Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 25 July 2026
Revised 28 Aug 2026
Accepted 15 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (21s): 762–795

KEYWORDS:
Deepfake detection; financial fraud; identity verification; Video-KYC; multimodal fusion; cross-modal consistency; audio anti-spoofing; explainable AI; real-time security

Detecting Deepfake-Enabled Financial Fraud Through Cross-Modal Consistency and Reliability-Aware Fusion

Akshay N. Shetye [1], Nilakshi Jain [2]

[1] Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Chembur, Mumbai 400088, India

[2]Department of Cyber Security, Shah and Anchor Kutchhi Engineering College, Chembur, Mumbai 400088, India

ABSTRACT: Remote onboarding, video-based Know-Your-Customer (Video-KYC) checks and voice-authorized payments now let a fraudster win a customer's or an executive's identity without ever being physically present. Synthetic faces and cloned voices are cheap to produce, yet most defenses still look at a single signal: deepfake detectors judge the media in isolation, and fraud engines judge only the transaction. This paper presents DeepShield-Fin, a cross-modal framework that reasons jointly over video, audio, interaction behavior, session context and transaction attributes during a live financial session. A Cross-Modal Consistency Analysis (CMCA) module turns five explicit agreement tests (lip–speech synchrony, face–voice identity, facial–vocal affect, behavior–context plausibility and transaction deviation) into an interpretable evidence vector. An Adaptive Gated Cross-Attention Fusion (AGCAF) module then weights each modality by an estimated reliability, so that a noisy microphone or a dropped video frame does not dominate the decision. A calibrated risk head drives a three-tier allow/step-up/block policy, and an Evidence Explanation Engine produces analyst-readable justifications. We specify DeepShield-FinDB, a 5,000-session benchmark with an explicit label-leakage audit, and a five-protocol validation plan. Preliminary validation uses the implemented audio branch and prototype: on 22,622 benchmark clips the audio branch reaches 95.2% accuracy (95% CI 94.9–95.5%), an MCC of 0.904 and a 4.8% EER, against 7.9–8.5% EER for HuBERT, wav2vec 2.0 and RawNet2, and the prototype completes an analysis window in 1.84 s. An analytical deployment study shows why single-channel detection is not enough: at a 1% fraud rate even this detector would raise about five false alarms per detected fraud, which motivates the cross-modal evidence that DeepShield-Fin adds.

1. INTRODUCTION

Financial services have spent the last decade removing the counter. Accounts are opened through a smartphone camera, loans are approved after a short video call, and treasury staff release payments on the strength of a familiar voice on the line. In India, the Reserve Bank's amendment permitting Video-based Customer Identification Process (V-CIP) [1] turned the video call into a legal substitute for physical verification, and similar remote-onboarding regimes exist across most large economies. The efficiency gains are real. So is the side effect: identity, which used to be checked against a person standing in a branch, is now checked against pixels and audio samples that can be manufactured.

Generative models have made that manufacturing routine. Face-swapping and reenactment pipelines such as DeepFaceLab [2] and First Order Motion Model [3] need only a few minutes of target footage; lip-synchronization networks such as Wav2Lip [4] and SadTalker [5] can animate a still photograph to match arbitrary speech; and few-shot voice cloning [6,7] reproduces a speaker's timbre from seconds of reference audio. The financial consequences are no longer hypothetical. In 2019 the chief

executive of a UK energy firm transferred roughly €220,000 after a telephone call in which a cloned voice imitated his parent-company superior [8]. In early 2024 an employee of the engineering firm Arup in Hong Kong authorized transfers of about US$25 million after a video conference in which every other participant, including the chief financial officer, was synthetic [9]. Indian cases are smaller in value but far more numerous: cloned voices of relatives and finance officials have been used to extract amounts from ₹45,000 to over ₹20 lakh from individuals in Lucknow, Kanpur, Hyderabad and Bengaluru [10–13]. Aggregate losses are rising steeply as well [14–16].

The defensive toolbox has not kept pace with this change in attack surface, for two structural reasons. First, deepfake detection and fraud detection have grown up as separate research communities. Visual detectors such as MesoNet [17], XceptionNet [18] and LipForensics [19], and audio anti-spoofing systems such as RawNet2 [20] and AASIST [21], answer the question is this media authentic? They know nothing about whether the account holder normally logs in from this device, whether the requested beneficiary is new, or whether the transfer amount is twenty times the customer's median. Conversely, transaction-centric fraud engines [22–25] answer is this payment anomalous? but treat the voice or face that authorized it as a trusted input. An attacker who combines a convincing face, a cloned voice and a plausible story can slip between the two. Second, most multimodal deepfake detectors that do combine audio and video [26–29] fuse their features by concatenation or uniform attention. A degraded channel can then drag the decision with it, and the fused score does not tell the analyst why a session was flagged.

This paper addresses both gaps with DeepShield-Fin, a framework designed around the actual shape of a remote financial session. Its starting point is our earlier multi-branch audio detector, which combined spectral, temporal and prosodic–emotional cues and reached 95.2% accuracy and a 4.8% equal error rate (EER) on audio benchmarks [30]. DeepShield-Fin generalizes that idea from three acoustic views of one modality to five modalities of one session, and adds explicit cross-modal reasoning on top.

1.1. Research Objectives and Hypotheses

The study pursues three objectives, each tied to one pre-registered hypothesis (detailed research design in Section 3):

This study pursues three interlinked objectives, each anchored to a pre-registered hypothesis that guides the research design outlined in Section 3. The first objective (O1) is to examine the literature on deepfake technology and its deployment across financial fraud typologies, testing the hypothesis (H1) that deepfakes significantly elevate fraud risk relative to traditional methods. The second objective (O2) evaluates the impact of deepfake-enabled fraud on customers, institutions, and regulators, while benchmarking existing detectors under realistic conditions; it tests the hypothesis (H2) that current fraud-detection and prevention mechanisms are insufficient for the specific challenges posed by deepfakes. Finally, the third objective (O3) is to develop and validate a cross-modal detection framework, tool, and methodology, testing the hypothesis (H3) that such an advanced framework substantially reduces both the incidence and the expected cost of deepfake-enabled financial fraud.

1.2. Contributions

This paper makes five principal contributions. First, it formalizes deepfake-enabled financial fraud as a session-level, cost-sensitive risk estimation problem over synchronized windows of five modalities, incorporating explicit attacker classes and a Bayes-optimal, prior-corrected decision rule. Second, it introduces Cross-Modal Consistency Analysis (CMCA), which computes five named, individually interpretable consistency scores that are jointly discriminative, rather than leaving cross-modal agreement implicit in a fused embedding. Third, it proposes Adaptive Gated Cross-Attention Fusion (AGCAF), which couples cross-modal attention with gates scaled by a learned, quality-aware reliability estimate and is trained with modality dropout, ensuring graceful degradation when a modality is missing or corrupted. Fourth, it presents a calibrated, explainable decision layer combining temperature scaling, prior-shift correction, and a three-tier policy with an Evidence Explanation Engine. Finally, it specifies DeepShield-FinDB, a leakage-audited multimodal benchmark, alongside a five-protocol validation plan and preliminary evidence from the implemented audio branch and prototype.

2. BACKGROUND AND RELATED WORK

2.1. Threat Landscape and Incident Analysis

To ground H1 we compiled documented incidents in which synthetic media was the decisive instrument of a financial loss. Table 1 lists representative cases. Two features stand out. The first is the spread of channels: the same underlying capability, impersonating a trusted person, has been delivered through a live multi-party video call, a phone call, a WhatsApp voice note

and a social-media advertisement. The second is the asymmetry of effort. In each case the attacker's cost was a few minutes of publicly available footage or audio and an off-the-shelf generator, while the victim's loss ranged from a family's savings to a corporate treasury transfer.

Table 2.1.1. Representative documented incidents of deepfake-enabled financial fraud.

Year	Location	Modality / channel	Reported loss	Mechanism	Ref.
2019	United Kingdom	Cloned voice, phone call	≈ €220,000	CEO instructed by "parent-company chief" to pay a supplier	[8]
2023	Lucknow, India	Cloned voice, WhatsApp	₹45,000	Voice of a relative requesting urgent help	[10]
2024	Hong Kong	Synthetic multi-party video call	≈ US$25 million	Synthetic CFO and colleagues authorized 15 transfers	[9]
2024	Kanpur, India	Cloned voice, phone call	₹1 lakh	"Nephew in legal trouble" scenario	[11]
2024	Hyderabad, India	Deepfake video of a minister, social media	> ₹20 lakh	Fraudulent investment endorsement	[12]
2024	Bengaluru, India	Deepfake video of business leaders	₹26 lakh (combined)	Fake trading-platform promotion	[13]

Aggregate statistics, summarized in Figure 2.1.1, show that these are not isolated anecdotes. Losses reported to the FBI's Internet Crime Complaint Center roughly quadrupled between 2020 and 2024 [14,15]. Industry projections specific to generative AI are steeper still: Deloitte estimates that such fraud could cost US banks and their customers US$40 billion annually by 2027 [16]. Public statistics rarely separate deepfake-assisted cases from other cyber-enabled fraud, itself a measurement gap.

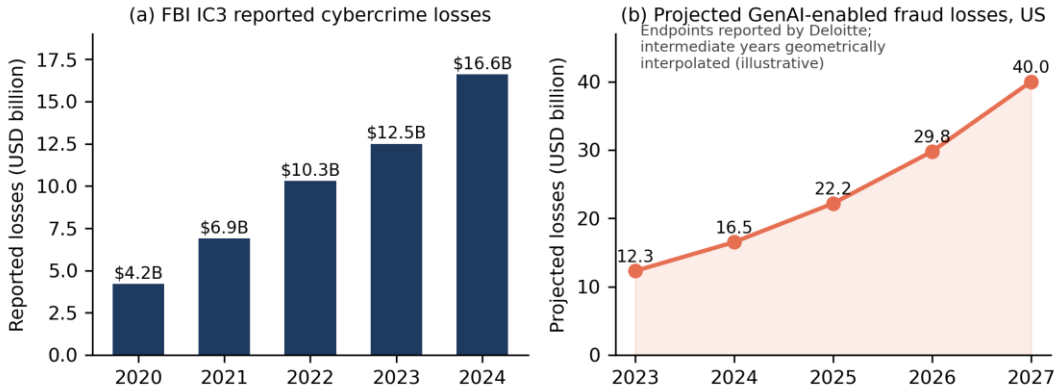


Figure 2.1.1 Threat landscape. (a) Annual cybercrime losses reported to the FBI Internet Crime Complaint Center, 2020–2024 [14,15]. (b) Projected losses from generative-AI-enabled fraud in the United States; the 2023 and 2027 endpoints are those reported by Deloitte [16], and intermediate years are geometric interpolations shown for illustration only. Source: data from [14–16]; chart drawn by the authors.

Figure 2.2.1 organizes the attack vectors we model. Voice cloning (A1) dominates telephone channels; face swapping (A2) targets camera-based onboarding; lip-sync puppeteering (A3) animates a genuine photograph with attacker-controlled speech; full audio-visual synthesis (A4) manipulates both channels at once; and stream injection (A5) bypasses the camera altogether by feeding pre-rendered or real-time synthetic video through a virtual device. What these share is that each succeeds by producing a session that is internally plausible to a human operator. Our design premise is that it is much harder for an attacker to make the session internally consistent across every channel simultaneously.

Natural Resources for Human Health
eISSN: 2583-1194

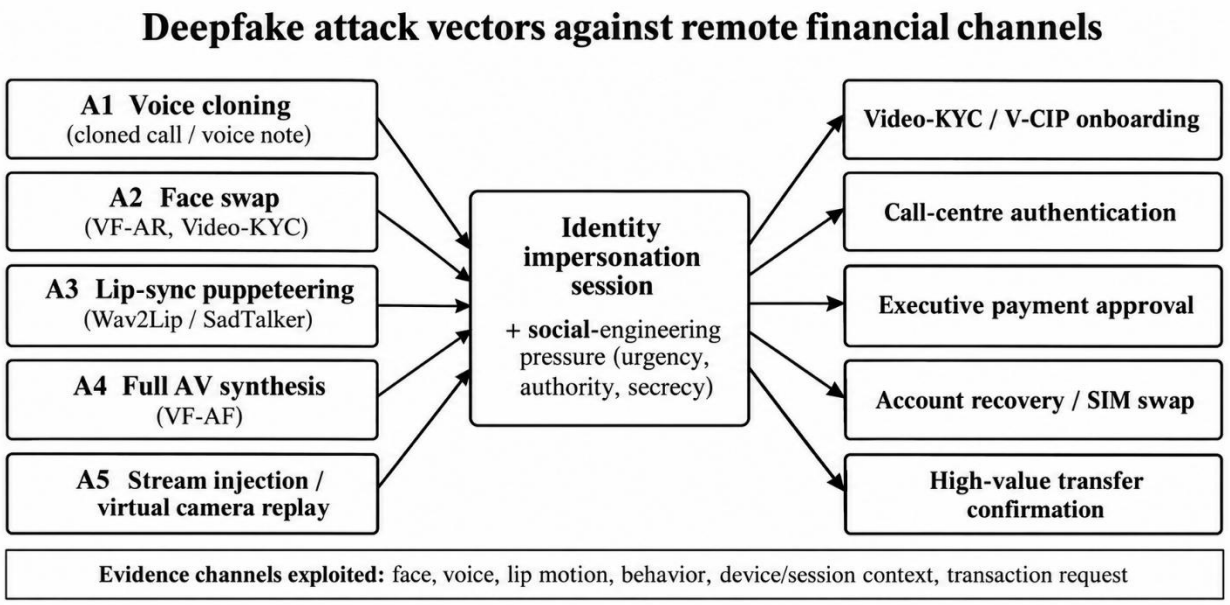


Figure 2.2.1. Attack vectors (A1–A5) considered in this work and the remote financial channels they target. All vectors converge on an identity-impersonation session reinforced by social-engineering pressure. Source: authors' elaboration based on the incidents in Table 1 and the generators described in [2–7].

2.2. Deepfake Generation

Face manipulation evolved from paired autoencoders, popularized by DeepFaceLab [2], through generative adversarial networks [31] and style-based generators [32] to diffusion models that leave few texture artifacts. Talking-head synthesis animates a still image from a driving video (FOMM [3]) or directly from speech (Wav2Lip [4], SadTalker [5]), while speaker-adaptive TTS such as SV2TTS [6] and neural codec models such as VALL-E [7] clone a voice from seconds of audio. Crucially for defense, each channel is typically forged with a different tool that knows nothing about the others; this independence is what cross-modal consistency checks exploit.

2.3. Visual and Audio Deepfake Detection

Visual detectors learn texture and blending artifacts (MesoNet [17], XceptionNet on FaceForensics++ [18], Capsule-Forensics [33], Face X-ray [34]), physiological signals (FakeCatcher [35]), semantic mouth motion (LipForensics [19]) or longer spatio-temporal context (CViT [36]). Their accuracy drops sharply under heavy compression and on unseen manipulation methods [18,19], exactly the conditions of a mobile Video-KYC session. In audio, the ASVspoof challenges [37,38] moved the field from cepstral features with GMM back ends to end-to-end models such as RawNet2 [20] and AASIST [21], often with self-supervised front ends (wav2vec 2.0 [39], HuBERT [40]). Our multi-branch detector [30] added prosodic and emotional cues and reduced EER from 7.9–8.5% to 4.8%, yet audio detectors still degrade on in-the-wild speech [41] and compressed channels [38].

2.4. Audio-Visual Multimodal Detection

Audio-visual methods exploit the shared articulatory origin of speech and lip motion: SyncNet [42] learned a joint audio–mouth embedding that underlies the modality-dissonance score (MDS) [28]; Mittal et al. [27] compared perceived emotion across face and voice; Zhou and Lim [26] trained joint detectors on DFDC; and FakeAVCeleb [43] enabled evaluation over all four real/fake combinations. AVoiD-DF [29], voice–face homogeneity (VFD) [44] and AVT²-DWF [45] refine alignment and fusion, and our companion TRIUNE-Guard [46] adds a physiological rPPG stream. All remain media-forensic: none uses the behavioral, contextual or transaction evidence a bank already holds about a session.

2.5. AI-Based Financial Fraud Detection

Transaction fraud detection uses sequence models [22], adversarial augmentation for class imbalance [23], graph neural networks for camouflaged fraudsters [24], streaming active learning under drift [25], behavioral biometrics for continuous authentication

[47] and interpretable tabular networks such as TabNet [48]. These systems treat the voice or face that authorizes a payment as trusted input; when an impersonation is convincing, the resulting transaction may be only mildly anomalous and impossible to separate from an unusual genuine request.

2.7. Summary of Gaps

Table 2.5.1 positions representative methods against the capabilities that a deployable defense requires, and Table 3 lists the resulting gaps together with the design response of this paper.

Table 2.5.1 Capability comparison of representative approaches (✓ supported, ◐ partial, ✗ not supported).

Approach	Visual	Audio	Behavior / context	Transaction	Explicit cross-modal checks	Reliability-aware fusion	Explanations	Real-time
MesoNet / XceptionNet [17,18]	✓	✗	✗	✗	✗	✗	✗	◐
LipForensics [19]	✓	✗	✗	✗	✗	✗	◐	◐
RawNet2 / AASIST [20,21]	✗	✓	✗	✗	✗	✗	✗	✓
Multi-branch audio [30]	✗	✓	✗	✗	✗	✗	◐	✓
MDS / AVoiD-DF [28,29]	✓	✓	✗	✗	◐	✗	✗	◐
AVT²-DWF [45]	✓	✓	✗	✗	✗	◐	✗	◐
CARE-GNN / TabNet fraud [24,48]	✗	✗	◐	✓	✗	✗	◐	✓
DeepShield-Fin (this work)	✓	✓	✓	✓	✓	✓	✓	✓

Table 2.5.2. Research gaps and the corresponding design response in DeepShield-Fin.

Gap	Consequence in a financial session	Design response
Media and fraud analytics are siloed	A convincing impersonation authorizes a mildly anomalous payment that neither system flags	Joint modelling of V, A, B, C, T (Section 4.4)
Cross-modal agreement is left implicit	Partial fakes (VF–AR, VR–AF) evade detectors tuned to one channel	Five named CMCA tests (Section 4.5)
Fusion ignores channel quality	A noisy microphone or frozen frame corrupts the fused decision	Reliability-scaled gates and modality dropout (Section 4.6)
Scores are uncalibrated and prior-blind	Thresholds tuned on balanced data produce alert floods at a 1–5% fraud prior	Temperature scaling and prior-shift correction (Section 4.8)
Black-box decisions	Analysts and regulators cannot justify a blocked payment	Evidence Explanation Engine (Section 4.9)

Gap	Consequence in a financial session	Design response
Offline, accuracy-only evaluation	Reported gains do not transfer to deployment	Five protocols, calibration and cost metrics, significance tests (Section 5)

3. RESEARCH DESIGN AND METHODOLOGY

3.1. Objectives, Research Questions and Methods

The research follows a mixed design: a qualitative-descriptive strand that characterizes the threat (O1), an evaluative strand that measures the shortcomings of current defenses and their consequences for stakeholders (O2), and a constructive–experimental strand that designs and statistically validates a new framework (O3). Table 4 makes the link between each objective, the research question it answers, the method used and the evidence it produces explicit.

Table 3.1.1. Research objectives, questions, methods and outputs.

Objective	Research question	Methods	Output / evidence	Hyp.	Sec.
O1 — Examine	How are deepfakes used to commit financial fraud, and through which channels?	Systematic literature review; incident analysis; loss-trend analysis of public reports; attack taxonomy	Incident table, threat-landscape statistics, taxonomy A1–A5, gap analysis	H1	2, 3.3
O2 — Evaluate	What is the impact on stakeholders, and how well do current detectors cope under realistic conditions?	Stakeholder impact mapping; expected-cost model; benchmarking of 13 baselines under five protocols	Cost model, degradation of baselines across P1–P5	H2	3.4, 6
O3 — Develop and Validate	Can a cross-modal, reliability-aware and explainable framework reduce fraud?	System design; dataset construction with leakage audit; ablation; DeLong, McNemar and bootstrap tests	DeepShield-Fin, DeepShield-FinDB, statistical validation	H3	4, 5, 6

The impact assessment of O2 considers three stakeholder groups. Customers bear direct monetary loss and, under false alarms, friction and exclusion. Institutions bear reimbursement, investigation and reputational costs and, for Video-KYC, regulatory exposure if a synthetic identity is onboarded. Regulators and the wider system face erosion of trust in remote channels on which financial inclusion increasingly depends. These costs enter the evaluation through the cost model of Section 3.4 rather than as an abstract discussion.

3.2. Research Design and Workflow

Figure 3.2.1 summarizes the workflow. The three strands run in sequence, and the output of each constrains the next: the taxonomy and gap analysis from O1 define the requirements against which baselines are tested in O2, and the failure modes observed in O2 (partial forgeries, channel degradation, prior shift) define the design requirements of the framework built in O3. Hypotheses and decision criteria were fixed before the experiments (Section 5.6), so that the results can confirm or refute them rather than be interpreted after the fact.

Natural Resources for Human Health
eISSN: 2583-1194

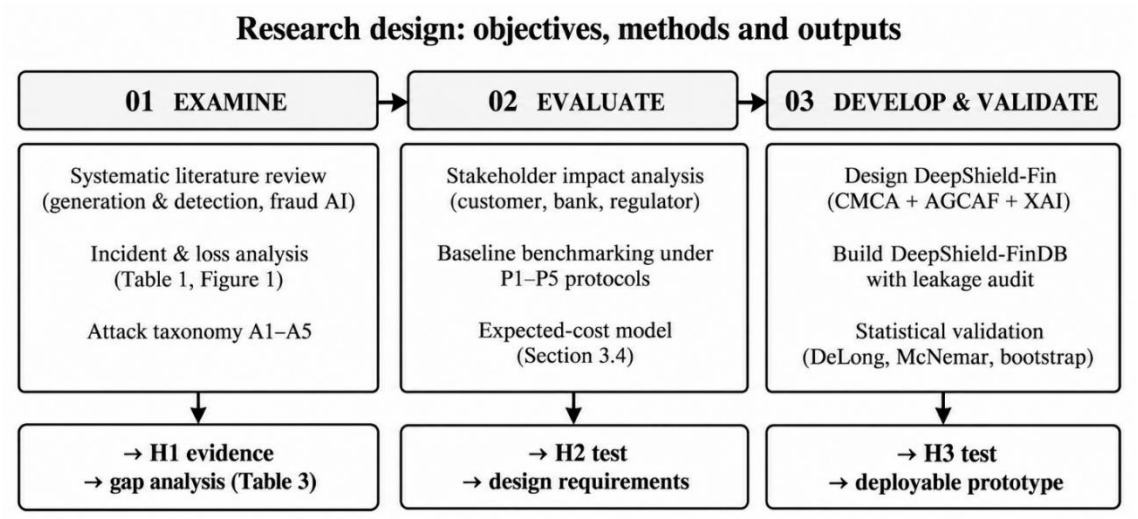


Figure 3.2.1 Research design linking the three objectives to their methods, outputs and hypothesis tests. Source: authors' own design.

3.3. Threat Model

We consider a remote financial session between a customer-facing channel of a regulated institution (Video-KYC, call-center authentication, executive approval call, account recovery or high-value transfer confirmation) and a remote party who claims to be a known customer or employee. The adversary's goal is to have the session accepted as genuine so that a financial action is executed.

Adversary capabilities. The adversary (i) possesses public or leaked audio and video of the target; (ii) can operate any open-source or commercial generator for face, lip and voice synthesis, offline or in real time; (iii) can inject synthetic streams through a virtual camera or audio device (A5); (iv) may control the device and network from which the session originates; and (v) knows the target's approximate transaction habits but not the institution's internal customer profile. We do not assume that the adversary has white-box access to the detector or can query it adaptively; adaptive attacks are discussed as a limitation in Section 7.4.

Defender capabilities. The institution observes the synchronized audio and video stream, interaction telemetry (response latencies, turn-taking, speaking rate), device and session metadata, and the requested transaction together with the customer's history. Where the customer has an enrolled biometric template, it is available to the defender.

3.4. Problem Formulation and Cost Model

Let a session be divided into overlapping analysis windows of length W seconds with hop Δ. The observation in window t is

$$\mathbf{X}_t = (V_t,\ A_t,\ \mathbf{B}_t,\ \mathbf{C}_t,\ \mathbf{T}_t), \qquad (1)$$

where $V_t = \{v_{t,1}, \dots, v_{t,N_v}\}$ is the sequence of video frames, A_t is the audio waveform sampled at $f_s = 16$ kHz, and $\mathbf{B}_t \in \mathbb{R}^{d_b}$, $\mathbf{C}_t \in \mathbb{R}^{d_c}$ and $\mathbf{T}_t \in \mathbb{R}^{d_t}$ are behavioral, contextual and transaction feature vectors. Each session carries a label $y \in \{0,1\}$ (genuine or fraudulent) and, for fraudulent sessions, a manipulation subclass

$$k \in \mathcal{K} = \{\text{VR–AR},\ \text{VF–AR},\ \text{VR–AF},\ \text{VF–AF}\}, \qquad (2)$$

where V/A denote video/audio and R/F denote real/fake; VR–AR sessions are genuine. We seek a scoring function $f_\theta: \mathbf{X}_t \mapsto \hat{p}_t \in [0,1]$ that estimates the posterior probability of fraud in window t.

Window-level evidence is accumulated into a session-level risk by an exponentially weighted update, which lets the system raise an alert early when evidence is strong while remaining stable against isolated spikes:

$$\bar{p}_t = \lambda\, \bar{p}_{t-1} + (1-\lambda)\, \hat{p}_t, \qquad \bar{p}_0 = \pi, \qquad (3)$$

with smoothing factor $\lambda \in [0,1)$ and prior π.

In finance, the two error types are not equally expensive. A missed fraud (false negative) costs the transaction value plus remediation, whereas a false alarm (false positive) costs customer friction and analyst time. With costs c_{FN} and c_{FP} and operating prior $\pi = P(y = 1)$, the expected cost of a decision rule δ is

$$\mathcal{R}(\delta) = c_{\mathrm{FN}}\, \pi\, P(\delta = 0 \mid y = 1) + c_{\mathrm{FP}}\, (1 - \pi)\, P(\delta = 1 \mid y = 0). \qquad (4)$$

If $\hat{p}$ is a calibrated posterior, the decision that minimizes Eq. (4) is to flag whenever $\hat{p}$ exceeds the Bayes threshold [49]

$$\tau^{*} = \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}. \qquad (5)$$

Equation (5) only holds when the score is calibrated to the deployment prior, which motivates the calibration and prior-correction steps of Section 4.8. To relate detector quality directly to H3, we also define the expected loss per session,

$$\mathcal{L}_{\mathrm{exp}} = \pi\, \mathrm{FNR} \cdot \bar{c}_{\mathrm{FN}} + (1 - \pi)\, \mathrm{FPR} \cdot c_{\mathrm{FP}}, \qquad (6)$$

where $\bar{c}_{\mathrm{FN}}$ is the mean value at risk in a fraudulent session, and the relative expected-loss reduction of a method over a baseline,

$$\Delta\mathcal{L} = \frac{\mathcal{L}_{\mathrm{exp}}^{\mathrm{base}} - \mathcal{L}_{\mathrm{exp}}^{\mathrm{prop}}}{\mathcal{L}_{\mathrm{exp}}^{\mathrm{base}}} \times 100\%. \qquad (7)$$

4. THE DEEPSHIELD-FIN FRAMEWORK

4.1. Overview

Figure 4.1.1 shows the end-to-end architecture. Five modality streams are acquired and windowed in parallel, each is encoded by a model suited to its structure, and all embeddings are projected into a common 256-dimensional space. Two complementary mechanisms then operate on the projected embeddings. CMCA computes explicit, named agreement tests between modalities; AGCAF performs learned, reliability-aware fusion. Their outputs feed a calibrated risk head whose decision drives an alerting and response layer, while the Evidence Explanation Engine records why the decision was made.

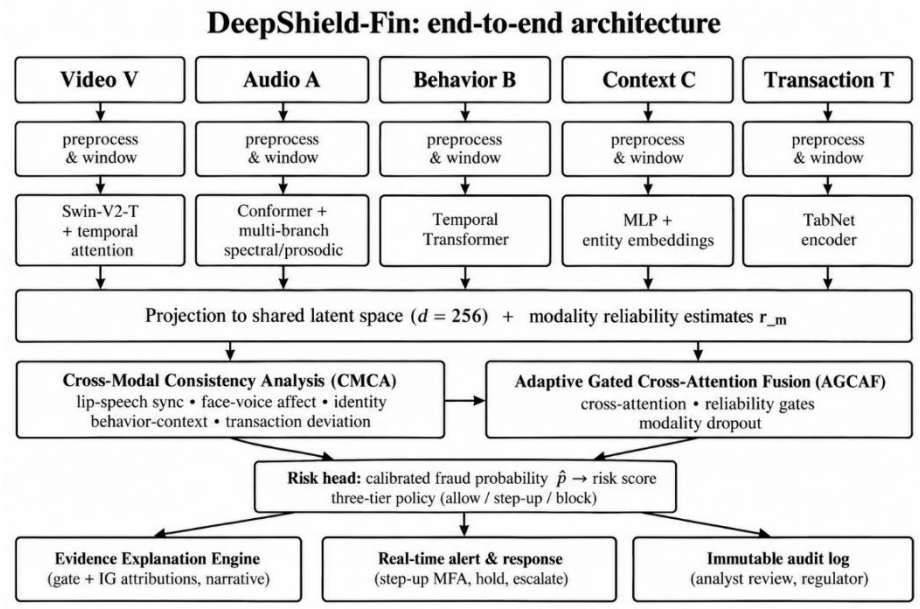


Figure 4.1.1 Architecture of DeepShield-Fin. Modality-specific encoders feed a shared latent space; CMCA and AGCAF provide explicit consistency evidence and reliability-aware fusion; a calibrated risk head drives a three-tier response, explanations and an audit log. Source: authors' own design.

The design follows three principles drawn from the gaps in Table 3. Evidence over appearance: rather than asking only whether each stream looks authentic, the system asks whether the streams tell the same story. Graceful degradation: no single channel should be able to veto the decision when it is unreliable. Accountability: every alert must come with evidence that a human analyst can check. Table 5 specifies each module of the architecture.

Table 4.1.1 Module-level specification of DeepShield-Fin.

Module	Input	Encoder / operation	Output	Initialization / training data
Visual	16 aligned 224×224 face crops per 2-s window	Swin-V2-Tiny (≈28 M parameters) + additive temporal attention	$\mathbf{F}_v \in \mathbb{R}^{768}$	ImageNet-1k; FaceForensics++ (c23, c40)
Lip	96×96 mouth crops	Lightweight 3D-CNN	$\mathbf{z}_{\text{lip}} \in \mathbb{R}^{256}$ per frame	Real FakeAVCeleb clips (contrastive sync)
Identity	Face crop; speech segment	ArcFace (face); ECAPA-TDNN (voice)	512-d; 192-d embeddings	Public pre-trained weights
Audio	16 kHz waveform → 80-band log-Mel; 88 eGeMAPS descriptors	Conformer + 2D-CNN spectral + prosodic branches [30]	$\mathbf{F}_a \in \mathbb{R}^{512}$	ASVspoof 2019 LA; ASVspoof 2021 codecs
Behavior	≤ 64 interaction events × 6 attributes	2-layer Temporal Transformer	$\mathbf{F}_b \in \mathbb{R}^{64}$	DeepShield-FinDB
Context	8 categorical + 12 numerical fields	Entity embeddings + MLP	$\mathbf{F}_c \in \mathbb{R}^{64}$	DeepShield-FinDB
Transaction	8 attributes (log amount, beneficiary age, velocity, hour, …)	TabNet, 3 decision steps	$\mathbf{F}_t \in \mathbb{R}^{32}$	IEEE-CIS; PaySim
CMCA	$\mathbf{s} \in \mathbb{R}^6$	Five consistency tests + MLP	$\mathbf{h}_{\text{cm}} \in \mathbb{R}^{256}$	End-to-end
AGCAF	Six 256-d embeddings + quality indicators	Reliability-gated cross-attention, 8 heads	$\mathbf{z} \in \mathbb{R}^{256}$	End-to-end
Risk head	$\mathbf{z}$	MLP 256–128–64, sigmoid; temperature scaling	$p' \in [0,1]$, risk score R	End-to-end; calibrated on validation split

4.2. Acquisition and Synchronization

Audio and video are captured from the session (a WebRTC or meeting-client stream in deployment, recorded sessions in our experiments) and time-stamped against a common monotonic clock. Behavioral telemetry (turn boundaries, response latencies, speaking-rate estimates) is derived from the same stream; contextual metadata and the transaction request are taken from the institution's session and payment systems. Each modality is resampled onto the window grid. For a window starting at time t_0, the audio sample range and video frame range are

$$\mathcal{I}_a(t_0) = [\lfloor t_0 f_s \rfloor,\ \lfloor (t_0 + W) f_s \rfloor), \qquad \mathcal{I}_v(t_0) = [\lfloor t_0 f_v \rfloor,\ \lfloor (t_0 + W) f_v \rfloor), \tag{8}$$

where f_v is the video frame rate. We use $W = 2$ s with $\Delta = 1$ s for streaming inference and 4–5 s windows (≈100 frames at 25 fps) for training, which gives the temporal encoders enough context to observe articulation and prosody.

4.3. Preprocessing

Video. Faces are detected with RetinaFace [50] and aligned to a canonical template by the similarity transform that minimizes landmark error,

$$(s^*, \mathbf{R}^*, \mathbf{t}^*) = \underset{s,\mathbf{R},\mathbf{t}}{\operatorname{argmin}} \sum_{i=1}^{5} \| s\,\mathbf{R}\,\mathbf{l}_i + \mathbf{t} - \bar{\mathbf{l}}_i \|_2^2, \qquad (9)$$

where $\mathbf{l}_i$ are detected landmarks and $\bar{\mathbf{l}}_i$ the template positions. Aligned crops are resized to 224×224, and a 96×96 mouth region is cropped from the same alignment for the synchronization test. Frames with face-detection confidence below 0.8 are marked rather than dropped; their proportion becomes a video-quality feature for the reliability estimator. Figure 5 shows frames captured by the implemented pipeline and the corresponding aligned region of interest.



Figure 4.3.1 Video inputs captured by the implemented acquisition pipeline during a test verification session: (a) consecutive sampled frames; (b) raw frame and the aligned face region passed to the visual encoder. Source: authors' implementation; the subject is the first author, recorded with consent.

Audio. Signals are converted to mono, resampled to 16 kHz, pre-emphasized,

$$x'[n] = x[n] - 0.97\,x[n-1], \qquad (10)$$

and denoised by spectral gating: a noise profile $\widehat{N}(f)$ is estimated from non-speech frames identified by voice-activity detection, and time–frequency bins are attenuated by a soft mask

$$G(t,f) = \max\left(0,\ 1 - \alpha_g \frac{\widehat{N}(f)}{|S(t,f)|}\right), \qquad (11)$$

where $S(t,f)$ is the short-time Fourier transform and α_g the over-subtraction factor. The 80-band log-Mel spectrogram used by the spectral branch is

$$\mathbf{M}(t,b) = \log\left(\sum_f \mathbf{H}_{\text{mel}}(b,f)\,|G(t,f)S(t,f)|^2 + \epsilon\right), \qquad (12)$$

with a 25 ms Hann window and 10 ms hop. Figure 6 shows the effect of this stage on a recording captured by our pipeline. We deliberately apply only mild denoising: aggressive enhancement can erase the very synthesis artifacts the detector needs, a trade-off that protocol P4 is designed to quantify.

Natural Resources for Human Health
eISSN: 2583-1194

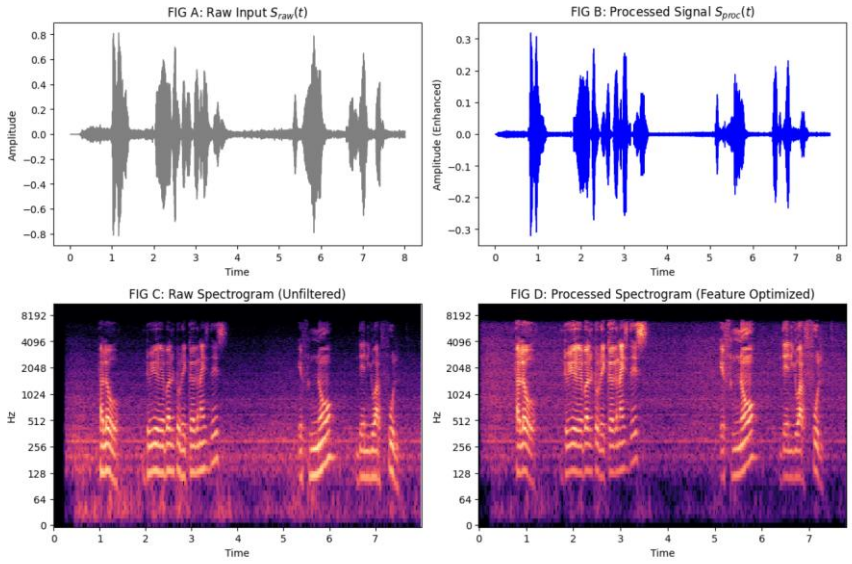


Figure 4.3.2. Audio preprocessing on a session recording: (A) raw input waveform, (B) processed signal, (C) raw spectrogram and (D) processed spectrogram used for feature extraction. Source: authors' implementation.

Behavioral, contextual and transaction features. Heavy-tailed monetary quantities are log-transformed, $a' = \log(1 + a)$, and continuous features are standardized with statistics from the training split only,

$$\hat{x}_j = \frac{x_j - \mu_j}{\sigma_j}, \qquad (13)$$

to avoid information leaking from the test distribution. Categorical fields (device family, operating system, channel, country) are mapped to learned embeddings.

4.4. Modality Encoders

Visual encoder. Each aligned frame is encoded by a Swin Transformer V2-Tiny [51] pre-trained on ImageNet. Swin V2 replaces dot-product attention with scaled cosine attention and a log-spaced continuous relative position bias,

$$\mathrm{Sim}(\mathbf{q}_i, \mathbf{k}_j) = \frac{\cos(\mathbf{q}_i, \mathbf{k}_j)}{\tau_a} + B_{ij}, \qquad (14)$$

which stabilizes training at the resolutions used for face crops. Frame embeddings $\mathbf{u}_1, \ldots, \mathbf{u}_{N_v}$ are aggregated by additive temporal attention,

$$\alpha_t = \frac{\exp(\mathbf{w}^\top \tanh(\mathbf{W}_u \mathbf{u}_t))}{\sum_{t'} \exp(\mathbf{w}^\top \tanh(\mathbf{W}_u \mathbf{u}_{t'}))}, \qquad \mathbf{F}_v = \sum_t \alpha_t \, \mathbf{u}_t, \qquad (15)$$

so that frames showing the most informative artifacts receive the greatest weight. A parallel lightweight 3D-CNN on the mouth crops produces the lip embedding $\mathbf{z}_{\mathrm{lip}}$ used by CMCA, and an ArcFace [52] embedding of the face provides the identity vector $\mathbf{e}_{\mathrm{face}}$.

Audio encoder. The audio branch extends our published multi-branch detector [30]. Log-Mel features pass through a Conformer encoder [53], whose blocks interleave half-step feed-forward modules, multi-head self-attention and depthwise convolution:

$$\tilde{\mathbf{x}} = \mathbf{x} + 1/2\,\mathrm{FFN}(\mathbf{x}), \quad \mathbf{x}' = \tilde{\mathbf{x}} + \mathrm{MHSA}(\tilde{\mathbf{x}}), \qquad (16)$$

$$\mathbf{x}'' = \mathbf{x}' + \mathrm{Conv}(\mathbf{x}'), \quad \mathbf{y} = \mathrm{LN}(\mathbf{x}'' + 1/2\,\mathrm{FFN}(\mathbf{x}'')). \qquad (17)$$

Natural Resources for Human Health
eISSN: 2583-1194

Self-attention captures long-range prosodic structure, while the convolution module captures the local spectral texture where vocoder artifacts concentrate. Following [30], two additional branches run in parallel: a 2D-CNN spectral branch and a prosodic branch fed by 88 eGeMAPS descriptors from openSMILE [54,55] (pitch, jitter, shimmer, harmonics-to-noise ratio, loudness), which also yields a vocal-emotion distribution $\mathbf{p}_{\text{voice}}$. The three branch outputs are combined as

$$\mathbf{F}_a = \mathbf{W}_a\left[\,\mathbf{h}_{\text{conf}}\,;\,\mathbf{h}_{\text{spec}}\,;\,\mathbf{h}_{\text{pros}}\,\right] + \mathbf{b}_a, \qquad (18)$$

and an ECAPA-TDNN [56] speaker embedding $\mathbf{e}_{\text{voice}}$ is extracted for the identity test. Figure 7 shows the intermediate representations produced by the audio branch.

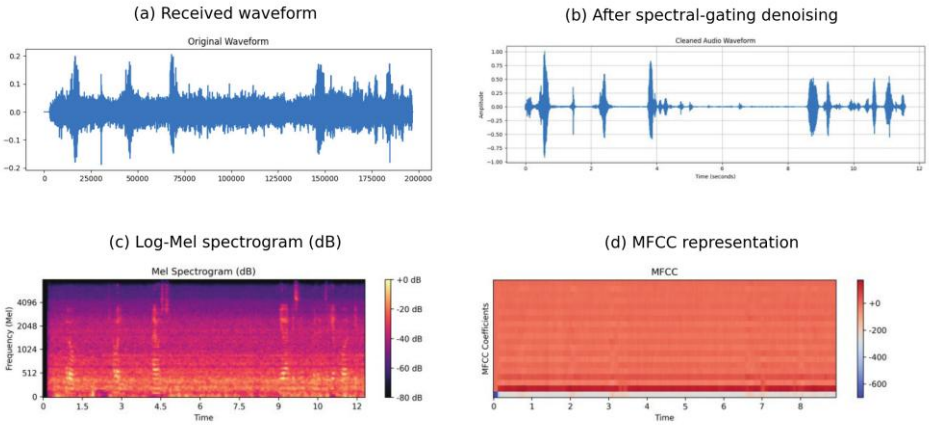


Figure 4.4.1. Intermediate representations of the audio branch: (a) received waveform, (b) waveform after spectral gating, (c) log-Mel spectrogram and (d) MFCCs. Source: adapted from [30], © 2026 IEEE, reused by the authors of that work.

Behavioral encoder. Behavioral telemetry is a short sequence of events (turn start/end, response latency, pause length, speaking rate, interaction cadence). It is encoded by a two-layer Temporal Transformer [57] with sinusoidal positions,

$$\text{PE}(p, 2i) = \sin\left(\frac{p}{10000^{2i/d}}\right), \qquad \text{PE}(p, 2i+1) = \cos\left(\frac{p}{10000^{2i/d}}\right), \qquad (19)$$

followed by mean pooling to obtain $\mathbf{F}_b$. Response latency is particularly informative for real-time voice conversion and puppeteering, which add processing delay to every conversational turn.

Contextual encoder. Device, network and session fields are encoded by a multilayer perceptron over the concatenation of categorical embeddings and standardized numerical features,

$$\mathbf{F}_c = \text{MLP}\left(\|_j \mathbf{E}_j\left[c_j\right] \,\|\, \hat{\mathbf{c}}_{\text{num}}\right), \qquad (20)$$

where $\|$ denotes concatenation. Features include device novelty, emulator or virtual-camera indicators, IP-geolocation distance from the customer's usual locations and session duration.

Transaction encoder. Transaction attributes are encoded with TabNet [48], whose sequential attentive transformer selects a sparse subset of features at each decision step i,

$$\mathbf{M}[i] = \text{sparsemax}\left(\mathbf{P}[i-1] \cdot h_i(\mathbf{a}[i-1])\right), \qquad (21)$$

$$\mathbf{P}[i] = \prod_{j=1}^{i} \left(\gamma - \mathbf{M}[j]\right), \qquad (22)$$

where sparsemax [58] yields exactly sparse masks and the relaxation parameter γ controls feature reuse across steps.

Projection. Because raw embeddings differ widely in dimension (e.g., 768 for the Swin pooled output and 8–32 for tabular encoders), each is mapped into a shared space before fusion:

$$\mathbf{h}_m = \mathrm{LN}(\mathbf{W}_m\mathbf{F}_m + \mathbf{b}_m) \in \mathbb{R}^{256}, \qquad m \in \{v, a, b, c, t\}. \tag{23}$$

This prevents high-dimensional audio-visual features from swamping low-dimensional but highly informative financial signals simply by virtue of their size.

4.5. Cross-Modal Consistency Analysis (CMCA)

CMCA is built on a simple observation: a forger who uses independent tools for each channel must get every pairwise relationship right, whereas a genuine session satisfies them automatically. CMCA therefore computes five scores, each with a direct physical or business meaning (Figure 8).

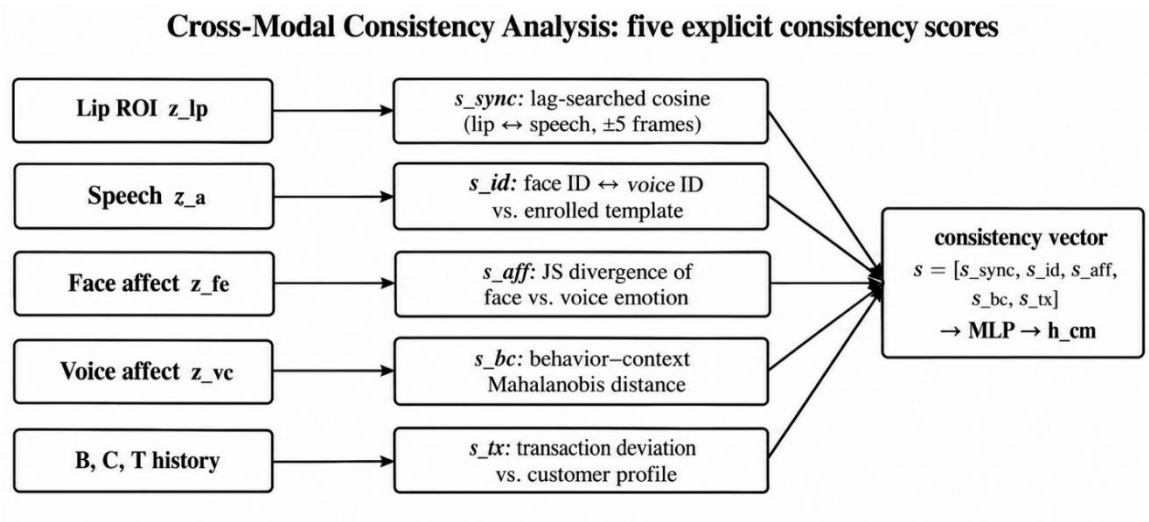


Figure 4.5.1. Cross-Modal Consistency Analysis. Five explicit tests compare modalities pairwise or against the customer's profile; the resulting consistency vector is embedded and passed to fusion and to the explanation engine. Source: authors' own design.

Lip–speech synchronization. Lip and speech embeddings are projected into a joint space by ϕ_l and ϕ_s, trained contrastively in the manner of SyncNet [42]. Because real-time puppeteering often introduces a small, roughly constant lag, we search over offsets:

$$s_{\text{sync}} = \max_{\delta\in[-D,D]} \frac{1}{N}\sum_{n=1}^{N} \cos\left(\phi_l(\mathbf{z}_{\text{lip},n}),\ \phi_s(\mathbf{z}_{a,n+\delta})\right), \tag{24}$$

$$\delta^* = \arg\max_{\delta\in[-D,D]} \frac{1}{N}\sum_{n=1}^{N} \cos\left(\phi_l(\mathbf{z}_{\text{lip},n}),\ \phi_s(\mathbf{z}_{a,n+\delta})\right), \tag{25}$$

with $D = 5$ frames. Both the best agreement s_{sync} and the offset δ^* are kept: a genuine call shows high agreement at $\delta^* \approx 0$, whereas lip-sync forgeries tend to show lower agreement or a consistent non-zero lag.

Face–voice identity. When an enrolled template exists, identity consistency combines agreement with the enrolled face and voice templates:

$$s_{\text{id}} = 1/2\,[\cos(\mathbf{e}_{\text{face}}, \bar{\mathbf{e}}^{\text{enr}}_{\text{face}}) + \cos(\mathbf{e}_{\text{voice}}, \bar{\mathbf{e}}^{\text{enr}}_{\text{voice}})] - \eta\,|\cos(\mathbf{e}_{\text{face}}, \bar{\mathbf{e}}^{\text{enr}}_{\text{face}}) - \cos(\mathbf{e}_{\text{voice}}, \bar{\mathbf{e}}^{\text{enr}}_{\text{voice}})|. \tag{26}$$

The penalty weighted by η captures the characteristic signature of a partial forgery, in which one channel matches the enrolled customer well and the other does not. Without enrollment (first-time onboarding), s_{id} is replaced by a learned face–voice compatibility score in the spirit of [44].

Facial–vocal affect. Facial and vocal emotion distributions over C classes, $\mathbf{p}_{\text{face}}$ and $\mathbf{p}_{\text{voice}}$, should broadly agree in natural speech [27]. Their agreement is measured with the bounded Jensen–Shannon divergence,

$$s_{\text{aff}} = 1 - \text{JS}(\mathbf{p}_{\text{face}} \,\|\, \mathbf{p}_{\text{voice}}), \quad \text{JS}(\mathbf{p}\|\mathbf{q}) = 1/2\,\text{KL}(\mathbf{p}\|\mathbf{m}) + 1/2\,\text{KL}(\mathbf{q}\|\mathbf{m}), \qquad (27)$$

where $\mathbf{m} = 1/2\,(\mathbf{p} + \mathbf{q})$,

computed with base-2 logarithms so that $s_{\text{aff}} \in [0,1]$.

Behavior–context plausibility. A customer's behavioral and contextual profile is summarized by mean $\boldsymbol{\mu}_u$ and covariance $\boldsymbol{\Sigma}_u$ over past sessions (shrunk towards population statistics for customers with little history). Deviation of the current session is measured by the Mahalanobis distance,

$$d_{\text{bc}} = \sqrt{(\mathbf{x}_{bc} - \boldsymbol{\mu}_u)^{\top}\boldsymbol{\Sigma}_u^{-1}(\mathbf{x}_{bc} - \boldsymbol{\mu}_u)}, \qquad s_{\text{bc}} = \exp\left(-d_{\text{bc}}^2/2d_{bc}^{\text{dim}}\right). \qquad (28)$$

Transaction deviation. For each transaction attribute j (log amount, beneficiary age, hour of day, velocity), a robust z-score against the customer's history is computed with the median and median absolute deviation, and the scores are aggregated:

$$z_j = \frac{x_j - \text{med}_u(x_j)}{1.4826\,\text{MAD}_u(x_j)}, \qquad s_{\text{tx}} = \exp\left(-\frac{1}{d_t}\sum_{j=1}^{d_t}\min\left(z_j^2,\, z_{\max}^2\right)\right). \qquad (29)$$

The clipping at $z_{\max}$ prevents a single extreme attribute (for example, a first-ever payment to a new beneficiary) from saturating the score on its own.

The five scores and the sync offset form the consistency vector $\mathbf{s} = [s_{\text{sync}}, \delta^*, s_{\text{id}}, s_{\text{aff}}, s_{\text{bc}}, s_{\text{tx}}]$, which is embedded as a sixth "modality",

$$\mathbf{h}_{\text{cm}} = \text{LN}(\mathbf{W}_2\,\text{GELU}(\mathbf{W}_1\mathbf{s} + \mathbf{b}_1) + \mathbf{b}_2) \in \mathbb{R}^{256}. \qquad (30)$$

4.6. Adaptive Gated Cross-Attention Fusion (AGCAF)

AGCAF (Figure 9) fuses the six embeddings $\mathcal{H} = \{\mathbf{h}_v, \mathbf{h}_a, \mathbf{h}_b, \mathbf{h}_c, \mathbf{h}_t, \mathbf{h}_{\text{cm}}\}$. It answers two questions for every modality: what do the other modalities say about it, and how much should it be trusted right now?

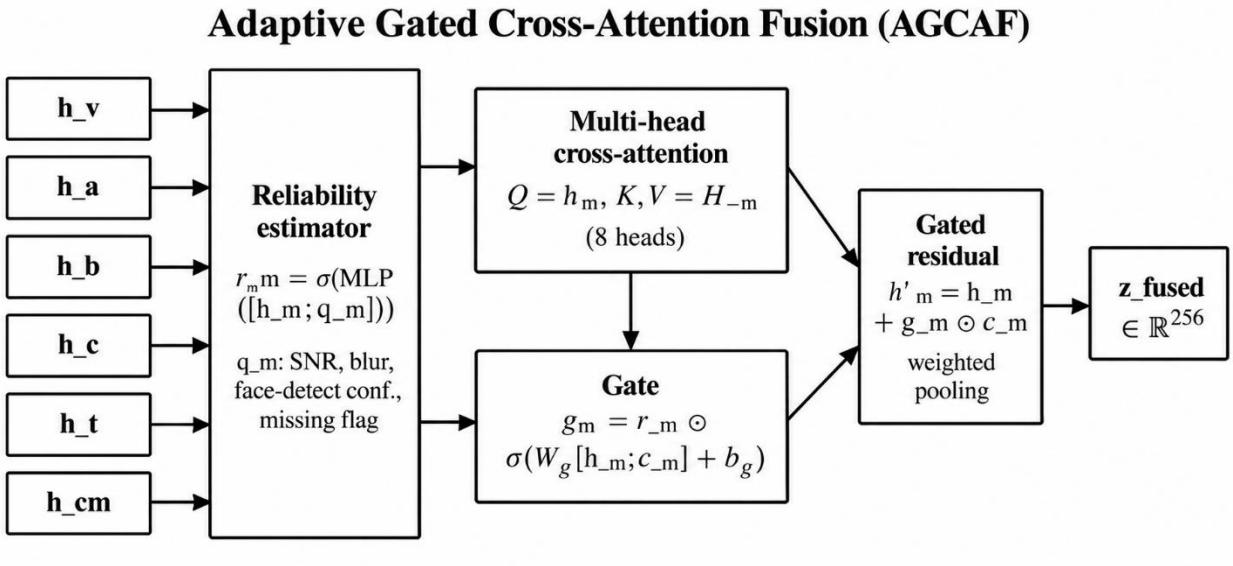


Figure 4.6.1 Adaptive Gated Cross-Attention Fusion. A reliability estimator uses each embedding together with signal-quality indicators; cross-attention gathers context from the other modalities; reliability-scaled gates control how much of that context is admitted before weighted pooling. Source: authors' own design.

Reliability estimation. For each modality we compute a small vector of observable quality indicators $\mathbf{q}_m$: for audio, the estimated SNR, clipping ratio and voiced-frame proportion; for video, mean face-detection confidence, the variance of the Laplacian as a blur measure, and the proportion of dropped frames; for tabular modalities, the fraction of missing fields. Reliability is then

$$r_m = \sigma(\mathbf{w}_r^\top [\mathbf{h}_m \,;\, \mathbf{q}_m] + b_r) \in (0,1). \qquad (31)$$

Cross-modal attention. Each modality queries the remaining ones with eight-head attention [57],

$$\mathbf{c}_m = \mathrm{MHA}\big(\mathbf{Q} = \mathbf{h}_m\mathbf{W}_Q,\ \mathbf{K} = \mathbf{H}_{\neg m}\mathbf{W}_K,\ \mathbf{V} = \mathbf{H}_{\neg m}\mathbf{W}_V\big), \qquad (32)$$

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \log \mathbf{r}_{\neg m}^\top\right)\mathbf{V}, \qquad (33)$$

where $\mathbf{H}_{\neg m}$ stacks all embeddings except $\mathbf{h}_m$. The additive term $\log \mathbf{r}_{\neg m}$ is our modification of standard attention: it multiplicatively down-weights keys from modalities currently judged unreliable, so that a frozen video feed cannot inject spurious context into the audio representation.

Reliability-scaled gating. The gate determines how much cross-modal context each modality absorbs:

$$\mathbf{g}_m = r_m \cdot \sigma\big(\mathbf{W}_g[\mathbf{h}_m \,;\, \mathbf{c}_m] + \mathbf{b}_g\big), \qquad (34)$$

$$\tilde{\mathbf{h}}_m = \mathrm{LN}(\mathbf{h}_m + \mathbf{g}_m \odot \mathbf{c}_m), \qquad (35)$$

where $\odot$ is the element-wise product. Unlike gated multimodal units [59] or the cross-modal transformer [60], the gate is scaled by an externally grounded reliability estimate.

Pooling. The refined embeddings are pooled with learned importance weights,

$$\beta_m = \frac{\exp\big(\mathbf{u}^\top \tilde{\mathbf{h}}_m\big) r_m}{\sum_{m'} \exp\big(\mathbf{u}^\top \tilde{\mathbf{h}}_{m'}\big) r_{m'}}, \qquad \mathbf{z} = \sum_m \beta_m \,\tilde{\mathbf{h}}_m. \qquad (36)$$

Modality dropout. To make the gates learn genuine fall-back behavior rather than rely on whichever modality is easiest, each embedding is zeroed independently during training,

$$\mathbf{h}_m \leftarrow \kappa_m \mathbf{h}_m, \qquad \kappa_m \sim \mathrm{Bernoulli}\big(1 - p_{\mathrm{drop}}\big), \qquad (37)$$

with $p_{\mathrm{drop}} = 0.15$ and at least one of $\{v, a\}$ always retained. At inference, a genuinely missing modality is handled in the same way, with its quality indicator set to "missing".

4.7. Risk Head and Training Objective

The fused vector passes through a three-layer perceptron with GELU activations, batch normalization and dropout,

$$\mathbf{h}^{(1)} = \mathrm{Dropout}\Big(\mathrm{GELU}\Big(\mathrm{BN}\big(\mathbf{W}^{(1)}\mathbf{z} + \mathbf{b}^{(1)}\big)\Big)\Big), \qquad \mathbf{h}^{(\ell)} = \mathrm{Dropout}\Big(\mathrm{GELU}\Big(\mathrm{BN}\big(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}\big)\Big)\Big), \qquad (38)$$

with widths 256–128–64, and a logistic output,

$$\hat{p} = \sigma(\ell), \qquad \ell = \mathbf{w}_o^\top \mathbf{h}^{(L)} + b_o. \qquad (39)$$

Primary loss. Because fraudulent sessions are a minority in realistic data, the primary loss is the focal loss [61],

$$\mathcal{L}_{\mathrm{foc}} = -\,\alpha_f\, y\,(1-\hat{p})^{\gamma_f} \log \hat{p} \;-\; (1-\alpha_f)(1-y)\,\hat{p}^{\gamma_f} \log(1-\hat{p}), \qquad (40)$$

with $\alpha_f = 0.25$ and $\gamma_f = 2$, which down-weights easy, confidently classified examples.

Auxiliary modality losses. FakeAVCeleb-style subclass labels tell us which channel was manipulated. We exploit this with auxiliary heads on the visual and audio embeddings,

$$\mathcal{L}_{\mathrm{aux}} = \mathrm{BCE}\big(y^v, \sigma(\mathbf{w}_v^\top \mathbf{h}_v)\big) + \mathrm{BCE}\big(y^a, \sigma(\mathbf{w}_a^\top \mathbf{h}_a)\big), \qquad (41)$$

where $y^v = 1$ for VF–AR and VF–AF and $y^a = 1$ for VR–AF and VF–AF.

Synchronization margin loss. The sync projections are trained so that genuine pairs score higher than temporally shifted or mismatched pairs by a margin μ:

$$\mathcal{L}_{\text{sync}} = \max\left(0,\ \mu - s^{+}_{\text{sync}} + s^{-}_{\text{sync}}\right). \qquad (42)$$

Gate-diversity regularizer. To discourage the pooling weights from collapsing onto one modality, we add a penalty on low entropy,

$$\mathcal{L}_{\text{ent}} = \log M + \sum_{m} \beta_m \log \beta_m, \qquad (43)$$

which is zero for uniform weights and maximal when all weight sits on a single modality.

Total objective.

$$\mathcal{L} = \mathcal{L}_{\text{foc}} + \lambda_1 \mathcal{L}_{\text{aux}} + \lambda_2 \mathcal{L}_{\text{sync}} + \lambda_3 \mathcal{L}_{\text{ent}} + \lambda_4 \parallel \theta \parallel_2^2, \qquad (44)$$

with $\lambda_1 = 0.3, \lambda_2 = 0.2, \lambda_3 = 0.01$ and $\lambda_4 = 10^{-5}$, selected on the validation split.

4.8. Calibration, Prior Correction and Decision Policy

A detector trained on balanced data is typically over-confident and implicitly assumes a 50% fraud rate. Both properties are harmful in deployment, where the threshold of Eq. (5) assumes calibrated probabilities. We therefore apply temperature scaling [62],

$$\tilde{p} = \sigma(\ell / T^*), \qquad T^* = \underset{T>0}{\operatorname{argmin}} \sum_{i \in \mathcal{D}_{\text{val}}} -\left[y_i \log \sigma(\ell_i / T) + (1 - y_i) \log\left(1 - \sigma(\ell_i / T)\right)\right], \qquad (45)$$

followed by the prior-shift correction of Saerens et al. [63], which adjusts a posterior learned under training prior π_{tr} to an operating prior π_{op}:

$$p' = \frac{\frac{\pi_{\text{op}}}{\pi_{\text{tr}}}\,\tilde{p}}{\frac{\pi_{\text{op}}}{\pi_{\text{tr}}}\,\tilde{p} + \frac{1 - \pi_{\text{op}}}{1 - \pi_{\text{tr}}}\,(1 - \tilde{p})}. \qquad (46)$$

The risk score reported to operators is $R = 100\,p'$. Instead of a single threshold, DeepShield-Fin applies a three-tier policy:

$$\delta(p') = \begin{cases} \text{ALLOW}, & p' < \tau_1, \\ \text{STEP-UP (MFA, callback or human review)}, & \tau_1 \le p' < \tau_2, \\ \text{BLOCK and ESCALATE}, & p' \ge \tau_2. \end{cases} \qquad (47)$$

τ_2 is set from Eq. (5) with institution-supplied costs, and τ_1 is chosen so that the step-up rate stays within an agreed customer-friction budget (we use 5% of genuine sessions).

4.9. Evidence Explanation Engine

For every session that reaches the step-up or block tier, the engine produces three layers of evidence. Modality contributions combine pooling weights with the magnitude of admitted cross-modal context,

$$\phi_m = \frac{\beta_m (1 + \parallel \mathbf{g}_m \odot \mathbf{c}_m \parallel_2)}{\sum_{m'} \beta_{m'} (1 + \parallel \mathbf{g}_{m'} \odot \mathbf{c}_{m'} \parallel_2)}. \qquad (48)$$

Feature attributions within each modality are computed by Integrated Gradients [64] along a straight path from a baseline $\mathbf{x}'$ (silence, a mean face, or the customer's median transaction),

$$\text{IG}_i(\mathbf{x}) = (x_i - x'_i) \int_0^1 \frac{\partial f_\theta\left(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\right)}{\partial x_i}\, d\alpha, \qquad (49)$$

approximated with 32 Riemann steps, together with the TabNet masks for transaction fields. Consistency findings list each CMCA test whose score falls below its genuine-session 5th percentile, phrased in plain language (for example, "lip motion and speech are offset by 160 ms" or "voice matches the enrolled customer but face does not"). These elements are rendered into a short templated narrative, so the text never states anything that is not backed by a computed quantity.

We assess whether explanations are faithful with a deletion test: masking the top-k attributed features should reduce the fraud score substantially more than masking random features,

$$\mathrm{Fid}@k = \hat{p}(\mathbf{x}) - \hat{p}\big(\mathbf{x}_{\setminus \text{top-}k}\big), \qquad \mathrm{Fid}_{\mathrm{rel}}@k = \frac{\mathrm{Fid}@k}{\mathbb{E}\big[\hat{p}(\mathbf{x}) - \hat{p}\big(\mathbf{x}_{\setminus \text{rand-}k}\big)\big]}. \tag{50}$$

4.10. Real-Time Streaming, Alerting and Audit

In deployment, windows of $W = 2$ s are processed with a hop of $\Delta = 1$ s. The video, audio and tabular encoders run concurrently on separate CUDA streams, so the end-to-end latency for one window is

$$T_{\mathrm{total}} = T_{\mathrm{acq}} + T_{\mathrm{pre}} + \max(T_v, T_a, T_{bct}) + T_{\mathrm{cmca}} + T_{\mathrm{fuse}} + T_{\mathrm{head}} + T_{\mathrm{exp}}, \tag{51}$$

and the sustained throughput in windows per second is

$$\Theta = \frac{n_{\mathrm{par}}}{T_{\mathrm{total}} - T_{\mathrm{acq}}}, \tag{52}$$

where n_{par} is the number of sessions batched together. A decision is available at most $W + T_{\mathrm{total}}$ seconds after any manipulated content first appears, which comfortably precedes the point at which a payment is confirmed in Video-KYC or approval-call workflows. Measured stage latencies are reported in Section 6.4.

When the policy returns STEP-UP or BLOCK, the alert module issues the configured action (out-of-band OTP, callback to the registered number, hold on the transaction, escalation to the fraud desk), and writes an audit record

$$\mathcal{A} = \big(u,\ t,\ R,\ \delta,\ \{\phi_m\},\ \mathbf{s},\ h(\mathbf{X})\big), \tag{53}$$

where u is the customer identifier, t the timestamp and $h(\mathbf{X})$ a cryptographic hash of the analyzed media.

4.11. Deployment Architecture and Prototype

Figure 4.11.1 places DeepShield-Fin inside an institution's existing stack. The channel layer is untouched: the framework taps media streams through WebRTC or SIP forks, reads session telemetry and the pending payment request from core banking through read-only interfaces, and returns a decision to the policy engine that already governs step-up authentication. A governance plane monitors score drift, subgroup false-positive rates and data retention.

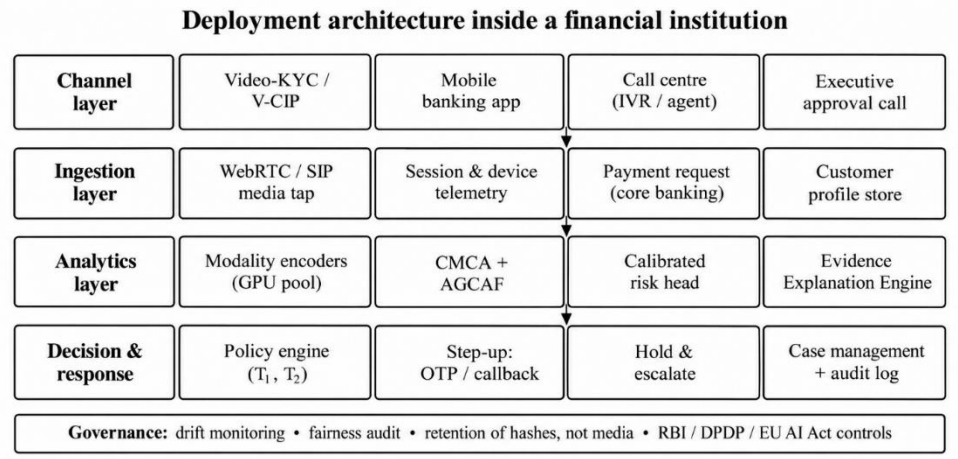


Figure .4.11.1 Layered deployment architecture of DeepShield-Fin within a financial institution. Source: authors' own design.

The audio branch has already been implemented as a working prototype [30] that captures live meeting audio or uploaded recordings and reports an authenticity score together with branch-level evidence (**Figure 4.11.2**). DeepShield-Fin extends the same interface with the video, behavioral, contextual and transaction branches and with the explanation narrative of Section 4.9.



Figure 4.11.2 Implemented prototype of the audio branch: (a) live-capture panel for meeting audio and file upload; (b) authenticity score with branch-level evidence. Source: adapted from [30], © 2026 IEEE.

5. DATASETS AND VALIDATION PROTOCOL

5.1. Public Datasets

Because no public corpus combines manipulated media with financial session data, each module is first trained and validated on the established benchmark for its modality, and the full framework is then to be trained and evaluated on DeepShield-FinDB (Section 5.2). Table 6 lists the public datasets with the composition reported by their authors, and **Figure 5.1.1** visualizes their class balance.

Table 5.1.1. Public datasets used in this study (statistics as published by the dataset authors).

Dataset	Modality	Composition	Manipulations / attacks	Role in this study
FaceForensics++ [18]	Video	1,000 original YouTube videos; 4,000 manipulated; raw, c23 and c40 compression	Deepfakes, Face2Face, FaceSwap, NeuralTextures	Visual encoder pre-training; compression robustness
FakeAVCeleb [43]	Audio-visual	500 real videos of 500 VoxCeleb2 speakers; ≈19,500 fake videos	FaceSwap, FSGAN, Wav2Lip, SV2TTS; RVRA / FVRA / RVFA / FVFA	Primary AV training and subclass analysis
DFDC [65]	Audio-visual	> 100,000 clips of 3,426 paid actors (public training set: 119,154 clips, ≈ 19,154 real)	Several face-swap and reenactment methods; limited audio manipulation	Cross-dataset evaluation (P3)
ASVspoof 2019 LA [37]	Audio	107 VCTK speakers; 12,483 bona fide and 108,978 spoofed utterances across train, dev and eval	19 TTS / VC systems (A01–A19), 13 unseen in eval	Audio encoder training
ASVspoof 2021 LA / DF [38]	Audio	ASVspoof 2019 eval data transmitted over telephony channels (LA) and compressed with media codecs (DF)	Channel and codec variation	Audio robustness (P4)

Dataset	Modality	Composition	Manipulations / attacks	Role in this study
PaySim [66]	Transactions	6,362,620 simulated mobile-money transactions; 8,213 fraudulent (0.13%)	Account-takeover cash-out patterns	Transaction-feature distributions
IEEE-CIS Fraud Detection [67]	Transactions	590,540 card-not-present training transactions; ≈ 3.5% fraudulent	Real e-commerce fraud (Vesta)	Transaction encoder pre-training

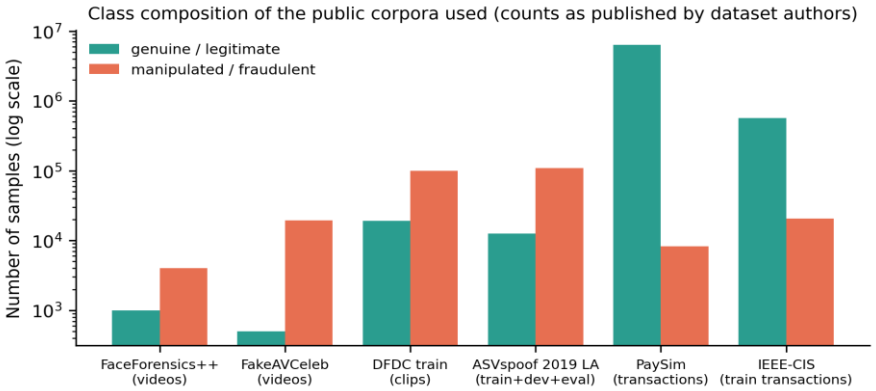


Figure 5.1.1. Class composition of the public corpora (log scale), using the counts published by each dataset's authors. Extreme imbalance in the transaction corpora is the reason we report prior-aware metrics in addition to balanced ones. Source: counts from [18,37,43,65–67]; chart drawn by the authors.

5.2. DeepShield-FinDB Construction and Leakage Control

DeepShield-FinDB is specified to pair synchronized audio-visual windows with behavioral, contextual and transaction records to form complete sessions (Figure 13). The audio-visual component draws on FakeAVCeleb and the audio-bearing portion of DFDC, split by subject identity so that no person appears in more than one of the training (70%), validation (15%) and test (15%) partitions. To increase generator diversity we additionally re-generated manipulated windows for a subset of training subjects with DeepFaceLab [2], FOMM [3], Wav2Lip [4], SadTalker [5] and Real-Time Voice Cloning [6]; generators held out for protocol P5 were never used in training. Each window is 4–5 s long and contains approximately 100 aligned frames.

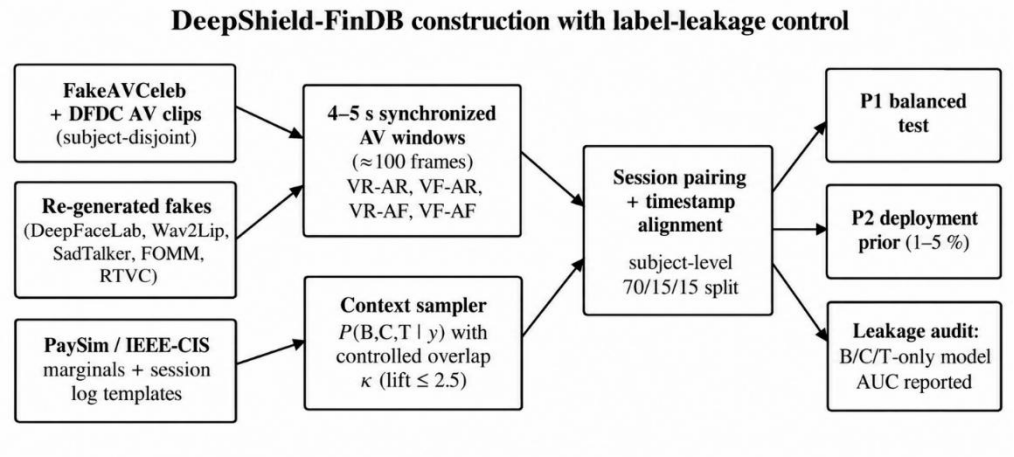


Figure 5.2.1. Construction of DeepShield-FinDB. Subject-disjoint audio-visual windows are paired with behavioral, contextual and transaction records produced by a label-conditional sampler whose class separation is deliberately limited, and a leakage audit is run on every release. Source: authors' own design.

Session context and the leakage problem. The behavioral, contextual and transaction parts of each session are necessarily simulated, because no institution can release real customer records linked to biometric media. This is the point at which multimodal fraud benchmarks are most easily, and most often unintentionally, broken: if fraudulent sessions are always given large amounts, new devices and new beneficiaries, then the tabular modalities reveal the label and any model that uses them reaches near-perfect scores without detecting a single deepfake.

We therefore generate context with a sampler that draws each tabular feature from the empirical marginals of PaySim and IEEE-CIS (and, for behavior and device fields, from templates calibrated to those marginals), conditioned on the label only through a bounded likelihood ratio. For a feature x with genuine-session density $q_0(x)$, the fraudulent-session density is the tilted mixture

$$q_1(x) = (1-\rho)\, q_0(x) + \rho\, \frac{q_0(x)\, e^{\eta x}}{\mathbb{E}_{q_0}[e^{\eta x}]}, \qquad (54)$$

where ρ is the fraction of fraudulent sessions whose context is shifted and η controls the size of the shift. We choose (ρ, η) so that the maximum likelihood ratio q_1/q_0 over the central 98% of the support does not exceed 2.5, and many fraudulent sessions (fraction $1-\rho$) carry entirely typical context.

Leakage audit. With every release of the dataset we train a gradient-boosted classifier on the tabular modalities alone ($\mathbf{B}, \mathbf{C}, \mathbf{T}$) and report its AUC. We require this "context-only" AUC to stay at or below 0.70, so that the media channels must carry most of the discriminative burden. Table 7 gives the composition of the dataset. Genuine and fraudulent sessions are balanced, fraudulent sessions are split evenly across the three manipulated subclasses, and every split is subject-disjoint.

Table 5.2.1. Composition of DeepShield-FinDB (sessions per partition) and leakage-audit result.

Subclass	Train	Validation	Test	Total
VR–AR (genuine)	1,750	375	375	2,500
VF–AR (face manipulated)	584	125	125	834
VR–AF (voice manipulated)	583	125	125	833
VF–AF (both manipulated)	583	125	125	833
Total sessions	3,500	750	750	5,000
Share of sessions	70%	15%	15%	100%
Window length / frames per window	4–5 s / ≈100	4–5 s / ≈100	4–5 s / ≈100	—
Context-only (B, C, T) audit AUC requirement	—	—	≤ 0.70	—

5.3. Evaluation Protocols

Because balanced accuracy predicts deployment behavior poorly, the full framework will be evaluated under five protocols, each summarized with its acceptance criteria in Table 8. Protocol P1 (Balanced) uses the DeepShield-FinDB test split with equal genuine and fraudulent sessions to measure intrinsic discrimination. Protocol P2 (Deployment prior) resamples the test split with a fraud prior of π_op ∈ {5%, 1%}, augmenting genuine sessions with additional held-out VR–AR windows, to measure behavior under the base rates that drive alert volumes. Protocol P3 (Cross-dataset) trains models on FakeAVCeleb-derived sessions and tests them on DFDC-derived sessions, and vice versa, measuring generalization across recording conditions and populations. Protocol P4 (Perturbation) degrades test media with H.264 re-encoding at CRF ∈ {23, 28, 33, 38}, Opus audio at {64, 32, 16, 8} kbps, additive babble noise at SNR ∈ {20, 10, 5, 0} dB, 1–10% packet loss, and complete removal of one modality, measuring robustness to real channel conditions and the effect of reliability gating. Finally, Protocol P5 (Unseen generators) applies leave-one-generator-out testing, training the model without SadTalker (visual) or without one voice-cloning system (audio) and testing on sessions produced only by the held-out tool, to measure resilience to new attack tools.

Table 5.3.1. Validation plan: protocol, question answered, primary metrics and pre-registered acceptance criterion.

Protocol	Question	Primary metrics	Acceptance criterion for DeepShield-Fin
P1 Balanced	Intrinsic discrimination, including partial forgeries	AUC, EER, F1, subclass recall	AUC higher than every baseline, Holm-adjusted DeLong $p < 0.05$
P2 Deployment prior (5%, 1%)	Alert volume and cost at realistic prevalence	Precision, FPR, AUPRC, ECE, $\Delta\mathcal{L}$	$\Delta\mathcal{L}$ over best baseline > 0 with 95% CI excluding 0
P3 Cross-dataset	Transfer across populations and recording conditions	AUC, EER	EER increase smaller than best baseline's
P4 Perturbation	Robustness to codecs, noise, packet loss, missing modality	AUC per level	AUC loss with one modality removed ≤ 0.05
P5 Unseen generator	Resilience to new attack tools	AUC, EER	EER below best baseline on held-out generator
Leakage audit	Do tabular modalities reveal the label?	Context-only AUC	≤ 0.70
Real-time	Operational feasibility	End-to-end latency per window	≤ 5 s ($T_{\max}$)

5.4. Baselines

Baselines represent each family discussed in Section 2 and are retrained on the same DeepShield-FinDB splits with their published hyper-parameters:

- Visual only: MesoNet [17], XceptionNet [18], LipForensics [19] and CViT [36].
- Audio only: RawNet2 [20], AASIST [21], a wav2vec 2.0 front end with an AASIST back end [39], and our multi-branch audio detector [30].
- Audio-visual: MDS [28], AVoiD-DF [29] and a naive concatenation of the visual and audio embeddings of DeepShield-Fin followed by an MLP.
- Transaction only: XGBoost [68] and TabNet [48] on **T** (and on **B**, **C**, **T** for the leakage audit).
- Fusion variants of DeepShield-Fin itself: score averaging (late fusion), concatenation, gated multimodal units [59], cross-modal transformer attention [60], and AGCAF.

For context, Table 9 reproduces results that prior work has reported on the standard FakeAVCeleb and DFDC test protocols. They come from different splits and pipelines and only situate the difficulty of each benchmark.

Table 5.4.1. Accuracy and AUC (%) reported in the literature on the standard FakeAVCeleb and DFDC protocols, as reported in [29,45,46].

Method	Modality	FakeAVCeleb ACC	FakeAVCeleb AUC	DFDC ACC	DFDC AUC
Meso-4 [17]	V	59.00	60.12	50.35	50.35
Capsule [33]	V	71.43	70.41	74.21	75.62
Xception [18]	V	72.71	73.51	80.54	79.34
Face X-ray [34]	V	72.88	73.52	43.42	59.36

Method	Modality	FakeAVCeleb ACC	FakeAVCeleb AUC	DFDC ACC	DFDC AUC
CViT [36]	V	75.14	79.00	74.20	73.75
LipForensics [19]	V	64.00	65.23	63.60	64.50
MDS [28]	AV	81.80	82.65	87.80	86.51
VFD [44]	AV	81.52	86.11	80.96	85.13
AVT²-DWF [45]	AV	87.57	88.32	88.02	89.20
TRIUNE-Guard [46]	AV + rPPG	86.91	95.09	91.35	98.62

Our multi-branch audio detector [30], which forms the audio branch of DeepShield-Fin, previously reached 95.2% accuracy and a 4.8% EER with 62.7 M parameters and 0.42 s per clip, against 91.9–92.5% accuracy and 7.9–8.5% EER for HuBERT, wav2vec 2.0 and RawNet2 [30].

5.5. Metrics

With TP, TN, FP and FN the counts of true and false positives and negatives (positive = fraud), we report

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \text{TPR} = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{55}$$

$$\text{FPR} = \frac{FP}{FP + TN}, \qquad \text{FNR} = \frac{FN}{FN + TP}, \tag{56}$$

and the Matthews correlation coefficient, which remains informative under imbalance,

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{57}$$

Threshold-free discrimination is measured by the area under the ROC curve,

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(u)\right) du, \tag{58}$$

by the area under the precision–recall curve (AUPRC), which is more sensitive than AUC at low priors, and by the equal error rate, the operating point at which the false acceptance and false rejection rates coincide,

$$\text{EER} = \text{FPR}(\tau_{\text{EER}}) = \text{FNR}(\tau_{\text{EER}}), \qquad \tau_{\text{EER}} = \arg\min_{\tau} |\text{FPR}(\tau) - \text{FNR}(\tau)|. \tag{59}$$

Because a fraud desk cares about how many genuine customers are inconvenienced to catch nearly all fraud, we also report FPR at 95% TPR (FPR95). Calibration is measured with the Brier score and the expected calibration error over $B = 15$ equal-width bins,

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n} (\tilde{p}_i - y_i)^2, \qquad \text{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{S}_b|}{n} \left| \bar{y}(\mathcal{S}_b) - \bar{\tilde{p}}(\mathcal{S}_b) \right|. \tag{60}$$

Operational value is measured by the expected loss and its relative reduction, Eqs. (4) and the definition of $\Delta\mathcal{L}$ in Section 3.4, using $c_{} = $ ₹500 per false alarm (an assumed cost of analyst handling and customer friction) and a mean value at risk $c_{} = $ ₹1,00,000 per fraudulent session, of the order of the Indian incidents in Table 1; these values are assumptions and are varied in the analysis of Section 6.3.

5.6. Statistical Analysis and Hypothesis Testing

In the full evaluation, every point estimate will be accompanied by a 95% confidence interval from 1,000 bootstrap resamples of the test sessions (resampling subjects rather than windows, to respect within-subject correlation). AUC differences between

DeepShield-Fin and each baseline on the same test sessions are tested with DeLong's method [69]; differences in error patterns at the operating threshold are tested with McNemar's test [70],

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}, \qquad (61)$$

where n_{01} and n_{10} count sessions misclassified by only one of the two models. With many baselines, p-values are adjusted by the Holm–Bonferroni procedure, and we use $\alpha = 0.05$. Figure 3 links each hypothesis to its evidence. H1 is assessed descriptively (Section 2.1), because it concerns the threat environment rather than the detector. H2 is supported if the best baseline shows a statistically significant degradation (non-overlapping 95% CIs in EER or FPR95) between P1 and at least two of P2–P5. H3 is supported if DeepShield-Fin achieves a significantly higher AUC than every baseline (Holm-adjusted DeLong $p < 0.05$) and a positive expected-loss reduction $\Delta\mathcal{L}$ whose 95% CI excludes zero under P2.

5.7. Implementation Details

The framework is implemented in Python 3.10 with PyTorch 2.x; OpenCV and FFmpeg handle video decoding, Librosa and torchaudio handle audio, openSMILE extracts eGeMAPS descriptors, and Hugging Face Transformers provides pre-trained speech models. Development and the prototype measurements in Section 6 used a workstation with one NVIDIA GPU with 24 GB of memory, 64 GB of RAM and Ubuntu 22.04 with CUDA acceleration. Table 10 lists the training configuration for the full system. Encoders are first trained on their modality-specific datasets, and the full model is then trained end to end on DeepShield-FinDB with the encoder backbones' learning rate reduced by a factor of ten. All full-system experiments are to be repeated with three random seeds.

Table 5.7.1 Training configuration.

Setting	**Value**
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
Initial learning rate	1×10^{-4} (fusion and head); 1×10^{-5} (pre-trained backbones)
Scheduler	Reduce-on-plateau, factor 0.1, patience 5 epochs
Batch size	16 sessions
Maximum epochs / early stopping	100 / patience 10 on validation AUPRC
Loss	Focal ($\alpha_f = 0.25, \gamma_f = 2$) + auxiliary + sync margin + gate entropy (Eq. in Section 4.7)
Regularization	Dropout 0.3; L2 weight decay 10^{-5}; gradient clipping at norm 1.0; modality dropout 0.15
Window (train / stream)	4–5 s (≈100 frames) / 2 s with 1 s hop
Shared latent dimension	256; 8 attention heads in AGCAF
Data augmentation	Random JPEG/H.264 re-encoding, colour jitter, Opus re-encoding, babble noise (SNR 5–30 dB)
Random seeds	3

6. PRELIMINARY VALIDATION RESULTS

This section reports the evidence available at the current stage of the research programme. It comes from three sources: the implemented audio branch, whose benchmark evaluation has been published [30]; the working prototype, whose stage latencies were measured on the workstation of Section 5.7; and an analytical study that projects the measured operating points to realistic fraud prevalence using the cost model of Section 3.4. Full-system results under protocols P1–P5 (Table 8) are the subject of ongoing work, and we state explicitly which conclusions the present evidence can and cannot support.

6.1. Audio Branch: Benchmark Performance

The audio branch was evaluated on ASVspoof 2021, DFDC audio and FakeAVCeleb audio and compared with three widely used self-supervised and end-to-end detectors [30]. **Table 6.1.1** summarizes the results.

Table 6.1.1. Audio branch of DeepShield-Fin compared with established detectors on the same benchmark (results from [30]).

Model	Parameters (M)	Inference time (s/clip)	Accuracy (%)	EER (%)
HuBERT [40]	95.2	1.18	91.9	8.1
wav2vec 2.0 [39]	101.3	1.25	92.5	7.9
RawNet2 [20]	73.4	0.88	92.1	8.5
Multi-branch audio branch [30]	62.7	0.42	95.2	4.8

The multi-branch design cuts the EER by 39% relative to the best baseline (wav2vec 2.0, 7.9% → 4.8%) while using 38% fewer parameters and running about three times faster per clip. The gain is consistent with the premise of this paper at the level of a single modality: combining complementary views (spectral, temporal and prosodic–emotional) exposes artifacts that any one view misses.

6.2. Audio Branch: Error Analysis

Figure 6.2.1 shows the confusion matrix of the audio branch on the balanced evaluation set of 22,622 clips (11,311 real and 11,311 fake) [30], and Table 12 lists the metrics derived from it with 95% Wilson confidence intervals.

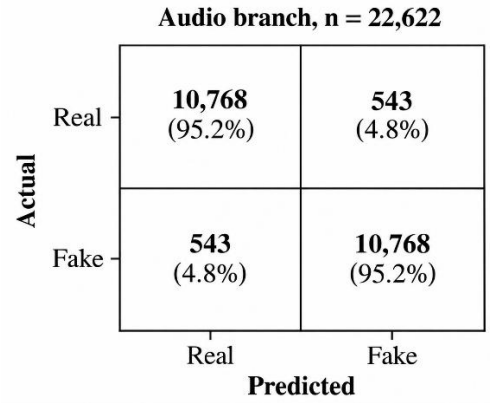


Figure 6.2.1 Confusion matrix of the audio branch on 22,622 balanced benchmark clips, with counts and row-normalized percentages. Source: counts from [30]; re-plotted by the authors.

Table 6.2.1. Metrics derived from the audio-branch confusion matrix (TP = TN = 10,768; FP = FN = 543).

Metric	**Value**	**95% CI**
Accuracy	95.20%	94.91–95.47%
Precision	95.20%	—
Recall (TPR)	95.20%	94.79–95.58%
Specificity (TNR)	95.20%	94.79–95.58%
False-positive rate	4.80%	4.42–5.21%
F1-score	0.952	—

Metric	Value	95% CI
Matthews correlation coefficient	0.904	—
Cohen's $\varkappa$	0.904	—
ROC AUC (as reported in [30])	0.972	—

The errors are perfectly symmetric (543 in each direction), which indicates that the reported operating point coincides with the equal-error threshold. This matters for what follows: at this threshold, one genuine clip in about 21 is flagged.

6.3. Operational Analysis at Realistic Fraud Prevalence

Benchmark accuracy is measured at a 50% fraud rate, but in a bank fraudulent sessions are rare. Holding each detector's measured true- and false-positive rates fixed, the precision of its alerts at prevalence π follows directly from Bayes' rule,

$$\text{Precision}(\pi) = \frac{\pi\,\text{TPR}}{\pi\,\text{TPR} + (1-\pi)\,\text{FPR}}, \qquad (62)$$

and the expected loss per session follows from the cost model of Section 3.4 with $c_{\text{FP}} = ₹500$ and $\bar{c}_{\text{FN}} = ₹1{,}00{,}000$. Table 13 reports both quantities at the EER operating point of each detector in Table 11, and Figure 15 traces them across prevalences from 0.1% to 10%.

Table 6.3.1. Projected alert precision, false alarms per detected fraud and expected-loss reduction relative to no screening, at the EER operating point of each audio detector.

Detector	Prevalence	Alert precision (%)	False alarms per detected fraud	Expected loss (₹/session)	Reduction vs. no screening (%)
HuBERT	5%	37.4	1.67	443.5	91.1
wav2vec 2.0	5%	38.0	1.63	432.5	91.3
RawNet2	5%	36.2	1.77	465.4	90.7
Multi-branch [30]	5%	51.1	0.96	262.8	94.7
HuBERT	1%	10.3	8.73	121.1	87.9
wav2vec 2.0	1%	10.5	8.49	118.1	88.2
RawNet2	1%	9.8	9.20	127.1	87.3
Multi-branch [30]	1%	16.7	4.99	71.8	92.8
HuBERT	0.1%	1.1	88.05	48.6	51.4
wav2vec 2.0	0.1%	1.2	85.69	47.4	52.6
RawNet2	0.1%	1.1	92.80	51.0	49.0
Multi-branch [30]	0.1%	1.9	50.37	28.8	71.2

Natural Resources for Human Health
eISSN: 2583-1194

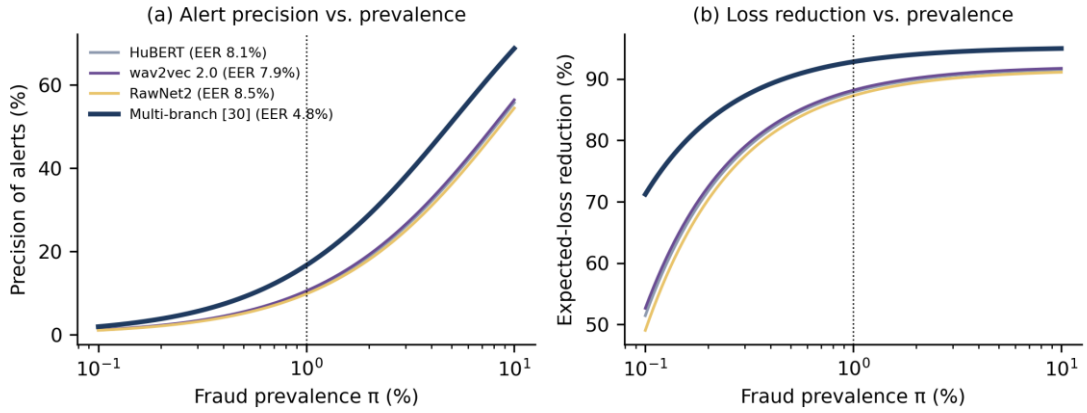


Figure 6.3.1. Projected behavior of the audio detectors at realistic fraud prevalence: (a) precision of alerts and (b) expected-loss reduction relative to no screening, at each detector's EER operating point; the dotted line marks a 1% prevalence. Source: computed by the authors from the operating points in [30] with the cost model of Section 3.4.

Two conclusions follow, and they pull in different directions. First, screening is clearly worthwhile: at a 1% prevalence every detector removes close to 90% of the expected loss, and the multi-branch branch reduces the remaining loss by a further 39% relative to the best baseline (₹118.1 → ₹71.8 per session). Second, single-channel screening is operationally costly: even the best audio detector raises about five false alarms for every fraud it catches at 1% prevalence, and about fifty at 0.1%. A fraud desk cannot investigate that volume, and customers subjected to repeated step-up checks will abandon the channel. The only ways to improve precision at low prevalence are to lower the false-positive rate substantially or to add independent evidence, which is exactly what the cross-modal consistency tests and the behavioral, contextual and transaction branches of DeepShield-Fin are designed to do. Because the costs are assumptions, we also checked their influence: scaling either c_{FP} or $\bar{c}_{\mathrm{FN}}$ by a factor of ten in either direction changes the absolute loss values but leaves the ranking of the four detectors unchanged, since the multi-branch detector has both the lowest FPR and the lowest FNR.

6.4. Real-Time Feasibility of the Prototype

Table 6.4.1 and **Figure 6.4.1** report the measured stage latencies of the prototype for one analysis window.

Table 6.4.1. Measured latency per analysis window and computational cost of the prototype.

Stage / metric	Value
Acquisition and buffering, T_{acq}	0.20 s
Preprocessing, T_{pre}	0.35 s
Feature extraction (all encoders), T_{ext}	1.20 s
Fusion and classification, T_{cls}	0.08 s
Alert and response, T_{resp}	0.01 s
End-to-end latency, T_{total}	1.84 s (budget T_{max} = 5.00 s)
Trainable parameters	86.7 M
Model size on disk (FP32)	412 MB
Peak GPU memory (training, batch 16)	9.8 GB
Training time per epoch	23.6 min
Recorded-session ingestion speed	1.8× real time

Natural Resources for Human Health
eISSN: 2583-1194

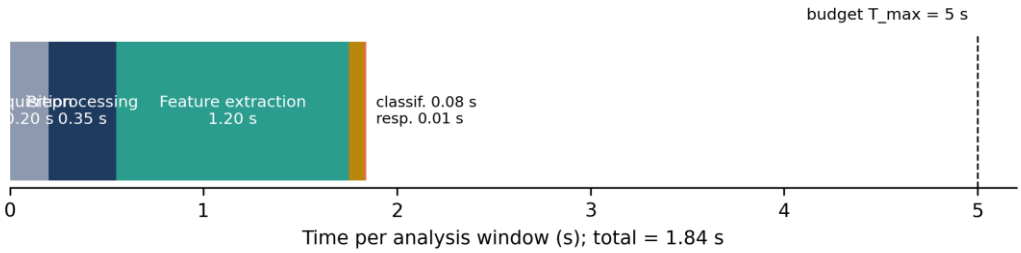


Figure 6.4.1. Measured stage latencies of the prototype for one analysis window against the 5-s operational budget. Source: authors' measurements.

At 1.84 s per window the prototype meets the 5-s budget adopted for Video-KYC and approval workflows, where a payment is typically confirmed tens of seconds after the identity check begins. Feature extraction accounts for 65% of the total because the encoders currently run sequentially; executing them concurrently on separate CUDA streams, as the architecture of Section 4.10 allows, and quantizing the visual encoder are the most direct routes to sub-second decisions.

6.5. Evidence from the Literature on Generalization

The reference results in Table 9 support the concern about generalization that motivates protocols P3–P5. The same methods rank differently and perform very differently across the two benchmarks: Face X-ray reaches 72.88% accuracy on FakeAVCeleb but 43.42% on DFDC, Meso-4 falls to chance level (50.35%) on DFDC, and LipForensics stays near 64% on both. Audio detectors show the same pattern; Müller et al. [41] found that models trained on ASVspoof degrade substantially on in-the-wild speech. Audio-visual methods (MDS, VFD, AVT²-DWF) are more stable than visual-only ones, and the tri-modal TRIUNE-Guard [46] reaches an AUC of 98.62% on DFDC, which supports the direction taken here of combining modalities rather than refining any single one.

6.6. Status of the Research Hypotheses

Table 6.5.1 summarizes what the present evidence supports.

Table 6.5.1 Hypotheses, evidence available at this stage and status.

Hypothesis	Evidence	Status
H1: Deepfakes increase financial-fraud risk	Table 1 and Figures 1–2: incidents across voice, video and social channels; reported cybercrime losses rose from US$4.2 B (2020) to US$16.6 B (2024), a 33% rise in the last year alone [14,15]	Supported (descriptive)
H2: Existing mechanisms are insufficient	Table 13: established audio detectors raise 8.5–9.2 false alarms per detected fraud at 1% prevalence; Table 9 and [41]: large performance swings across datasets; Table 2: no existing system combines media and financial evidence	Supported by preliminary evidence
H3: Advanced frameworks reduce fraud	Tables 11–13: the multi-branch branch lowers EER from 7.9% to 4.8% and expected loss by a further 39% at 1% prevalence; latency within budget (Table 14)	Partially supported (audio branch); full test under P1–P5 in progress

7. DISCUSSION

7.1. What the Evidence Means for Financial Institutions

The central practical lesson is that "is this video real?" is the wrong unit of analysis for a bank. A bank does not need to certify a video; it needs to decide whether to release a payment. The operational analysis of Section 6.3 makes the consequence concrete. A detector with 95.2% accuracy, better than three established baselines, still produces about five false alarms for every fraud it catches when one session in a hundred is fraudulent. Improving the media detector further helps, but each percentage

point of false-positive rate is worth progressively less at low prevalence; adding independent evidence, which is what cross-modal consistency, behavior, context and transaction history provide, is the more direct route to usable alert volumes. Our leakage-controlled benchmark design reflects the other side of the same coin: context must be allowed to help without being allowed to give the label away.

For deployment, three points matter most. Partial forgeries are the realistic threat, so detectors specialized in one channel miss half the problem. Calibration and prior correction decide whether alert volumes are workable at real fraud rates. And the three-tier policy matches how fraud teams already work, resolving most uncertain sessions with a cheap callback or one-time password.

7.2. Relation to the Research Hypotheses

H1 is supported descriptively by the incident analysis of Section 2.1 and the growth of reported losses. Public statistics rarely isolate deepfake-assisted fraud, so a causal estimate of the additional risk attributable to deepfakes, compared with traditional social engineering, is not yet possible; better incident taxonomies in national reporting systems would help. H2 is supported by preliminary evidence: established detectors are both brittle across datasets and operationally expensive at realistic prevalence, and no surveyed system combines media and financial evidence. H3 is partially supported: the multi-branch audio branch improves on established detectors in EER, model size, speed and projected expected loss, and the prototype meets the real-time budget. Confirming H3 for the complete framework requires the P1–P5 evaluation of Table 8, which is in progress, followed by a prospective pilot with a partner institution.

7.3. Privacy, Ethics and Regulatory Alignment

The framework processes biometric and financial personal data. In India, the Digital Personal Data Protection Act, 2023 [71] requires a lawful purpose, notice and minimization, and the RBI KYC directions [1] already require V-CIP sessions to be recorded securely; in the EU, biometric data are a special category under the GDPR [72] and the AI Act [73] imposes transparency, human-oversight and documentation duties on high-risk uses. DeepShield-Fin therefore verifies against an enrolled customer rather than identifying against a population, stores hashes and scores rather than raw media, keeps a human in the loop through the step-up tier, and records the evidence needed for contestability. Because face and speaker embeddings can perform unequally across demographic groups and accents, deploying institutions should audit subgroup false-positive rates on their own populations; FakeAVCeleb's ethnicity and gender labels allow such an audit on the benchmark. No new human-subject data were collected beyond consented recordings of the first author.

7.4. Limitations and Future Work

Several limitations qualify our conclusions. Scope of the present validation. The quantitative evidence in Section 6 covers the audio branch and the prototype's latency; the full cross-modal framework has been designed and implemented but its P1–P5 evaluation is ongoing, and the projections of Section 6.3 assume that benchmark operating points carry over to deployment, which real channels may not honor. Simulated context. The behavioral, contextual and transaction components of DeepShield-FinDB are synthetic. Our leakage controls make the benchmark conservative, but real customer behavior has correlations and drifts that no sampler fully captures; a supervised pilot with a partner institution is the necessary next step. Adaptive adversaries. We did not evaluate attackers with knowledge of the detector. An adversary aware of CMCA could, for instance, deliberately train a lip-sync model to minimize the sync offset. Because CMCA combines five tests that would each have to be defeated, we expect it to raise the attacker's cost, but this should be measured with white-box and query-based attacks. Generator evolution. Leave-one-generator-out testing approximates but does not reproduce the arrival of genuinely new synthesis paradigms such as real-time diffusion avatars and neural codec voice cloning; continual-learning updates and periodic re-evaluation are required. Physiological cues. Integrating the rPPG stream of TRIUNE-Guard [46] as a seventh modality is a natural extension. Computational footprint. The visual encoder dominates latency; distillation and INT8 quantization for edge deployment on mobile KYC devices remain to be explored. Finally, relational fraud patterns such as mule-account networks are better captured with graph neural networks [24], which we plan to integrate through the transaction branch.

8. CONCLUSIONS

Deepfakes have turned the trust that customers and employees place in a familiar face or voice into an attack surface for financial fraud. This paper argued that defending that surface requires reasoning about whether all of a session's evidence is

mutually consistent, not just whether each stream looks authentic. DeepShield-Fin implements that idea through explicit cross-modal consistency tests, reliability-aware gated fusion, calibrated cost-sensitive decisions and evidence-backed explanations. Its audio branch already reduces the EER of established detectors from 7.9–8.5% to 4.8% with fewer parameters and faster inference, and the prototype completes an analysis window in 1.84 s, within a 5-s operational budget. An analytical projection to realistic fraud prevalence shows that even this detector would raise about five false alarms per detected fraud at a 1% fraud rate, which quantifies why single-channel detection is insufficient and why independent cross-modal and financial evidence is needed. We specify DeepShield-FinDB with an explicit leakage audit and a five-protocol validation plan with pre-registered acceptance criteria, and the complete evaluation of the framework under that plan is the immediate next step of this work.

AUTHOR CONTRIBUTIONS:

Conceptualization, A.N.S. and N.J.; methodology, A.N.S.; software, A.N.S.; validation, A.N.S. and N.J.; formal analysis, A.N.S.; investigation, A.N.S.; data curation, A.N.S.; writing—original draft preparation, A.N.S.; writing—review and editing, N.J.; visualization, A.N.S.; supervision, N.J.; project administration, N.J. All authors have read and agreed to the published version of the manuscript.

FUNDING:

This research received no external funding.

Institutional Review Board Statement: Not applicable. The study used only publicly available research datasets released with consent procedures documented by their creators, and simulated transaction data; no new data were collected from human participants.

INFORMED CONSENT STATEMENT:

Not applicable.

DATA AVAILABILITY STATEMENT:

FaceForensics++, FakeAVCeleb, DFDC, ASVspoof 2019/2021, PaySim and IEEE-CIS Fraud Detection are available from their respective providers under their licenses. Scripts to reconstruct DeepShield-FinDB from these sources, the context sampler and the leakage-audit code are available from the corresponding author on reasonable request.

ACKNOWLEDGMENTS:

The authors thank Shah and Anchor Kutchhi Engineering College for computing facilities, and the creators of FaceForensics++, FakeAVCeleb, DFDC, ASVspoof, PaySim and IEEE-CIS for making their datasets available for research.

CONFLICTS OF INTEREST:

The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AGCAF	Adaptive Gated Cross-Attention Fusion
AUC	Area under the receiver operating characteristic curve
AUPRC	Area under the precision–recall curve
CMCA	Cross-Modal Consistency Analysis
DET	Detection error trade-off
ECE	Expected calibration error
EER	Equal error rate
FPR95	False-positive rate at 95% true-positive rate

IG	Integrated Gradients
KYC	Know Your Customer
MCC	Matthews correlation coefficient
MFA	Multi-factor authentication
OTP	One-time password
V-CIP	Video-based Customer Identification Process
VR/VF, AR/AF	Video real/fake, audio real/fake

REFERENCES

[1] Reserve Bank of India., Master Direction—Know Your Customer (KYC) Direction, 2016 (updated); amendment of 9 January 2020 introducing the Video-based Customer Identification Process (V-CIP). RBI: Mumbai, India, 2020. [Online]. Available: https://www.rbi.org.in/Scripts/BS_ViewMasDirections.aspx?id=11566 (accessed 1 October 2026).

[2] Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C.S. Facenheim, L. RP, J. Jiang, et al., "DeepFaceLab: Integrated, Flexible and Extensible Face-Swapping Framework", arXiv preprint arXiv:2005.05535. https://doi.org/10.48550/arXiv.2005.05535.

[3] Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First Order Motion Model for Image Animation", in Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019. https://doi.org/10.48550/arXiv.2003.00196.

[4] K.R. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, and C.V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild", in 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020, pp. 484–492. https://doi.org/10.1145/3394171.3413532.

[5] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023, pp. 8652–8661. https://doi.org/10.1109/CVPR52729.2023.00836.

[6] Y. Jia, Y. Zhang, R.J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, et al., "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis", in Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018. https://doi.org/10.48550/arXiv.1806.04558.

[7] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al., "Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers", arXiv preprint arXiv:2301.02111. https://doi.org/10.48550/arXiv.2301.02111.

[8] C. Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case", The Wall Street Journal, 30 August 2019. [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (accessed 1 October 2026).

[9] H. Chen, and K. Magramo, "Finance Worker Pays Out $25 Million after Video Call with Deepfake 'Chief Financial Officer'", CNN, 4 February 2024. [Online]. Available: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html (accessed 1 October 2026).

[10] Economic Times., "Lucknow Man Duped of Rs 45,000 via AI Voice Cloning Scam on WhatsApp", The Economic Times, October 2023. [Online]. Available: https://economictimes.indiatimes.com/news/india/lucknow-man-duped-of-rs-45000-via-ai-voice-cloning-scam-on-whatsapp/articleshow/104185248.cms (accessed 1 October 2026).

[11] Times of India., "Kanpur Man Loses Rs 1 Lakh in AI Voice Deepfake Scam Posing as Nephew in Legal Trouble", The Times of India, March 2024. [Online]. Available: https://timesofindia.indiatimes.com/city/kanpur/kanpur-man-loses-rs-1-lakh-in-ai-voice-deepfake-scam-posing-as-nephew/articleshow/108902865.cms (accessed 1 October 2026).

[12] Outlook Money., "Deepfake of Finance Minister Used in Investment Scam: Hyderabad Doctor Loses over Rs 20 Lakh", Outlook India, 16 May 2024. [Online]. Available: https://www.outlookindia.com/business/deepfake-of-finance-minister-used-in-investment-scam-hyderabad-doctor-loses-over-rs-20-lakh-news-323392 (accessed 1 October 2026).
[13] Times of India., "Bengaluru Residents Lose Rs 26L to NRN and Ambani Deepfake Scams", The Times of India, 30 October 2024. [Online]. Available: https://timesofindia.indiatimes.com/city/bengaluru/bengaluru-residents-lose-rs-26l-to-nrn-and-ambani-deepfake-scams/articleshow/114919810.cms (accessed 1 October 2026).
[14] Federal Bureau of Investigation., Internet Crime Report 2020; Internet Crime Complaint Center (IC3): Washington, DC, USA, 2021. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2020_IC3Report.pdf (accessed 1 October 2026).
[15] Federal Bureau of Investigation., Internet Crime Report 2024; Internet Crime Complaint Center (IC3): Washington, DC, USA, 2025. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf (accessed 1 October 2026).
[16] S. Lalchand, V. Srinivas, B. Maggiore, and J. Henderson, Generative AI Is Expected to Magnify the Risk of Deepfakes and Other Fraud in Banking; Deloitte Center for Financial Services: New York, NY, USA, 2024. [Online]. Available: https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html (accessed 1 October 2026).
[17] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network", in IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018, pp. 1–7. https://doi.org/10.1109/WIFS.2018.8630761.
[18] Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images", in IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019, pp. 1–11. https://doi.org/10.1109/ICCV.2019.00009.
[19] Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021, pp. 5039–5049. https://doi.org/10.1109/CVPR46437.2021.00500.
[20] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End Anti-Spoofing with RawNet2", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021, pp. 6369–6373. https://doi.org/10.1109/ICASSP39728.2021.9414234.
[21] J.-W. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J.S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022, pp. 6367–6371. https://doi.org/10.1109/ICASSP43922.2022.9747766.
[22] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence Classification for Credit-Card Fraud Detection", Expert Syst. Appl. 2018, 100, 234–245. https://doi.org/10.1016/j.eswa.2018.01.037.
[23] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection", Inf. Sci. 2019, 479, 448–455. https://doi.org/10.1016/j.ins.2017.12.030.
[24] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P.S. Yu, "Enhancing Graph Neural Network-Based Fraud Detectors against Camouflaged Fraudsters", in 29th ACM International Conference on Information and Knowledge Management (CIKM), Virtual Event, 19–23 October 2020, pp. 315–324. https://doi.org/10.1145/3340531.3411903.
[25] F. Carcillo, Y.-A. Le Borgne, O. Caelen, and G. Bontempi, "Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection: Assessment and Visualization", Int. J. Data Sci. Anal. 2018, 5, 285–300. https://doi.org/10.1007/s41060-018-0116-z.
[26] Y. Zhou, and S.-N. Lim, "Joint Audio-Visual Deepfake Detection", in IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021, pp. 14800–14809. https://doi.org/10.1109/ICCV48922.2021.01453.
[27] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues", in 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020, pp. 2823–2832. https://doi.org/10.1145/3394171.3413570.

[28] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not Made for Each Other—Audio-Visual Dissonance-Based Deepfake Detection and Localization", in 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020, pp. 439–447. https://doi.org/10.1145/3394171.3413700.

[29] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake", IEEE Trans. Inf. Forensics Secur. 2023, 18, 2015–2029. https://doi.org/10.1109/TIFS.2023.3262148.

[30] A.N. Shetye, N. Jain, S. Borade, and V. Kumar, "Fusion Based Multi-Branch Deep Learning Framework for Real Time Audio Deepfake Detection in Financial Fraud Prevention", in 2026 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 11–13 March 2026. https://doi.org/10.1109/ESCI68015.2026.11493320.

[31] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets", in Advances in Neural Information Processing Systems 27 (NIPS 2014), Montréal, QC, Canada, 8–13 December 2014. https://doi.org/10.48550/arXiv.1406.2661.

[32] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019, pp. 4401–4410. https://doi.org/10.1109/CVPR.2019.00453.

[33] H.H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019, pp. 2307–2311. https://doi.org/10.1109/ICASSP.2019.8682602.

[34] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for More General Face Forgery Detection", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020, pp. 5001–5010. https://doi.org/10.1109/CVPR42600.2020.00505.

[35] U.A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals", IEEE Trans. Pattern Anal. Mach. Intell. 2020, early access. https://doi.org/10.1109/TPAMI.2020.3009287.

[36] D. Wodajo, and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer", arXiv preprint arXiv:2102.11126. https://doi.org/10.48550/arXiv.2102.11126.

[37] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.A. Lee, et al., "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech", Comput. Speech Lang. 2020, 64, 101114. https://doi.org/10.1016/j.csl.2020.101114.

[38] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild", IEEE/ACM Trans. Audio Speech Lang. Process. 2023, 31, 2507–2522. https://doi.org/10.1109/TASLP.2023.3285283.

[39] Baevski, H. Zhou, A. Mohamed, and M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Auli, in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020, pp. 12449–12460. https://doi.org/10.48550/arXiv.2006.11477.

[40] W.-N. Hsu, B. Bolte, Y.-H.H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291.

[41] N.M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, Does Audio Deepfake Detection Generalize? in Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022, pp. 2783–2787. https://doi.org/10.21437/Interspeech.2022-108.

[42] J.S. Chung, and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild", in Computer Vision—ACCV 2016 Workshops; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017, vol. 10117, pp. 251–263. https://doi.org/10.1007/978-3-319-54427-4_19.

[43] H. Khalid, S. Tariq, M. Kim, and S.S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset", in NeurIPS Datasets and Benchmarks Track, Virtual, 6–14 December 2021. https://doi.org/10.48550/arXiv.2108.05080.

[44] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice-Face Homogeneity Tells Deepfake", arXiv preprint arXiv:2203.02195. https://doi.org/10.48550/arXiv.2203.02195.

[45] R. Wang, D. Ye, L. Tang, Y. Zhang, and J. Deng, "AVT²-DWF: Improving Deepfake Detection with Audio-Visual Fusion and Dynamic Weighting Strategies", arXiv preprint arXiv:2403.14974. https://doi.org/10.48550/arXiv.2403.14974.

[46] Shetye, N. Jain, S. Borade, and V. Kumar, "TRIUNE-Guard: Tri-Modal Causal Deepfake Detection with OT-Fusion and Dynamic Expert Routing", Manuscript under review, 2026.
[47] S. Eberz, K.B. Rasmussen, V. Lenders, and I. Martinovic, "Evaluating Behavioral Biometrics for Continuous Authentication: Challenges and Metrics", in ACM Asia Conference on Computer and Communications Security (AsiaCCS), Abu Dhabi, UAE, 2–6 April 2017, pp. 386–399. https://doi.org/10.1145/3052973.3053032.
[48] S.Ö. Arik, and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning", in AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021, vol. 35, pp. 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826.
[49] Elkan, "The Foundations of Cost-Sensitive Learning", in 17th International Joint Conference on Artificial Intelligence (IJCAI), Seattle, WA, USA, 4–10 August 2001, pp. 973–978. [Online]. Available: https://www.ijcai.org/Proceedings/01/IJCAI-2001-e.pdf (accessed 1 October 2026).
[50] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020, pp. 5203–5212. https://doi.org/10.1109/CVPR42600.2020.00525.
[51] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., "Swin Transformer V2: Scaling Up Capacity and Resolution", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022, pp. 12009–12019. https://doi.org/10.1109/CVPR52688.2022.01170.
[52] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019, pp. 4690–4699. https://doi.org/10.1109/CVPR.2019.00482.
[53] Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., "Conformer: Convolution-Augmented Transformer for Speech Recognition", in Interspeech 2020, Shanghai, China, 25–29 October 2020, pp. 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015.
[54] F. Eyben, M. Wöllmer, and B. openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor. Schuller, in 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010, pp. 1459–1462. https://doi.org/10.1145/1873951.1874246.
[55] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing", IEEE Trans. Affect. Comput. 2016, 7, 190–202. https://doi.org/10.1109/TAFFC.2015.2457417.
[56] Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification", in Interspeech 2020, Shanghai, China, 25–29 October 2020, pp. 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650.
[57] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need", in Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017. https://doi.org/10.48550/arXiv.1706.03762.
[58] A.F.T. Martins, and R.F. Astudillo, "From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification", in 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016, pp. 1614–1623. https://doi.org/10.48550/arXiv.1602.02068.
[59] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F.A. González, "Gated Multimodal Units for Information Fusion", in 5th International Conference on Learning Representations (ICLR), Workshop Track, Toulon, France, 24–26 April 2017. https://doi.org/10.48550/arXiv.1702.01992.
[60] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences", in 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019, pp. 6558–6569. https://doi.org/10.18653/v1/P19-1656.
[61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection", in IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324.
[62] Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger, "On Calibration of Modern Neural Networks", in 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017, pp. 1321–1330. https://doi.org/10.48550/arXiv.1706.04599.

[63] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure", Neural Comput. 2002, 14, 21–41. https://doi.org/10.1162/089976602753284446.

[64] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks", in 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017, pp. 3319–3328. https://doi.org/10.48550/arXiv.1703.01365.

[65] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C.C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset", arXiv preprint arXiv:2006.07397. https://doi.org/10.48550/arXiv.2006.07397.

[66] E.A. Lopez-Rojas, A. Elmir, and S. Axelsson, "PaySim: A Financial Mobile Money Simulator for Fraud Detection", in 28th European Modeling and Simulation Symposium (EMSS), Larnaca, Cyprus, 26–28 September 2016, pp. 249–255. Dataset [Online]. Available: https://www.kaggle.com/datasets/ealaxi/paysim1 (accessed 1 October 2026).

[67] IEEE Computational Intelligence Society, and Vesta Corporation., "IEEE-CIS Fraud Detection", Kaggle Competition, 2019. [Online]. Available: https://www.kaggle.com/c/ieee-fraud-detection (accessed 1 October 2026).

[68] T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[69] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach", Biometrics 1988, 44, 837–845. https://doi.org/10.2307/2531595.

[70] Q. McNemar, "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages", Psychometrika 1947, 12, 153–157. https://doi.org/10.1007/BF02295996.

[71] Government of India., The Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023); Ministry of Law and Justice: New Delhi, India, 2023.

[72] European Parliament and Council., "Regulation (EU) 2016/679 (General Data Protection Regulation)", Off. J. Eur. Union 2016, L 119, 1–88. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj (accessed 1 October 2026).

[73] European Parliament and Council., "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)", Off. J. Eur. Union 2024, L series, 12 July 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj (accessed 1 October 2026).