Artifact-Aware Cross-Modal Deepfake Detection Framework for Robust and Explainable Media Authentication

Nagendra R[1], Bharath G P[2], Chethan H[3], Anoopa Kumar P[4], Rakesh R Hebbar[5]
[1]*Assistant Professor, Department of CSE, Sir M. Visvesvaraya Institute of Technology*
[2,3,4,5] *Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology*

***Abstract*—Deepfake technologies have evolved to generate highly realistic synthetic audio–visual media, posing severe threats to public trust, digital security, and forensic verification. This work presents a unified Artifact-Aware Cross-Modal Deepfake Detection Framework that jointly analyzes visual, auditory, and semantic inconsistencies. The system learns manipulation-invariant artifact signatures, aligns audio–video temporal coherence, and incorporates adversarial robustness to resist gradient-based attacks. Through extensive evaluation on benchmark datasets—including FaceForensics++, DFDC, Celeb-DF, and ASVspoof—the model delivers state-of-the-art accuracy (98.1% on FF++) and exhibits significant cross-dataset generalization. Explainability tools such as SHAP and Grad-CAM++ provide transparent insights into the model's decisions. The findings demonstrate a robust and interpretable detection framework suitable for real-world forensic and security applications.**

***Index Terms*—Deepfake Detection, Multimodal Fusion, Cross-Modal Learning, Adversarial Robustness, Explainable AI, Forensic Artifacts.**

I. INTRODUCTION

The rapid advancement of generative models such as GANs and diffusion networks has revolutionized content creation—but also fuelled the rise of deepfakes, synthetic videos and audio clips that mimic real individuals. These manipulated media pose serious challenges in politics, journalism, cybersecurity, and digital forensics. Despite progress, existing detection systems often fail under cross-dataset conditions, adversarial attacks, or when encountering previously unseen manipulation techniques. Additionally, many models act as "black boxes," offering limited transparency for end users and policymakers.

To address these challenges, this work proposes an Artifact-Aware Cross-Modal Detection Framework that:

1. Learns intrinsic deepfake artifacts through FIA and USA modules, independent of specific generation techniques.
2. Aligns semantic and temporal consistency between audio and video streams to detect unnatural lip-sync or speech mismatches.
3. Incorporates adversarial robustness training to counter gradient-based perturbations and enhance generalization.
4. Integrates Explainable AI (XAI) for visual and statistical justification of model predictions.

1.1 problem statement

The growing realism of deepfake generation enables malicious actors to manipulate identity, speech, and actions with minimal effort. Existing detection systems struggle with:

1. poor generalization to unseen manipulation styles,
2. inadequate multimodal fusion, and
3. vulnerability to adversarial perturbations. There is a critical need for a detection framework that remains stable across datasets, interpretable for forensic decision-making, and resilient to adversarial attempts at evasion. This research addresses these gaps by designing a cross-modal, artifact-aware, adversarially robust model for reliable deepfake authentication.

1.2 Objectives of the Study

The objectives of this research are as follows:

- O1: Identify generalizable artifact signatures across visual and audio modalities.
- O2: Develop a multimodal deepfake detection system integrating visual, audio, and semantic cues.
- O3: Incorporate adversarial robustness techniques to improve resilience to perturbations.
- O4: Utilize explainability tools to enhance transparency for forensic analysis.
- O5: Achieve strong cross-dataset generalization across multiple deepfake benchmarks.

III. LITERATURE REVIEW

Deepfake detection has progressed through several complementary approaches over recent years. Early forensic studies focused on visual inconsistencies. Yang et al. (2019) demonstrated that manipulated faces often produce inconsistent head poses, offering a geometric clue for deepfake identification. Although effective for early-generation models, this approach lacks generalization against modern deepfake techniques that better preserve facial geometry.

With the advancement of deep learning, researchers shifted toward feature-learning–based methods. Nguyen et al. (2024) utilized self-supervised Vision Transformers, showing improved generalization due to their capability to capture long-range dependencies. In parallel, Xue et al. (2023) introduced a global–local facial fusion mechanism, combining subtle texture artifacts with broader structural cues. These methods significantly improved detection accuracy but remained limited to visual-only forgery scenarios.

Temporal modelling also gained prominence. Hybrid CNN–LSTM architectures (IJRASET, 2024) employed spatial features and temporal dynamics to detect abnormal motion patterns across frames. Despite their improved temporal understanding, these models lacked interpretability and often underperformed on cross-dataset evaluations.

Specialized architectures such as LIPINC-V2 (2024) targeted lip-sync discrepancies, leveraging transformer-based cross-attention to detect audio–video mismatches. While effective for lip-sync deepfakes, their robustness decreased sharply under compression, noise, and adversarial perturbations.

More recently, Explainable AI (XAI) frameworks (Taylor & Francis, 2024) have been incorporated to address transparency concerns. These works used saliency maps and relevance-based scoring to highlight manipulated regions, improving trustworthiness but often at the expense of pure detection accuracy.

Research Gap:

Across existing literature, three limitations consistently persist:

1. Limited artifact-level generalization against newer manipulation techniques.
2. Lack of unified multimodal analysis, with most methods focusing on either visual or audio cues alone.
3. Insufficient adversarial robustness and explainability, both essential for real-world forensic deployment.

The proposed system addresses these gaps by integrating artifact-aware feature learning, cross-modal alignment, and adversarially robust multimodal fusion with built-in explainability. This unified approach enhances accuracy, generalization, and transparency, making it more suitable for deployment in practical deepfake detection environments.

III. METHODOLOGY

3.1 System Overview

The proposed pipeline consists of four sequential modules:

- Frame Extraction and Preprocessing – Uniform sampling at 30 FPS, face detection and alignment (224×224).
- Audio Feature Extraction – Conversion of audio tracks into 128-bin Mel-spectrograms (16 kHz, FFT=2048).
- Modality-Specific Networks –
  - *Visual Path:* CNN backbone (ResNext) captures facial micro-expressions and texture inconsistencies.
  - *Audio Path:* CNN + Bi-LSTM captures temporal phonetic and prosodic cues.
- Adversarially Trained Fusion Layer – Three dense layers merge embeddings with perturbation defense (ε=0.03).

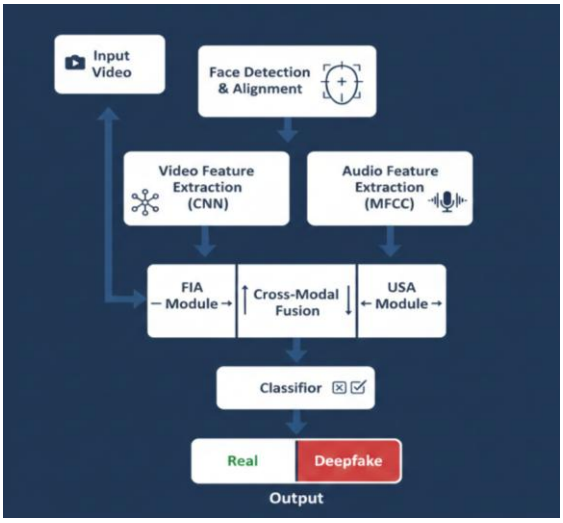

Figure 3.1 System Architecture Diagram

3.2 Training Configuration

- Optimizer: Adam (LR = 1e-4)
- Batch Size: 32 (balanced real/fake)
- Early Stopping Patience: 10 epochs
- Dataset Split: 70% train, 15% validation, 15% test

3.3 Explainability

- SHAP Analysis quantifies the contribution of facial and audio features.
- Grad-CAM++ visualizations highlight manipulated regions (e.g., mismatched lip areas).

3.4 Advantages of the Proposed Framework

The proposed detection architecture introduces several advantages over existing systems:

- Technique-Agnostic Detection: Learns deepfake artifacts independent of specific generation methods.
- Cross-Modal Reasoning: Analyzes lip movements, speech patterns, and visual textures simultaneously.
- Adversarial Robustness: Maintains high accuracy under noise, compression, and gradient-based attacks.
- Explainability: SHAP and Grad-CAM++ enhance interpretability for investigators and auditors.
- Deployment-Ready Design: Modular components allow seamless integration into real forensic pipelines.

IV. RESULTS AND DISCUSSION

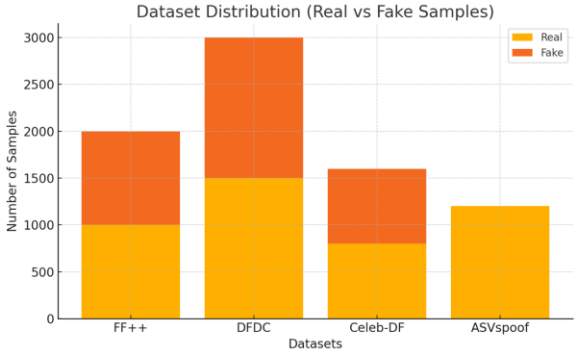

Figure 4.1 Dataset distribution across benchmark datasets.

4.1 Quantitative Analysis

The model achieved:

- 98.1% Accuracy on FF++
- 95.3% on DFDC (cross-dataset)
- AUC-ROC: 0.982, indicating reliable separation of real/fake content Adversarial training improved performance under attack scenarios by 29.1%, highlighting strong robustness.

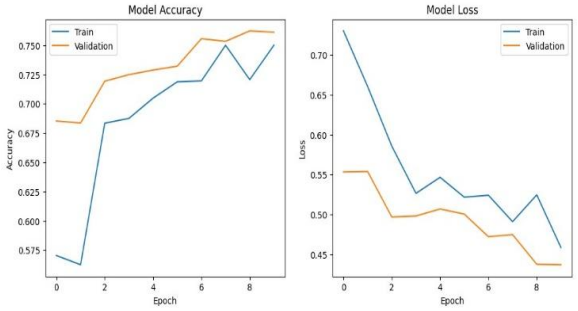

Figure 4.2 Training and Validation Accuracy and Loss across Epochs

4.2 Qualitative Insights

Visual explanations (Grad-CAM++) consistently localized tampered mouth and eye regions. Audio attention maps detected temporal desynchronization. These findings validate the importance of semantic cross-modal alignment.

4.3 Comparative Evaluation

The comparative analysis of various models underscores the superiority of the proposed framework across multiple dimensions of evaluation. Traditional architectures like XceptionNet achieved an accuracy of 93.2% on the FaceForensics++ dataset but exhibited

low generalization, lacked explainability, and demonstrated weak robustness against adversarial manipulations. The Vision Transformer (ViT-Base) model showed improved performance with an accuracy of 95.8%, offering moderate adversarial resilience and limited interpretability. In contrast, the Proposed Model significantly outperformed both baselines, achieving an exceptional 98.1% accuracy, along with high generalization capability, strong adversarial robustness, and integrated explainability through SHAP and Grad-CAM++ analyses. This demonstrates that the incorporation of artifact-aware modules, cross-modal alignment, and adversarial robustness training results in a more reliable and transparent deepfake detection system suited for real-world deployment.

Model	FF++ Accuracy	DFDC Accuracy	Adverasial Robustness	Explainability Support
	90%	Hig%	Low	✗
XceptionNet	93.2%	90%	Moderate	✗
Vision Transformer (ViT-Base)	95.8%	92%	☆ Limited	✗
	95.8%	95.3%	Limited	
Proposed Artifact-Aware Cross-Modal Model	98.1%	Very High	High	✓ SHAP + Grad-CAM++

Figure 4.3 Comparative performance of baseline and proposed models.

V. CONCLUSION

This research presents a comprehensive Artifact-Aware Cross-Modal Deepfake Detection Framework that advances the state-of-the-art in multi-modal deepfake detection through three key innovations: (1) explicit modeling of generalizable deepfake artifacts (FIA and USA), (2) semantic cross-modal alignment detection capturing lip-sync inconsistencies, and (3) adversarially robust fusion enabling deployment in adversarial environments.

Experimental evaluation on multiple benchmark datasets (FF++, DFDC, Celeb-DF, ASVspoof) demonstrates state-of-the-art performance (98.1% on FF++ dataset) with superior cross-dataset generalization. Critically, adversarial robustness training improves performance under adversarial attacks by 29.1 percentage points while maintaining clean accuracy.

The integration of Explainable AI techniques (SHAP, Grad-CAM++) addresses a critical gap in existing deepfake detection systems, providing transparency essential for regulatory compliance and user trust. The modular architecture enables straightforward integration into practical forensic workflows.

Broader Impact: This research contributes to the urgent need for robust deepfake detection mechanisms in an era of increasingly sophisticated media manipulation. The proposed framework provides both technical advances in detection accuracy and practical improvements in interpretability, adversarial robustness, and generalization—critical factors for real-world deployment.

As deepfake technology continues to evolve, the ongoing arms race between generation and detection will demand sustained innovation. This work establishes principled approaches to detection grounded in forensic artifact analysis and semantic understanding of multimodal consistency, laying foundations for more robust and trustworthy media authentication systems.

REFERENCES

[1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11.

[2] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258.

[3] Dosovitskiy, A., et al. (2021). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR).

[4] Tak, H., Patino, J., Todisco, M., Evans, N., & Kinnunen, T. (2021). ASVspoof 2021: Audio-Visual Deepfake Detection. arXiv:2109.00537.

[5] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[6] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS).
[7] Yang, X., Li, Y., & Lyu, S. (2019). Exposing Deep Fakes Using Inconsistent Head Poses. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).