Adaptive Multi-Modal Deepfake Detection for Safer Learning Environments

Sama, S. S., Chakraborty, S., & Mian, S.

Adaptive Systems | Deepfake Detection | Educational Integrity | learning engineering | Multi-modal AI

Extended Abstract

Introduction

Generative AI now enables highly convincing deepfake audio and video, creating risks for education such as impersonated instructors, falsified assignments, and compromised proctoring. Although many detectors exceed 95% accuracy in laboratory tests, their performance often drops to 45–70% in real-world conditions due to dataset shift and rapidly evolving generative methods, limiting trust and adoption. From a Learning Engineering perspective, these failures represent a system-level breakdown in which learning technologies do not reliably support instructional integrity, trust, and equity. Our work asks: How can deepfake detection remain accurate, explainable, and fair enough for real educational deployment?

Approach or Design

We present a dynamic multi-modal ensemble framework that treats deepfake detection as an adaptive, data-informed service within learning platforms. The system integrates three specialized modes: Visual Sentinel for image artifacts, Acoustic Guardian for speech cues, and Temporal Oracle for video dynamics, each combining multiple state-of-the-art models. A gating network adaptively weights their contributions based on content characteristics and historical performance. Framed as a nested Learning Engineering cycle, the approach begins with identifying an authentic educational risk, implements multimodal AI solutions, evaluates performance in real learning contexts, and iteratively adapts the system based on empirical evidence. The pipeline updates itself as usage data and error patterns indicate where retraining, reweighting, or fairness adjustments are

needed, and human-centered design principles guide the use of explainable Grad-CAM visualizations to help instructors and administrators understand why content is flagged.

Findings or Insights

Across DF40's 40 generation methods and several public benchmarks, individual modes achieve 96–98% accuracy, and the dynamic ensemble reaches 97.3% in laboratory settings while reducing false positives to approximately 1.3%. In real-world evaluations using Deepfake-Eval-2024, the system maintains 89.7% accuracy while baselines fall to 45–70%, reducing the laboratory-to-deployment performance gap from as high as 48% to 7.6%. Cross-modal verification identifies 73% of sophisticated deepfakes missed by single-modality detectors. Early fairness analysis shows 3–7% performance gaps across demographic groups, motivating further bias mitigation. We also received feedback from AAAI reviewers, which we plan to incorporate into the next stage of development. From a Learning Engineering standpoint, these findings illustrate how iterative evaluation across controlled and authentic contexts can reduce deployment risk and improve system trustworthiness in learning environments.

Table 1: Performance Metrics of Multi-Modal Ensemble Modes

Table 1 summarizes the performance metrics for each of the three specialized detection modes. By utilizing a dynamic weighted ensemble of transformer-based architectures, the system achieves high precision and AUC across all modalities. These high levels of accuracy provide a reliable foundation for maintaining educational integrity in digital learning environments where generative AI risks are prevalent.

Model Ensemble	Accuracy	Precision	Recall	F1-Score	AUC
Visual Sentinel (Image)	97.9%	98.2%	97.5%	97.8%	0.993
Acoustic Guardian (Audio)	96.2%	96.8%	95.7%	96.2%	0.984
Temporal Oracle (Video)	97.3%	97.8%	96.9%	97.3%	0.991

Implications

This work shows how adaptive ensemble architectures and cross-modal verification can turn deepfake detection into a continuously improving service suited to real educational needs. The framework can be integrated into learning management systems, virtual classrooms, and proctoring tools to verify the authenticity of student video submissions, provide explainable alerts rather than opaque classifications, and adapt as new generative models emerge. More broadly, the project reflects a learning-engineering cycle in which an educational risk is identified, multimodal AI methods are applied, real-world robustness and fairness are evaluated, and interpretable insights guide policy and instructional practice.

References

Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training

environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 conference proceedings: Solving for complexity at scale (pp. 115–132). https://doi.org/10.59668/2109.21735

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://arxiv.org/abs/2010.11929

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. Journal of Military Learning. https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/