IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4638225/
Vol. 2025 (34), Article ID 4638225

Deep Network Representations as Reliable Indicators of Synthetic Content in Audiovisual and Clinical Contexts*

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 introduces an interpretable framework for detecting synthetic audiovisual content using deep neural representations, applied to the DeepFake RealWorld (DFRW) dataset (46 371 clips; 77% with audio). Visual, acoustic, and cross-modal embeddings from ResNet, Vision Transformer, SlowFast, Wav2Vec2, and ECAPA-TDNN were evaluated with frequency-based metrics ($\Delta p \geq 0.15$, $PR \geq 1.5$). The strongest indicators were facial embedding variance ($\Delta p = 0.29$, $PR = 3.4$), Mahalanobis distance ($\Delta p = 0.25$), and audiovisual coherence ($\Delta p = 0.23$), all stable within 15% under compression and re-capture. In teledentistry and telemedicine, such explainable AI markers enhance authenticity verification of digital evidence and strengthen medico-legal reliability.

Keywords: deepfake detection; deep neural representations; multimodal coherence; audiovisual forensics; explainable AI; telemedicine; teledentistry; data integrity

Introduction

Advances in generative models have turned deepfakes into a real threat to data integrity and digital trust (Amodei, Hernandez and Sastry 2018; Chesney and Citron 2019; Hashmi et al. 2024). Beyond disinformation, synthetic audiovisual materials increasingly endanger healthcare, where authenticity of diagnostic images and teleconsultations is essential for medico-legal reliability (Schwendicke, Samek and Krois 2020). Existing detectors, based on opaque convolutional classifiers, degrade under compression or adversarial perturbations (Marcel 2018; Zou et al. 2024) and lack calibrated, interpretable outputs limiting clinical use (Guo et al. 2017). This study applies deep network embeddings: visual, acoustic, and multimodal, as measurable indicators of synthetic origin, validated on DFRW data. Emphasizing explainability and robustness, it links multimedia forensics with clinical reliability in verifying medical recordings.

Materials and Methods

The study used the DeepFake RealWorld (DFRW) dataset of 46 371 audiovisual clips (77% with audio) reflecting realistic conditions such as compression, resampling, and re-capture. All clips were standardized to MP4/H.264, 30 fps, 48 kHz. Embeddings were extracted from pretrained models (Wang and Huang 2024): ResNet50, Vision Transformer (ViT), and SlowFast for visual and temporal data, and Wav2Vec2 and ECAPA-TDNN for audio. Statistical descriptors included intra-identity variance, Mahalanobis distance, and audiovisual coherence.

A frequency-based evaluation without classification was applied: anomaly thresholds were defined on authentic samples, and indicators retained if $\Delta p \geq 0.15$ or $PR \geq 1.5$ with $p\_real \leq 20\%$. Stability under degradations was confirmed with ≤15% performance drop, 95% bootstrap confidence, and FDR control ($q < 0.05$). Explainable-AI

Cite this Article as: Karol JĘDRASIAK and Julia BIJOCH, Vol. 2025 (34) "Deep Network Representations as Reliable Indicators of Synthetic Content in Audiovisual and Clinical Contexts " Communications of International Proceedings, Vol. 2025 (34), Article ID 4638225, https://doi.org/10.5171/2025.4638225

tools (Grad-CAM, attention maps) ensured interpretability and traceability, supporting reproducible forensic and clinical validation of synthetic-content indicators.

Results and Discussion

Analysis of high-level embeddings confirmed clear separation between authentic and synthetic recordings. Facial identity variance ($Var_{(track)}$) reached $\Delta p = 0.29$ and PR = 3.4, indicating instability of reconstructed facial features. Mahalanobis distance achieved $\Delta p = 0.25$ and PR = 3.1, revealing systematic deviations in latent space. Audiovisual synchronization (Wav2Vec2–SyncNet) yielded $\Delta p = 0.23$, PR = 3.0, and audio–identity mismatch (ECAPA-TDNN–ArcFace) $\Delta p = 0.21$, PR = 2.8. All indicators remained stable under degradations with ≤15% loss (Fig. 1).

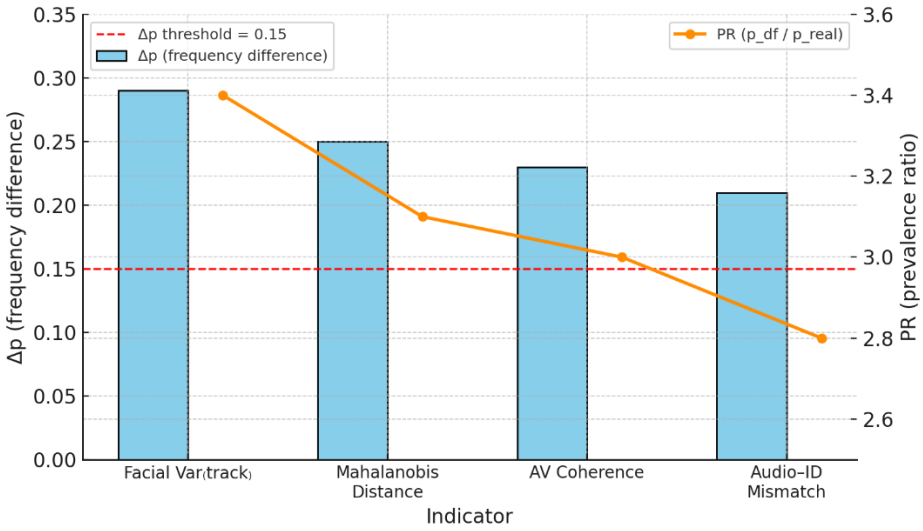

Fig. 1. Comparative performance of deep embedding indicators.

Cross-modal coherence features outperformed black-box detectors in cross-distribution tests (Verdoliva 2020; Nie et al. 2024; Khan, Khan and Ahmad 2025). The stability of Δp and PR across compression and re-capture levels supports their use as explainable forensic metrics. In teledentistry and telemedicine, these markers enable verification of audiovisual integrity, prevent tampering, and strengthen medico-legal confidence. Results confirm that deep embeddings are interpretable, resilient indicators of synthetic origin applicable to clinical data assurance.

Clinical and Forensic Relevance

Authenticity verification of audiovisual data is becoming essential in clinical environments, particularly in teledentistry and telemedicine, where recordings of examinations, diagnostic images, and treatment consultations may serve as legal or evidential material. Embedding-based indicators of synthetic origin can detect subtle inconsistencies in facial motion, lip synchronization, or acoustic patterns that may indicate manipulation. Their explainable structure and calibrated uncertainty allow transparent reporting in medico-legal workflows and compliance with data integrity standards such as ISO/IEC 42001. By integrating these interpretable deep features into clinical systems, healthcare institutions can enhance digital trust, protect patient records, and strengthen forensic reliability in medical documentation.

Conclusions and Future Work

Deep network embeddings constitute stable, interpretable, and reproducible indicators of synthetic origin in audiovisual data. Their robustness across compression, resampling, and re-capture conditions confirms their suitability for real-world medical and forensic applications. Integrating such explainable metrics into telemedicine and dental documentation workflows enhances data authenticity and legal credibility. Future research will focus on extending evaluation to DFRW v2 with ≥500,000 clips, multimodal calibration, and large-scale validation in clinical contexts to ensure reliable deployment in healthcare and evidence-based AI governance (Zhang et al. 2025).

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

- Ahmad, A., Khan, I. and Khan, K. (2025) 'A Comprehensive Survey of DeepFake Generation and Detection Techniques in Audio-Visual Media,' ICCK Journal of Image Analysis and Processing, 1(2), 73-95.
- Amodei, D., Hernandez, D. and Sastry, G. (2018) AI and compute. OpenAI Blog. [Online]. Available: https://openai.com/blog/ai-and-compute
- Chesney, R. and Citron, D. (2019) 'Deepfakes and the new disinformation war: The coming age of post-truth geopolitics,' Foreign Affairs, 98, 147-155.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) 'On calibration of modern neural networks,' Proceedings of the International Conference on Machine Learning, 1321-1330. PMLR.
- Hashmi, A., Shahzad, S.A., Lin, C.W., Tsao, Y. and Wang, H.M. (2024) 'Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights,' arXiv preprint arXiv:2411.07650.
- Marcel, P. (2018) DeepFakes: a new threat to face recognition. Assessment and detection. [Manuscript].
- Nie, F., Ni, J., Zhang, J., Zhang, B. and Zhang, W. (2024) 'DIP: diffusion learning of inconsistency pattern for general deepfake detection,' IEEE Transactions on Multimedia.
- Schwendicke, F.A., Samek, W. and Krois, J. (2020) 'Artificial intelligence in dentistry: Chances and challenges,' Journal of Dental Research, 99(7), 769-774.
- Verdoliva, L. (2020) 'Media forensics and deepfakes: An overview,' IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.
- Wang, Y. and Huang, H. (2024) 'Audio-visual deepfake detection using articulatory representation learning,' Computer Vision and Image Understanding, 248, 104133.
- Zhang, B., Cui, H., Nguyen, V. and Whitty, M. (2025) 'Audio deepfake detection: What has been achieved and what lies ahead,' Sensors (Basel, Switzerland), 25(7), 1989.
- Zou, H., Shen, M., Hu, Y., Chen, C., Chng, E.S. and Rajan, D. (2024) 'Cross-modality and within-modality regularization for audio-visual deepfake detection,' Proceedings of ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing, 4900–4904. IEEE.