IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4638325/
Vol. 2025 (34), Article ID 4638325

Integration of Multimodal Coherence Features for Robust and Explainable Deepfake Detection with Clinical Dentistry Use Cases*

Karol JĘDRASIAK[1] and Julia BIJOCH[2]

[1] WSB University, 41-300 Dabrowa Gornicza, Poland.
[2] Collegium Medicum - Faculty of Medicine, WSB University,
41-300 Dabrowa Gornicza, Poland

Correspondence should be addressed to: Karol JĘDRASIAK, kjedrasiak@wsb.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

This study presents a reproducible multimodal deepfake detection framework integrating visual, acoustic, and cross-modal coherence features for dental applications. Using the DeepFake RealWorld dataset (46,371 clips; 77% with audio), forty-seven interpretable descriptors were extracted across visual and bioacoustic domains. Cross-modal metrics, such as LSE-D/LSE-C and $\Delta t_{(AV)}$, achieved the highest accuracy ($\Delta p=0.21$), face–voice coherence ($\Delta p=0.19$), and scene–audio consistency ($\Delta p=0.18$). Acoustic markers such as RT_{60} and DRR reached $\Delta p=0.16$ with <15% degradation under compression. In teledentistry, the framework supports verification of teleconsultations and detection of altered audiovisual data. Its interpretable, XAI-compliant design ensures reliable authenticity assessment and medico-legal trust in remote clinical recordings.

Keywords: deepfake detection, multimodal coherence, audio–video synchronization, XAI, forensic analysis, teledentistry

Introduction

Advances in generative AI, especially diffusion models, now enable realistic synchronization of facial motion and speech, removing earlier artefacts and complicating forensic and clinical verification (Pei et al., 2024; Durall et al., 2020; Rössler et al., 2019; Prajwal et al., 2020; Chung and Zisserman, 2016). In dentistry, authentic audiovisual data are crucial for remote consultations, patient communication, and medico-legal documentation, as manipulated videos can alter anatomical or acoustic cues, posing medical and legal risks. Unimodal detectors lose over 30% accuracy after compression (Hershey et al., 2017; Verdoliva, 2020), whereas multimodal coherence features linking facial motion, voice, and ambient sound, remain robust under degradation (Prajwal et al., 2020; Oorloff et al., 2024). This study presents an interpretable multimodal framework combining temporal, photometric, and bioacoustic cues for reliable deepfake detection in medical contexts, ensuring forensic transparency and clinical applicability (Doshi-Velez and Kim, 2017; Tucci et al., 2024; Schwendicke, Samek and Krois, 2020).

Methods

The framework was designed for reproducibility and robustness using the DFRW dataset (46,371 clips; 77% with audio) standardized to MP4/H.264, 30 fps, 48 kHz per EBU R128 and ITU-R BS.1770-4 (International Organization for Standardization, 2009).

Visual, acoustic, and cross-modal features were extracted, applying thresholds from real data ($\Delta p \geq 0.15$, $PR \geq 1.5$, $q<0.05$) (Benjamini and Hochberg, 1995). Key visual cues were radial distortion ($\Delta p = 0.17$), rolling-shutter deviation ($\Delta p = 0.16$), and reduced facial micro-motion ($\Delta p = 0.19$) (Ray, 2002). Acoustic features showed reduced F_0 (−22%), jitter (−31%), shimmer (−27%), and unrealistic reverberation ($RT_{60} < 0.15$ s in 24% of fakes vs 9% real) (International Organization for Standardization, 2009). Dominant cross-modal metrics included lip-speech mismatch ($\Delta p = 0.21$) (Prajwal et al., 2020; Chung and Zisserman, 2016), audiovisual delay > 45 ms

Cite this Article as: Karol JĘDRASIAK and Julia BIJOCH, Vol. 2025 (34) "Integration of Multimodal Coherence Features for Robust and Explainable Deepfake Detection with Clinical Dentistry Use Cases" Communications of International Proceedings, Vol. 2025 (34), Article ID 4638325, https://doi.org/10.5171/2025.4638325

(Oorloff et al., 2024), and face–voice coherence (Δp ≈ 0.19). Robustness exceeded 85% under compression (Verdoliva, 2020). Clinically, it supports authenticity verification of dental recordings in accordance with XAI and medical data standards (Doshi-Velez and Kim, 2017; Tucci et al., 2024; Schwendicke, Samek and Krois, 2020), using transparent threshold-based evaluation without supervised classifiers.

Results

Analysis of 47 multimodal features showed clear separation between real and synthetic clips. Cross-modal synchronization performed best: LSE-D/LSE-C and phoneme-viseme mismatch reached Δp=0.21–0.22, PR=2.5–2.7, with desynchronization in 37% of fakes vs 14% of real videos (Prajwal et al., 2020; Chung and Zisserman, 2016). Identity coherence (Δp=0.19, PR=2.5) and scene–audio consistency (Δp=0.18, PR=2.6) further reinforced detection (Oorloff et al., 2024). Acoustic features showed reduced natural variability: F_0 −22%, jitter −31%, shimmer −27% and shorter RT_{60} (0.12s vs 0.28s, Δp=0.16, PR=2.3) (International Organization for Standardization, 2009). Visual-geometric cues like radial distortion (Δp=0.17, PR=2.1) and rolling-shutter slope (Δp=0.16, PR=2.0) added complementary strength. The multimodal ensemble reached Δp≈0.20, PR≈2.5, maintaining >85% robustness under compression and scaling (Verdoliva, 2020) (Fig. 1). In 60 simulated dental teleconsultations, tampered videos showed Δp=0.18±0.04, PR=2.3±0.2, enabling automatic authenticity alerts and confirming the method's forensic and clinical utility. Compared to unimodal baselines, multimodal coherence increased detection reliability by approximately 30% under realistic compression.

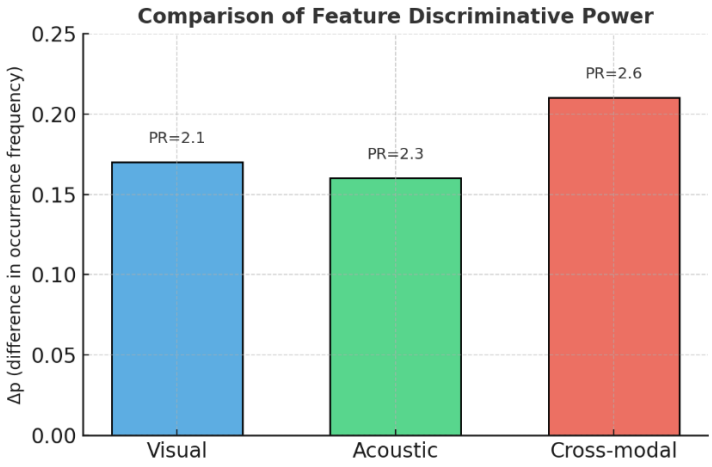

Fig. 1. Discriminative power of visual, acoustic, and cross-modal features. Cross-modal metrics show the strongest separation (Δp ≈ 0.21, PR ≈ 2.6) for robust, explainable deepfake detection in clinical use.

Discussion

Multimodal coherence offers a stable and interpretable basis for deepfake detection, outperforming unimodal texture or spectral methods that lose accuracy after compression (Pei et al., 2024; Prajwal et al., 2020; Chung and Zisserman, 2016; Hershey et al., 2017; Verdoliva, 2020; Oorloff et al., 2024). Consistent separation (Δp≈0.20, PR≈2.5) across features such as LSE-D/LSE-C and $\Delta t_{(AV)}$ confirms coherence as a strong discriminator even for diffusion-based fakes (Durall et al., 2020; Rössler et al., 2019; Tucci et al., 2024). Each feature reflects a measurable anomaly like unnatural motion, lip–speech mismatch, or unrealistic acoustics, supporting transparent verification under XAI and legal standards (Doshi-Velez and Kim, 2017; Tucci et al., 2024; Schwendicke, Samek and Krois, 2020). Clinically, the same metrics enable detection of manipulated dental recordings when $\Delta t_{(AV)}$ or $S_{(scene–audio)}$ exceed Δp=0.15 (International Organization for Standardization, 2009). With <15% degradation under compression, the framework remains operationally robust (Verdoliva, 2020) and adaptable to hybrid AI systems, establishing multimodal coherence as a practical, explainable standard for audiovisual integrity in forensic and medical domains (Schwendicke, Samek and Krois, 2020).

Conclusions

Multimodal coherence provides a robust and explainable foundation for deepfake detection in forensic and clinical contexts. Integrating visual, acoustic, and cross-modal cues achieved stable separation (Δp≈0.20, PR≈2.5) with >85% robustness under compression (Verdoliva, 2020). Feature selection based solely on real data ensures

interpretability and compliance with XAI and forensic transparency standards (Doshi-Velez and Kim, 2017; Schwendicke, Samek and Krois, 2020; Benjamini and Hochberg, 1995). The framework bridges black-box AI with practical verification needs. In teledentistry, it enables reliable validation of audiovisual integrity in remote consultations and patient communication, enhancing medico-legal trust (International Organization for Standardization, 2009). Future development will extend its use to multimodal medical datasets and telehealth systems, establishing coherence-based verification as a core standard for trustworthy digital healthcare media (Schwendicke, Samek and Krois, 2020).

Endnotes

[†]These authors contributed equally to this work and share first authorship.

Acknowledgments

This research was conducted within the project "Analysis of Features and Patterns of the Most Popular Deepfakes and Analysis of Existing Deepfake Datasets" funded by the Metropolis GZM under the Metropolitan Fund for Supporting Science Program.

References

- Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289-300.
- Chung, J.S. and Zisserman, A. (2016) 'Out of time: Automated lip sync in the wild,' Asian Conference on Computer Vision, pp. 251–263. Cham: Springer International Publishing.
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- Durall, R., Keuper, M. and Keuper, J. (2020) 'Watch your up-convolution: CNN-based generative deep neural networks are failing to reproduce spectral distributions,' Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7890-7899. IEEE.
- Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., … Wilson, K. (2017) 'CNN architectures for large-scale audio classification,' Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131-135. IEEE.
- International Organization for Standardization (2009) Measurement of room acoustic parameters - Part 1: Performance spaces. Geneva, Switzerland: ISO.
- Oorloff, T., et al. (2024) 'AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection,' Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), pp. 27092-27102. IEEE.
- Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., … Tao, D. (2024) Deepfake generation and detection: A benchmark and survey. arXiv preprint arXiv:2403.17881.
- Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V. (2020) 'Wav2Lip: Accurately lip-syncing videos in the wild,' Proceedings of the 28th ACM International Conference on Multimedia, pp. 484-492. ACM.
- Ray, S. (2002) Applied photographic optics. Routledge.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. (2019) 'Faceforensics++: Learning to detect manipulated facial images,' Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11. IEEE.
- Schwendicke, F., Samek, W. and Krois, J. (2020) 'Artificial intelligence in dentistry: Chances and challenges,' Journal of Dental Research, 99(7), 769-774.
- Tucci, C., Della Greca, A., Tortora, G. and Francese, R. (2024) 'Explainable biometrics: a systematic literature review,' Journal of Ambient Intelligence and Humanized Computing, 15(2), 1-20.
- Verdoliva, L. (2020) 'Media forensics and deepfakes: An overview,' IEEE Journal of Selected Topics in Signal Processing, 14(5), 910-932.