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 5, May 2026

DOI: 10.17148/IJARCCE.2026.155222

Explainable Multimodal AI for Deepfake Detection and Digital Content Authenticity

R S Geethanjali[1], Sneha KS[2], Vaidehi Vasudev Arundekar[3], Vinutha BR[4]

Assistant Professor,Dept. of Computer Science and Engineering,

K S School of Engineering and Management, Bengaluru, India[1]

Student, Dept. of Computer Science and Engineering, K S School of Engineering and Management, Bengaluru, India[2-4]

Abstract: Recent breakthroughs in AI and deep learning have enabled the creation of highly convincing synthetic media — commonly termed deepfakes — that are increasingly difficult to distinguish from authentic content. Techniques such as GANs, transformer-based architectures, and diffusion models introduce serious risks to cybersecurity, journalistic integrity, digital trust, and democratic processes. Existing solutions are predominantly limited to single-modality analysis, hindering effectiveness against coordinated multimedia misinformation spanning video, audio, and text simultaneously. This paper surveys existing literature on deepfake detection, explainable AI, and multimodal misinformation analysis, identifying key research gaps and limitations. Based on these findings, we propose an Explainable Multimodal AI Framework that unifies ResNet18, CNNs, and DistilBERT with SHAP, LIME, and RAG-based contextual reasoning for simultaneous detection and authenticity verification of manipulated multimedia content. The proposed framework is expected to deliver improved accuracy, transparency, and real-time inference capability over existing unimodal and non-explainable approaches when evaluated on standard benchmarks such as FaceForensics++, DFDC, ASVspoof 2019, and FakeNewsNet.

Keywords: Deepfake Detection; Explainable AI; Multimodal Learning; Digital Authenticity; Generative Adversarial Networks; CNN; ResNet18; DistilBERT; Retrieval-Augmented Generation; SHAP; LIME; Real-Time Detection.

I. INTRODUCTION

Advances in AI and deep learning have fundamentally reshaped how synthetic media is produced and distributed. Powerful generative models — including GANs [6], Variational Autoencoders (VAEs), transformer-based systems, and diffusion-based architectures — now enable the synthesis of photorealistic images, videos, audio, and text at unprecedented scale. While such capabilities hold legitimate value in entertainment, education, and virtual collaboration, their exploitation has given rise to deepfakes: fabricated media that fuel disinformation, enable impersonation, facilitate financial fraud, and destabilize political discourse.

As synthetic media generation grows more sophisticated, distinguishing fabricated content from authentic material has become increasingly untenable through manual review alone. Detection systems operating on a single modality are ill-equipped to handle coordinated attacks that simultaneously alter video, audio, and textual narratives. Most deployed AI models function as opaque black-boxes, offering no interpretive insight into their outputs, eroding user confidence and limiting forensic accountability.

A growing body of research has sought to close these gaps. Architectural comparisons spanning ResNet18, ResNet50, EfficientNet, InceptionV3, and Vision Transformers validate strong representational capacity for image-based deepfake classification. Interpretable approaches using DenseNet121 with Integrated Gradients, Grad-CAM, and LIME demonstrate that high accuracy need not sacrifice transparency. RAG-based pipelines cross-referencing extracted claims against verified knowledge sources have shown notable gains in fake news detection robustness.

These contributions notwithstanding, critical shortcomings persist: high computational overhead, weak generalization across datasets, vulnerability to adversarial perturbations, shallow cross-modal fusion, and inadequate throughput for real-time deployment. Motivated by these unresolved challenges, this paper surveys the existing literature, characterises the open research gaps, and proposes an Explainable Multimodal AI Framework that couples lightweight deep learning with XAI attribution methods and RAG-based evidence retrieval to deliver transparent, real-time deepfake detection across all three modalities simultaneously.

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 5, May 2026

DOI: 10.17148/IJARCCE.2026.155222

II. LITERATURE REVIEW

Rao et al. [1] developed a Hybrid Multimodal Verification System (HMVS) combining NLP, computer vision, GAN fingerprinting, blockchain-based credibility ledgers, and reinforcement learning. Benchmarked across LIAR, FakeNewsNet, DFDC, and Celeb-DF, the system suffers from prohibitive computational overhead from on-chain verification and lacks XAI integration.

Mule et al. [2] benchmarked ResNet18, ResNet50, EfficientNet-B0, Vision Transformers, InceptionV3, and custom CNNs on FaceForensics++ for image-level deepfake classification, applying Integrated Gradients post-hoc to localize manipulated facial regions. The study was limited to the visual modality with no mechanism for cross-modal verification. Patel et al. [3] surveyed XAI methods in deepfake detection, cataloguing SHAP, LIME, Grad-CAM, saliency maps, attention visualization, and adversarial robustness auditing, and documented shared shortcomings including inconsistent dataset generalization, inference overhead, and absence of unified evaluation benchmarks.

Bai and Fu [4] addressed textual misinformation through the Full-Context Retrieval and Verification (FCRV) framework, chaining LLM-based claim extraction with RAG-driven evidence retrieval. Its principal limitation is exclusive focus on text, rendering it unsuitable for multimedia deepfake scenarios.

Abinash et al. [5] combined DenseNet121 with Integrated Gradients, Grad-CAM, Occlusion Sensitivity, Vanilla Gradients, and LIME on a large corpus of images drawn from FaceForensics++ and Celeb-DF, achieving strong detection performance. However, the system is constrained to the image modality and lacks cross-modal fusion capability.

Collectively, the reviewed works reveal that no unified solution adequately combines multimodal detection with transparency, contextual reasoning, and real-world deployment throughput. Rao et al. [1] cover video and text modalities but provide no XAI integration, no RAG support, and no real-time capability, with the primary limitation being the prohibitive overhead of blockchain-based verification. Mule et al. [2] are restricted to the video modality alone, offer only partial XAI through post-hoc Integrated Gradients, and provide no cross-modal fusion. Patel et al. [3], as a survey contribution, catalogue XAI techniques but identify the absence of a unified benchmark and the lack of real-time applicability as the field's most pressing open problems. Bai and Fu [4] incorporate RAG-based evidence retrieval but are entirely limited to textual misinformation, with no extension to audio or video. Abinash et al. [5] achieve strong image-level detection with XAI support but remain constrained to the image modality, with no provision for audio or text. The proposed framework is designed to overcome all of these limitations simultaneously by covering video, audio, and text modalities within a single architecture, integrating both SHAP and LIME for full XAI attribution, incorporating RAG-based contextual verification, and targeting real-time inference throughput.

The proposed framework directly addresses these gaps by integrating all three modalities with explainability and real-time inference.

III. PROPOSED FRAMEWORK

The proposed Explainable Multimodal AI Framework consists of five integrated modules: data collection and preprocessing, deep learning feature extraction, multimodal fusion, XAI attribution, and RAG-based contextual verification.

A. Dataset and Preprocessing

The framework is designed for training and evaluation on four benchmark datasets: FaceForensics++ [7] and DFDC [8] for deepfake video detection; ASVspoof 2019 [9] for synthetic speech detection; and FakeNewsNet [10] for textual misinformation. An 80/10/10 train/validation/test split is applied uniformly across all modalities. Video frames are extracted at 25 fps using OpenCV; facial regions are localised via MTCNN, resized to 224×224 pixels, and z-score normalised. Audio files are converted to Mel spectrograms (128 mel bins, hop length 512, window size 2048) and denoised with spectral subtraction. Textual data undergoes tokenisation, stop-word removal, and DistilBERT sub-word encoding with a maximum sequence length of 512 tokens.

This multi-dataset design is expected to provide broad coverage across manipulation types and compression conditions, addressing the generalisation weaknesses identified in prior single-dataset studies.

B. Deep Learning Feature Extraction

Three modality-specific components operate in parallel. For video, ResNet18 and EfficientNet-B0 are pre-trained on ImageNet and fine-tuned with a classification head (512 to 128 units, ReLU activation, 0.3 dropout). For audio, a

DOI: 10.17148/IJARCCE.2026.155222

spectrogram-based CNN with five convolutional blocks (64, 128, 256, 256, 128 filters each followed by batch normalisation and max-pooling) processes Mel spectrograms. For text, DistilBERT with a classification head (768 to 256 to 128 units) captures contextual semantics. All components are trained using the Adam optimiser (beta-1=0.9, beta-2=0.999, learning rate=1e-4) with cosine annealing decay and early stopping at patience of 5 epochs.
The use of lightweight architectures such as ResNet18 and DistilBERT, rather than heavier alternatives, is expected to reduce inference latency substantially compared to prior works that relied on computationally intensive models without real-time constraints.

C. Multimodal Fusion
Feature vectors from each modality-specific network (ResNet18: 512-d; EfficientNet-B0: 1280-d; speech CNN: 256-d; DistilBERT: 768-d) are concatenated into a 2816-dimensional joint representation. A three-layer fusion network (1024, 512, 256 units with ReLU activations, 0.4 dropout, and batch normalisation) maps this joint representation to a binary authenticity decision. An attention-weighted late fusion mechanism additionally computes modality-specific confidence scores via Monte Carlo dropout, dynamically weighting each stream's contribution based on input uncertainty estimates. This attention-weighted fusion strategy is expected to outperform simple concatenation baselines, as the dynamic modality weighting should allow the framework to compensate when one modality provides unreliable or missing signals — a scenario common in real-world multimedia content.

D. Explainable AI Integration
SHAP (SHapley Additive exPlanations) [11] and LIME (Local Interpretable Model-agnostic Explanations) [12] are integrated post-hoc. SHAP DeepExplainer is applied to the video and audio CNN branches to produce pixel-level and spectrogram-bin-level attribution scores, highlighting the features most responsible for each prediction. LIME perturbs input text segments and trains local surrogate models to identify which words or phrases drive the DistilBERT classification decision. Both methods produce human-readable explanation artefacts logged alongside each prediction for forensic review.
Integrating dual XAI methods is anticipated to improve expert trust and forensic utility compared to systems that offer no interpretive output, addressing the transparency gap documented across the surveyed literature.

E. RAG-Based Contextual Verification
A Retrieval-Augmented Generation (RAG) module [15] embeds extracted claims using a sentence-transformer model and retrieves evidence from a FAISS-indexed knowledge base of verified news articles and fact-check databases. The top-5 retrieved documents are passed to a DistilBERT-based entailment classifier that identifies contradictions between retrieved evidence and the input claim, augmenting the textual classification branch and improving robustness against AI-generated misinformation.
The inclusion of RAG-based verification is expected to improve precision on textual misinformation detection over the DistilBERT-only baseline, by grounding predictions against factual documentary evidence rather than relying solely on learned parametric knowledge.

F. Social Media and Bot Behaviour Analysis
A gradient-boosted classifier trained on labelled bot-activity datasets analyses user behaviour signals — including posting frequency, engagement patterns, account age, and semantic similarity across posts — to identify automated accounts. This module operates independently of the multimedia pipeline and contributes a provenance signal to the overall authenticity decision, offering an additional layer of verification not present in any of the surveyed systems.

IV. DISCUSSION

A. Expected Modality-Specific Performance
Each modality-specific component is expected to demonstrate competitive performance on its respective benchmark. The ResNet18 and EfficientNet-B0 video classifiers, leveraging fine-tuned ImageNet representations and MTCNN-based facial localisation, are anticipated to achieve detection accuracy on FaceForensics++ that improves upon VGG16 and comparable baseline architectures documented in the literature. The spectrogram-based CNN for audio is expected to reduce the Equal Error Rate on ASVspoof 2019 relative to shallow baselines, given the representational richness of Mel spectrogram features combined with multi-stage convolutional processing. The DistilBERT text classifier, augmented with RAG-based evidence retrieval, is expected to surpass LSTM-based baselines on FakeNewsNet in terms of F1-score, consistent with the broader trend of transformer models outperforming recurrent architectures on natural language tasks.

DOI: 10.17148/IJARCCE.2026.155222

B. Expected Multimodal Fusion Performance

The fused multimodal framework is expected to outperform the best individual unimodal component, as cross-modal fusion contributes complementary discriminative signals unavailable to any single modality. The audio stream is anticipated to provide strong evidence in cases where video-only analysis may be deceived by high-quality face-swaps, while the text stream is expected to capture misinformation embedded in captions that visual analysis cannot detect. The attention-weighted fusion mechanism, by dynamically adjusting modality contributions based on uncertainty estimates, is expected to yield further improvements over simple feature concatenation.

C. Expected XAI and RAG Performance

SHAP and LIME explanations are anticipated to consistently highlight manipulated facial regions, spectrogram anomalies, and semantically suspicious text segments, providing forensically actionable attribution outputs. Based on findings from prior XAI-enabled deepfake detection systems [2, 5], expert evaluators are expected to rate the generated explanations as informative, thereby improving practitioner trust relative to black-box detection models. The RAG module is expected to improve precision on text authenticity classification by grounding predictions in retrieved factual documents, echoing the gains demonstrated by Bai and Fu [4] in a text-only context.

D. Inference Latency and Scalability Considerations

The framework is designed to satisfy real-time processing constraints. The choice of lightweight architectures — ResNet18 over heavier alternatives, DistilBERT over full BERT — is deliberately motivated by latency requirements. Modality-specific preprocessing steps, including MTCNN localisation and Mel spectrogram conversion, are expected to account for a manageable share of the overall pipeline latency. The RAG retrieval stage, while introducing additional latency, is expected to remain within acceptable bounds given the use of FAISS approximate nearest-neighbour indexing. Overall, the framework is anticipated to operate well within the real-time processing threshold commonly cited in deployment literature, enabling practical use in content moderation and digital forensics contexts.

E. Ablation Study Design

A planned ablation study will quantify the contribution of each framework component. Conditions of interest include: removing the RAG module to isolate the contribution of evidence-grounded verification; replacing the attention-weighted fusion mechanism with simple concatenation to measure the value of dynamic modality weighting; and removing SHAP and LIME outputs to quantify their effect on expert trust ratings independent of classification accuracy. Such an ablation is expected to confirm that each component contributes meaningfully, consistent with the incremental design rationale outlined in Section III.

F. Comparison with State-of-the-Art

The proposed framework is positioned against leading multimodal deepfake detection systems. HMVS [1] operates on video and text modalities but without XAI integration or real-time support, and its blockchain-based verification introduces prohibitive overhead. DenseNet121 with XAI [5] offers strong image-level detection with interpretability but is restricted to a single modality. FCRV [4] supports text-only fact verification with RAG but lacks multimedia coverage or explainability. The proposed framework is designed to be the most comprehensive among these, combining all three modalities, full XAI attribution, RAG-based evidence grounding, and real-time inference within a single unified system. On these design dimensions, it is expected to represent a meaningful advance over each of the compared systems.

V. CONCLUSION AND FUTURE DIRECTIONS

The proliferation of AI-synthesised media poses mounting risks to digital trust, journalistic integrity, cybersecurity infrastructure, and informed public discourse. This paper has surveyed existing approaches to deepfake detection, multimodal misinformation analysis, and explainable AI, revealing a consistent set of limitations across the literature: single-modality coverage, lack of transparency, absence of contextual grounding, and insufficient throughput for real-time deployment. No existing system simultaneously addresses all four of these dimensions.

To address these gaps, we have proposed an Explainable Multimodal AI Framework that integrates ResNet18, CNNs, and DistilBERT for video, audio, and text analysis respectively, unified through an attention-weighted fusion mechanism and augmented with SHAP, LIME, and RAG-based contextual verification. The conceptual framework is grounded in established building blocks from the surveyed literature, and each design choice is motivated by documented limitations in prior work. The proposed system is expected to improve upon current state-of-the-art approaches in detection coverage, interpretability, and deployment feasibility.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155222

Several directions remain open for future investigation. Incorporating Vision Transformer (ViT) architectures may capture longer-range spatial dependencies in video frames beyond the capacity of standard CNNs. Extending the NLP component to multilingual contexts would broaden applicability across global disinformation ecosystems. Blockchain-anchored provenance tracking, edge-optimised model deployment via quantisation and pruning, and live webcam integration represent promising engineering extensions. Federated learning approaches could enable privacy-preserving collaborative training across institutions without centralising sensitive data. Systematic adversarial stress-testing would be essential to evaluate the framework's resilience against evasion attempts, a vulnerability noted across multiple surveyed systems. Empirical validation of the proposed framework on the target benchmarks constitutes the immediate next step, and will be the subject of subsequent work.

REFERENCES

[1]. D. N. Rao, Y. J. N. Kumar, K. Mouneshwari, A. Soy, P. R. Kiran, and T. Srihari, "AI-Powered Real-Time Misinformation Detection: A Deep Learning Framework for Combating Fake News and Deepfakes," in Proc. 2025 Int. Conf. Metaverse and Current Trends in Computing (ICMCTC), IEEE, 2025.
[2]. V. Mule, P. Prasad, D. Bhangale, A. Yenkikar, R. Mirajkar, A. Sayyad, and N. Sable, "Seeing Beyond the Fake: Comparative Deepfake Detection with Integrated Explainability," in Proc. 2025 IEEE Pune Section Int. Conf. (PuneCon), IEEE, 2025.
[3]. K. Patel, M. Sutariya, and P. Parmar, "A Comprehensive Survey on Explainable Deepfake Detection: Techniques, Challenges, and Future Directions," in Proc. 2025 IEEE Conf. on AI and Machine Vision (AIMV), IEEE, 2025.
[4]. Y. Bai and K. Fu, "A Large Language Model-based Fake News Detection Framework with RAG Fact-Checking," in Proc. 2024 IEEE Int. Conf. on Big Data, IEEE, 2024.
[5]. Abinash P., Sushmitha P., Sanjay Kumar B., and Niveditha M., "Interpretable Deepfake Detection: Enhancing Real vs. Fake Face Classification with Explainable AI and Transfer Learning," in Proc. 2025 Int. Conf. on Intelligent Information Technologies (ICIIT), IEEE, 2025.
[6]. Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, and V. Vimal, "A Comprehensive Survey on Deepfake Generation and Detection Using Machine Learning and Deep Learning," IEEE Access, vol. 11, pp. 45688-45724, 2023.
[7]. D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, "A Robust Approach to Multimodal Deepfake Detection," J. Imaging, vol. 9, no. 6, p. 122, 2023.
[8]. B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.
[9]. X. Yi et al., "ADD 2023: The Second Audio Deepfake Detection Challenge," in Proc. ICASSP, IEEE, 2023.
[10]. H. Wan, S. Feng, Z. Tan, H. Wang, Y. Tsvetkov, and M. Luo, "DELL: Generating Reactions and Explanations for LLM-based Misinformation Detection," in Findings of ACL 2024, ACL, 2024.
[11]. F. Anagnostopoulos, S. Papadopoulos, and I. Kompatsiaris, "Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection," in Proc. ACM Workshop on Multimedia AI against Disinformation (MAD), 2024.
[12]. A. Govindu, P. Kale, A. Hullur, A. Gurav, and P. Godse, "Deepfake Audio Detection and Justification with Explainable AI (XAI)," Research Square preprint, 2023.
[13]. C. Budati and A. Jadam, "Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics," Semantic Scholar, 2024.
[14]. H. Liu, W. Wang, and H. Li, "Interpretable Multimodal Misinformation Detection with Logic Reasoning," in Findings of ACL 2023, pp. 9781-9796, 2023.
[15]. Y. Bai and K. Fu, "Fake News Detection with Retrieval Augmented Generative Artificial Intelligence," in Proc. IEEE Int. Conf. on Big Data, IEEE, 2024.

