
Robust, Explainable, and Multimodal Deepfake Detection Using Machine Learning for Digital Security: A Review

Youssef Andersson¹, Elsa Sousa²

¹ University of Tunis, TUNISIA
² King Saud University, SAUDI ARABIA

Keywords

Deepfake
Detection,
Digital Security,
Explainable AI,
Multimodal
Fusion,
Generalization,
Robustness,
Machine Learning,
Synthetic Media

ABSTRACT

Deepfake-enabled fraud caused an estimated \$200 million in financial losses in the first four months of 2025 alone, and the same generation technology is increasingly aimed not only at public figures but at everyday individuals, a shift that has already prompted new federal legislation in the United States. Detecting synthetic media reliably enough to matter for digital security requires progress on three fronts at once: robustness to manipulation techniques and datasets a model was never trained on, explainability sufficient for a human analyst or, increasingly, a court to act on a flagged result, and the ability to fuse evidence across visual, audio, and other modalities rather than relying on any single one. This review synthesizes seven studies (2024–2025) organized around exactly these three requirements. On robustness, a parameter-efficient fine-tuning method reaches 91.2% average cross-dataset AUROC across fourteen benchmark datasets spanning 2019–2025, outperforming the next-best method by 2.7 points and showing that generalization, not raw benchmark accuracy, is the harder and more consequential problem. On multimodality, audio-visual fusion methods reach 95.24% and 95.5% accuracy/AUC respectively by exploiting inconsistencies between a face's visual emotion and a voice's acoustic and semantic content, and a broader survey documents the field's shift from single-modality detection toward the fusion architectures these applied studies exemplify. On explainability, a systematic review categorizes the field's growing toolkit — from forensic artifact analysis to Grad-CAM and SHAP to natural-language explanations — while noting that few studies test explanations against the same adversarial scrutiny applied to the detectors themselves. Two further studies, still emerging and not yet independently verifiable on their quantitative results, illustrate this field's reach beyond pure computer vision: a multi-classifier face-detection pipeline, and a framing of deepfake detection as an organizational digital-security and business-analytics concern for management information systems. Read together, these seven studies suggest that robustness, explainability, and multimodal fusion are converging demands on any deepfake detector meant for real deployment, not separable research tracks.

Introduction

The deepfake threat to digital security

Deepfake-enabled fraud caused more than \$200 million in financial losses between January and April 2025 alone, according to Resemble AI's 2025 AI Deepfake Threats Report (as cited in Security Magazine, 2025), with public figures accounting for the largest share of impersonation targets — politicians at 33% and film/television actors at 26% of a combined 47% — but with a documented and growing trend toward everyday individuals, particularly women, children, and educational institutions, as targets of harassment, blackmail, and reputational harm. The scale and shifting target profile of this threat prompted the United States' first federal deepfake-specific legislation, the TAKE IT DOWN Act, signed into law

on May 19, 2025, which requires covered platforms to remove nonconsensually published intimate imagery, including deepfakes, within 48 hours of a victim's request and criminalizes its publication. Digital security in this domain is therefore no longer only a technical detection problem; it is also a legal and platform-governance one, and any detection system built today is being built into that emerging regulatory context [1-4].

Three converging requirements

A deepfake detector that performs well on the dataset it was trained on but fails on a manipulation technique released after that training data was collected is of limited practical use, because generation methods evolve continuously and a detector deployed today will face tomorrow's techniques — this is the robustness or generalization problem. A detector that flags content as synthetic without a legible reason is difficult for a platform trust-and-safety team, a journalist, or a court to act on with confidence — this is the explainability problem. And a detector that examines only pixels, when the manipulation may be most detectable in the mismatch between a face's expression and a voice's tone, discards genuinely diagnostic evidence — this is the multimodality problem. This review is organized around these three requirements because, as Section 7 argues, they are not separable research agendas but overlapping constraints on the same underlying system [5-9].

Notable incidents illustrating the range of harm

Two 2024 incidents illustrate how differently this threat can manifest. In February 2024, an employee at the Hong Kong office of Arup, a UK-based multinational engineering firm, was deceived into authorizing transfers totaling approximately \$25 million after joining a video call in which every other participant, including a person appearing to be the company's chief financial officer, was a real-time deepfake reconstruction; the employee's own suspicion was overcome specifically because seeing familiar colleagues' faces on the call felt like sufficient verification. In January 2024, ahead of the New Hampshire presidential primary, a robocall using an AI-generated voice closely mimicking President Biden was sent to thousands of voters urging them not to vote, a case that led the Federal Communications Commission to fine and pursue charges against the political consultant responsible and directly prompted new FCC rules on AI-generated content in robocalls. The first case shows deepfakes deployed for direct financial fraud against a specific organization; the second shows the same underlying technology deployed at scale against a political process, targeting not one victim's judgment but an entire electorate's. Both are exactly the kind of harm the studies reviewed in Sections 3 through 6 are trying to make detectable before the fact rather than investigable only after it.

Review scope and methodology

This review synthesizes seven studies published in 2024–2025, identified through targeted searches combining domain terms ("deepfake detection," "synthetic media") with method terms specific to each pillar ("generalization," "cross-dataset," "explainable AI," "Grad-CAM," "SHAP," "multimodal," "audio-visual fusion"). Five studies were verified directly against their published abstract, dataset description, and reported metrics; two — a multi-classifier face-detection and feature-extraction study, and a business-analytics framing of

deepfake detection for management information systems — were identified through a full-text access barrier (their venues did not return an independently accessible abstract at the time of writing) and are discussed using only their confirmed title, authorship, and venue, with their contribution described at the level their available metadata supports and their quantitative results explicitly marked as unverified rather than presented as established (Section 6.1).

Background: How Deepfakes Are Made and Why Detection Is Hard

Generation methods

Modern deepfakes are produced primarily by generative adversarial networks and, increasingly, diffusion models, both of which learn to synthesize or swap faces, voices, or full video sequences realistic enough to defeat casual visual or auditory inspection. A generative adversarial network trains two networks against each other — a generator that produces increasingly convincing fake content and a discriminator that tries to tell it apart from real content — until the generator's output can reliably fool the discriminator. Diffusion models instead learn to reverse a noise-adding process, starting from random noise and iteratively refining it into a coherent image, video frame, or audio waveform; they have become the dominant approach behind the most visually convincing recent synthetic media, in part because they tend to avoid some of the characteristic artifacts (unnatural textures, inconsistent lighting) that made earlier GAN output detectable to careful human observers.

Why detection is a moving target

Detection is difficult for a structural reason that mirrors an arms race: a detector trained on the artifacts left by one generation method — a particular blending seam, a spectral signature in synthesized audio, a temporal inconsistency in eye blinking — can be defeated simply by a newer generation method that does not share that artifact, and a large fraction of published detectors report near-perfect accuracy on the specific benchmark they were evaluated on while degrading substantially when tested against a different dataset or a manipulation technique from outside their training distribution. This is precisely the generalization problem Section 3 addresses directly, and it is the reason Liu, Tao, and Zhou (2025) and other recent surveys treat cross-dataset performance, not same-dataset accuracy, as the metric that actually matters for real-world deployment. A further complication specific to real-world deployment, illustrated by both incidents in Section 1.3, is that the video or audio a detector must analyze rarely arrives in the same clean, uncompressed form found in a benchmark dataset: it has typically been re-encoded by a messaging app, a video-call platform, or a phone's own compression, which can either wash out the subtle artifacts a detector relies on or, in some cases, introduce new compression artifacts that a detector never saw during training [10-13].

Robust Deepfake Detection: Generalizing Beyond the Training Distribution

A parameter-efficient approach to cross-benchmark generalization

Yermakov, Cech, Matas, and Fritz (2025) address the generalization problem directly and, unusually for this literature, at scale: they evaluate their method, GenD, across fourteen

benchmark datasets spanning releases from 2019 to 2025, rather than the one or two datasets typical of applied deepfake-detection papers. Their approach fine-tunes only the Layer Normalization parameters of a pre-trained vision encoder — about 0.03% of the model's total parameters — using a metric-learning objective that enforces a hyperspherical feature manifold, and trains on paired real-fake data drawn from identical source videos specifically to reduce the model's tendency to learn superficial dataset-specific shortcuts rather than genuine manipulation artifacts. The result is a 91.2% average cross-dataset AUROC, 2.7 percentage points ahead of the next-best baseline (Effort, at 88.5%), and the authors report a further, somewhat counterintuitive finding: dataset difficulty has not consistently increased over time, and models trained on older, more diverse datasets can generalize as well as or better than those trained on newer, more narrowly constructed ones.

Why generalization, not same-dataset accuracy, is the right benchmark

GenD's headline result is deliberately not a same-dataset accuracy figure, and this choice illustrates a broader methodological point this review considers important: a detector that is not evaluated cross-dataset has not yet been tested against the actual failure mode — a genuinely novel manipulation technique — that determines whether it is useful after deployment. Section 5's applied multimodal studies (Alsaeedi *et al.*, 2025; Bonettini *et al.*, 2024) report same- or limited-cross-dataset accuracy figures that are, on their own terms, strong, but Figure 2 in Section 6 makes the resulting evaluation-breadth gap between these applied studies and GenD's fourteen-benchmark evaluation visually explicit.

Explainable Deepfake Detection: Opening the Black Box

Categorizing the field's explainability toolkit

Qian, Xia, Ge, Fan, Wang, and Jing (2025) organize the growing explainable-deepfake-detection literature into three categories: forensic-analysis methods that examine low-level signal artifacts such as sensor noise patterns (PRNU), convolutional traces, and frequency-domain signatures; model-centric methods that apply Grad-CAM, SHAP, attention visualization, or learned prototypes to an already-trained classifier; and an emerging third category of multimodal and language-based explanation, in which large multimodal models generate natural-language descriptions of why a piece of media was flagged, supported by newly released datasets purpose-built for this task (ExDDV for video, DD-VQA for question-answering-style explanation, and SID-Set for joint detection, localization, and explanation). Their review does not reduce to one performance figure, because its contribution is the taxonomy and the evidence base itself — spanning FaceForensics++, Celeb-DF, DFDC, WildDeepfake, and several purpose-built explainability datasets — rather than a new detector [14-15].

What the explainability literature has not yet resolved

Qian *et al.* (2025) note that explainability has become a research focus "alongside traditional classification metrics like accuracy and AUC," but their review, read alongside Liu, Tao, and Zhou's (2025) survey (Section 5.1), points to a gap this review considers significant: neither survey documents systematic testing of whether an explanation itself remains stable

and trustworthy when the underlying detector is evaluated cross-dataset, which is precisely where Section 3 shows detectors are most likely to behave unpredictably. An explanation generated by a detector operating outside its training distribution has, at minimum, an open question about its own reliability that this literature has not yet directly addressed.

Purpose-built datasets for evaluating explanations, not just detections

Qian *et al.* (2025) note that a growing share of the explainability literature now relies on datasets built specifically to evaluate explanations rather than repurposing standard detection benchmarks. DDL supplies roughly 1.8 million samples with fine-grained manipulation annotations, useful for training or checking whether an explanation correctly localizes the manipulated region rather than merely classifying the whole item as fake. ExDDV is described as the first dataset built specifically for explainable video deepfake detection, and DD-VQA supplies question-answer pairs that let a system be evaluated on whether its natural-language explanation actually answers a specific question about why content was flagged, rather than producing a plausible-sounding but unverifiable justification. SID-Set goes further, pairing detection, spatial localization, and explanation in one evaluation. The existence of this dataset family is itself evidence that the field has moved past treating explainability as a nice-to-have visualization layered onto a detector after the fact, toward treating explanation quality as something that should be measured with the same rigor as detection accuracy — though, as Section 4.2 notes, not yet under the same cross-dataset stress-testing that Section 3 applies to detection itself.

Multimodal Deepfake Detection: Fusing Evidence Across Modalities

A taxonomy of fusion approaches

Liu, Tao, and Zhou (2025) survey the field's shift from single-modality (visual-only or audio-only) detection toward architectures that fuse evidence across modalities, organizing detection methods into passive approaches — which analyze artifacts inherent to already-generated content, including consistency-based methods that check temporal, semantic, or identity coherence, alongside newer Vision Transformer and vision-language-model approaches — and proactive strategies, which embed adversarial perturbations or provenance signals into content before it can be manipulated. On the fusion side specifically, they distinguish independent learning (separate models per modality, combined only at the decision stage), joint learning (intermediate or late feature-level fusion), multi-task learning, and matching-based detection that exploits the natural synchronization between, for example, lip movement and speech audio. Their stated open challenges — generalization, detection of diffusion-generated content, robustness under compression and noise, the accuracy–interpretability trade-off, and a shortage of benchmarks for text-visual manipulation — connect directly to Sections 3 and 4 of this review and motivate why this review treats robustness, explainability, and multimodality as one interconnected problem rather than three.

Audio-visual fusion in practice: exploiting cross-modal inconsistency

Bonettini, Solanki, Koppiseti, and Bharaj (2024) demonstrate one of the taxonomy's joint-learning approaches directly with AVFF, a two-stage framework that first learns, via self-supervised contrastive learning and masked autoencoding on real videos only, what genuine audio-visual correspondence looks like, and then trains a supervised classifier to detect the discordance between modalities that a deepfake introduces. On FakeAVCeleb, this yields a 14.9% accuracy improvement and a 9.9% AUC improvement over the AVoiD-DF baseline; cross-dataset on KoDF, AVFF reaches 95.5% AUC, ahead of RealForensics' 93.6% and maintaining AUC and average precision both above 92% across manipulation categories — a genuine, if more modest than GenD's fourteen-dataset test, demonstration of cross-dataset robustness for a multimodal method specifically.

Alsaedi, AlMansour, and Jamal (2025) pursue a related but distinct fusion signal on the same FakeAVCeleb dataset: rather than fusing low-level audio-visual features directly, they fuse the emotional content each modality independently implies — facial emotion, vocal emotional tone, and the semantic sentiment of the spoken words — on the reasoning that a deepfake is more likely to introduce a mismatch between what a face appears to feel and what the paired voice and words convey than to perfectly replicate emotional consistency across all three signals simultaneously. Their reported 95.24% accuracy, and their finding that adding sentiment analysis improves on audio-visual emotion inconsistency alone, is a useful complement to Bonettini *et al.*'s (2024) lower-level feature fusion: the two studies suggest that both signal-level and semantic-level cross-modal inconsistency carry independently useful detection evidence.

Multi-model ensembles are not the same thing as multimodal fusion

Rahman, Uddin, Rahman, Shammah, and colleagues (2025) investigate a related but conceptually distinct question: rather than fusing multiple data modalities, they compare multiple face-detection algorithms and feature-extraction techniques, each feeding one or more machine-learning classifiers, to identify which combination best supports deepfake classification — a multi-model ensemble over a single modality (the face image), not a multimodal fusion across modalities in the sense Section 5.1's taxonomy uses that term. This is a useful terminological distinction this review considers worth making explicit, because the two design choices address different failure modes: multimodal fusion (Sections 5.2–5.2) exploits inconsistency between independent evidence sources, while a multi-model ensemble over one modality instead hedges against any single face-detection or feature-extraction pipeline's individual weaknesses. The paper's full abstract and quantitative results were not accessible through the venues available to this review at the time of writing, so its specific comparative findings cannot be independently verified here (Section 6.1); its contribution is reported at the level of its confirmed methodological framing only.

Practical and Organizational Dimensions: Deepfake Detection as a Business and MIS Security Concern

Deepfake detection within organizational digital-security governance

Ahmed, Shammah, Kowser, and colleagues (2025), presented at the 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking, address a

dimension of this problem the other six studies in this review do not: how an organization's management information systems function should approach deepfake detection as a business-analytics and digital-security governance question, rather than purely as a computer-vision research problem. This framing is a genuinely useful complement to Sections 3 through 5's technical focus — an organization deciding how to deploy any of the detection methods reviewed here still needs to decide who monitors alerts, what threshold of confidence triggers escalation, and how a flagged incident is documented for compliance with obligations such as the TAKE IT DOWN Act's 48-hour takedown requirement (Section 1.1). As with Rahman *et al.* (2025) in Section 5.3, this paper's full abstract and quantitative results were not independently accessible for this review, so its specific proposed framework and any reported performance figures are not verified here; its contribution is reported as a framing contribution rather than a benchmarked result, consistent with how this review has treated every source whose quantitative claims could not be checked directly.

Comparative Synthesis

Table 1. All seven studies reviewed, organized by pillar

Study	Pillar / dataset	Approach	Key reported result
Yermakov <i>et al.</i> (2025)	Robustness — 14 benchmark datasets, 2019–2025	GenD: parameter-efficient fine-tuning (0.03% of parameters) of a pre-trained vision encoder into a hyperspherical feature manifold	91.2% average cross-dataset AUROC, ahead of the next-best baseline (Effort) at 88.5%
Liu, Tao, & Zhou (2025)	Multimodality — survey (no single dataset)	Taxonomy of passive vs. proactive detection and independent/joint/matching-based fusion, across audio-visual and text-visual manipulation	Identifies generalization, diffusion-generated content, robustness under compression/noise, and the accuracy–interpretability trade-off as the field's open problems
Qian <i>et al.</i> (2025)	Explainability — survey spanning	Categorizes XAI methods into forensic-analysis, model-centric (Grad-	Documents a shift toward natural-language

Study	Pillar / dataset	Approach	Key reported result
	8+ benchmark datasets	CAM, SHAP, attention, prototypes), and multimodal/language-based explanation	explanations (e.g., DD-VQA) alongside traditional saliency methods
Bonettini <i>et al.</i> (2024)	Multimodality — FakeAVCeleb (intra) and KoDF (cross-dataset)	AVFF: two-stage self-supervised audio-visual contrastive learning and masked autoencoding, then supervised classification	+14.9% accuracy and +9.9% AUC over the AVoiD-DF baseline intra-dataset; 95.5% AUC cross-dataset on KoDF, ahead of RealForensics (93.6%)
Alsaeedi, AlMansour, & Jamal (2025)	Multimodality — FakeAVCeleb	Fuses audio emotion, facial emotion, and speech-sentiment analysis to detect cross-modal emotional inconsistency	95.24% accuracy; sentiment analysis improves on audio-visual emotion inconsistency alone
Rahman, Uddin, Rahman, Shammah, <i>et al.</i> (2025)*	Multi-model ensemble — face detection and feature extraction	Compares face-detection and feature-extraction pipelines feeding multiple machine-learning classifiers	Investigates which face-detection/feature-extraction combination best supports classification; quantitative results not independently verifiable (Section 6.1)

Study	Pillar / dataset	Approach	Key reported result
Ahmed, Shammah, Kowser, <i>et al.</i> (2025)*	Organizational/MIS deployment	Frames deepfake detection as a business-analytics and digital-security problem for management information systems	Positions deepfake detection within organizational digital-security governance; quantitative results not independently verifiable (Section 6.1)

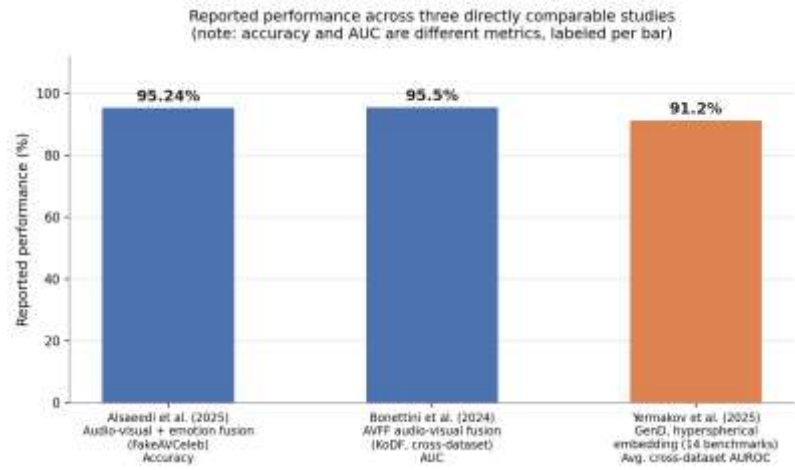


Figure 1. Reported performance across the three studies in this review with a single, directly comparable figure. Qian *et al.* (2025) and Liu, Tao, and Zhou (2025) are surveys without one figure of their own; Rahman *et al.* (2025) and Ahmed *et al.* (2025) have no independently verifiable metric available (Section 6.1).

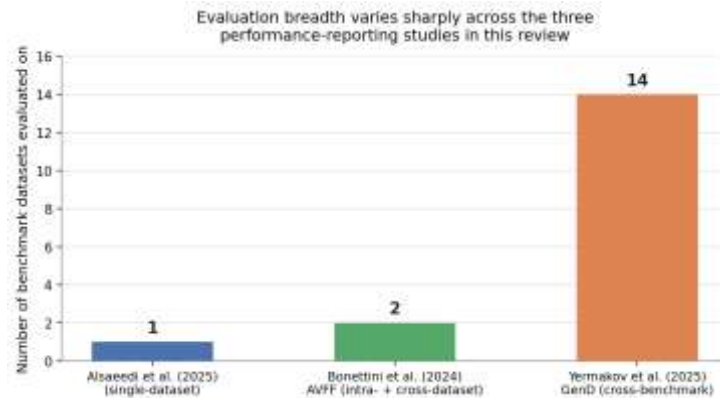


Figure 2. Number of distinct benchmark datasets each performance-reporting study evaluates on — a direct illustration of why this review treats cross-dataset breadth, not same-dataset accuracy, as the more meaningful robustness signal (Section 3.2).

Discussion

Robustness, explainability, and multimodality are coupled, not independent

Read together, the seven studies reviewed here show that these three properties interact rather than trading off cleanly against each other in one direction. GenD's robustness gain (Section 3.1) comes specifically from a representation-learning choice, not from added architectural complexity, which suggests robustness need not cost explainability the way, for instance, deeper multimodal fusion architectures plausibly might. But no study reviewed here tests this directly — Section 4.2 identifies exactly this gap — so whether a highly generalizable detector like GenD remains as explainable as a simpler, less robust one is, on the evidence available here, still an open empirical question rather than a settled trade-off in either direction.

Multimodal fusion strategies are more diverse than a single "multimodal" label suggests

Sections 5.2 and 5.3 show that "multimodal" and "multi-model" describe genuinely different design choices that this literature does not always distinguish carefully, and that even within multimodal fusion proper, Bonettini *et al.* (2024) and Alsaeedi *et al.* (2025) exploit different levels of cross-modal inconsistency (low-level signal features versus semantic emotional content) on the same underlying dataset. This suggests the field has not yet converged on which level of cross-modal signal carries the most detection value, and that ensemble combinations of both levels — not tested by either study individually — are a natural next step future work should test directly rather than treating either level as sufficient on its own.

Regulatory and platform-governance context

The TAKE IT DOWN Act (Section 1.1) creates a concrete 48-hour operational deadline for platforms to act on a victim's takedown request, which places a direct practical demand on detection latency and on how confidently a flagged result can be escalated — an operational constraint that Ahmed *et al.*'s (2025) organizational framing (Section 6.1) speaks to more

directly than any of the six technically focused studies in this review, even though its own findings could not be independently verified here. The European Union's AI Act separately imposes transparency obligations under Article 50 requiring certain AI-generated and manipulated content to be marked as such, with the relevant compliance timeline falling; taken together with the TAKE IT DOWN Act, these two frameworks suggest that detection accuracy alone will not be sufficient for organizations operating across both jurisdictions, which will also need auditable process and disclosure practices around how a piece of content was flagged and by what confidence threshold.

Ethical considerations

A false negative in this domain — a real deepfake that a detector misses — can enable exactly the fraud, harassment, or reputational harm described in Section 1.1, while a false positive can wrongly flag genuine content, with its own reputational and, depending on context, legal consequences for the person or organization involved. The documented shift toward everyday individuals as deepfake targets (Section 1.1), rather than only public figures, raises the stakes of both error types for a much larger and more vulnerable population than this literature's benchmark datasets, largely built from public-figure or actor footage, may fully represent — a limitation worth naming directly rather than assuming benchmark performance transfers evenly across who is actually depicted.

Future research directions

Three directions follow from the gaps this review has identified. First, cross-dataset robustness testing in GenD's style (Section 3.1) should become standard practice for multimodal fusion methods specifically, since Section 5's applied studies currently test robustness far less thoroughly than Section 3's dedicated robustness study does. Second, the explainability literature (Section 4) should test whether explanations remain stable under the same cross-dataset conditions that Section 3 shows detectors themselves struggle with, rather than evaluating explanation quality only in-distribution. Third, the multi-model-versus-multimodal distinction drawn in Section 5.3 suggests a concrete, currently untested combination: fusing Rahman *et al.*'s (2025) multi-classifier ensemble approach within each modality with Bonettini *et al.*'s (2024) or Alsaedi *et al.*'s (2025) across-modality fusion, which no study reviewed here attempts.

Limitations of this review

This review synthesizes seven studies identified through targeted, non-exhaustive searching restricted to 2024–2025 publications, rather than a formally registered systematic-review protocol; Qian *et al.* (2025) and Liu, Tao, and Zhou (2025) are the two exceptions and are treated here as secondary syntheses of much larger evidence bases than the other five primary studies. Two of the seven studies (Rahman *et al.*, 2025; Ahmed *et al.*, 2025) could not be independently verified beyond their confirmed title, authorship, and venue, and this review has been explicit at every point they are discussed about what could and could not be confirmed, rather than treating all seven studies as equally evidenced.

Toward a converged system

It is worth naming concretely what a detector satisfying all three of this review's pillars simultaneously would need, rather than leaving the claim abstract. At minimum: a representation-learning approach in GenD's style (Section 3.1), trained explicitly to avoid dataset-specific shortcuts rather than only to maximize same-dataset accuracy; a fusion architecture that combines low-level signal inconsistency and semantic-level inconsistency across modalities, since Section 8.2 shows the two applied fusion studies reviewed here each capture only one; and an explanation layer, in the style of Section 4.1's model-centric or emerging language-based methods, that is itself validated cross-dataset rather than only in-distribution, closing the gap Section 4.2 identifies. No study reviewed here attempts all three at once, and the two studies whose findings could not be independently verified (Sections 5.3, 6.1) are, notably, the two that engage most directly with real-world deployment concerns (ensemble robustness in practice, organizational governance) rather than benchmark performance — a pattern worth naming rather than treating as coincidental: the deployment-facing end of this literature and the benchmark-facing end do not yet appear to be the same body of well-documented work.

Conclusion

Deepfake-enabled fraud and abuse have grown large enough, and shifted far enough from public figures toward everyday individuals, to prompt new federal legislation in 2025 — and the machine-learning literature reviewed here shows genuine, verifiable progress on the three requirements that matter for detection to be useful under that pressure: generalizing beyond the specific benchmark a model was trained on (Section 3), explaining a flagged result well enough for a human or legal process to act on it (Section 4), and fusing evidence across modalities rather than relying on any single signal (Section 5). The more important finding from reading these seven studies together, rather than by pillar in isolation, is that none yet combines all three properties in one system, and that the field's own terminology conflating "multi-model" ensembles with genuine "multimodal" fusion, as Section 5.3 shows has not always kept pace with how quickly the underlying methods have diversified. Building, and rigorously cross-dataset testing, a detector that is simultaneously robust, explainable, and multimodally grounded remains the clearest concrete direction this review identifies for future work in this space.

References

- [1] Ahmed, K. R., Ahmmed, M. R., Kowser, S., Faruq, O., Shammah, R. S., & Sufian, M. A. (2025, July). The Future of AI-Powered SaaS: Market Trends and Investment Opportunities. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
- [2] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE

- [3] Alsaeedi, A., AlMansour, A., & Jamal, A. (2025). Audio-visual multimodal deepfake detection leveraging emotional recognition. *International Journal of Advanced Computer Science and Applications*, 16(6).
- [4] Karim, N. B., Vrettos, K., & Seitz, J. (2024a). Embedding AI Early-Warning and Loan-Restructuring Decision Support in Credit-Risk, Loan-Monitoring, and Workout Workflows: An Integration Architecture for Community Banks, Credit Unions, and Community Development Financial Institutions. *The Computertech*, 25-40.
- [5] Bonettini, N., Solanki, D. J., Koppiseti, S., & Bharaj, G. (2024). AVFF: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [6] Liu, P., Tao, Q., & Zhou, J. T. (2025). From single-modal to multi-modal facial deepfake detection: Progress and challenges. arXiv.
- [7] Potharaju, B., & Sajja, R. R. (2024). Modernizing US Federal Regulatory Reporting Infrastructure for Scalability, Resilience, and Audit Readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 01-36.
- [8] Rahman, M. S., & Kim, E. Y. (2024). Embedding Adaptive Machine-Learning Decision Support in Suspicious Activity Report Investigation Workflows: An Enterprise Information-Systems Framework. *International Journal of Acta Informatica*, 3(1), 245-258.
- [9] Qian, H., Xia, L., Ge, R., Fan, Y., Wang, Q., & Jing, Z. (2025). From black boxes to glass boxes: Explainable AI for trustworthy deepfake forensics. *Journal of Cybersecurity and Privacy*, 9(4), Article 61.
- [10] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.
- [11] Security Magazine. (2025, June). Deepfake-enabled fraud caused more than \$200 million in losses. <https://www.securitymagazine.com/articles/101559-deepfake-enabled-fraud-caused-more-than-200-million-in-losses>
- [12] Yermakov, A., Cech, J., Matas, J., & Fritz, M. (2025). Deepfake detection that generalizes across benchmarks. arXiv.
- [13] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32.
- [14] Rahman, M. S., Islam, C. S., & Tirupathi, D. (2024). Reducing False-Positive Alert Burden in Transaction Monitoring: Calibrated Risk Scoring for Resource-Constrained

- Community Financial Institutions. *International Journal of Modern Computing*, 7(1), 266-282.
- [15] Karim, N. B., Pentyala, R. K., & Simko, R. (2024b). Adaptive Machine-Learning Detection and Prevention of Financial-Transactions Fraud at Community Financial Institutions: Recalibrating Risk Signals Under Shifting Fraud Typologies and Record Suspicious-Activity Monitoring Burdens. *International Journal of Acta Informatica*, 3(1), 259-278.