

# RELIABILITY AWARE MULTIMODAL AI SYSTEM FOR REAL-TIME SMART CLASSROOM ATTENDANCE MONITORING

OMAR IMAMVERDIYEV<sup>1,2</sup>, JAMAL TEYMUROV<sup>3</sup>, SABINA MAHMUDOVA<sup>4</sup>,  
NAZILE MAHMUDOVA<sup>4</sup>, AYISHA AHMADOVA<sup>4</sup>, AFET KARIMOVA<sup>5</sup>, TARANA  
NURUBEYLI<sup>5,6,7</sup>

1. School of Information Technologies and Engineering, ADA University, Baku, Azerbaijan  
2. Department of Computer Science, School of Engineering and Applied Science, The George Washington University, Washington, DC, USA Landau School, Baku, Azerbaijan, Nakhchivan State University, Nakhchivan, Azerbaijan 5. Institute of Physics PLE, Ministry of Science and Education of Republic of Azerbaijan, Baku, Azerbaijan, 6. Azerbaijan University of Architecture and Construction, Baku, AZ1073, Azerbaijan

7. Khazar University, Mahsati Str. 41, Baku, AZ 1096, Azerbaijan.

O. Imamverdiyev- ORCID: 0009-0009-3252-7150

S. Mahmudova - ORCID: 0009-0000-2694-1448

J. Teymurov - ORCID: 0009-0001-6165-8506

A. Ahmadova - ORCID: 0009-0004-9657-5649

N. Mahmudova - ORCID: 0009-0000-6834-5812

A. Kerimova- ORCID: 0009-0009-9631-7072

T. Nurubeyli- ORCID: 0000-0001-9696-2199

Email: omarimamverdiyev13@gmail.com, jamal@jamalteymurov.com, ayishaahmadova@ndu.edu.az, nazile.mahmudova4@gmail.com, t.nurubeyli@physics.science.az., omartarana@gmail.com

## ABSTRACT

Accurate and scalable attendance monitoring is a practical challenge in digitally supported higher education, particularly in large classrooms where manual registration is time-consuming, error-prone, and vulnerable to proxy attendance. This study proposes a reliability-aware multimodal artificial intelligence system for real-time smart classroom attendance monitoring. The proposed framework integrates visual recognition, audio-based participation analysis, and text/database-level identity verification through a probabilistic fusion mechanism. The system employs lightweight and edge-deployable models, including YOLOv5 for student detection, MobileNet for face recognition, DistilBERT for identity matching, and CNN-LSTM or wav2vec 2.0 for audio-based activity analysis. To improve decision reliability under non-ideal classroom conditions, the fusion layer dynamically adjusts modality contributions according to the quality of visual, audio, and database evidence. In addition, uncertain cases are redirected to an abstention or manual-review state instead of being automatically accepted. A pilot deployment scenario was evaluated in a smart classroom at Mingachevir State University using recognition accuracy, inference latency, multimodal consistency, and computational efficiency as the main performance criteria. The integrated system achieved an average attendance recognition accuracy of  $93.2 \pm 0.7\%$  with a decision latency of 45–55 ms, demonstrating its suitability for real-time classroom operation. The results indicate that reliability-aware multimodal fusion provides more stable attendance decisions than single-modality configurations and can reduce the risk of incorrect registration under occlusion, illumination variability, background noise, and incomplete identity evidence. The proposed system provides a practical basis for reliable AI-supported attendance monitoring in smart educational environments.

**Keywords:** *Multimodal Artificial Intelligence; Smart Classroom; Attendance Monitoring; Edge AI; Reliability-Aware Fusion; Real-Time Recognition*

## 1. INTRODUCTION

The digital transformation of higher education has increased the demand for intelligent information systems capable of supporting teaching,

monitoring, and administrative processes in a reliable and efficient manner. Smart classrooms are increasingly equipped with sensing devices, artificial intelligence modules, database-connected

services, and real-time monitoring tools. Within this context, attendance monitoring remains an important operational task, especially in large academic groups where manual registration is time-consuming, error-prone, and vulnerable to proxy attendance [1], [2], [3].

Conventional attendance monitoring methods are commonly based on manual roll calls, paper lists, card-based identification, QR codes, RFID-based identification, mobile applications, or semi-automated database records. Although these approaches are simple and inexpensive, they provide limited reliability under practical classroom conditions. Manual roll calls interrupt the teaching process, while card-, RFID-, QR-, and mobile-based systems mainly confirm the presentation of an identifier rather than the actual physical presence of a student. As a result, there is a need for automated attendance monitoring systems that can verify presence more reliably while remaining computationally feasible for real-time classroom deployment [1], [4], [3].

Recent advances in artificial intelligence, computer vision, speech processing, and natural language processing have created new opportunities for developing automated recognition systems. Deep learning models such as convolutional neural networks, lightweight visual architectures including MobileNet and EfficientNet, object detection models such as YOLO, and transformer-based models such as BERT, DistilBERT, and wav2vec 2.0 have demonstrated strong potential for visual, audio, and text-based data processing. However, the deployment of these models in classroom environments requires careful consideration of computational cost, inference latency, sensor quality, and system-level reliability. Therefore, an attendance monitoring system should not be evaluated only as an isolated recognition model, but as an applied AI-based information system operating under real classroom constraints [5], [6], [7], [8], [9].

Many existing attendance systems rely mainly on a single modality, most commonly face recognition, RFID, QR codes, mobile-device authentication, or database-based identification. These systems may perform well under controlled conditions, but their reliability often decreases in non-ideal environments. In practical classroom settings, visual recognition can be affected by partial face occlusion, side-view poses, motion blur, non-uniform illumination, or crowded seating arrangements. Audio-based evidence may be degraded by background noise, overlapping speech, microphone distance, and room acoustics.

Database-level identity verification may also be incomplete, delayed, or insufficient for confirming actual physical presence. Consequently, single-modality systems may generate uncertain or incorrect attendance decisions when one source of evidence becomes unreliable [10], [11], [4].

Multimodal artificial intelligence provides a promising solution by combining complementary evidence from visual, audio, and text/database channels. In an attendance monitoring system, visual recognition can provide the primary identity evidence, audio analysis can support participation- or activity-related cues, and text/database verification can act as a consistency check against institutional records. Nevertheless, simple multimodal aggregation is not sufficient for reliable deployment. The contribution of each modality should depend on the quality and reliability of the corresponding input signal. For example, visual evidence should be down-weighted when the face is occluded or blurred, while audio evidence should be down-weighted when the classroom environment is noisy. This makes reliability-aware fusion an important component of practical multimodal attendance monitoring systems [12], [11], [13].

The problem is particularly relevant for educational institutions where smart classroom infrastructure is developing and locally adaptable solutions are required. Imported or platform-dependent systems may not fully reflect local classroom layouts, institutional workflows, database structures, language conditions, and available computational resources. Therefore, there is a practical need for a deployable AI-based attendance monitoring framework that can operate under realistic classroom conditions, support real-time decision-making, and reduce the risk of incorrect registration. Such systems should also consider transparency, privacy, personal data protection, and responsible use of AI in educational environments [2], [14], [15], [16], [17].

To address this need, the present study proposes a reliability-aware multimodal artificial intelligence system for real-time smart classroom attendance monitoring. The proposed framework integrates visual recognition, audio-based participation analysis, and text/database-level identity verification within a probabilistic fusion model. The system employs lightweight and edge-deployable AI components, including YOLOv5-based student detection, MobileNet-based face recognition, DistilBERT-assisted identity matching, and CNN-LSTM or wav2vec 2.0-based audio analysis. These components are selected with respect to recognition performance, computational

feasibility, and real-time deployment requirements [5], [6], [18], [8].

The main novelty of this study lies in the reliability-aware decision framework rather than in the isolated use of individual AI models. The proposed system dynamically adjusts the contribution of visual, audio, and text/database evidence according to modality quality. In addition, the system does not force an automatic positive attendance decision when the available evidence is weak or contradictory. Instead, uncertain cases are redirected to an abstention or manual-review state. This decision logic is intended to reduce false attendance records and improve the operational reliability of AI-supported classroom monitoring.

The main contributions of this study are as follows. First, a multimodal attendance monitoring framework is developed by combining visual, audio, and text/database evidence within a unified reliability-aware probabilistic fusion architecture. Second, a real-time smart classroom deployment scenario is proposed using edge-compatible AI models selected according to computational load, latency, and operational feasibility. Third, the integrated system is evaluated using attendance recognition accuracy, inference latency, multimodal consistency, and computational efficiency as the main performance criteria. Finally, the study analyzes failure cases, robustness conditions, and scalability limitations, showing how reliability-aware fusion can improve attendance decision stability under non-ideal classroom conditions.

## 2. RELATED WORK

### 2.1. Multimodal Artificial Intelligence in Edge-Based Recognition Systems

Over the past decade, artificial intelligence systems have increasingly shifted from isolated recognition models toward integrated multimodal frameworks capable of processing heterogeneous data sources in real time. This transition is particularly important for smart classroom environments, where an attendance monitoring system must process camera input, microphone signals, and database-linked identity information under practical constraints such as limited computational resources, variable sensor quality, and low tolerance for delayed decisions [1], [11], [3].

In the visual domain, convolutional neural networks remain widely used because of their strong feature extraction capability and mature deployment ecosystem. Lightweight visual architectures such as MobileNet and EfficientNet are especially relevant for edge-oriented

applications because they provide a practical balance between recognition accuracy and computational efficiency. More complex architectures, including Vision Transformers, can achieve high benchmark performance, but their higher memory requirements and inference cost may limit their direct use in real-time classroom systems unless additional optimization is applied [19], [20], [6], [9].

Text-related processing also plays an important role in AI-supported attendance monitoring, especially when identity recognition must be linked to institutional records, student metadata, QR/ID inputs, or database-level consistency checks. Transformer-based models such as BERT and compact variants such as DistilBERT and MobileBERT provide effective contextual representations for semantic matching. However, their use in edge or near-edge environments requires attention to latency, memory consumption, and model compression techniques such as pruning, quantization, or distillation [21], [8], [22].

Speech and audio analysis provide an additional source of evidence in classroom environments. Traditional speech-processing approaches relied on statistical models such as Hidden Markov Models, while recent systems increasingly use deep neural architectures, including CNN-LSTM models and wav2vec 2.0. These models can support activity- or participation-related detection, but their reliability is strongly affected by background noise, overlapping speech, microphone distance, and room acoustics [5], [23], [24], [13].

Although visual, text-based, and audio-based recognition technologies are well developed as individual modules, many studies still evaluate them separately rather than as components of a unified attendance monitoring system. For practical classroom deployment, recognition accuracy alone is not sufficient. A reliable system must also address inference latency, computational feasibility, uncertainty handling, and the quality-dependent contribution of each modality.

### 2.2. Visual Recognition Models for Smart Classroom Monitoring

Visual recognition is one of the main components of automated attendance systems because it can directly support student detection and face-based identity recognition. High-capacity architectures such as VGG-16 and Vision Transformers may provide strong recognition performance, but their memory demand, inference

cost, and energy consumption can make them less suitable for latency-sensitive classroom applications [19], [20], [25].

For edge-based smart classroom systems, lightweight visual models are usually more practical. MobileNet-family architectures reduce computational cost through efficient convolutional design, while EfficientNet improves model scaling by balancing network depth, width, and input resolution. Compact YOLO variants are also useful for student or face detection because they provide a favorable balance between detection accuracy and real-time inference speed. Therefore, such models are suitable for attendance systems where recognition decisions must be generated during classroom sessions rather than after offline processing [6], [18], [7], [9].

Table 1 compares representative visual models considered for the design of an edge-oriented attendance monitoring system. The comparison focuses on the expected role of each model, reported accuracy range, latency range, and rationale for inclusion. The reported values should be interpreted as literature-based indicative ranges used for design justification rather than as final performance results of the proposed system..

*Table 1. Literature-based comparison of candidate visual models for edge-oriented attendance system design.*

| Model            | Role in system           | Reported accuracy range (%) | Reported latency range (ms) | Rationale for inclusion                                   |
|------------------|--------------------------|-----------------------------|-----------------------------|-----------------------------------------------------------|
| MobileNet        | Face identification      | 88–91                       | 30–50                       | Lightweight visual branch suitable for edge deployment    |
| YOLOv5 Distilled | Face/object detection    | 88–91                       | ≈45                         | Balanced detection quality and inference speed            |
| EfficientNet     | Alternative visual model | 90–94                       | 45–70                       | Higher visual accuracy with moderate additional cost      |
| VGG-16 / ViT     | Heavier reference models | 92–96 / 93–96               | 120–150 / 100–130           | Strong benchmark performance but limited edge suitability |

The comparison shows that MobileNet and compact YOLO-based models provide the most practical combination of speed and computational feasibility for classroom-scale inference. Heavier models may remain useful as reference architectures, but they are less appropriate for real-time deployment unless supported by stronger hardware or additional optimization techniques.

### 2.3. Text and Database-Level Identity Verification

Text recognition and natural language processing play a supporting role in multimodal attendance systems. In smart classroom environments, text-aware modules can be used to verify student identity against institutional records, classroom databases, QR/ID inputs, or structured metadata. This function is important because visual recognition alone may be insufficient when students have similar appearances, when image quality is low, or when visual evidence is incomplete.

Transformer-based models are widely used for contextual representation and semantic matching. However, the original BERT architecture is often too resource-intensive for real-time classroom deployment. DistilBERT and MobileBERT offer more practical alternatives because they reduce model size and inference time while preserving acceptable matching quality. In an attendance monitoring system, these models do not need to operate as standalone natural language understanding engines. Their main function is to support identity consistency and reduce ambiguity in the final decision.

Table 2 summarizes the main characteristics of selected lightweight transformer-based models suitable for identity matching in constrained environments.

*Table 2. Literature-based comparison of candidate text-processing modules for mobile or edge-oriented inference.*

| Model      | Role in system               | Reported accuracy range (%) | Reported latency range (ms) | Rationale for inclusion                                          |
|------------|------------------------------|-----------------------------|-----------------------------|------------------------------------------------------------------|
| BERT       | Contextual semantic matching | 90–95                       | 80–120                      | Strong text modeling but too heavy for efficient edge deployment |
| DistilBERT | Text/database                | 85–90                       | 40–60                       | Better                                                           |

|             |                        |       |       |                                                              |
|-------------|------------------------|-------|-------|--------------------------------------------------------------|
|             | se matching            |       |       | trade-off between semantic quality and real-time suitability |
| MobileBE RT | Compact text inference | 84–89 | 35–55 | Alternative lightweight NLP module for constrained devices   |

In the proposed system, the text/database branch is considered a consistency-supporting component rather than the dominant recognition source. This role is important because it allows automatic recognition to be checked against institutional information while preventing database records alone from being treated as proof of physical classroom presence.

#### 2.4. Audio-Based Participation and Activity Cues

Audio analysis can provide useful complementary evidence in smart classrooms, especially when visual recognition is temporarily degraded. For example, speech activity, participation cues, or voice-related patterns may help support the presence of a student when camera-based recognition is affected by occlusion, pose changes, or poor lighting. Modern deep learning approaches such as wav2vec 2.0 provide robust speech representations, while CNN-LSTM models remain attractive for real-time systems because they combine temporal modeling with moderate computational complexity.

At the same time, audio is one of the most context-sensitive modalities in classroom environments. Its reliability may decrease because of overlapping speech, background noise, microphone distance, hallway sound leakage, air-conditioning noise, or reverberation. For this reason, audio evidence should not be treated as equally reliable under all operating conditions. A practical multimodal attendance system should be able to reduce the contribution of the audio branch when acoustic quality is poor.

Table 3 presents a comparison of candidate speech-processing models considered for classroom-scale deployment.

*Table 3. Literature-based comparison of candidate speech-processing modules for edge-oriented classroom applications.*

| Model       | Role in system                 | Reported accuracy range (%) | Reported latency range (ms) | Rationale for inclusion                                                    |
|-------------|--------------------------------|-----------------------------|-----------------------------|----------------------------------------------------------------------------|
| CNN-LSTM    | Audio activity analysis        | 85–90                       | 40–50                       | Suitable compromise between temporal modeling and practical inference      |
| wav2vec 2.0 | Advanced speech representation | 88–93                       | 40–60                       | More robust audio branch when acoustic evidence is informative             |
| HMM         | Classical lightweight baseline | 70–75                       | 30–40                       | Computationally efficient but less robust under realistic noise conditions |

The comparison indicates that compact deep audio models are more suitable than classical HMM-based approaches for multimodal classroom systems. However, their contribution should remain adaptive because the acoustic environment of a classroom is not stable over time.

#### 2.5. Reliability-Aware Multimodal Fusion

Multimodal artificial intelligence systems combine evidence from visual, audio, and text/database channels to improve decision reliability and reduce ambiguity. In attendance monitoring, each modality has a different role. The visual branch provides the main identity evidence, the audio branch contributes supplementary activity-related information, and the text/database branch checks identity consistency against institutional records. The combination of these channels can make the system more robust than a single-modality configuration.

However, multimodal fusion is not only a matter of combining several model outputs. In real classroom environments, the reliability of each modality changes according to input quality. Visual evidence may become unreliable when a face is occluded, blurred, or poorly illuminated. Audio evidence may become unreliable when the classroom is noisy or when several students speak simultaneously. Text/database evidence may become unreliable when identity records are incomplete, delayed, or inconsistent. Therefore, a practical fusion mechanism should assign modality

weights according to the quality of available evidence.

The relationship between the modalities can be represented as a general multimodal recognition pipeline, where visual, audio, and text/database evidence are processed separately and then integrated within a fusion and decision layer, as shown in Figure 1.

Such a pipeline is useful for smart classroom environments because it supports both automatic attendance logging and uncertainty management. When the fused evidence is strong, the system can record attendance automatically. When the evidence is weak or contradictory, the decision can be redirected to an abstention or manual-review state. This mechanism reduces the risk of false attendance records and improves the operational reliability of the system under non-ideal classroom conditions [26], [11], [27].

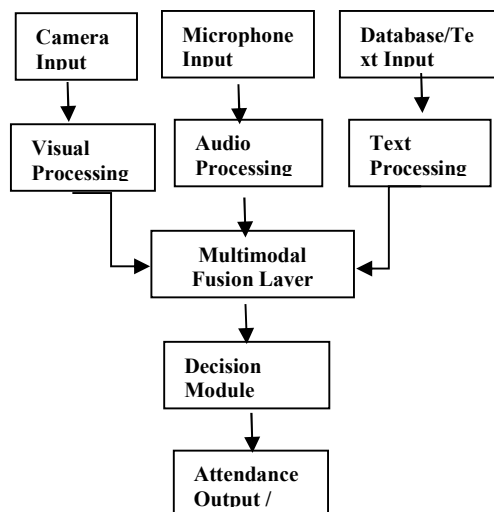


Figure 1. General multimodal recognition pipeline with visual, audio, and text/database branches followed by a fusion and attendance decision layer.

## 2.6. Limitations of Existing Systems and Research Gap

Existing attendance-monitoring systems can be grouped into three main categories: visual-only systems, card/QR/database-based systems, and multimodal systems. Visual-only systems are contactless and intuitive, but their reliability decreases under occlusion, shadowing, pose variation, motion blur, or crowded seating conditions. Card-, QR-, and database-based systems are computationally simple, but they mainly confirm identifier presentation rather than continuous physical presence. Multimodal systems

can reduce these limitations, but they also require careful design of fusion logic, latency control, and uncertainty handling.

Several research gaps remain. First, many existing systems are evaluated offline or under controlled conditions that do not reflect continuous classroom operation. Second, some systems depend on comparatively heavy hardware, which reduces their suitability for real-time edge deployment. Third, many studies report overall recognition accuracy without clearly separating literature-level model characteristics from the final performance of the integrated system. Finally, uncertainty handling and abstention mechanisms are often insufficiently discussed, although they are essential for practical and trustworthy classroom deployment.

To address these limitations, the present study separates two analytical levels: literature-based module selection and pilot-scale system-level evaluation. The proposed system uses lightweight visual, audio, and text/database modules as candidate components, but the final contribution is evaluated at the level of integrated attendance decisions. This distinction makes it possible to analyze not only recognition performance but also latency, reliability, uncertainty handling, and deployment feasibility in a smart classroom environment.

Within this scope, the study contributes to the literature in four main ways. First, it proposes a reliability-adaptive probabilistic fusion rule with an abstention option for uncertain identities. Second, it develops a pilot deployment scenario for a 22-seat smart classroom using edge-oriented AI modules. Third, it explicitly separates candidate-module comparison from final integrated-system evaluation. Fourth, it discusses environmental stressors, failure modes, and scalability limits relevant to realistic classroom operation [28-30].

Table 4 summarizes the candidate AI modules selected for the design stage of the proposed system.

Table 4. Literature-based comparison of candidate AI modules used for system design.

| Module    | Role in system      | Reported accuracy range (%) | Reported latency range (ms) | Rationale for inclusion                                |
|-----------|---------------------|-----------------------------|-----------------------------|--------------------------------------------------------|
| MobileNet | Face identification | 88–91                       | 30–50                       | Lightweight visual branch suitable for edge deployment |
| YOLOv5    | Face/object         | 88–91                       | ≈45                         | Provides                                               |

|                           |                             |       |        |                                                                 |
|---------------------------|-----------------------------|-------|--------|-----------------------------------------------------------------|
| Distilled                 | detection                   |       |        | balanced detection quality and inference speed                  |
| DistilBERT                | Text/database matching      | 85–90 | 40–60  | Supports context-aware identity verification                    |
| CNN-LSTM / wav2vec 2.0    | Audio activity analysis     | 85–93 | 40–50  | Useful when visual evidence is partially degraded               |
| EfficientNet / ViT / BERT | Alternative stronger models | 90–96 | 45–130 | Higher computational cost; less suitable for real-time edge use |

This comparison clarifies that the reported ranges describe candidate module characteristics used for system design. They should not be interpreted as the final performance of the proposed system. The final evaluation of the proposed framework is reported at the integrated attendance-decision level rather than at the isolated model level.

### 3. PROPOSED METHOD AND RELIABILITY-AWARE DECISION FRAMEWORK

#### 3.1. Real-Time Processing Pipeline of the Proposed System

The proposed attendance monitoring system is designed as a reliability-aware multimodal artificial intelligence framework for smart classroom environments. Its main objective is to combine visual, audio, and text/database-level evidence in order to generate reliable attendance decisions under practical classroom conditions. Unlike conventional attendance tools that only record an identifier or require manual confirmation, the proposed system performs automatic recognition, evaluates the reliability of each modality, and redirects uncertain cases to abstention or manual review.

The system workflow consists of five main stages: multimodal data acquisition, preprocessing and normalization, modality-specific recognition, reliability-aware multimodal fusion, and attendance logging with optional manual review. During data acquisition, the system receives synchronized inputs from classroom cameras, microphones, and institutional database records. Visual frames are prepared for student detection and face-oriented identification, audio segments are denoised and

segmented for activity analysis, and text/database records are converted into structured identity-related information.

The visual branch performs student detection and face-based identification. The audio branch extracts participation- or activity-related cues from classroom sound signals, while the text/database branch verifies identity consistency using institutional records or ID-linked information. These branch-level outputs are not treated as equally reliable in all situations. Instead, they are passed to a reliability-aware fusion layer, where the contribution of each modality is adjusted according to the estimated quality of the available evidence.

The final attendance decision is generated through a threshold-and-consensus mechanism. If the fused evidence is sufficiently reliable, the student's attendance is recorded automatically and synchronized with the electronic attendance register. If the evidence is weak, incomplete, or contradictory, the system redirects the case to an abstention or manual-review state. This mechanism reduces the risk of forced incorrect recognition and improves the operational reliability of AI-supported attendance monitoring. Figure 2 illustrates how camera, microphone, and database inputs are processed through separate recognition branches and then combined in a reliability-aware decision layer.

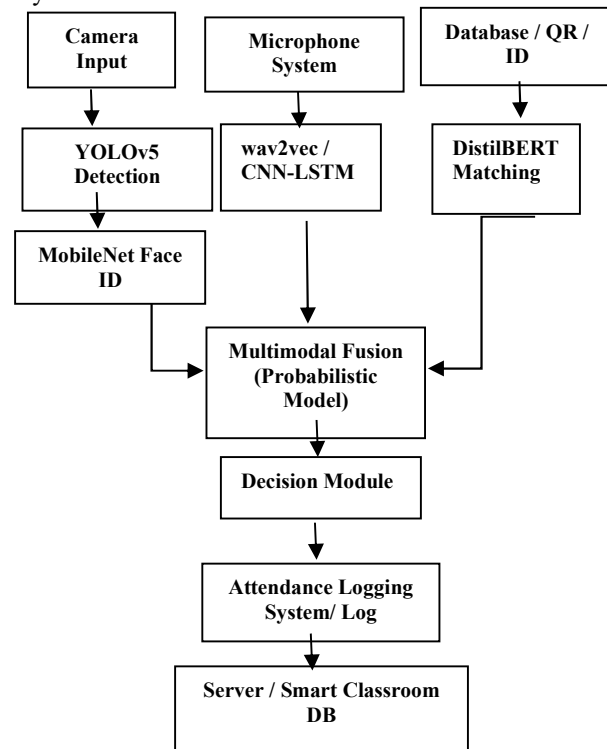


Figure 2. Proposed reliability-aware multimodal AI system for smart classroom attendance monitoring, integrating visual, audio, and text/database branches within a unified fusion and decision framework.

As shown in the architecture, the system is not limited to automatic recognition. It also includes reliability estimation, decision validation, attendance logging, and manual review when the available evidence is insufficient.

### 3.2. Reliability-Aware Multimodal Decision Rule

Let  $x_v$ ,  $x_a$ , and  $x_t$  denote the visual, audio, and text/database observations collected for a candidate student identity  $i$  during a short decision window. The proposed system does not use a simple unweighted average of modality outputs. Instead, it computes a reliability-aware posterior score that combines branch-specific confidence values with adaptive modality weights:

$$P(i | x_v, x_a, x_t) \propto \pi_i \cdot [p_v(i | x_v)]^{w_v} \cdot [p_a(i | x_a)]^{w_a} \cdot [p_t(i | x_t)]^{w_t}$$

where  $\pi_i$  is the prior probability associated with identity  $i$ , and  $p_v$ ,  $p_a$ , and  $p_t$  are the confidence scores produced by the visual, audio, and text/database branches, respectively. These confidence scores are normalized to the interval  $[0, 1]$ . The parameters  $w_v$ ,  $w_a$ , and  $w_t$  represent adaptive modality weights and satisfy the following condition:

$$w_v + w_a + w_t = 1$$

This formulation allows the system to increase or decrease the influence of each modality depending on its reliability. For example, the visual modality should have lower influence when the face is blurred, occluded, or affected by poor lighting. Similarly, the audio modality should have lower influence when background noise or overlapping speech reduces signal quality. The text/database modality should also be down-weighted when identity information is incomplete, delayed, or inconsistent.

Each modality weight is calculated using a branch-quality indicator  $q_m$ :

$$w_m = \frac{q_m}{q_v + q_a + q_t + \epsilon}, m \in \{v, a, t\}$$

where  $q_v$ ,  $q_a$ , and  $q_t$  represent the estimated quality of the visual, audio, and text/database branches, respectively. The parameter

$\epsilon$  is a small stabilization constant used to avoid numerical instability when the total signal quality is low.

The final attendance decision is not based only on the highest fused score. Let  $S_i$  denote the fused score of candidate identity  $i$ . Attendance is recorded only when the following condition is satisfied:

$$S_i \geq \tau$$

In addition, at least two modalities must exceed their local confidence thresholds. If these conditions are not satisfied, the observation is transferred to an abstention or manual-review state. This decision logic reduces the risk of false acceptance and prevents the system from recording attendance when the evidence is weak or contradictory.

Thus, the proposed decision framework combines three principles: probabilistic evidence aggregation, reliability-adaptive weighting, and threshold-based consensus validation. Together, these principles make the system more suitable for non-ideal classroom conditions, where one or more modalities may be temporarily unreliable.

### 3.3. Computational Complexity and Real-Time Feasibility

The proposed system is intended for real-time classroom operation. Therefore, computational feasibility is an important part of the system design. The approximate computational cost of one decision cycle can be expressed as follows:

$$C_{total} = C_{det} + N(C_{face} + C_{audio} + C_{text}) + C_{fuse}$$

where  $C_{det}$  denotes the cost of scene-level student or face detection,  $N$  is the number of candidate students in the decision window,  $C_{face}$  is the cost of face-oriented identity inference,  $C_{audio}$  is the cost of audio processing,  $C_{text}$  is the cost of text/database matching, and  $C_{fuse}$  is the cost of multimodal fusion.

This decomposition shows that the main computational burden is related to detection and candidate-level recognition rather than to the fusion rule itself. As the number of students increases, the cumulative cost of face, audio, and text/database processing becomes more significant. In contrast, the fusion stage introduces only a small additional overhead. For this reason, the use of lightweight

models such as YOLOv5, MobileNet, DistilBERT, and compact audio models is important for maintaining real-time operation in a smart classroom environment.

From a system implementation perspective, latency is not only a technical parameter but also an operational requirement. If attendance decisions are delayed, the system may interrupt the classroom workflow or reduce user trust. Therefore, the proposed architecture is designed to balance recognition accuracy, computational cost, and response time.

### 3.4. Pilot Deployment Scenario and Practical Implementation Context

The proposed system is designed for pilot deployment in a smart classroom equipped with cameras, microphones, and a database-linked attendance infrastructure. In this study, the deployment scenario corresponds to a 22-seat classroom operating under practical educational conditions, including different student positions, partial visibility, non-uniform lighting, and moderate acoustic interference.

Visual monitoring, audio activity capture, and database-level verification are synchronized through a local processing pipeline. The resulting attendance decisions are transmitted to the electronic registration system for storage and reporting. In cases where the fused evidence is insufficient, the system does not automatically register attendance. Instead, the case is marked for abstention or manual review. This function improves decision safety and reduces the likelihood of incorrect attendance records.

The purpose of the pilot-scale setup is not to claim full institutional validation or universal generalization. Rather, it is intended to demonstrate the operational feasibility of the proposed reliability-aware multimodal decision framework under realistic smart classroom conditions. Future large-scale deployment would require validation across more classrooms, different seating layouts, larger student groups, and longer academic periods.

To make the operational workflow of the proposed system more explicit, Table 5 summarizes the main functional stages, their input data, processing operations, and expected outputs.

This staged representation shows that the system is designed as an end-to-end applied AI pipeline, from multimodal data acquisition to final attendance logging or manual review.

Table 5. Functional stages of the proposed multimodal attendance monitoring pipeline.

| Stage                | Input                                  | Main operation                                  | Output                                  |
|----------------------|----------------------------------------|-------------------------------------------------|-----------------------------------------|
| Data acquisition     | Camera, microphone, and database input | Synchronous multimodal capture                  | Raw multimodal observations             |
| Preprocessing        | Images, audio, and text records        | Normalization, denoising, and structuring       | Clean branch-specific inputs            |
| Visual branch        | Image frames                           | Student detection and face-oriented recognition | Visual confidence score                 |
| Audio branch         | Audio segments                         | Participation/activity analysis                 | Audio confidence score                  |
| Text/database branch | Institutional records or ID data       | Identity matching and consistency checking      | Text/database confidence score          |
| Fusion               | Branch confidence values               | Reliability-adaptive probabilistic aggregation  | Fused decision score                    |
| Decision and logging | Fused score and modality thresholds    | Threshold-and-consensus validation              | Attendance record or manual-review case |

## 4. EXPERIMENTAL SETUP, EVALUATION PROTOCOL AND MULTIMODAL DECISION RESULTS

### 4.1. Pilot Classroom Deployment Scenario

The pilot deployment was designed for the Cybersecurity smart classroom at Mingachevir State University. The classroom has an area of approximately 40.88 m<sup>2</sup> and includes 22 student workstations. The operational setup consists of fixed classroom cameras, a microphone subsystem, a local or campus-connected attendance database, and a GPU-supported processing server used for synchronized inference and attendance logging.

The purpose of the pilot deployment was not to establish a large-scale institutional benchmark. Instead, it was designed to examine whether the proposed reliability-aware multimodal AI system can support stable attendance decisions under realistic classroom conditions. These conditions include partial face occlusion, non-uniform illumination, moderate acoustic disturbance, seating variability, and temporary uncertainty in identity evidence.

From a system evaluation perspective, the pilot environment is important because the attendance decision is not treated as an isolated recognition output. The system is designed to record reliable attendance decisions automatically when the available evidence is sufficient and to redirect uncertain cases to an abstention or manual-

review state when the evidence is weak or contradictory. This design allows the system to reduce forced incorrect recognition and improve the reliability of attendance logging under practical classroom constraints.

#### 4.2. Evaluation Protocol and Reproducibility-Oriented Clarifications

The evaluation protocol separates branch-level evidence from final decision-level performance. This distinction is important because the purpose of the study is not to report isolated benchmark results for individual AI models, but to evaluate the integrated attendance monitoring system under pilot classroom conditions.

The visual branch was evaluated using a classroom image set containing 5,000 images. These images were used to support student detection and face-oriented identification. The audio and text/database branches were evaluated as synchronized components of the same classroom-session workflow. The audio branch provided participation- or activity-related evidence, while the text/database branch supported identity consistency checking against institutional records.

The final system-level evaluation was repeated five times under the same pilot conditions. The integrated attendance accuracy was reported as mean  $\pm$  standard deviation. The final metrics refer to attendance decision outcomes rather than only to isolated branch outputs. This reporting approach is more appropriate for an applied AI-based attendance monitoring system because it reflects both recognition performance and decision reliability.

The main conditions of the pilot deployment and evaluation procedure are summarized in Table 6 in order to clarify how the system-level results were obtained.

Table 6. Pilot deployment and evaluation protocol used in this study

| Component           | Reported condition                                                    | Role in evaluation                       | Clarification                                            |
|---------------------|-----------------------------------------------------------------------|------------------------------------------|----------------------------------------------------------|
| Classroom scenario  | Cybersecurity smart classroom, 40.88 m <sup>2</sup> , 22 workstations | Operational pilot environment            | Supports the case-study component and deployment realism |
| Visual data         | 5,000 classroom images                                                | Face detection and identification branch | Used for visual recognition under classroom conditions   |
| Audio/text channels | Synchronized classroom-session                                        | Support and disambiguate uncertain       | Evaluated as components of the same                      |

|                       | evidence                                      | visual cases                                     | pilot workflow                                                         |
|-----------------------|-----------------------------------------------|--------------------------------------------------|------------------------------------------------------------------------|
| Repetition protocol   | Five repeated runs under identical conditions | Statistical stability of final decision accuracy | Provides the basis for mean $\pm$ SD reporting                         |
| Final reporting level | Attendance decision level                     | System-wide validation                           | Separates final system performance from literature-level module ranges |

These conditions define the scope of the reported results and show that the evaluation was performed at the integrated attendance-decision level rather than only at the isolated model level.

#### 4.3. Algorithmic Workflow of the Proposed Attendance System

The algorithmic workflow of the proposed attendance monitoring system is illustrated in Figure 3.

The process begins with synchronized acquisition of visual, audio, and text/database-linked inputs. These inputs are then processed through preprocessing and quality-control stages. Visual frames are normalized for detection and face recognition, audio signals are denoised or segmented, and database-linked identity information is structured for matching.

After preprocessing, the visual branch performs student detection and face-oriented recognition. The audio branch extracts participation- or activity-related cues, while the text/database branch verifies identity consistency. The outputs of these branches are then passed to the reliability-aware fusion mechanism described in Section 3.

The final decision is produced through a threshold-and-consensus rule. If the fused score and local modality confidence values satisfy the required decision conditions, attendance is recorded automatically. If the evidence is weak, incomplete, or contradictory, the system does not force a positive recognition decision. Instead, the case is redirected to an abstention or manual-review state. This workflow improves decision safety and prevents the system from recording attendance when the available multimodal evidence is insufficient.

To further clarify the decision process, Figure 3 presents the algorithmic workflow of the proposed attendance monitoring system from sensor activation to attendance logging and report generation.

The workflow highlights the role of threshold checking and manual review, which

prevents the system from recording attendance when the recognition evidence is weak or contradictory.

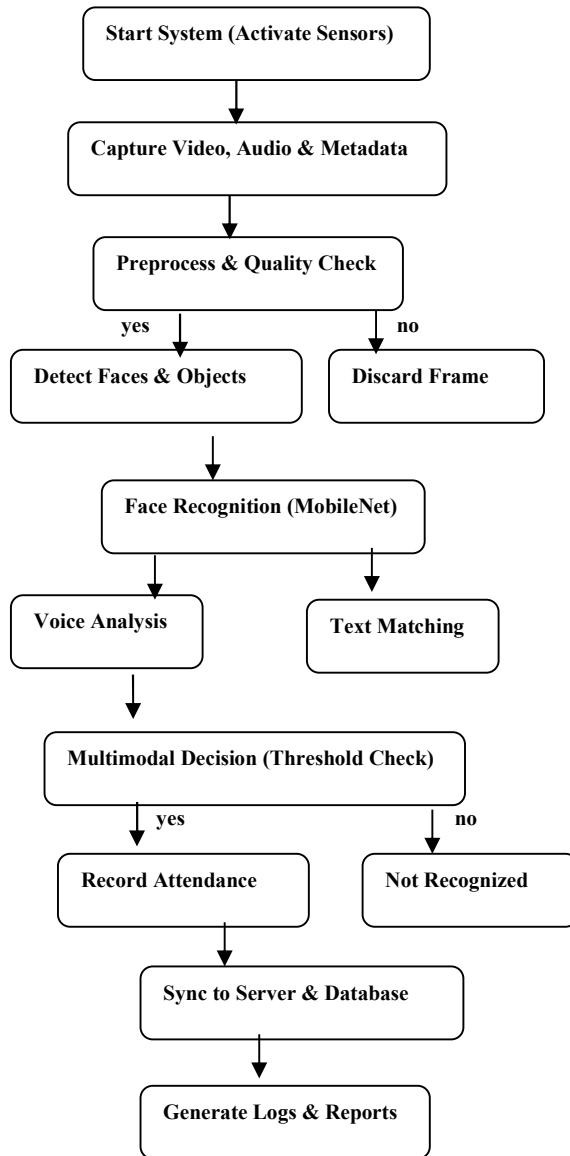


Figure 3. Algorithmic workflow of the proposed multimodal attendance monitoring system, from synchronized data acquisition and preprocessing to reliability-aware fusion, threshold validation, attendance logging, and manual review.

#### 4.4. Multimodal Decision-Level Results

The proposed system was evaluated using four main metric groups: overall attendance accuracy, inference latency, misidentification rate, and abstention/manual-review rate. Attendance accuracy measures the proportion of correct final

attendance decisions. Inference latency represents the average time required to produce one attendance decision window. Misidentification rate indicates cases where one student is incorrectly attributed to another. Abstention or manual-review rate reflects unresolved cases intentionally withheld by the system when the available evidence is insufficiently reliable.

This metric set is more appropriate for an applied real-time attendance monitoring system than model-centric accuracy alone. In practical deployment, a useful attendance system must not only recognize students accurately, but also manage uncertainty, avoid false acceptance, and maintain responsiveness during classroom operation.

The integrated multimodal system achieved a mean attendance accuracy of  $93.2 \pm 0.7\%$  across five repeated runs under pilot classroom conditions. The final decision latency was in the range of 45–55 ms, indicating that the proposed approach is suitable for real-time attendance monitoring in a medium-size smart classroom. The fused decision result was higher than the representative branch-level ranges of the individual modalities, showing that reliability-aware multimodal fusion improves the stability of attendance decisions compared with single-modality configurations.

The use of an abstention mechanism is also important from a deployment perspective. When the system receives weak or contradictory evidence, it avoids making a forced automatic decision. Such cases are redirected to manual review, which reduces the risk of false attendance records and improves the operational reliability of the system.

Overall, the pilot results demonstrate that the proposed reliability-aware multimodal AI framework can provide stable attendance decisions under non-ideal classroom conditions. At the same time, the reported results should be interpreted as pilot-scale evidence rather than as a large-scale institutional benchmark. Broader validation across different classroom layouts, student group sizes, lighting conditions, and acoustic environments is required for future deployment.

## 5. RESULTS AND DISCUSSION

The pilot results are interpreted from a deployment-oriented and applied AI system perspective. The study does not claim universal generalization across all classroom environments. Instead, it demonstrates the operational feasibility of a reliability-aware multimodal attendance

monitoring system under the conditions of a medium-size smart classroom.

The main result is that the integrated decision layer achieved an attendance accuracy of  $93.2 \pm 0.7\%$  over five repeated runs, with final decision latency in the 45–55 ms range. These values indicate that the system can operate in real time without introducing a significant delay into the classroom workflow. This is important because an attendance monitoring system should support classroom administration without interrupting teaching activities.

The benefit of the proposed fusion strategy is not limited to accuracy improvement. Its main advantage is greater decision stability when one modality becomes unreliable. Under normal seating and lighting conditions, the visual branch provides the primary identity evidence. When visual confidence decreases because of face occlusion, side-view pose, motion blur, or uneven lighting, the audio and text/database branches provide complementary support. Similarly, when acoustic evidence becomes unreliable because of background noise or overlapping speech, the system can reduce the weight of the audio branch and rely more strongly on visual and database-linked evidence.

The fusion layer should therefore be understood as a decision stabilizer rather than a replacement for individual recognition branches. The visual branch remains the main source of identity evidence, the audio branch provides supplementary participation- or activity-related information, and the text/database branch acts as a consistency constraint that reduces identity ambiguity. This division of roles improves the interpretability of the final decision and supports more reliable attendance logging under non-ideal classroom conditions.

The quantitative pilot observations are summarized in Table 7, where the branch-level ranges are compared with the final multimodal decision result.

The comparison indicates that the fused decision layer provides the most stable attendance output, while the individual branches remain important as complementary sources of evidence. The final multimodal decision achieves higher stability than any single branch alone because the reliability-aware fusion mechanism adapts modality weights according to signal quality. This result supports the main assumption of the study: attendance monitoring in practical classroom environments benefits from adaptive multimodal evidence

aggregation rather than from a single recognition source.

At the same time, the reported values should be interpreted as pilot-scale results rather than as a large-scale institutional benchmark. The evaluation was conducted in a 22-seat smart classroom under controlled pilot conditions. Therefore, broader validation across different classrooms, lighting conditions, acoustic environments, student group sizes, and longer academic periods would be required before making stronger generalization claims.

Table 7. Pilot-scale results of the proposed attendance monitoring system.

| Decision branch                       | Observed accuracy | Latency (ms) | Decision role                   | Status in this study                     |
|---------------------------------------|-------------------|--------------|---------------------------------|------------------------------------------|
| MobileNet-based visual identification | 88–91%            | 30–50        | Primary face-based evidence     | Branch-level pilot range                 |
| DistilBERT text/database verification | 85–90%            | 40–60        | Identity consistency check      | Branch-level pilot range                 |
| CNN-LSTM / wav2vec 2.0 audio analysis | 85–93%            | 40–50        | Supplementary activity evidence | Branch-level pilot range                 |
| Final multimodal fusion decision      | $93.2 \pm 0.7\%$  | 45–55        | Attendance logging output       | Final pilot result reported in this work |

## 6. FAILURE CASES, ROBUSTNESS AND SCALABILITY LIMITS

A structured discussion of failure cases is necessary because the advantage of multimodal fusion cannot be established by accuracy alone. In real classroom environments, recognition errors may occur for different reasons, and each modality has its own limitations. For this reason, the proposed framework is designed not only to improve recognition performance, but also to manage uncertainty when one or more sources of evidence become unreliable.

The main visual failure cases observed in the pilot scenario are partial face occlusion, side-view head pose, motion blur during student entry, and non-uniform illumination near windows or projectors. These conditions reduce detector and identifier confidence and may lead to uncertain visual recognition. In such cases, the proposed

fusion framework lowers the contribution of the visual modality and requires support from other sources of evidence or redirects the case to manual review.

The audio branch is sensitive to overlapping speech, background noise, air-conditioning noise, hallway sound leakage, and room reverberation. Because of these factors, audio evidence should not be treated as equally reliable in all classroom situations. The proposed framework reduces the audio weight when acoustic quality is low, preventing the system from making audio-dominant decisions under noisy conditions.

The text/database branch is computationally efficient and useful for identity consistency checking, but it cannot independently verify continuous physical presence. For this reason, it should not be used as the dominant attendance indicator. Instead, it acts as a supporting component that helps reduce ambiguity when visual or audio evidence is incomplete.

Scalability remains an important limitation. A 22-seat classroom is suitable for pilot validation, but larger lecture halls may require additional cameras, multi-view fusion, distributed edge inference, or synchronized capture during peak entry periods. Multi-classroom deployment would also introduce additional requirements related to network stability, storage, privacy governance, model updates, and system maintenance. These issues are outside the scope of the present pilot and should be considered in future work.

To support a more practical interpretation of the system limitations, Table 8 summarizes the main failure conditions, affected modalities, expected impact, and mitigation logic used in the proposed framework.

This analysis shows that the proposed system does not treat all errors uniformly. Instead, it adapts the fusion weights or redirects uncertain cases to review depending on the type of failure. Within these limits, the proposed system should be interpreted as an edge-oriented proof of concept for medium-size smart classrooms. The pilot results indicate that reliability-aware multimodal fusion can improve attendance decision stability, but further testing is required before deployment in larger and more heterogeneous educational environments.

Table 8. Critical failure modes and mitigation logic in the proposed framework.

| Failure condition                               | Most affected modality | Expected impact                            | Mitigation in the proposed framework                                 |
|-------------------------------------------------|------------------------|--------------------------------------------|----------------------------------------------------------------------|
| Face partially blocked or turned away           | Visual                 | Reduced detector and identifier confidence | Lower $w_v$ , require support from text/audio, or redirect to review |
| High background noise or overlapping speech     | Audio                  | Unstable speech/activity evidence          | Lower $w_a$ , avoid audio-dominant decisions                         |
| Database mismatch or delayed identity lookup    | Text/database          | Weak consistency check                     | Allow visual-audio agreement to proceed only with review support     |
| Many students entering simultaneously           | System-wide            | Temporary latency increase                 | Use batching, staggered capture, or additional edge nodes            |
| Large lecture hall with limited camera coverage | Visual and system-wide | Missed detections and blind zones          | Use additional cameras and multi-view fusion                         |

## 7. CONCLUSION

This study presented a reliability-aware multimodal artificial intelligence framework for autonomous attendance monitoring in smart classroom environments. The proposed system integrates visual recognition, audio-based participation or activity cues, and text/database-level identity verification within a unified probabilistic fusion model. Unlike single-modality attendance systems, the proposed approach adjusts the contribution of each modality according to signal quality and redirects uncertain cases to an abstention or manual-review state. This design improves decision safety and reduces the risk of forced incorrect attendance registration under practical classroom conditions.

The pilot deployment scenario demonstrated the feasibility of the proposed system in a medium-size smart classroom. The integrated multimodal decision layer achieved an attendance accuracy of  $93.2 \pm 0.7\%$  with a final decision latency of 45–55 ms, indicating that the system can support real-time

classroom operation. The results show that reliability-aware fusion can improve decision stability when one modality is degraded by occlusion, illumination variability, background noise, or incomplete identity evidence. In this context, the fusion layer acts as a decision stabilizer that combines the strengths of visual, audio, and text/database branches while reducing the influence of unreliable evidence.

The study also clarified the distinction between literature-based module characteristics and pilot-scale integrated-system performance. This distinction is important because the proposed system should be interpreted as a deployment-oriented proof of concept rather than as a universal benchmark solution. The obtained results are promising for medium-size smart classrooms, but they remain limited to the specific pilot environment, classroom configuration, and evaluation conditions considered in this study.

Future research should focus on broader validation across different classroom layouts, lighting conditions, acoustic environments, and student group sizes. Longer-term deployment studies are also needed to evaluate system stability over full academic periods. In addition, future work should investigate privacy-aware multi-camera deployment, distributed edge inference, secure attendance data management, and improved interfaces for reviewing uncertain cases. These developments would help transform the proposed pilot framework into a more mature and scalable AI-supported attendance monitoring system for smart educational environments.

## Declarations

### Availability of Data and Materials

The data and materials supporting the findings of this study are available from the corresponding author upon reasonable request. Due to privacy and ethical restrictions related to classroom images, audio recordings, biometric information, and student identity records, raw visual and audio data cannot be made publicly available. The manuscript provides the system architecture, evaluation protocol, metric definitions, decision framework, and aggregated results required to understand the research design and reproduce the methodological rationale.

### Ethics Statement and Data Privacy

This study was conducted as a pilot technical evaluation of a smart classroom attendance monitoring system. The pilot evaluation was conducted as an internal technical assessment,

and the collected data were used only for aggregated system-level analysis.

The study involved classroom-related visual, audio, and database-linked information for evaluating recognition accuracy, decision reliability, and system feasibility. No personally identifiable classroom images, audio recordings, biometric templates, or student records are published in this manuscript. Access to raw data was restricted to the research team. All reported results are presented only in aggregated form, and no individual student is identified or analyzed separately. Raw visual and audio data are not made publicly available due to privacy, ethical, and data-protection restrictions. Future large-scale deployment of the proposed system will require formal institutional ethics approval, informed consent procedures, privacy impact assessment, and a detailed data protection protocol in accordance with applicable institutional and legal requirements.

## Declaration of AI-Assisted Writing

During the preparation of this manuscript, ChatGPT was used only for language refinement, academic style improvement, grammar correction, and structural editing. The tool was not used to generate research data, perform experiments, conduct statistical analysis, create results, or make scientific conclusions.

All AI-assisted revisions were reviewed, edited, and approved by the authors. The authors take full responsibility for the scientific content, methodology, results, interpretations, references, and conclusions of the manuscript. No AI tool was listed as an author, and no AI-generated content was included without human verification and author approval.

## Authors' Contributions

### O. Imamverdiyev:

Conceptualization, investigation, data collection, pilot deployment, formal analysis, writing – original draft.

### J. Teymurov:

Software, system architecture, technical implementation, AI model integration, data processing.

### S. Mahmudova:

Methodology, system architecture, writing – review and editing.

**N. Mahmudova:**

Data curation, validation, formal analysis, visualization, interpretation of results.

**A. Ahmadova:**

Methodology, literature review, system design, evaluation protocol, writing – original draft.

**A. Karimova:**

Methodology, validation, interpretation of results, writing – review and editing.

**T. Nurubeyli:**

Conceptualization, methodology, supervision, project administration, validation, writing – review and editing.

**Competing Interests**

The authors declare that they have no competing interests.

**Funding**

This research received no external funding.

**Consent for Publication**

All authors have read and approved the final version of the manuscript and agree with its submission to this journal.

**REFERENCES:**

- [1] E. Al Hajri, F. Hafeez, and A. N. V. Azhar, “Fully automated classroom attendance system”, *International Journal of Interactive Mobile Technologies*, Vol. 13, No. 8, 2019, pp. 95–106.  
<https://doi.org/10.3991/ijim.v13i08.10100>
- [2] T. M. Alam, G. A. Stoica, and Ö. Özgöbek, “Asking the classroom with technology: A systematic literature review”, *Smart Learning Environments*, Vol. 12, Article 7, 2025.  
<https://doi.org/10.1186/s40561-024-00348-z>
- [3] M. Zhao, G. Zhao, and M. Qu, “College smart classroom attendance management system based on Internet of Things”, *Computational Intelligence and Neuroscience*, Vol. 2022, Article 4953721, 2022.  
<https://doi.org/10.1155/2022/4953721>
- [4] A. S. Lateef and M. Y. Kamil, “Face recognition-based automatic attendance system in a smart classroom”, *Iraqi Journal for Electrical and Electronic Engineering*, Vol. 20, No. 1, 2024, pp. 37–47.  
<https://doi.org/10.37917/ijeee.20.1.4>
- [5] L. C. Suleymanova, T. K. Nurubeyli, and S. Z. Babayev, “Structure and degradation behavior of polymeric dielectrics under partial discharges in SF<sub>6</sub> and O<sub>3</sub> environments,” *International Journal on Technical and Physical Problems of Engineering*, vol. 17, no. 3, pp. 326–335, 2025.
- [6] A. G. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, “Searching for MobileNetV3”, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.  
<https://doi.org/10.48550/arXiv.1905.02244>
- [7] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement”, *arXiv preprint arXiv:1804.02767*, 2018.  
<https://doi.org/10.48550/arXiv.1804.02767>
- [8] G. Z. Abbasov, G. N. Mammadova, T. K. Nurubeyli et.al., “Intelligent maintenance strategies for transport systems: A machine learning approach to thermomechanical stress reduction,” *International Journal on Technical and Physical Problems of Engineering*, vol. 17, no. 2, pp. 152–162, 2025.
- [9] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks”, *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114.  
<https://proceedings.mlr.press/v97/tan19a.html>
- [10] İ. Ardic, B. Yıldızhan, and K. Uyar, “Smart Classroom Attendance and Management System with Deep Learning”, *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, Vol. 40, No. 3, 2024, pp. 487–497.  
<https://izlik.org/JA89TB35MG>
- [11] J. Samatha and M. Gudavalli, “Smart attendance monitoring system using multimodal biometrics”, *Sigma Journal of Engineering and Natural Sciences*, Vol. 43, No. 1, 2025, pp. 168–188.  
<https://doi.org/10.14744/sigma.2024.00030>
- [12] C. Gupta, N. S. Gill, P. Gulia, A. Kumar, H. Karamti, D. M. Moges, and I. Safra, “A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features”, *Journal of Big Data*, Vol. 12, Article 256, 2025.  
<https://doi.org/10.1186/s40537-025-01300-9>
- [13] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, “Multimodal fusion for audio-image and video action recognition”, *Neural Computing and Applications*, Vol. 36, 2024, pp. 5499–5513.  
<https://doi.org/10.1007/s00521-023-09186-5>

- [14] R. D. Alfredo, V. Echeverria, Y. Jin, L. Yan, Z. Swiecki, D. Gašević, and R. Martinez-Maldonado, "Human-centred learning analytics and AI in education: A systematic literature review", *Computers and Education: Artificial Intelligence*, Vol. 6, Article 100215, 2024. <https://doi.org/10.1016/j.caeai.2024.100215>
- [15] K. T. Kotsis, "Issues between artificial intelligence and personal data in education", *International Research in Education*, Vol. 13, No. 1, 2025, pp. 45–65. <https://doi.org/10.5296/ire.v13i1.22850>
- [16] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics", *British Journal of Educational Technology*, Vol. 45, No. 3, 2014, pp. 438–450. <https://doi.org/10.1111/bjet.12152>
- [17] L. Tang and Y. S. Su, "Ethical implications and principles of using artificial intelligence models in the classroom: A systematic literature review", *International Journal of Interactive Multimedia and Artificial Intelligence*, 2024. <https://doi.org/10.9781/ijimai.2024.02.010>
- [18] G. Jocher, A. Chaurasia, A. Stoller, and J. Qiu, "YOLOv5 by Ultralytics (Version 7.0)", *GitHub Computer Software*, 2020. <https://github.com/ultralytics/yolov5>
- [19] C. Alippi, S. Disabato, and M. Roveri, "Moving convolutional neural networks to embedded systems: The AlexNet and VGG-16 case", *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2018.
- [20] G. Z. Abbasov, A. G. Mustafayev, G. N. Mammadova, et al., "AI-driven digital twin optimization for reducing thermomechanical stress in transport components," *International Journal on Technical and Physical Problems of Engineering*, vol. 17, no. 62, pp. 365–374, 2025.
- [21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186. <https://doi.org/10.48550/arXiv.1810.04805>
- [22] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2158–2170.
- [23] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol. 77, No. 2, 1989, pp. 257–286. <http://dx.doi.org/10.1109/5.18626>
- [24] D. Ryumin, D. Ivanko, and E. Ryumina, "Audio-visual speech and gesture recognition by sensors of mobile devices", *Sensors*, Vol. 23, No. 4, Article 2284, 2023. <https://doi.org/10.3390/s23042284>
- [25] S. Naveen and M. R. Kounte, "Optimized convolutional neural network at the IoT edge for image detection using pruning and quantization", *Multimedia Tools and Applications*, Vol. 84, 2025, pp. 5435–5455. <https://doi.org/10.1007/s11042-024-20523-1>
- [26] S. Amershi, D. Weld, M. Vorvoreanu, A. Fournery, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13. <https://doi.org/10.1145/3290605.3300233>
- [27] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe and trustworthy", *International Journal of Human-Computer Interaction*, Vol. 36, No. 6, 2020, pp. 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- [28] M. Yadav, K. Singh, V. H. Abdullayev, F. Moreira, D. D. Barak, R. K. Sharma, S. Kumar, and D. Kumar, "A comprehensive analysis of machine learning-based retention of staff members", *Scientific Works*, No. 1, 2026, pp. 92–100. <https://doi.org/10.58225/sw.2026.1-92-100>
- [29] H. Varum, A. Kasımozade, and A. Aliyeva, "First-level seismic risk assessment of reinforced concrete public buildings considering hilltop and slope effects in Samsun, Türkiye", *Scientific Works*, No. 1, 2026, pp. 101–115. <https://doi.org/10.58225/sw.2026.1-101-115>
- [30] D. Miroshnichenko and M. Abbasov, "Modeling convergence rate estimates of the Heavy Ball algorithm via a random forest with self-attention", *Scientific Works*, No. 1, 2026, pp. 53–61. <https://doi.org/10.58225/sw.2026.1-53-61>