

Development of an Intelligent Multimodal Biometric Identification System for Students in A Digital Educational Environment

Akmoldina AI*, Mukhiyaidin AU, Abdrakhmanov DM, Kenebayeva DB and Zeken A

ESIL UNIVERSITY Astana, Kazakhstan

***Corresponding author:** Akmoldina A.I, ESIL UNIVERSITY Astana, Kazakhstan

ARTICLE INFO

Received: 📅 October 09, 2026

Published: 📅 October 26, 2026

Citation: Akmoldina AI, Mukhiyaidin AU, Abdrakhmanov DM, Kenebayeva DB and Zeken A. Development of an Intelligent Multimodal Biometric Identification System for Students in A Digital Educational Environment. Biomed J Sci & Tech Res 67(1)-2026. BJSTR. MS.ID.010441.

ABSTRACT

Online examinations require reliable evidence that the registered learner is the person completing an assessment throughout the session, rather than only at login. Passwords and possession-based credentials cannot independently establish this link, while unimodal face verification remains vulnerable to changes in illumination, pose, and presentation attacks. This paper presents the design and pilot evaluation of a multimodal biometric verification framework that combines face embeddings, speaker verification, a layered face presentation-attack detection pipeline, and continuous session monitoring. The prototype is implemented in Python and uses a standard RGB webcam and microphone. Face representations are extracted with FaceNet512 and compared by cosine distance. Voice verification is described in the source implementation through a speaker-embedding pathway based on the Generalized End-to-End (GE2E) approach and Resemblyzer; a legacy MFCC-Gaussian mixture model (GMM) pathway is also documented and must be distinguished during replication. At score level, face and voice scores are combined using weights of 0.6 and 0.4, respectively. Presentation-attack detection combines Laplacian texture statistics and HSV skin-colour features with eye-blink analysis based on Eye Aspect Ratio (EAR) and a randomized head-turn challenge evaluated using solvePnP. Continuous monitoring uses intersection-over-union (IoU) tracking to maintain face-track identity and a voice activity detection (VAD) component to trigger periodic voice checks. In the reported pilot involving four participants, the system achieved 0% false acceptance and 0% false rejection under the stated normal-condition test, while the reported presentation-attack detection rate was 100% for the tested photograph and replay scenarios. A sequence of 15 repeated checks produced a reported fusion-score standard deviation of 0.019. Performance deteriorated under poor illumination, where the source reports a false rejection rate of 75%. These results demonstrate technical feasibility, not population-level performance: the small sample, limited attack diversity, and incomplete threshold calibration restrict generalization. The paper contributes an integrated, reproducible architecture and a transparent account of its evaluation boundaries, and outlines a validation protocol for larger studies.

Keywords: multimodal biometrics; online proctoring; face verification; speaker verification; presentation attack detection; liveness detection; continuous authentication; score-level fusion; digital education

Abbreviations: GE2E: Generalized End-to-End, GMM: Gaussian Mixture Model, EAR: Eye Aspect Ratio, IOU: Intersection-Over-Union, VAD: Voice Activity Detection, MFCCs: Mel-Frequency Cepstral Coefficients, APCER: Attack Presentation Classification Error Rate, TTL: Time-To-Live, BPCER: Bona Fide Presentation Classification Error Rate, FAR: False Acceptance Rate, FRR: False Rejection Rate, EER: Equal Error Rate, ROC: Receiver Operating Characteristic

Highlights

- A practical face-voice verification architecture is designed for standard laptop webcams and microphones.
- A layered presentation-attack detection pipeline combines texture, blink, and randomized head-turn checks.
- IoU-based tracking and periodic voice checks extend verification beyond the login event.
- Pilot results are reported together with a major illumination-related failure mode and sample-size limitations.
- The manuscript distinguishes prototype observations from claims that require larger, independent evaluation.

Introduction

Digital and blended learning have become established components of higher education. Remote examinations, however, create a verification problem: an institution must determine whether the enrolled learner is completing an assessment and whether that identity remains consistent throughout the session. A successful login does not prevent a different person from taking over later, and a password or one-time code proves possession of a credential rather than the identity of the person using it. The problem is especially important in high-stakes assessments, professional certification, and remote admission procedures, where confidence in the assessment process affects the credibility of the resulting qualification. Biometric verification offers a way to associate a session with measurable physiological or behavioural characteristics. Face verification can be performed with an ordinary RGB camera, and speaker verification can use a built-in microphone. These modalities are attractive because they require little additional hardware, but neither is reliable in every condition. Face images change with pose, illumination, camera quality, and occlusion. Voice recordings vary with microphone placement, background noise, speaking duration, and the acoustic environment. Both modalities can also be attacked: a camera may be shown a printed photograph or replayed video, while a microphone may receive a recording or synthetic speech. A multimodal approach can reduce dependence on a single signal source. It can combine independently generated similarity scores and apply a decision rule that accounts for the relative reliability of each channel. Nevertheless, combining modalities does not automatically guarantee security. Fusion can propagate errors if the scores are poorly calibrated, if both channels are affected by correlated conditions, or if a presentation attack bypasses the acquisition stage. The design must therefore integrate biometric matching, presentation-attack detection, session-state management, and secure handling of templates. This paper presents a research prototype for online educational assessment that integrates face verification, voice verification, layered face liveness checks, and continuous monitoring. The implementation described in the supplied source uses a modular Python desktop application and consumer-grade acquisition devices. It is designed to process camera frames locally, reducing the need to transmit raw video to a remote service. The contribution is architectural and empirical: the paper documents how the modules interact, reports the available pilot observations, and identifies the additional evidence required before operational deployment.

a. Research Problem and Questions

The central research problem is how to combine two biometric channels with presentation-attack detection and continuous session monitoring in a resource-constrained online examination environment. The study is guided by three questions:

1. How can face and voice verification be integrated at score level while retaining independent modality scores for diagnosis?
2. Can a layered liveness pipeline detect the limited classes of

photograph and video-replay attacks considered in the prototype?

3. What operational limitations emerge when the prototype is evaluated under changes in illumination, pose, voice level, and repeated verification?

b. Contributions

The work makes four bounded contributions. First, it describes a modular architecture that separates acquisition, face processing, voice processing, score fusion, security services, and persistence. Second, it combines passive texture analysis with active blink and randomized head-turn checks. Third, it extends the initial identity check through IoU-based face tracking and periodic voice activity-triggered verification. Fourth, it reports pilot results and adverse conditions rather than treating a small demonstration as conclusive evidence of general accuracy. The novelty claim is limited to this specific integration and workflow; broader claims of being the first system of its kind require a systematic literature search and should not be made without further verification.

Background and related work

a. Biometric Verification in Remote Assessment

Biometric systems generally perform either identification or verification. Identification searches for a person among a set of enrolled identities, whereas verification tests a claimed identity against a stored template. Online proctoring normally uses one-to-one verification after the learner has authenticated to the learning platform. Continuous authentication repeats or maintains this verification during the assessment, addressing the possibility that the person in front of the camera changes after the initial check. Biometric characteristics are often grouped into physiological and behavioural categories. Face geometry and iris patterns are physiological; voice contains both anatomical and behavioural information. The choice of modality in a remote assessment is constrained by hardware availability, accessibility, privacy, and environmental variability. Fingerprint and iris systems may deliver strong performance in specialized settings but generally require sensors that are not present on every student's device. Face and voice are therefore operationally attractive, though their performance must be evaluated under realistic capture conditions.

b. Face Representations and Metric Comparison

Classical face-recognition methods such as Eigenfaces, Fisherfaces, and local binary patterns rely on engineered representations and can be sensitive to lighting and pose. Deep convolutional networks learn hierarchical image features and map faces into an embedding space. FaceNet introduced metric learning with a triplet loss, encouraging embeddings from the same identity to be closer than embeddings from different identities. The supplied prototype uses FaceNet512 through the DeepFace framework and compares embeddings with cosine distance. A five-view enrollment procedure—frontal, left, right, upward, and downward—aims to reduce

sensitivity to pose by averaging the extracted representations. For embeddings e_a and e_b , cosine distance can be expressed as $d_{cos} = 1 - (e_a \cdot e_b) / (\|e_a\| \|e_b\|)$. Smaller distances indicate greater directional similarity. A threshold converts the distance into a verification decision. Threshold selection is application-specific: a permissive threshold can reduce false rejections but increase false acceptances, whereas a strict threshold can have the opposite effect. Published benchmark performance should not be interpreted as a guarantee for a local webcam system, since image quality, demographic composition, and acquisition conditions differ.

c. Speaker Verification

Speaker verification estimates whether an audio sample was produced by a claimed speaker. Earlier systems commonly combined mel-frequency cepstral coefficients (MFCCs) with Gaussian mixture models (GMMs). MFCCs summarize the spectral envelope of speech, while a GMM models the distribution of acoustic feature vectors. Neural embedding approaches, including x-vectors and GE2E-trained representations, seek to produce compact vectors that preserve speaker-discriminative information across utterances. The supplied document describes a Resemblyzer/GE2E path that generates 256-dimensional speaker embeddings. It also includes a legacy MFCC-GMM implementation, with 40 MFCC coefficients and 16 Gaussian components, and a score normalization rule for log-likelihood values. These are distinct approaches and should not be conflated. For a reproducible evaluation, the final implementation must state which voice backend was active in each experiment, the exact model version, sample duration, audio resampling settings, and the threshold applied. In this manuscript, the GE2E embedding path is treated as the intended primary architecture, while the GMM path is noted as a legacy implementation described in the source.

d. Presentation-Attack Detection and Liveness

A presentation attack attempts to deceive a biometric sensor by presenting an artefact or manipulated signal. Face presentation attacks include printed photographs, replayed video, and three-dimensional masks. Voice attacks include replayed speech, text-to-speech synthesis, and voice conversion. ISO/IEC 30107-3 provides a framework for testing and reporting presentation-attack detection. Relevant measures include the attack presentation classification error rate (APCER) and bona fide presentation classification error rate (BPCER), which should be reported by attack species and test condition. Passive face presentation-attack detection examines signal properties without asking the user to perform a task. Texture measures, including local patterns and Laplacian variance, can help distinguish a live face from a displayed image, while colour features in HSV space can provide additional cues. Such measures are not universal indicators of liveness: thresholds depend on camera focus, display properties, compression, lighting, and skin appearance. Active methods ask the user to perform an unpredictable action. Blink detection based on facial landmarks and Eye Aspect Ratio can help reject static photo-

graphs, while a randomized head-turn challenge makes a fixed replay less likely to satisfy the request. These methods remain vulnerable to sophisticated, responsive, or generated attacks and therefore need systematic testing.

e. Multimodal Fusion

Multimodal biometric systems can combine signals at the sensor, feature, score, or decision level. Feature-level fusion joins representations before matching but may be difficult when modalities have different dimensions and statistical properties. Decision-level fusion combines separate accept/reject outputs, which can discard useful confidence information. Score-level fusion retains a scalar confidence measure from each matcher and combines the scores before applying a final threshold. For a two-channel system, a simple weighted rule is

$$S_{fusion} = w_f S_{face} + w_v S_{voice}, \quad \text{where } w_f + w_v = 1.$$

The prototype uses $w_f = 0.6$ and $w_v = 0.4$, reflecting the design assumption that face matching is the primary channel while voice adds corroborating evidence. These weights should not be considered universally optimal. A rigorous deployment would calibrate each score on a representative development set, normalize the scales, select operating thresholds for an explicit security/usability objective, and report results on a separate test set.

f. Research Gap

The supplied document's comparison of three commercially described proctoring platforms—OES, Proctor.kz, and ProctorEdu—suggests that publicly available product descriptions emphasize face detection, behavioural monitoring, and noise detection. The document does not identify public technical documentation establishing independent speaker verification, a randomized challenge-response protocol, or transparent score-level fusion for those products. This should be interpreted as a limitation of publicly accessible documentation, not proof that the commercial systems do not implement these capabilities internally. The research opportunity is therefore a transparent, inspectable prototype whose mechanisms and limitations can be evaluated independently.

System Design and Methods

Design Principles

(Figure 1) The prototype follows a modular architecture implemented in Python 3.10. The modules separate user interaction, audio/video acquisition, face verification, voice verification, fusion, security, and database operations. This separation allows individual components to be tested and replaced without redesigning the complete application. A desktop architecture was selected in the source project to avoid network latency in the real-time camera loop and to process biometric signals locally. The implementation uses OpenCV for video capture, DeepFace with FaceNet512 for face embeddings, MediaPipe facial landmarks for geometry, audio-processing librar-

ies for voice analysis, SciPy for cosine similarity/distance, and cryptographic functions for protecting stored templates. The architecture can be described as six logical layers: [1] user interface and enrollment workflow; [2] acquisition and preprocessing; [3] parallel face and voice verification; [4] score-level fusion and decision; [5] security controls, liveness and audit logging; and [6] database persistence. The

face and voice paths operate independently until fusion. Template encryption and access controls apply to persistence operations. The architecture is a research prototype and requires further engineering before it can be considered suitable for high-stakes production use [7] (Figure 2 & Table 1).

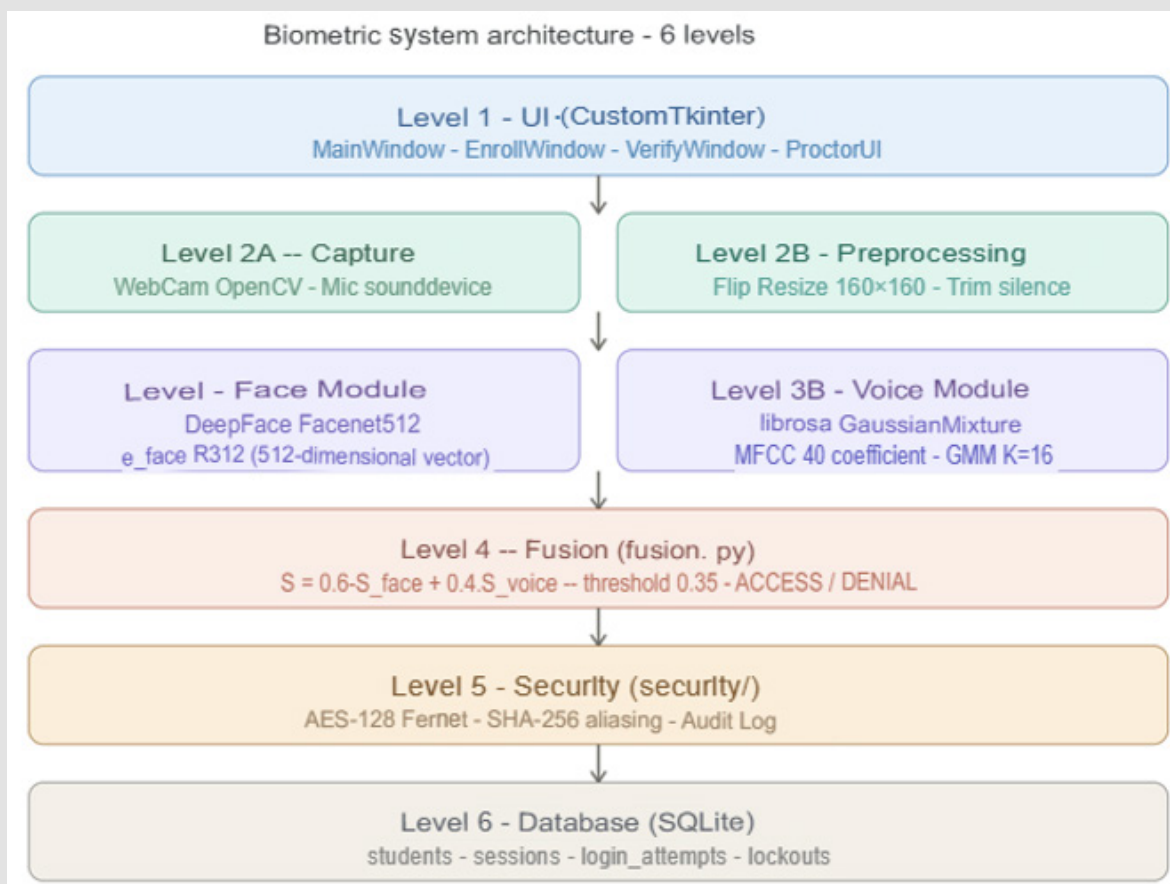


Figure 1: Overall architecture of the multimodal biometric verification system.



Figure 2: Face-enrollment examples from the source prototype.

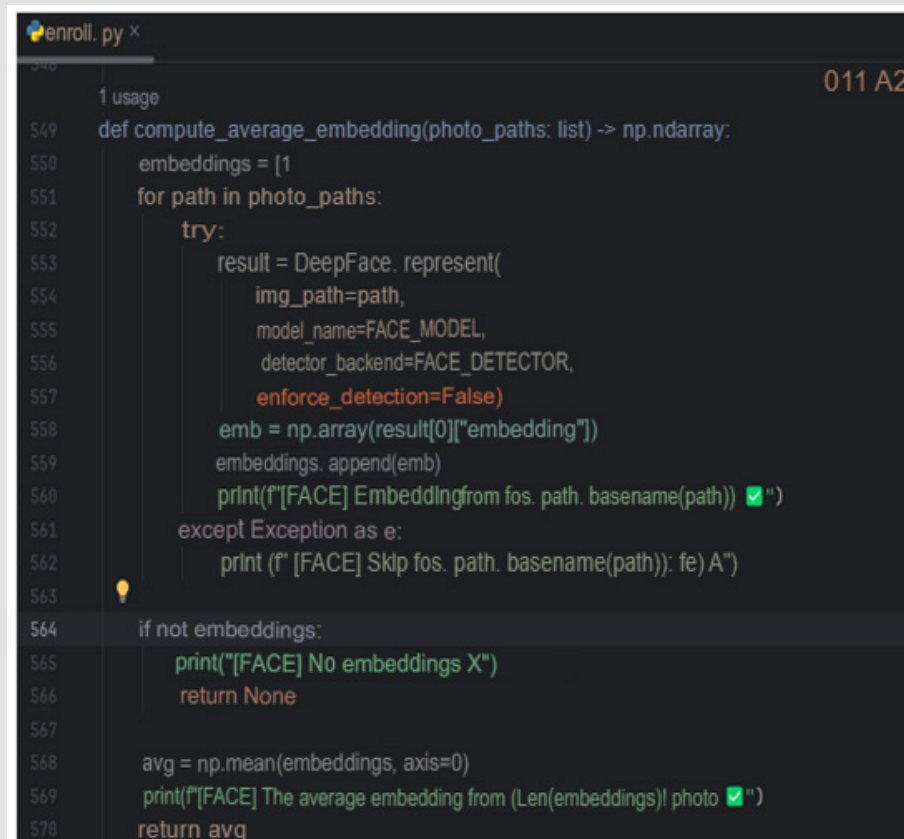
Table 1

Layer	Main functions	Principal output
1. User interface	Enrollment, verification workflow, prompts and status	User action and workflow state
2.Acquisition	Webcam frames, microphone samples, preprocessing	Image and audio segments
3A. Face module	Face detection, FaceNet512 embedding, cosine distance	Face score and liveness cues
3B. Voice module	Voice activity detection and speaker representation	Voice score
4. Fusion	Weighted score combination and threshold decision	Accept, reject, or review state
5. Security	Presentation-attack checks, encryption, audit events	Security state and event log
6. Persistence	Student/session records and encrypted templates	Stored verification artefacts

Enrollment and Face Template Generation

The face enrollment workflow captures five views: frontal, left profile, right profile, upward tilt, and downward tilt. Each image is processed to obtain a 512-dimensional FaceNet512 embedding. The template is computed as the component-wise mean of the available embeddings, $e_template = (1/N)\sum e_i$, where N is the number of valid enrollment views. Averaging can reduce view-specific variation, but it does not remove the need for quality control. A production implementation should reject blurred, overexposed, occluded, or poor-

ly aligned images and should assess whether averaging normalized embeddings is more appropriate than averaging raw outputs for the selected model. During verification, a live frame is acquired after the liveness procedure. Its embedding is compared with the registered template using cosine distance. The match decision is obtained by comparing the distance with a threshold [8]. The supplied source reports approximate same-person distances of 0.20–0.35 and different-person distances of 0.50–0.80 on its stated model context, but such ranges are not universal and must not be substituted for threshold calibration on the target population [9]



```

1 usage
549 def compute_average_embedding(photo_paths: list) -> np.ndarray:
550     embeddings = []
551     for path in photo_paths:
552         try:
553             result = DeepFace.represent(
554                 img_path=path,
555                 model_name=FACE_MODEL,
556                 detector_backend=FACE_DETECTOR,
557                 enforce_detection=False)
558             emb = np.array(result[0]["embedding"])
559             embeddings.append(emb)
560             print(f"[FACE] Embedding from fos. path. basename(path))")
561         except Exception as e:
562             print(f"[FACE] Skip fos. path. basename(path): {e}")
563
564     if not embeddings:
565         print("[FACE] No embeddings X")
566         return None
567
568     avg = np.mean(embeddings, axis=0)
569     print(f"[FACE] The average embedding from (Len(embeddings))! photo")
570     return avg
  
```

Figure 3: Voice-processing implementation excerpt from the source prototype.

Voice Enrollment and Verification

(Figure 3) The intended primary voice path uses Resemblyzer with a GE2E-trained encoder to produce speaker embeddings. Enrollment collects several short samples and stores an identity representation derived from the recordings [10]. Verification captures a test utterance and compares its embedding with the enrolled representation using a similarity measure. The source also documents an alternative GMM pipeline: the audio is resampled to 16 kHz, silence is trimmed, 40 MFCC coefficients are extracted, and a 16-component Gaussian mixture is fitted to the speaker's feature distribution. A log-likelihood score is then normalized using a sigmoid transformation [11]. Because the two paths have different score distributions, they cannot be compared or combined without calibration. For publication-quality reproducibility, each reported trial should record the backend identifier, model weights, audio length, sample rate, preprocessing settings, and score calculation. If the GMM path remains in the codebase, it should be treated as an alternative baseline rather than described as identical to the GE2E system [13].

Score-Level Fusion

The prototype's fusion score is computed as $S_{fusion} = 0.6 S_{face} + 0.4 S_{voice}$. The final decision is made by comparing the fusion score with an acceptance threshold. In the supplied architecture diagram, a threshold of approximately 0.35 is shown, while other tables report normalized scores and acceptance decisions. Since score conventions can differ—cosine distance is lower-is-better, whereas similarity scores are typically higher-is-better—the implementation must explicitly transform each modality into a consistent direction before fusion [14]. A safer generalized formulation is $S_{fusion} = w_f g_f(s_f) + w_v g_v(s_v)$, where g_f and g_v are calibration mappings learned on development data. The mappings can be logistic calibration, empirical quantile normalization, or another validated transformation. The current fixed weights are a prototype design choice, not evidence that the fusion rule is statistically optimal. The calibration set must be separated from the final test set to avoid optimistic estimates [15] (Figure 4).

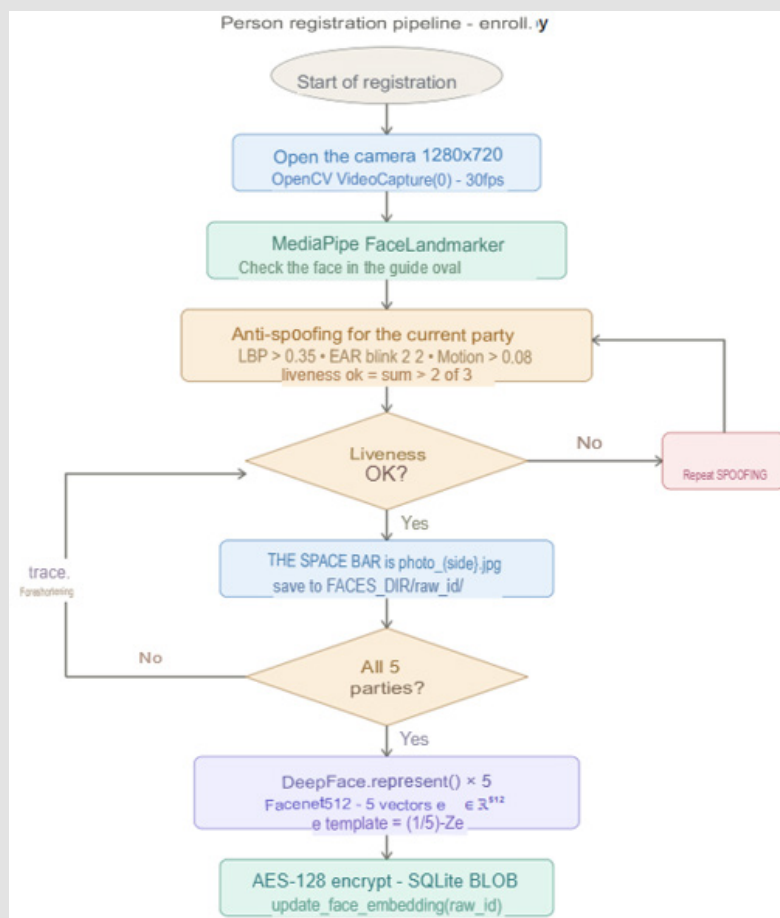


Figure 4: Decision workflow for presentation-attack detection.

Three-Stage Presentation-Attack Detection

Passive Texture and Colour Analysis: The passive stage uses Laplacian variance as a rough sharpness/texture descriptor and the proportion of pixels falling within a selected skin-colour range in HSV space. These cues are combined as a screening signal [16]. The ratio-

nale is that a display or printed image may introduce texture, edge, and colour characteristics that differ from a live face. However, Laplacian variance is also affected by focus, compression, motion blur, and camera resolution; HSV skin ratios vary with illumination and skin tone. The values should therefore be treated as features rather than universal physiological rules (Figure 5).

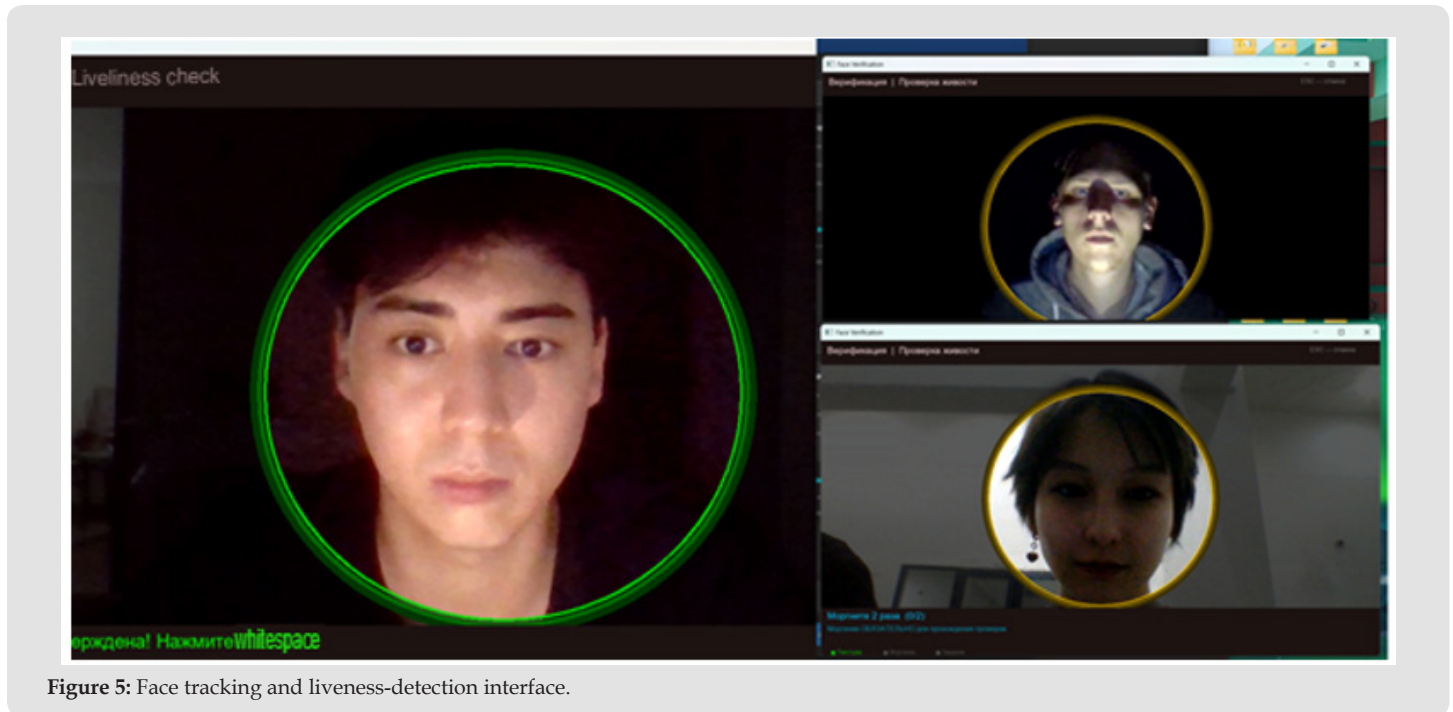


Figure 5: Face tracking and liveness-detection interface.

Blink Detection Using Eye Aspect Ratio: The active stage uses facial landmarks to estimate the Eye Aspect Ratio. Given six landmarks around an eye, EAR is computed as the sum of two vertical distances divided by twice the horizontal distance. A sustained drop below a calibrated threshold can indicate eye closure. The detector should use temporal windows and confidence checks rather than a single frame, because blinks may be missed due to glasses, occlusion, low frame rate, or landmark errors. Blink evidence contributes to liveness assessment but should not alone determine the identity decision. [18].

Randomized Challenge-Response: (Figure 6) The system issues a random head-turn instruction, such as turning left or right. The response is evaluated using facial landmarks and solvePnP-based head-pose estimation. Randomization reduces the effectiveness of a fixed pre-recorded video because the replay cannot anticipate a newly generated instruction. It does not guarantee resistance to real-time deepfake systems, responsive human assistance, or sophisticated manipulation. Evaluation should record challenge type, response latency, pose-estimation confidence, success rate, and false rejection rate for genuine participants. [19].

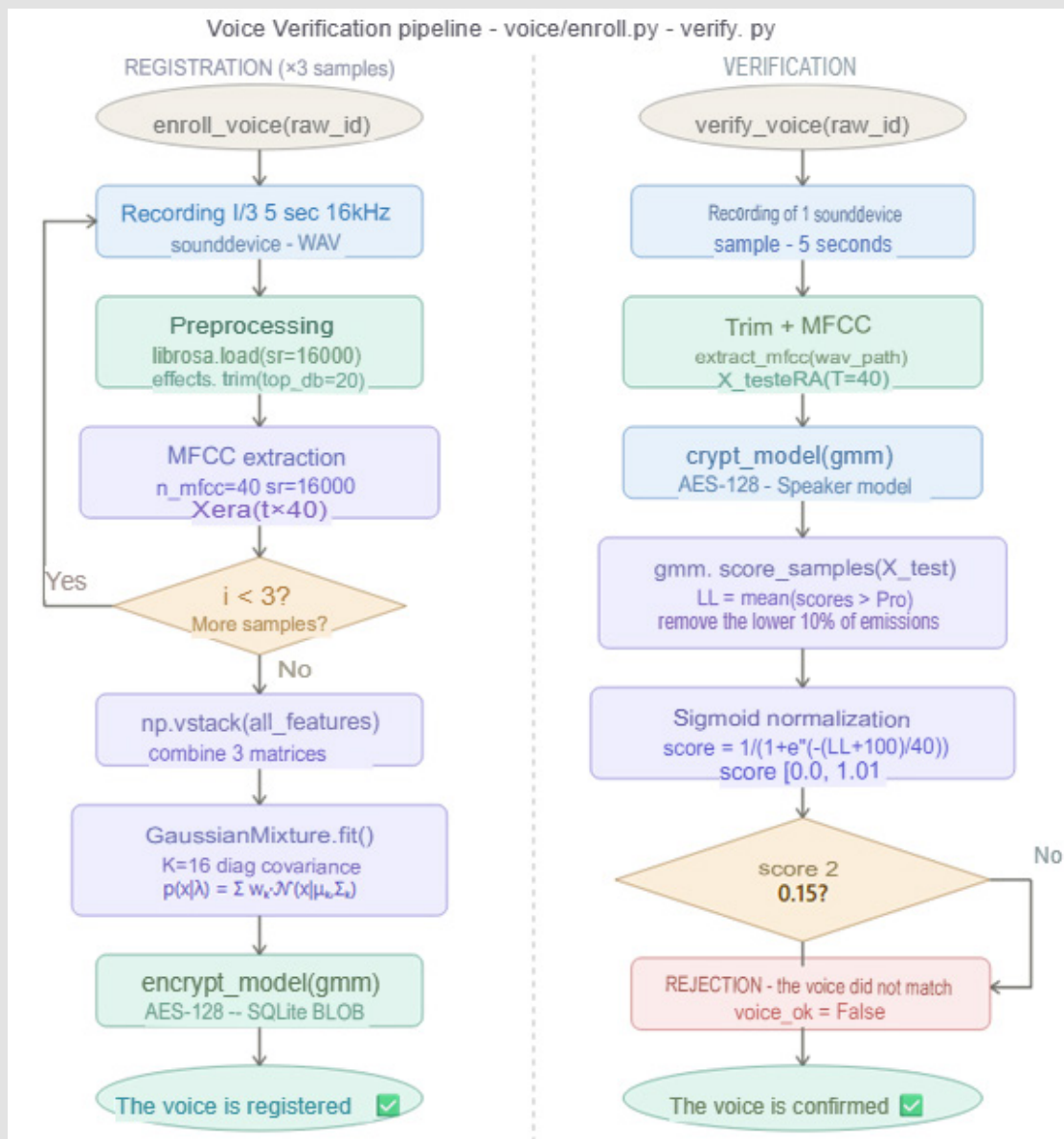


Figure 6: Verification and challenge-response workflow.

Continuous Monitoring

(Figure 7) After the initial verification, the system tracks the face across frames using intersection-over-union (IoU) between successive bounding boxes. The tracking state is retained for a limited time-to-live (TTL), allowing the system to avoid treating every temporary detection failure as a new identity. A state machine represents transitions such as verified, temporarily absent, multiple faces detected, liveness required, and verification failed [20]. The source also describes a background voice monitor that uses voice activity detection

to identify suitable moments for periodic speaker checks. Continuous monitoring should be designed to minimize unnecessary interruptions and avoid treating every anomaly as proof of misconduct. Camera occlusion, network or hardware problems, disability-related movement differences, and environmental noise can all generate alerts [21]. A responsible workflow should distinguish technical uncertainty from confirmed identity mismatch, preserve an auditable event history, and provide a review path before any consequential academic decision is made.

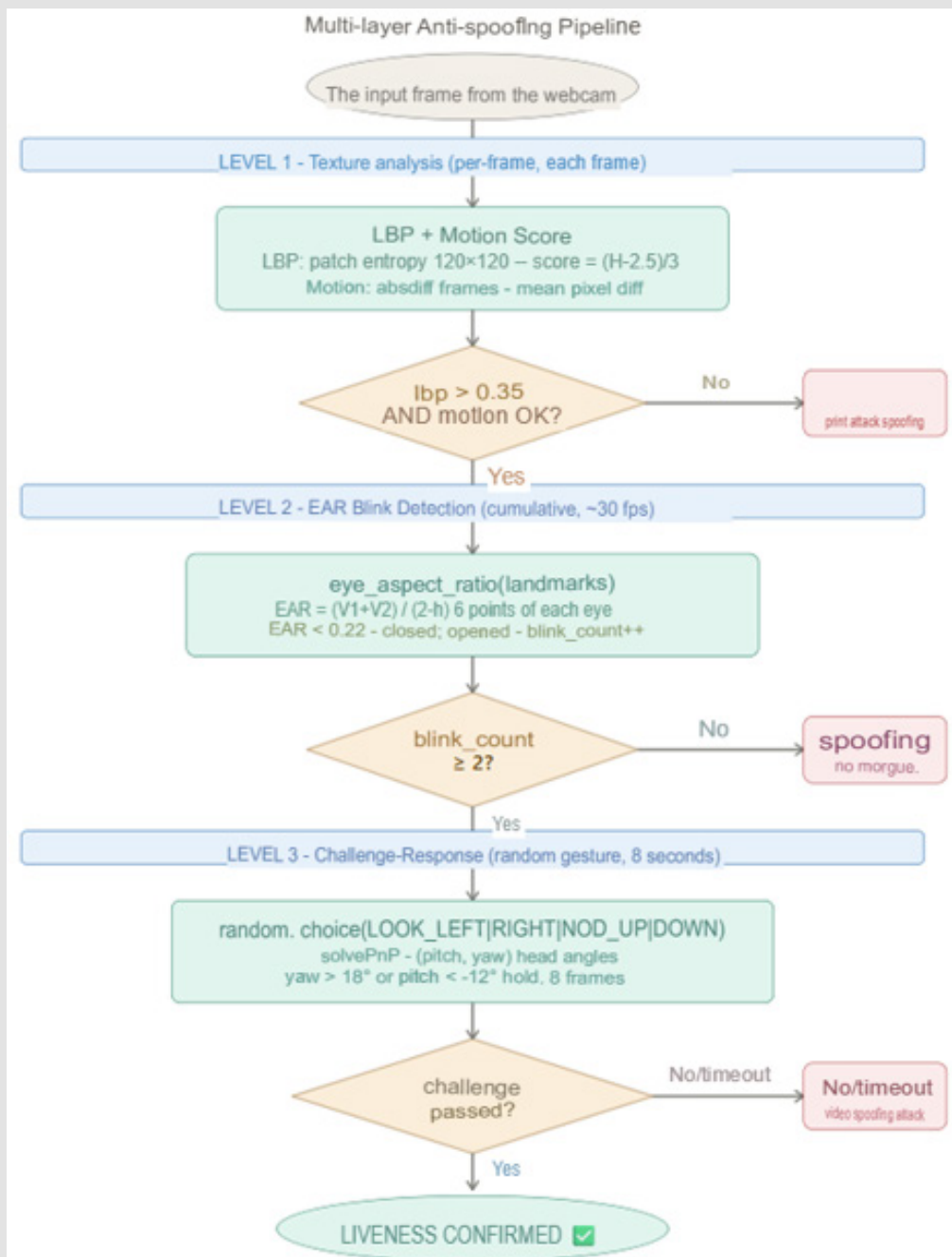


Figure 7: Continuous-monitoring workflow.

Data Protection and System Security

(Figure 8) The source describes encrypted storage of biometric templates using AES-128/Fernet-related mechanisms, hashing for identifiers or integrity-related values, and a local SQLite database. Encryption at rest is useful but does not by itself establish compliance or comprehensive security. Key management, least-privilege access,

backup protection, retention limits, secure deletion, dependency updates, logging controls, and incident response are also required [22]. Hashing is not a substitute for encryption when a value must later be recovered, and biometric templates should not be logged in plaintext. Biometric data are sensitive because a person cannot simply replace their face or voice in the way they can replace a password. Enrollment should therefore use informed consent, a clear purpose limitation,

minimal collection, restricted retention, and an accessible alternative for users who cannot or do not wish to use biometric verification. The legal basis and applicable obligations must be checked against cur-

rent Kazakhstan law and institutional policy before deployment. The present paper describes technical controls; it does not constitute a legal compliance certification [23].

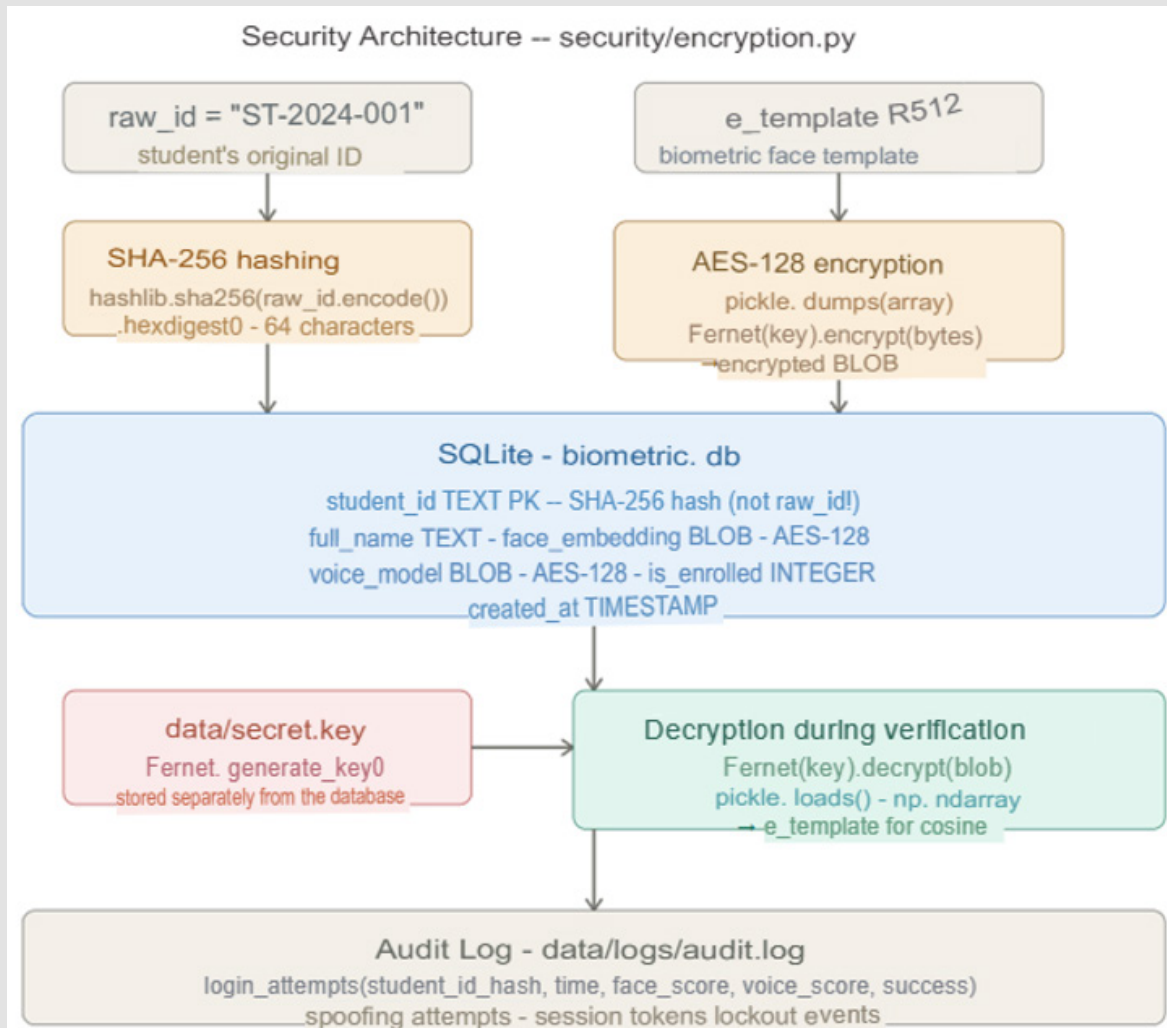


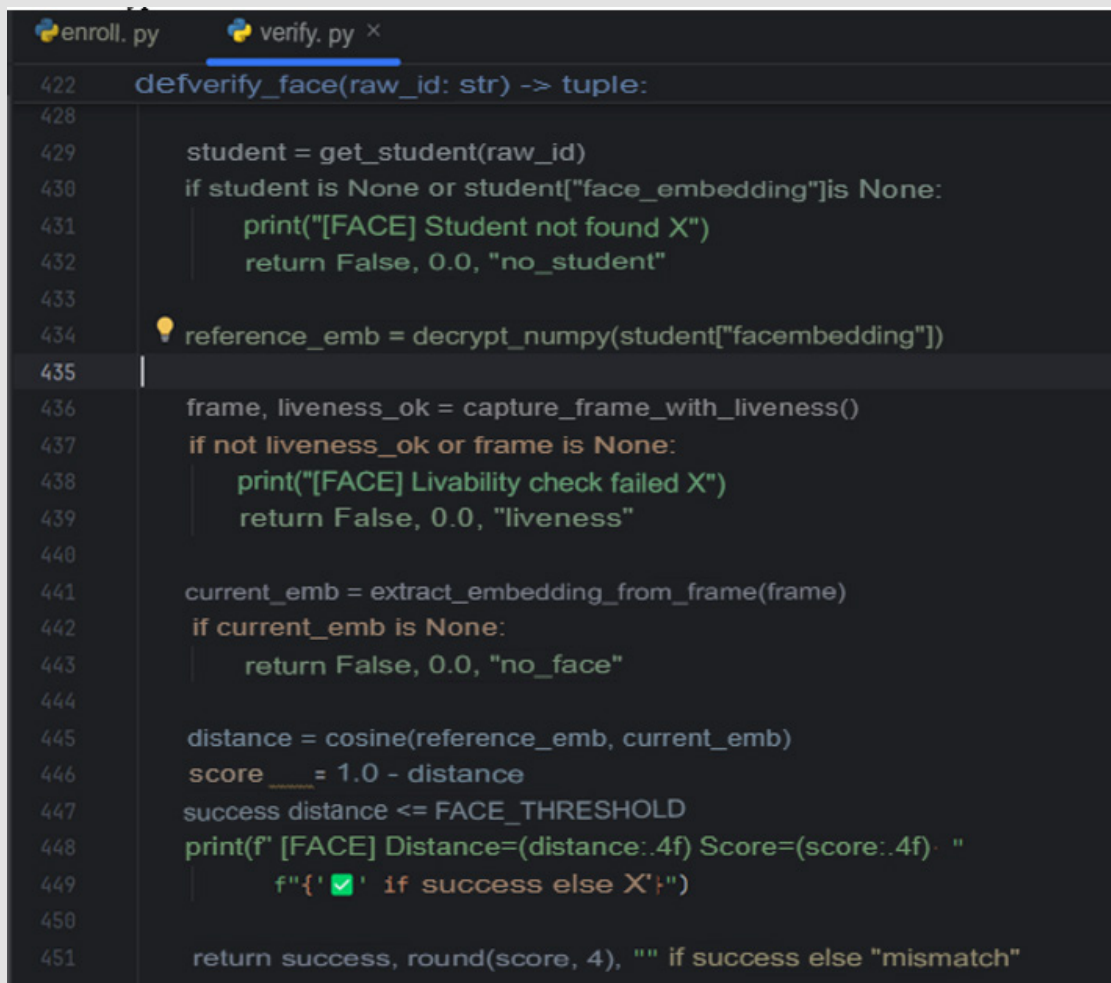
Figure 8: Security architecture and data-protection controls.

Experimental Design and Evaluation Metrics

Prototype Environment

(Figure 9) The system is described as a Python 3.10 desktop prototype using a standard 720p webcam and a built in microphone, with local processing and no dedicated GPU requirement stated in the

source. The software stack includes OpenCV, DeepFace/FaceNet512, MediaPipe, audio-processing components, SciPy, cryptographic utilities, and SQLite. Exact hardware specifications, operating system, library versions, runtime measurements, and model hashes were not fully available in the source document and should be added to a reproducibility package [24].



```

422 def verify_face(raw_id: str) -> tuple:
428
429     student = get_student(raw_id)
430     if student is None or student["face_embedding"] is None:
431         print("[FACE] Student not found X")
432         return False, 0.0, "no_student"
433
434     reference_emb = decrypt_numpy(student["facembedding"])
435
436     frame, liveness_ok = capture_frame_with_liveness()
437     if not liveness_ok or frame is None:
438         print("[FACE] Livability check failed X")
439         return False, 0.0, "liveness"
440
441     current_emb = extract_embedding_from_frame(frame)
442     if current_emb is None:
443         return False, 0.0, "no_face"
444
445     distance = cosine(reference_emb, current_emb)
446     score ____ = 1.0 - distance
447     success distance <= FACE_THRESHOLD
448     print(f" [FACE] Distance=(distance:.4f) Score=(score:.4f) "
449           f"{'✅' if success else 'X'}")
450
451     return success, round(score, 4), "" if success else "mismatch"

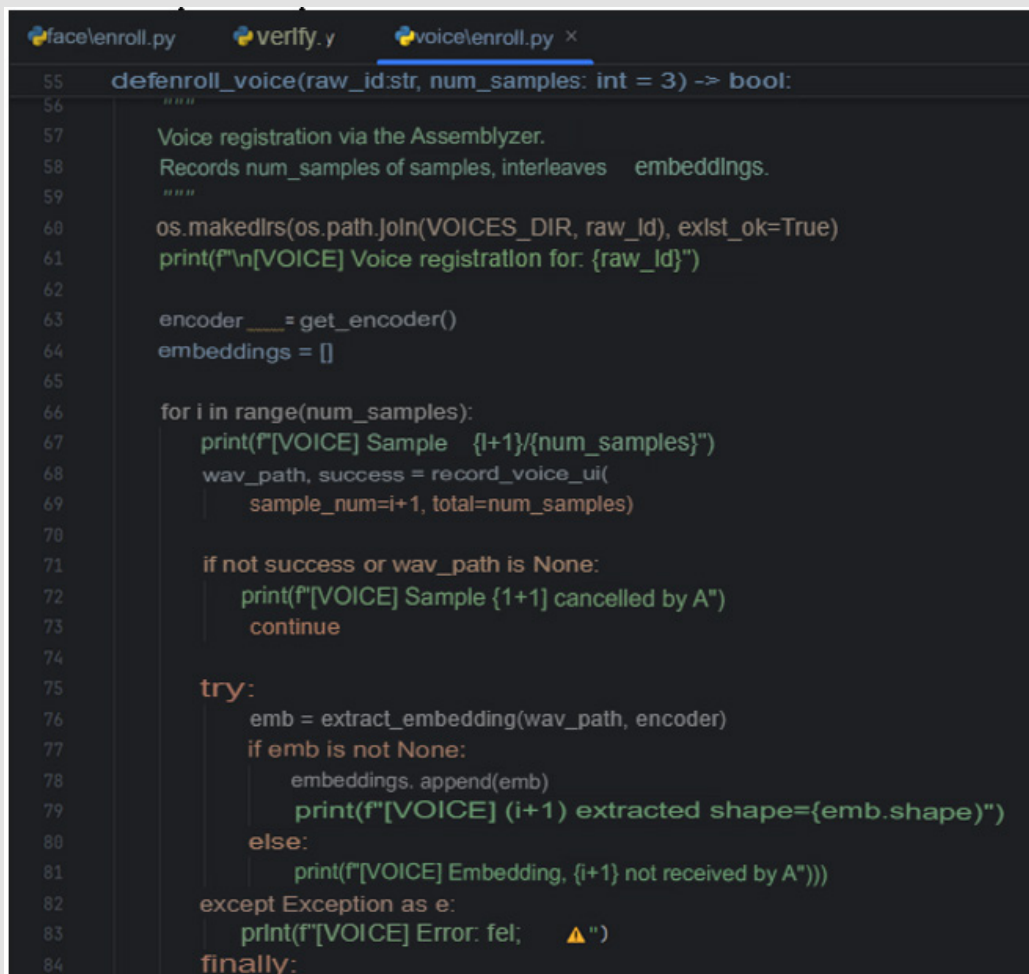
```

Figure 9: Source-code excerpt illustrating a prototype verification routine.

Evaluation Protocol Reported in the Source

(Figure 10) The source reports a pilot with four participants, a repeated verification series of 15 trials for one test identity, impostor checks involving three impostor participants, and presentation-attack scenarios involving photographs and replayed video. It also reports

condition changes including distance from the camera, head rotation, quiet speech, and poor illumination. The tables are descriptive pilot records rather than a statistically powered benchmark. The paper preserves the reported values but does not infer confidence intervals or population-level performance from them [25].



```

55 defenroll_voice(raw_id:str, num_samples: int = 3) -> bool:
56     """
57     Voice registration via the Assemblyzer.
58     Records num_samples of samples, interleaves embeddings.
59     """
60     os.makedirs(os.path.Join(VOICES_DIR, raw_id), exist_ok=True)
61     print(f"\n[VOICE] Voice registration for: {raw_id}")
62
63     encoder ____ = get_encoder()
64     embeddings = []
65
66     for i in range(num_samples):
67         print(f"[VOICE] Sample {i+1}/{num_samples}")
68         wav_path, success = record_voice_ui(
69             sample_num=i+1, total=num_samples)
70
71         if not success or wav_path is None:
72             print(f"[VOICE] Sample {i+1} cancelled by A")
73             continue
74
75         try:
76             emb = extract_embedding(wav_path, encoder)
77             if emb is not None:
78                 embeddings.append(emb)
79                 print(f"[VOICE] (i+1) extracted shape={emb.shape}")
80             else:
81                 print(f"[VOICE] Embedding, {i+1} not received by A"))
82         except Exception as e:
83             print(f"[VOICE] Error: {e}; ⚠")
84     finally:

```

Figure 10: Source-code excerpt from the reported evaluation workflow.

Metrics

False Acceptance Rate (FAR) is the proportion of impostor attempts incorrectly accepted; False Rejection Rate (FRR) is the proportion of genuine attempts incorrectly rejected. Equal Error Rate (EER) is the operating point where FAR and FRR are equal or approximately equal as the threshold varies. Accuracy is the proportion of correctly classified trials, but it can be misleading when genuine and impostor trials are imbalanced. Receiver Operating Characteristic (ROC) curves show the relationship between true acceptance and false acceptance across thresholds [26]. For presentation-attack detection, APCER should be computed separately for each attack species and BPCER for bona fide presentations. A single overall attack detection rate can hide differences between printed photos, screen replays, masks, and generated media. With very small denominators, percentages are unstable: a result of four detections out of four trials is a 100% observed

rate, but it does not establish a true detection probability of 100%. A future study should report the number of trials, confidence intervals, per-attack outcomes, and the exact decision thresholds.

Reproducibility and Validity Considerations

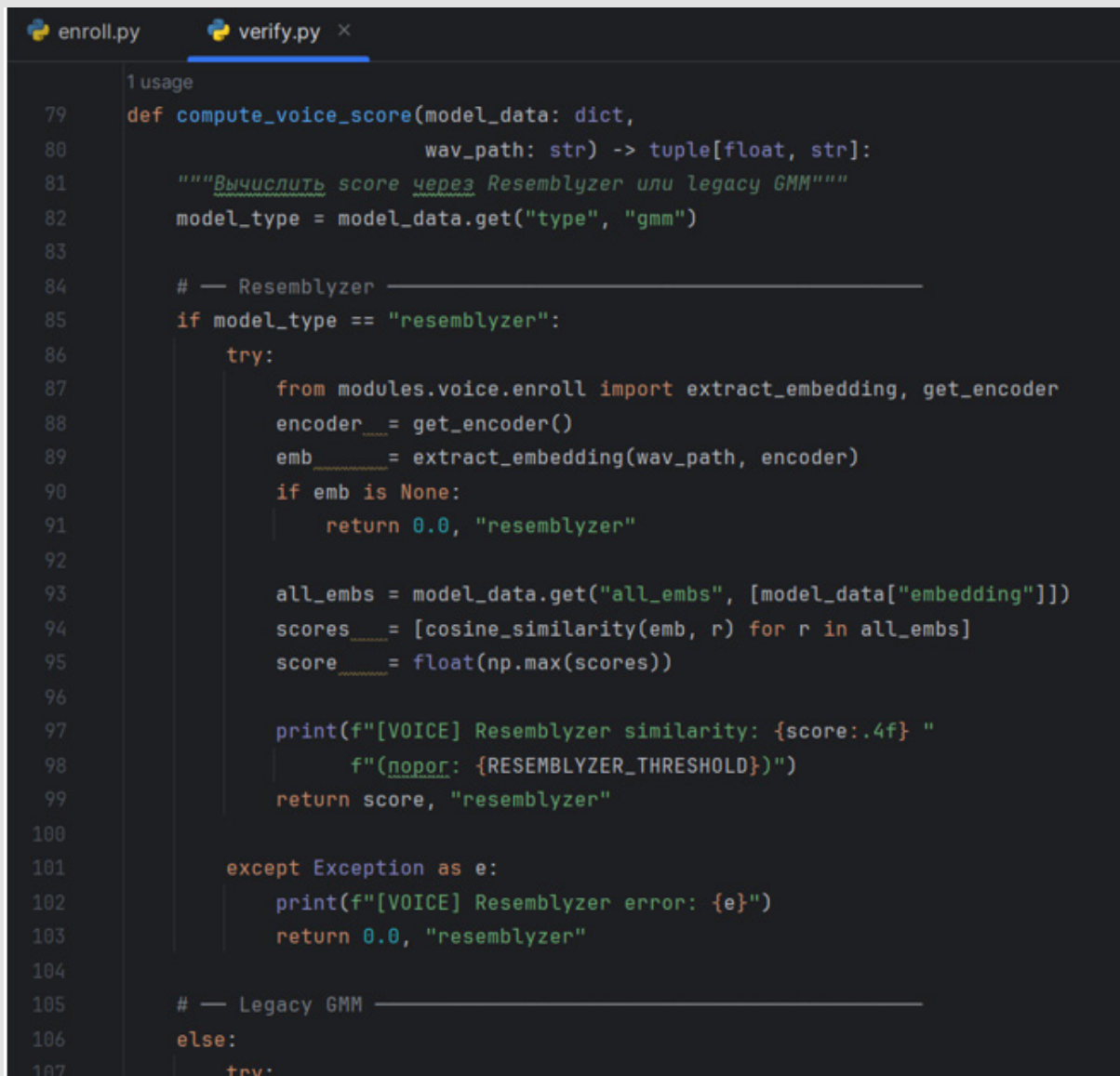
Three limitations affect interpretation. First, four participants are insufficient for reliable estimates of false acceptance and false rejection across demographic and environmental variation. Second, the source uses both GE2E/Resemblyzer and a legacy GMM voice path; the exact backend must be tied to each reported result [27]. Third, some result tables contain inconsistencies—for example, a reported fusion score does not always appear to follow the stated weighted equation exactly. Before submission, the authors should recompute all tables from raw trial-level logs, specify whether face scores represent similarity or distance, and align thresholds and score directions.

Results

Face and Voice Scores Under Reported Conditions

(Figure 11) The source provides small descriptive score tables for several test conditions. Across these records, the voice scores are generally more stable than face scores, while face scores are more sen-

sitive to changes in pose and illumination. This pattern is consistent with the design rationale for multimodal fusion, but the small number of observations prevents a general claim that voice is inherently more reliable than face verification. The result should be interpreted as a property of the current prototype, sample, microphone, and environment [28] (Table 2).



```

enroll.py  verify.py x
1 usage
79 def compute_voice_score(model_data: dict,
80                          wav_path: str) -> tuple[float, str]:
81     """Вычислить score через Resemblyzer или legacy GMM"""
82     model_type = model_data.get("type", "gmm")
83
84     # — Resemblyzer —
85     if model_type == "resemblyzer":
86         try:
87             from modules.voice.enroll import extract_embedding, get_encoder
88             encoder = get_encoder()
89             emb = extract_embedding(wav_path, encoder)
90             if emb is None:
91                 return 0.0, "resemblyzer"
92
93             all_embs = model_data.get("all_embs", [model_data["embedding"]])
94             scores = [cosine_similarity(emb, r) for r in all_embs]
95             score = float(np.max(scores))
96
97             print(f"[VOICE] Resemblyzer similarity: {score:.4f} "
98                   f"(nopor: {RESEMBLYZER_THRESHOLD})")
99             return score, "resemblyzer"
100
101         except Exception as e:
102             print(f"[VOICE] Resemblyzer error: {e}")
103             return 0.0, "resemblyzer"
104
105     # — Legacy GMM —
106     else:
107         try:

```

Figure 11: Source-code excerpt associated with score processing.

Table 2

Reported condition / series	Participants or trials	Reported observation	Interpretation
Normal conditions	4 participants	FAR 0%; FRR 0%; accuracy 100%	Pilot feasibility only; very small sample
Repeated verification series	15 checks for one test identity	Fusion-score standard deviation 0.019; reported accuracy 100%	Indicates within-series score stability in this limited case
Impostor testing	3 impostor participants	Reported FAR 0%	No false acceptance observed in the tested attempts; uncertainty remains high
Poor illumination	4 participants in source table	Reported FRR 75%	Major usability and robustness limitation
Head rotation	4 participants in source table	Source reports severe rejection in the tested condition	Pose and enrollment coverage require further calibration
Quiet voice	4 participants in source table	Reported accuracy 100% in the tested condition	Limited evidence; should be repeated with controlled acoustic levels

Presentation-Attack Scenarios

(Figure 12) The source reports detection of four scenario categories: a participant leaving the camera frame, two people appearing in the frame, a photograph presented to the camera, and a video replay. Each category lists four participants or trials and a 100% observed

detection rate. These observations support the feasibility of the layered pipeline for the exact scenarios tested. They do not demonstrate resistance to three-dimensional masks, high-quality replay displays, real-time face replacement, deepfake video, voice conversion, or synthetic speech. The source itself identifies several of these attack classes as untested (Table 3).

```

enroll.py  verify.py  voice_analyzer.py  fusion.py x
60  def verify_student(raw_id: str) -> tuple:
113      f"Voice verification failed (score=(voice_score))"
114      fail_count = increment_fail(raw_id, LOCKOUT_MINUTES)
115      log_attempt(raw_id, face_score, voice_score, fuscure: 0.0,
116                  success: False, details["fail_reason"])
117
118      #Save it to CSV
119      _save_to_dissertation(raw_id, face_score, voice_score, fusion_score: 0.0,
120                          success: False, details["fail_reason"])
121
122      print(f"[FUSION] X The voice did not pass. Loss: (fail_count)/(had_LOGIN_ATTEMPTS) ")
123      return False, details
124
125      print(f"[FUSION] Voice confirmed (score=(voice_score))\n")
126
127      # - War 4: Fusion
128      fusion_score = fuse_scores(face_score, voice_score)
129      details["fusion_score"] = fusion_score
130      success = fusion_score >= Fusion_threshold
131      details["success"] = success
132
133      print(f"[FUSION] Calculation: ")
134      print(f"Face: (face_score:.4f) (FACE_WEIGHT) = (face_score*FACE_WEIGHT: .4f3")
135      print(f"Voice: (voice_score:.4f) (VOICE_WEIGHT) = (voice_score*VOICE_WEIGHT: .4f1")
136      print(f"Result: (fusion_score:.4f) (threshold: (FUSION_THRESHOLD))")
137
138      Step 5: Logging from
139      security_photo_log import save_verification_photo
140      save_verification_photo(raw_id, success)

```

Figure 12: Source-code excerpt associated with attack-detection scenarios.

Table 3

Scenario in the source	Reported trials	Observed outcome	Evidence boundary
Student leaves the frame	4	4 detected; 100% reported SDR	Only the tested disappearance events
Two people in frame	4	4 detected; 100% reported SDR	Does not quantify all multi-person/occlusion conditions
Printed photograph	4	4 detected; 100% reported SDR	No evidence for all print/display qualities
Video replay	4	4 detected; 100% reported SDR	Does not establish resistance to adaptive or generated video

Comparison with Publicly Described Proctoring Systems

The source compares the prototype with OES, Proctor.kz, and ProctorEdu using information presented as publicly available product descriptions. It reports that these descriptions include face recognition and behavioural or audio monitoring, while algorithmic

details about speaker identity verification, score-level fusion, and challenge-response were not publicly described. Since vendors may use proprietary mechanisms that are not disclosed, the comparison is limited to documented features and should not be interpreted as a definitive reverse-engineering assessment (Table 4).

Table 4

Feature	Public descriptions reviewed in source	Proposed prototype
Face verification	Reported across the reviewed platforms	FaceNet512 embedding and cosine comparison
Voice processing	Noise/audio monitoring described; independent speaker verification not established in reviewed materials	Speaker embedding verification intended
Presentation-attack detection	Details limited or not described in reviewed materials	Texture + blink + randomized head-turn challenge
Continuous monitoring	Partial or platform-specific features described	IoU tracking with retained state and periodic voice checks
Score-level face-voice fusion	Not established from reviewed public descriptions	Weighted fusion with 0.6/0.4 weights
Algorithm transparency	Proprietary/limited public technical detail	Research code structure and pipeline are inspectable

Interpretation of the Illumination Result

The reported FRR of 75% under poor illumination is the clearest adverse result in the source. It shows that the face channel can become the bottleneck even when the voice channel is available. Weighted fusion can compensate only if the voice score remains valid and the score scales are calibrated. If the face module is unable to extract a usable embedding, the system must define an explicit fallback or retry policy; silently treating a missing face score as a low score can create systematic rejection. Future work should compare controlled illumination levels, camera exposure settings, distance, pose, glasses, and skin-tone variation, while documenting the effect on both false acceptance and false rejection.

Discussion

What the Pilot Results Support

The evidence supports a narrow conclusion: a modular face-voice verification prototype with layered liveness checks and continuous monitoring can be implemented using commonly available laptop sensors, and it can detect the limited presentation scenarios

included in the pilot. The reported score stability across 15 repeated checks is a useful engineering observation. The absence of observed false acceptance in the small impostor test is encouraging but statistically weak. It cannot be used as proof that the system has a zero false-acceptance probability. The main contribution is not a claim of record-setting biometric accuracy. It is the explicit integration of several functions that are often discussed separately: initial enrollment, two-channel matching, active liveness, persistent tracking state, periodic speaker verification, and template protection. This integration offers a foundation for controlled evaluation and provides a clearer audit trail than a black-box decision service, provided that implementation logs and configuration are documented.

Fusion is Only as Reliable as Calibration

The weighted fusion equation is simple and computationally inexpensive, which suits a CPU-based prototype. However, raw scores from different matchers are not necessarily commensurate. A face similarity score and a speaker cosine similarity can have different distributions and different sensitivity to nuisance variables. Fixed weights should therefore be treated as provisional. Calibration should use genuine and impostor trials collected from a development cohort

that is distinct from the final test cohort. The operating point should be chosen according to the consequences of false acceptance and false rejection, and the paper should report results at that fixed point. A useful follow-up experiment would compare face-only, voice-only, and fused configurations under the same trial set. It should include ablation of each liveness stage and compare equal weights, the current 0.6/0.4 rule, and weights estimated on a development set. Such experiments would reveal whether each component adds measurable value and whether improvements are robust across conditions.

Security Limitations and Adaptive Attacks

The three-stage liveness pipeline addresses complementary cues: passive texture and colour features, a temporal blink cue, and an unpredictable head-turn challenge. The layers are useful because an attack that defeats one stage may fail another. Nevertheless, the reported 100% detection rate is restricted to the small attack set. The approach does not establish protection against 3D masks, high-resolution replay, real-time face reenactment, camera virtualisation, injected video streams, voice cloning, or replayed and converted speech. A production threat model must include both physical presentation attacks and digital injection attacks. Security testing should include adaptive adversaries who know the challenge protocol, as well as ordinary failures such as motion blur and poor lighting. Thresholds and random challenge sequences should be protected against predictable reuse. Audit logs should record the event category and confidence without unnecessarily retaining raw biometric material. Security controls should be evaluated independently from academic misconduct decisions.

Privacy, Fairness, and Educational Governance

Biometric monitoring raises issues beyond technical accuracy. Students may have disabilities, facial differences, speech differences, accents, temporary illness, limited bandwidth, or environments unsuitable for continuous camera and microphone use. These conditions can produce unequal error rates. Institutions should provide reasonable alternatives, explain the system's purpose, limit access to the collected data, establish retention periods, and allow human review of disputed events. The system should not infer cheating solely from gaze, movement, or an uncertain biometric score. The implementation's local processing design can reduce exposure of raw video and audio in transit, but local processing does not eliminate privacy risks. The application still captures sensitive data and stores templates and event records. A privacy impact assessment, threat model, access-control review, and legal assessment are needed before deployment. Consent must be meaningful and should not be treated as a substitute for an appropriate legal basis or institutional safeguards.

Practical Deployment Implications

For a university pilot, the system should initially operate in a non-punitive mode: it should produce verification events for research review rather than automatically invalidate an examination. The in-

stitution should collect performance data across devices and environmental conditions, examine false-alert causes, and obtain feedback from students and invigilators. Operational readiness also requires documentation, versioned models, dependency management, secure key storage, incident handling, and an accessible manual fallback.

Limitations and Future Work

The limitations of the current study are substantial and should be explicit. The pilot includes only four participants and therefore cannot establish reliable population-level FAR, FRR, EER, or demographic fairness. The reported attack scenarios are limited to a small number of photographs and video replays; three-dimensional masks, deep-fake videos, voice conversion, and synthetic speech were not tested. The source does not provide sufficient trial-level information to calculate confidence intervals or fully reconstruct ROC curves. In addition, the source describes both GE2E/Resemblyzer and MFCC-GMM voice processing, and the precise active backend for every result must be verified. Some table values may not align perfectly with the stated fusion formula and should be recomputed from raw logs before journal submission.

Future work should follow a staged validation plan:

- Recruit a substantially larger and more diverse cohort, with informed consent and documented inclusion criteria.
- Separate enrollment, calibration, and test data at participant level to prevent identity leakage.
- Report FAR, FRR, EER, ROC curves, APCER, and BPCER with confidence intervals and fixed, disclosed thresholds.
- Evaluate face-only, voice-only, and fused systems under identical conditions, including ablation of each anti-spoofing stage.
- Test a broader attack set, including multiple display qualities, printed media, 3D masks, replayed and synthesized audio, face reenactment, and digital injection.
- Measure latency, CPU and memory consumption, battery impact, and performance on low-cost devices and slow connections.
- Assess subgroup error rates and accessibility impacts, and provide non-biometric alternatives.
- Publish a reproducibility package containing configuration files, dependency versions, threshold calibration procedure, anonymized trial summaries, and implementation details that do not expose personal data.

Conclusion

This paper has described a multimodal biometric verification framework for online education that combines face embeddings, speaker verification, layered presentation-attack detection, and con-

tinuous session monitoring. The prototype uses FaceNet512 for face representations, an intended GE2E/Resemblyzer speaker-embedding path, score-level fusion with weights of 0.6 and 0.4, texture and colour cues, blink detection, randomized head-turn challenges, IoU-based tracking, and periodic voice checks. The design is intended to run with standard webcam and microphone hardware and to keep the principal processing workflow local. The pilot results reported in the source show successful operation in the limited normal-condition trials and detection of the tested photograph and replay scenarios. The reported deterioration under poor illumination, with FRR reaching 75%, demonstrates that the current system remains sensitive to acquisition conditions. Most importantly, the four-person sample and limited attack set do not support broad claims of general accuracy or production readiness. The results should be treated as feasibility evidence and as a basis for a larger, independently controlled evaluation. The practical value of the work lies in an inspectable architecture that joins multimodal matching with liveness checks and session-state monitoring. The next research stage should prioritize score calibrationss, replication of all reported results from raw logs, expanded attack testing, statistically defensible metrics, privacy safeguards, and independent validation. These steps are necessary before the prototype can be considered for high-stakes educational assessment. The source document does not specify a public dataset or repository. The authors should state whether code, anonymized trial-level results, and evaluation scripts can be shared, and provide a repository link if one is created. Biometric images, audio recordings, and identifiable templates should not be published. The source document does not provide a complete ethics approval identifier or a detailed participant-consent protocol. Before submission, the authors should insert the relevant institutional ethics approval or exemption statement, describe informed consent, and explain how biometric data were stored, accessed, and deleted. Do not claim formal ethics approval unless it was actually obtained. The authors should declare any financial, commercial, or institutional conflicts of interest. If none exist, insert the journal's required no-conflict declaration after confirming that it is accurate. Funding information was not specified in the source document. Insert the relevant funding agency and grant number, or the journal's required statement that no external funding was received, only after verification.

References

1. UNESCO (2020) Education: From disruption to recovery.
2. OECD. Education at a Glance 2024: OECD Indicators. Paris: OECD Publishing; 2024.
3. Bureau of National Statistics of the Republic of Kazakhstan. Education statistics 2024. [Full report title, URL, and access date should be verified before submission].
4. International Center for Academic Integrity. The State of Academic Integrity Report 2024. [Publisher and URL should be verified before submission].
5. World Bank. Digital learning and education technology resources. <https://www.worldbank.org/en/topic/edutech>. [Verify the exact report title and publication year cited in the source document].
6. ISO/IEC 2382-37:2022 (2022) Information technology-Vocabulary-Part 37: Biometrics. Geneva: International Organization for Standardization.
7. ISO/IEC 19795-1:2021 (2021) Information technology-Biometric performance testing and reporting-Part 1: Principles and framework. Geneva: International Organization for Standardization.
8. ISO/IEC 30107-3:2023 (2023) Information technology-Biometric presentation attack detection-Part 3: Testing and reporting. Geneva: International Organization for Standardization.
9. Firmansyah AM, Herutomo A, Setyanto T, Prihatmanto A. Comparison of face recognition accuracy of ArcFace, FaceNet and FaceNet512 models on DeepFace framework. Proceedings of IEEE ICCoSITE; 2023. [Verify complete proceedings metadata and DOI].
10. Wan L, Wang Q, Papir A, Lopez Moreno I (2018) Generalized end-to-end loss for speaker verification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
11. Turk M, Pentland A (1991) Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1): 71-86.
12. Taigman Y, Yang M, Ranzato M, Wolf L (2014) DeepFace: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition p. 1701-1708.
13. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition arXiv:1503.03832.
14. Deng J, Guo J, Xue N, Zafeiriou S (2019) ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; arXiv:1801.07698.
15. Huang GB, Ramesh M, Berg T, Learned Miller E (2007) Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. University of Massachusetts Amherst Technical Report 07-49.
16. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10(1-3): 19-41.
17. Snyder D, Garcia Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: Robust DNN embeddings for speaker recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.
18. Desplanques B, Thienpondt J, Demuynck K (2020) ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Proceedings of Interspeech p. 3830-3834.
19. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: A large-scale speaker identification dataset. In: Proceedings of Interspeech p. 2616-2620.
20. Chung JS, Nagrani A, Zisserman A. VoxCeleb2: Deep speaker recognition. In: Proceedings of Interspeech; 2018. p. 1086-1090.
21. Soukupová T, Čech J. Real-time eye blink detection using facial landmarks. In: Proceedings of the Computer Vision Winter Workshop; 2016.
22. Boulkenafet Z, Komulainen J, Hadid A (2016) Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security 11(8): 1818-1830.
23. Yamagishi J, et al. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. [Verify full author list, volume, pages, and DOI].

24. Labayen M, Vea R, Flórez J, Aginako N, Sierra B (2021). Online student authentication and proctoring system based on multimodal biometrics technology. *IEEE Access* 9: 72398-72411.
25. Ross A, Nandakumar K, Jain AK (2006) *Handbook of Multibiometrics*. New York: Springer.
26. Europol. (2022) *Facing Reality? Law Enforcement and the Challenge of Deepfakes*. The Hague: Europol.
27. Kartynnik Y, Ablavatski A, Grishchenko I, Grundmann M (2019) Real-time facial surface geometry from monocular video on mobile GPUs. *arXiv* 1907.06724.
28. Bazarevsky V, Kartynnik Y, Vakunov A, Raveendran K, Grundmann M (2019) *BlazeFace: Sub-millisecond neural face detection on mobile GPUs*. *arXiv* 1907.05047.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2026.67.010441

Akmoldina A.I . Biomed J Sci & Tech Res



This work is licensed under Creative Commons Attribution 4.0 License

Submission Link: <https://biomedres.us/submit-manuscript.php>**Assets of Publishing with us**

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

<https://biomedres.us/>