Robust Multimodal Deepfake Detection

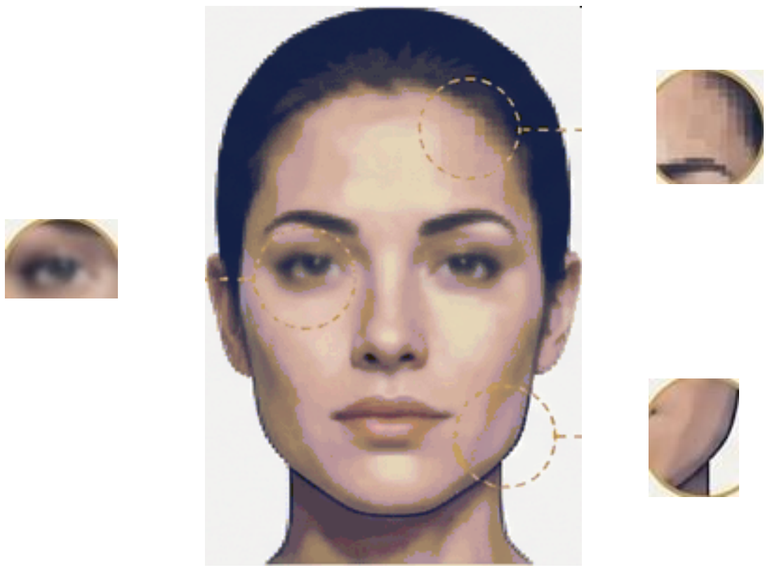
A Statistically Grounded Framework Integrating Audio-Visual, Lip-Sync, and Metadata Cues

Dr.Faisal Bukhari ,Muhammad Yasir

Rao Huzaifa javed

The Deepfake Arms Race: Detection Is Struggling to Keep Pace

The accessibility and realism of deepfake generation tools are outpacing the robustness of existing detection mechanisms, posing serious threats to digital trust and information integrity.



Early Generation Artifacts

Modern Generative Realism

Visual-Only Detectors Are Vulnerable: Strong performance on benchmarks, but significant degradation under real-world conditions (compression, lighting, pose variations). Modern GANs produce fewer detectable visual artifacts.

Audio-Only Detectors Are Insufiicient: Fail when audio is authentic but vfdeo is manipulated, Dr when attackers attackers jointly optimize audio and visual realism.

The Critical Challenge: A lack of generalization across different datasets and manipulation techniques severely limits real-world deployment readiness.

Our Approach: A Robust, Efficient, and Statistically Validated Framework

1. A Robust Multimodal Framework

We integrate four distinct modalities—visual, audio, lip-sync, and metadata—using an effective late fusion strategy.

2. Eflicient Lip-Sync Analysis

We introduce a novel strategy that analyzes lip-sync consistency at inference time **without** requiring additional, costly model training.

3. Statistically Validated Metadata Contribution

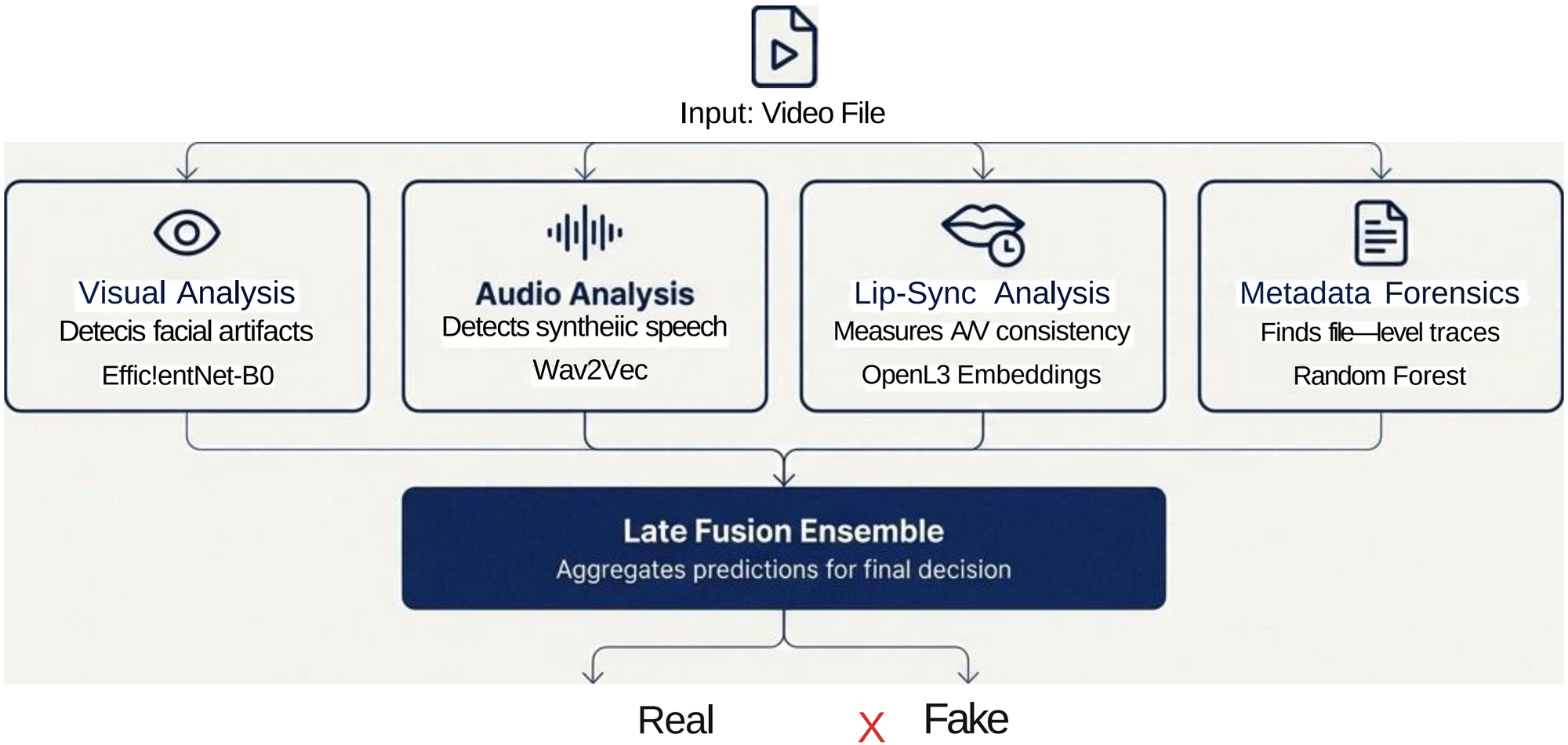
Unlike prior work, we use McNemar's test to rigorously assess whether metadata provides a statistically significant improvement, moving beyond simple accuracy metrics.

4. Proven Generalization

We demonstrate robustness through rigorous cross-dataset evaluation on the unseen LAV-DF benchmark, proving real-world applicability.

Assembling an Elite Team: The Multimodal Detection Architecture

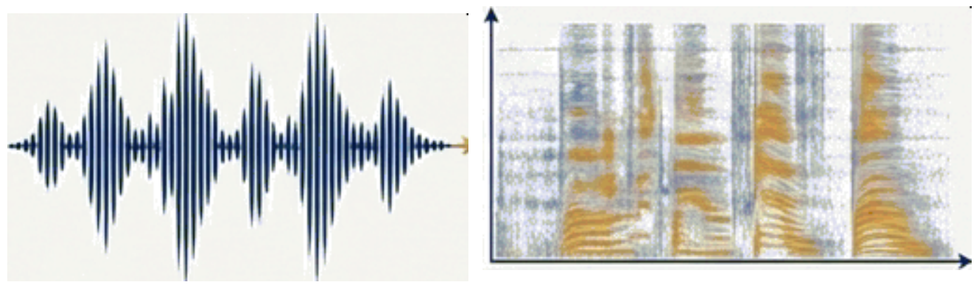
Our framework processes each modality independently using a specialized model. The predictions are then aggregated at the decision level via late fusion, allowing for flexible and powerful integration of heterogeneous forensic cues.



The Content Specialists: Detecting Visual and Auditory Artifacts

Visual Deepfake Detection

- Model: EfficientNet-B0, chosen for its strong performance-efficiency trade-off.
- Input: Face-aligned video frames.(20)
- Target: Captures spatial artifacts and subtle inconsistencies from face manipulation.
- Training Data: Fine-tuned on a subset of the FakeAVCeleb vI.2 dataset.
  99 %accuracy on validation set



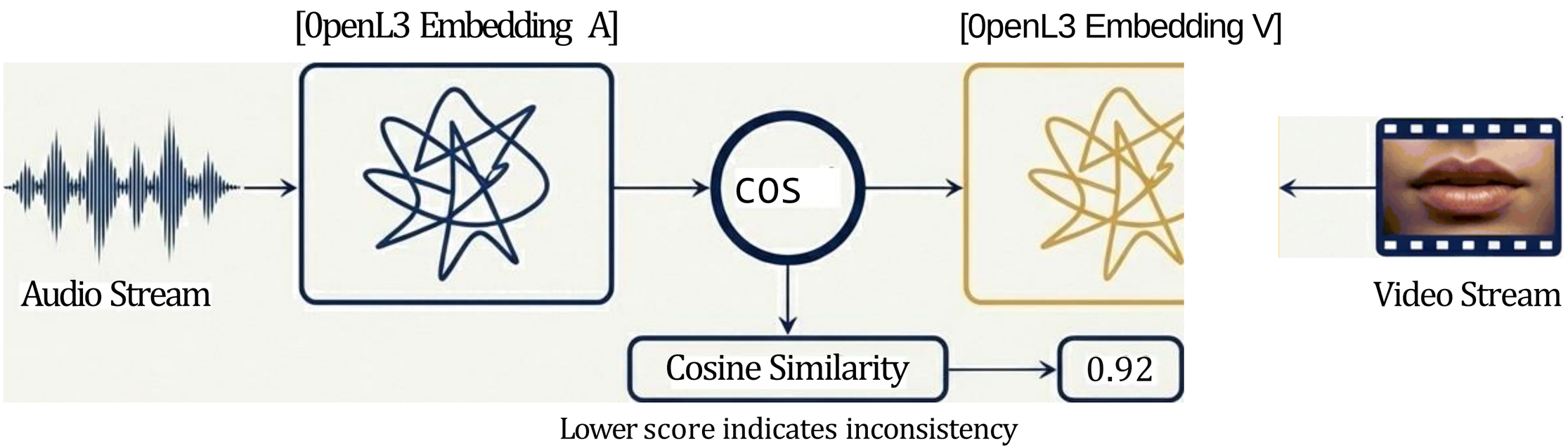
Audio Deepfake Detection

- Model: Wav2Vec, a self—supervised model learning rich features from raw audio.
- Input: Audio waveforms.
- Target: Detects artifacts from voice cloning and text-to-speech systems.
- Training Data: Trained on the full FakeAVCeleb vl.2 audio set (21,000 samples) for robustness.
- Accuracy:99

The Synchronicity Expert: Efficient Lip-Sync Analysis at Inference Time

The Challenge: Traditionallip-sync models require specialized architectures and extensive, costly fine-tuning.

Our Solution: We bypass extra training by performing analysis directly at the fusion stage.



Lower score indicates inconsistency

Key Benefit: This design adds a powerful, complementarymodality with minimal computational overhead, enhancing both efficiency and scalability.

The Forensic Analyst: Uncovering Clues in Video Metadata

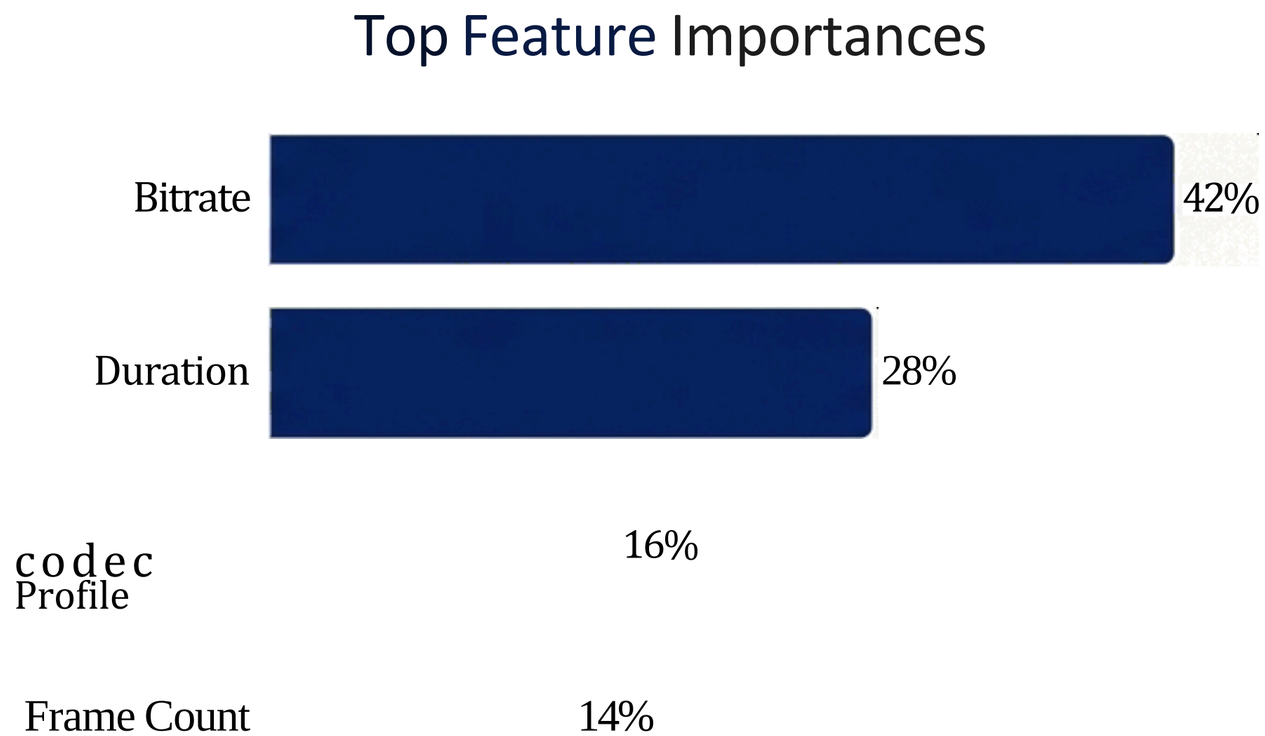
Video metadata can reveal traces of re-encoding and tampering that are not visible at the pixel or audio level.

Features Extracted

- Codec Information: Codec name and profile
- File Attributes: Bitrate, duration, resolution, frame rate
- Container Properties: Number of audio/video frames, encoding attributes

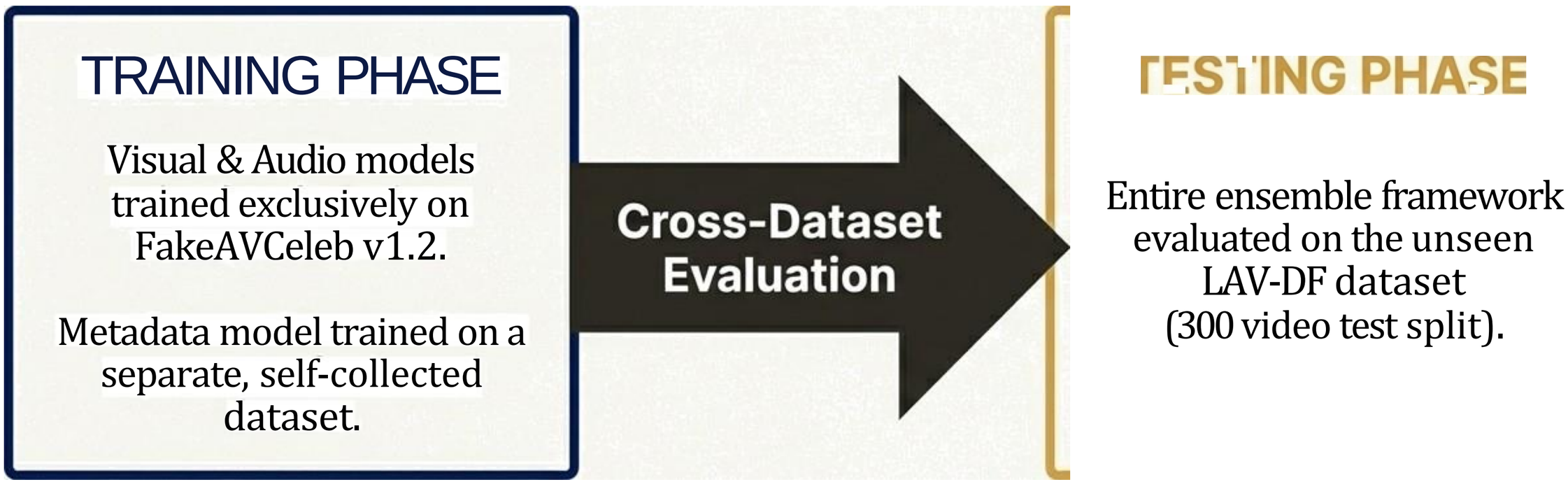
Methodology

- A Random Forest classifier is trained on these heterogeneous features.
- The model learns to distinguish metadata patterns of real vs. manipulated content.



The Gauntlet: A Rigorous Cross-Dataset Evaluation Protocol

To assess real-world robustness and prevent dataset-specific overfltting.



Why This Matters

This strict separation ensures our results reflect the model's ability to generalize to new manipulations, demographics, and recording conditions not seen during training.

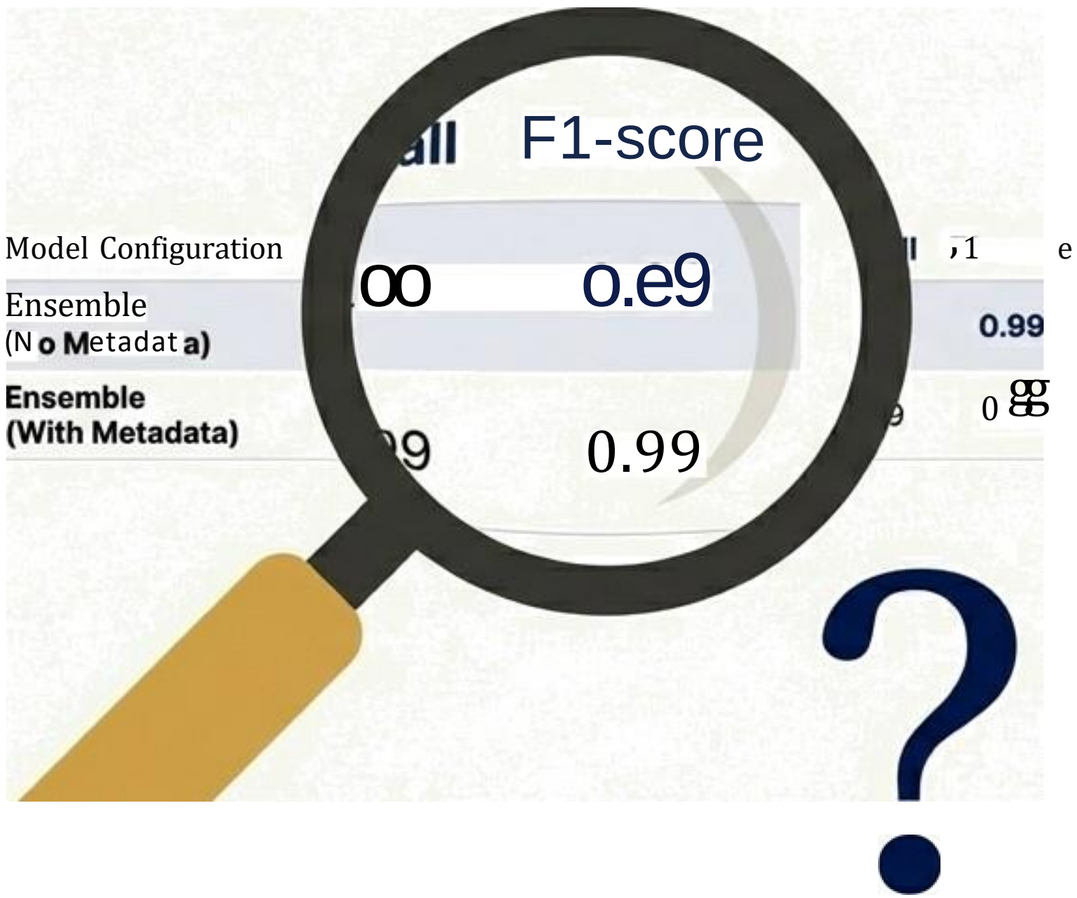
Ensemble Performance: Multimodal Fusion Achieves Excellent Generalization

When tested on the unseen LAV-DF dataset, the late fusion ensemble significantly outperforms individual modality models, confirming its robustness.

Model Configuration	Modalities Used	Accuracy	Precision	Recall	F1-score
Ensemble (No Metadata)	Visual + Audio + Lip-sync	1.00	0.99	1.00	0.99
Ensemble (With Metadata)	Visual + Audio + Lip-sync + Metadata	.99	0.99	0.99	0.99

Main Takeaway: Combining complementary audio-visual and lip-sync cues provides powerful generalization against unseen deepfake manipulations.

The Central Question: Does Metadata Provide a Meaningful Advantage?



The Investigation

We observed a marginal difference in performance when metadata was included. But is this difference real, or just statistical noise?

- Simply comparing aggregate Fl-scores is not enough.
- Yfe need to determine if the change in the model's predictions is statistically significant.

Ablation Study Setup

- Configuration 1: Ensemble without metadata (Visual + Audio + Lip-sync)
- Configuration 2: Ensemble with metadata (Visual + Audio + Lip-sync + Metadata)
- Test: We apply McNemar'é test to the paired predictions from both configurations on the same 300 test videos.

The Verdict: Metadata's Contribution is Not Statistically Significant

McNemar's Test Results

Null Hypothesis (Ho): The inclusion of metadata does not result in a statistically significant difference in performance.

p = 1.0

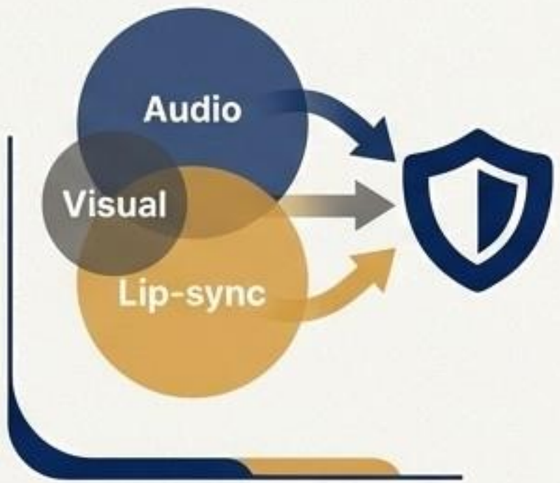
a = 0.05
Significance Threshold

2x2 Contingency Table

	Correct w/ Meta	Incorrect w/ Meta
Correct w/o Meta	296	1
Incorrect w/o Meta	0	3

Since p > 0.05, we do not reject the null hypothesis. The observed performance difference can be be attributed to random variation. The inclusion of metadata provides no statistically significant advantage when combined with strong audio-visual and lip-sync features.

Key Insights and Implications



1. Multimodal Fusion s Essential for Robustness

Our cross-dataset results prove that combining audio-visual and lip-sync cues is a highly effective strategy for creating generalizable deepfake detectors.

2. Strong Content Cues Dominate

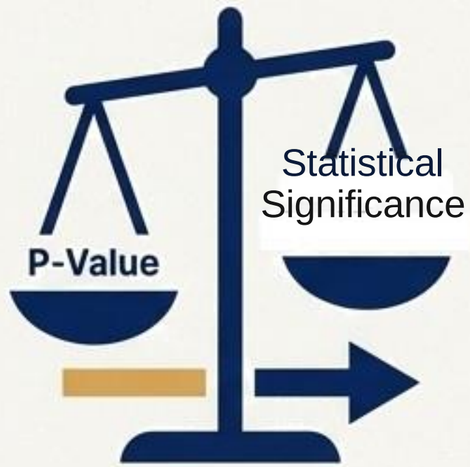
The high performance of the core audio-visual ensemble suggests that these features are powerful enough to make near-perfect classifications on their own in many cases.

3. The Limits of Metadata

While metadata can provide forensic clues in isolation, its contribution becomes negligible when paired with a strong content-based detection system. Not all additional modalities guarantee a meaningful performance gain.

« Statistical Rigor

This work demonstrates the necessity of using statistical tests like McNemar's to validate contributions, providing a more transparent and honest assessment of a system, s components.



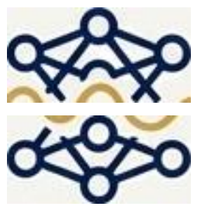
A Practical Framework for Today, with a Roadmap for Tomorrow

Conclusion

We developed a robust multimodal framework that achieves excellent generalization. We proved, through rigorous statistical analysis, that strong audio-visual and lip-sync cues are sufficient for state-of-the-art detection, and that metadata's contribution is not statistically significant in this context.

Our work provides a practical, scalable, and transparently evaluated solution for detecting sophisticated deepfakes.

Future Directions



Temporal Modeling: Exploring transformer architectures to capture long-range dependencies in video and audio.

Adaptive Fusion: Developing confidence-aware strategies to dynamically weight modalities based on content quality.

Expanded Forensics: Integrating social media metadata and provenance information for misinformation detection.

Thank You!