISON

Study	Method emphasis	Relevance and distinction
Study	Method emphasis	Relevance and distinction
[2] Zafar et al. (2025)	EfficientNet-B0 and temporal CNN	Spatial-temporal video baseline; different feature extractor and temporal model.
[3] Khormali and Yuan (2024)	Self-supervised ViT, graph CNN, transformer	Visual generalization and relevancy mapping; not the same audiovisual design.
[4] Javed et al. (2025)	Local-global audiovisual fusion and diffusion preprocessing	Multimodal benchmark study; more complex architecture.
[5] Tian et al. (2025)	Unsupervised intra-/cross-modal inconsistency discovery	Motivates cross-modal evidence; different unsupervised objective.
[6] Kaur et al. (2024)	Survey of deepfake-video detection	Identifies generalization, data, cost, and reliability challenges.

A. *Positioning of the Proposed System*

The proposed system is positioned as a practical, modular audiovisual classifier that combines frame-level spatial features, temporal sequence features, and audio-spectrum features. Its research value depends on demonstrating that each branch contributes useful information and that fusion improves performance over visual-only and audio-only baselines under an identical evaluation protocol. The project materials do not yet establish this empirical result.

III. RESEARCH PROBLEM AND OBJECTIVES

Let V denote an input video and A its associated audio waveform. The target is a binary label y {0,1}, where 0 denotes real and 1 denotes fake. The system estimates p(y=1 | V,A) and returns a class decision together with a confidence value. The problem is challenging because the visual and audio streams have different sampling rates, temporal resolutions, and feature spaces.

A. *Objectives*

The objectives are to detect facial regions in sampled video frames; extract spatial facial representations; model temporal

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510155

variation across ordered frames; transform the audio waveform into a Mel-spectrogram and classify audio features; combine modality representations using the documented attention/late-fusion design; and present a prediction and confidence score through a user interface.

B. *Research Questions*

RQ1: Does temporal modelling improve over frame-level visual classification on the same held-out videos? RQ2: Does adding the audio branch improve performance over the visual-only branch? RQ3: How does the system behave under compression, frame-rate changes, audio degradation, and unseen manipulation methods? RQ4: Can explanations and confidence estimates be made sufficiently stable and calibrated for human review? These questions require experimental evidence and are not answered by architecture description alone.

IV. PROPOSED METHODOLOGY

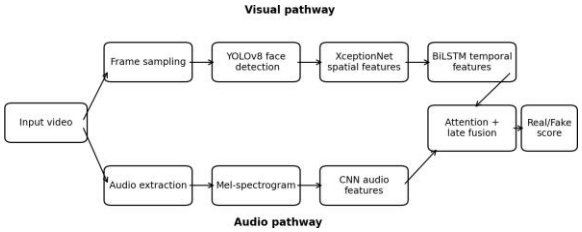


Fig. 1. Multimodal audio-visual deepfake detection architecture.

A. *Input Processing*

The system accepts a video file, validates that it can be decoded, samples a sequence of frames, and extracts its audio track. Let the sampled frame sequence be $F = \{f_1, f_2, \ldots, f_T\}$, where T is the number of selected frames. The audio signal is represented as A[n]. The frame-sampling interval, image resolution, audio sample rate, and clip-duration policy must be fixed and recorded for both training and evaluation. These exact settings are not specified in the supplied project presentation.

B. *Visual Branch: Face Detection and Spatial Features*

YOLOv8 is used to detect faces in sampled frames. For each frame f_t, the detector returns a set of bounding boxes and confidence values. A selected face region is cropped and resized to the input dimensions expected by XceptionNet. XceptionNet maps the face crop to a feature vector x_t R^d. The feature sequence $X = (x_1, x_2, \ldots, x_T)$ is passed to a BiLSTM to model forward and backward temporal dependencies. The use of a face detector can reduce background variation, but missed detections, multiple faces, occlusion, and crop misalignment must be handled consistently.

A BiLSTM computes hidden representations by combining forward and reverse recurrent states: $h_t = [\rightarrow h_t ; \leftarrow h_t]$. A sequence representation z_v may be obtained by a documented pooling or final-state operation. The exact pooling operation and feature dimensions must be confirmed from the implementation; they are not specified in the project slides.

Audio Branch: Mel-Spectrogram and CNN

The audio track is converted to a time-frequency representation. For a discrete waveform A[n], the short-time Fourier transform is $S(m,k) = \Sigma_n A[n]w[n-mH]e^{(-j2\pi kn/N)}$, where w is the analysis window, H is the hop size, and N is the transform length. The power spectrum is passed through Mel filter banks to obtain $M(m,r) = \Sigma_k H_r(k)|S(m,k)|Z$. A logarithmic compression step may be applied before the Mel-spectrogram is supplied to the CNN. The exact sample rate, window length, hop size, number of Mel bands, and normalization must be reported in the final experimental version.

The CNN maps the Mel-spectrogram to an audio representation z_a and an audio classification score. The project presentation confirms Mel-spectrogram generation and CNN-based audio feature extraction/classification, but does not specify the CNN architecture, its layer configuration, or whether the audio classifier is trained jointly with the fusion module.

C. *Multimodal Attention and Late Fusion*

Let z_v and z_a denote the visual and audio representations. The project presentation specifies an attention layer and late fusion, followed by binary Real/Fake classification [1]. A generic fusion formulation is $z_f = \varphi(W_v z_v \; W_a z_a)$, where denotes concatenation, W_v and W_a are learned projections, and φ is a nonlinear mapping. A classifier then produces $\hat{p} = \sigma(w^T z_f + b)$. If attention weights are used, they may be represented as $\alpha = \text{softmax}(g(z_v, z_a))$ and applied to the modality representations before fusion. This is a formal abstraction of the documented design, not a claim about the exact code-level operator. The actual attention function, fusion dimensions, missing-audio policy, and synchronization method must be verified against the implementation.

DOI: 10.17148/IJARCCE.2026.1510155

The predicted label is $\hat{y} = 1$ when $\hat{p} \geq \tau$ and $\hat{y} = 0$ otherwise, where τ is the decision threshold. The default threshold may be 0.5 only if the classifier output is calibrated and the project implementation uses that threshold; otherwise, τ must be selected on validation data and disclosed. The confidence score displayed by the interface should not be described as calibrated probability unless calibration has been measured.

D. *Explainability*

Grad-CAM is identified in the project presentation for highlighting image regions that influence a prediction [1]. For a target class c and convolutional feature maps A^k, Grad-CAM computes channel weights $\alpha_k^c = (1/Z)\Sigma_i\Sigma_j \partial y^c/\partial A^k_{ij}$ and a localization map $L^c = \mathrm{ReLU}(\Sigma_k \alpha_k^c A^k)$. The target layer and whether Grad-CAM is generated for the visual branch alone or for a fused decision must be confirmed. A heatmap is an explanatory visualization, not proof that a prediction is correct or that the highlighted region is causally responsible.

E. *Training Objective*

For binary classification, a common objective is binary cross-entropy: $L = -(1/B)\Sigma_i[y_i \log(\hat{p}_i)+(1-y_i)\log(1-\hat{p}_i)]$, where B is the batch size. This equation describes a suitable binary objective; the project materials do not specify the actual loss, optimizer, learning rate, batch size, epochs, class weighting, or training schedule. These details must be taken from source code and training logs rather than assumed.

V. IMPLEMENTATION AND TECHNICAL DETAILS

The Phase-2 presentation lists the following components as completed: video input and preprocessing; YOLOv8 face detection; face extraction; XceptionNet spatial feature extraction; BiLSTM temporal analysis; audio extraction; Mel-spectrogram generation; CNN audio feature extraction and audio classification; multimodal late fusion; Real/Fake prediction; confidence-score calculation; and a Gradio interface for video upload and result display [1]. The presentation also identifies Grad-CAM as the explanation/visualization module.

A. *Inference Procedure*

Algorithm 1: Audiovisual deepfake inference

Input: video file V; trained visual, audio, and fusion models

1. Decode V and validate video/audio streams.
2. Sample T frames and extract the audio waveform A.
3. For each frame f_t, detect face regions using YOLOv8.
4. Crop/resize the selected face and extract x_t using XceptionNet.
5. Pass $(x_1, \ldots, x_T)$ through BiLSTM to obtain z_v.
6. Compute the Mel-spectrogram M from A.
7. Apply the audio CNN to obtain z_a.
8. Apply the implemented attention and late-fusion operation to (z_v, z_a).
9. Compute $\hat{p}$ and compare it with the validated threshold τ.
10. Return class, score, and available explanation to the interface.

Output: Real/Fake label, confidence score, optional Grad-CAM visualization.

B. *Computational Considerations*

For T sampled frames, visual inference cost is approximately $O(T(C_{det} + C_x))$, where C_{det} is the per-frame face-detection cost and C_x is the per-face XceptionNet feature-extraction cost. The BiLSTM cost depends on sequence length T, hidden dimension h, and input dimension d; for a conventional recurrent layer it scales approximately as $O(T(dh+hZ))$ per direction, up to constant factors. Audio CNN cost depends on the spectrogram dimensions and convolutional layer configuration. Fusion cost is generally small relative to the feature extractors when the fused vectors are low-dimensional. These are structural complexity descriptions; no runtime or memory measurements are available in the supplied project evidence.

C. *Implementation Details Requiring Confirmation*

For reproducibility, the final paper must state the dataset and licensing/provenance; number of real and manipulated clips; identity/source-disjoint split; frame-sampling rate and resolution; face-selection rule; audio sample rate and Mel parameters; exact CNN and BiLSTM dimensions; attention and fusion operator; training loss and optimizer; learning rate, batch size, epochs, random seeds; hardware/software environment; and inference-time measurement procedure. These values are not provided by the presentation and are intentionally left unspecified.

VI. RESULTS AND DISCUSSION

The project presentation reports that the principal processing modules and Gradio interface have been implemented, but it also states that module-level/system-level testing and final accuracy/performance evaluation are in progress [1]. The supplied materials do not establish a named dataset, sample count, class distribution, train/validation/test split, model-training configuration, or measured evaluation scores. Therefore, this paper does not report numerical accuracy,



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.1510155

precision, recall, F1-score, ROC-AUC, confusion-matrix counts, training curves, or inference latency.

A. Required Quantitative Evaluation

The completed system should be evaluated on a held-out test set with no identity or source-video overlap between training and testing. Report accuracy, per-class precision and recall, macro-F1, ROC-AUC, confusion matrix, and confidence intervals where feasible. Because class imbalance can make accuracy misleading, macro-F1 and per-class results should be emphasized. Report threshold-selection procedure and calibration metrics if the interface presents confidence as probability.

B. Ablation and Baseline Comparisons

To test whether multimodal fusion contributes measurable value, evaluate (i) frame-level visual features without temporal modelling, (ii) visual features with BiLSTM, (iii) audio-only CNN, and (iv) the complete audiovisual fusion model. All variants must use the same data partitions and comparable training conditions. Additional ablations should remove attention, vary the number of sampled frames, and test the effect of audio-video desynchronization. Cross-dataset testing should assess generalization to manipulation methods not represented in training.

C. Interpretation

The architecture provides a plausible route for combining spatial, temporal, and audio cues. However, architecture alone does not establish improved reliability. Any claim that fusion outperforms a single branch, that the system generalizes to unseen deepfakes, or that it operates in real time must be supported by measured results. Published values from other studies [2], [4] are not direct baselines unless the same dataset, split, preprocessing, and evaluation procedure are used.

VII. LIMITATIONS AND FUTURE WORK

The current evidence has several limitations. First, dataset identity and split design are not documented in the project presentation. Second, training hyperparameters and exact model configurations are unavailable. Third, the fusion and attention operations are named but not defined at implementation level. Fourth, audio-video synchronization and missing-modality behavior are not specified. Fifth, final performance testing is still in progress. These omissions prevent reproducibility and quantitative comparison.

Future work should complete a subject/source-disjoint evaluation, conduct visual-only/audio-only/fused ablations, test cross-dataset generalization, and evaluate robustness to compression, resizing, frame-rate changes, noise, and unseen manipulation methods. Confidence calibration and threshold selection should be measured before presenting scores as probabilities. The Grad-CAM output should be validated on representative examples and reviewed by humans. Inference latency, memory use, and throughput should be measured on a clearly specified hardware platform before claiming real-time suitability. Model optimization may include frame sampling, batching, quantization, or lightweight feature extractors, but such changes should be reported as future work until implemented and evaluated.

VIII. CONCLUSION

This paper has presented a technically structured description of the "Unmasking Deepfakes: Audio-Visual Detection Using AI" project based on its Phase-2 documentation. The system combines YOLOv8 face detection, XceptionNet spatial representation, BiLSTM temporal modelling, Mel-spectrogram-based CNN audio analysis, attention-assisted late fusion, binary prediction, confidence-score display, Grad-CAM visualization, and a Gradio interface. The documented implementation status indicates that the principal processing modules are completed, while testing and final performance evaluation remain ongoing. The manuscript therefore provides the architecture, mathematical foundations, inference procedure, and evaluation requirements without fabricating experimental results. A defensible empirical contribution requires complete dataset and training documentation, leakage-resistant testing, modality ablations, cross-dataset validation, and quantitative reporting.

ACKNOWLEDGMENT

The authors may acknowledge project supervision and institutional support in the final version. The project presentation identifies Mrs. Haritha Y. as the project guide; formal acknowledgment wording should be approved by the authors and institution before submission.

REFERENCES

[1] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, "A hybrid deep learning framework for deepfake detection using temporal and spatial features," IEEE Access, vol. 13, pp. 79560–79570, 2025, doi: 10.1109/ACCESS.2025.3566008.

[2] A. Khormali and J.-S. Yuan, "Self-supervised graph transformer for deepfake detection," IEEE Access, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3392512.

[3] M. Javed, Z. Zhang, F. H. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local–global feature integration and diffusion models," Signal, Image and Video Processing, vol. 19, Art. no. 400, 2025, doi:



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.1510155

10.1007/s11760-025-03970-7.

[4] M. Tian, M. Khayatkhoei, J. Mathai, and W. AbdAlmageed, "Unsupervised multimodal deepfake detection through explicit intra-modal and cross-modal inconsistency discovery," in Proc. British Machine Vision Conference (BMVC), 2025.

[5] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: challenges and opportunities," Artificial Intelligence Review, vol. 57, Art. no. 159, 2024, doi: 10.1007/s10462-024-10810-6.