DIDAKTIKA
Vol. 6 No. 3, September 2026, pp. 300-314
https://ejournal-serang.upi.edu/didaktika

Development of a Computer Vision-Based Instrument for Assessing Mindful, Meaningful, and Joyful Learning in Elementary Schools

Tegar Baradika Putra Karunia[1], Muhammad Hanif[2]✉, Dian Asmarajati[3], Iman Ahmad Ihsannuddin[4], Budi Setiawan[5], & Aan Khosihan[6]

[1] Universitas Sains Al-Quran Jawa Tengah, tegarbrdka@gmail.com, Orcid ID: 0009-0001-4656-0244
[2]✉Universitas Pendidikan Indonesia, muhammadhanif@upi.edu, Orcid ID: 0000-0003-4132-9788
[3] Universitas Sains Al-Quran Jawa Tengah, asmarajati@unsiq.ac.id, Orcid ID: 0009-0005-1026-475X
[4] Universitas Sains Al-Quran Jawa Tengah, ahmadihsan@unsiq.ac.id, Orcid ID: 0000-0001-8482-1823
[5] Universitas Pendidikan Indonesia, budi_setiawan@upi.edu, Orcid ID: 0000-0001-5371-8129
[6] Universitas Pendidikan Indonesia, aankhosihan@upi.edu, Orcid ID: 0000-0002-7483-6527

Article Info

History Articles
Received:
09 August 2026
Accepted:
10 September 2026
Published:
20 September 2026

Abstract

The Deep Learning approach (Mindful, Meaningful, Joyful) is a student-centered pedagogical framework adopted in Indonesia's Merdeka Belajar curriculum. However, the lack of objective, scalable assessment instruments remains a significant challenge. This study aimed to develop and validate EduClassify, a Computer Vision-based assessment instrument for deep learning in elementary schools. Using the ADDIE R&D model, the system was evaluated across 28 classroom observation sessions. EduClassify integrates four Computer Vision models (YOLOv8, MediaPipe, DeepFace, and gaze tracking) to detect indicators across all 3M dimensions automatically. Results showed that 66.7% of classes met the 'Fairly Good' category (score ≥ 50). Meaningful Learning showed the strongest correlation with the overall Deep Learning score (r = 0.847). A strong positive correlation was also found between student talk time and Meaningful score (r = 0.910), confirming student-centered interaction as the primary driver of meaningful learning. Reliability testing using Cronbach's Alpha yielded α = 0.722 (Mindful) and α = 0.748 (Meaningful), whereas Joyful scored α = 0.031 due to rear-angle camera limitations. This study demonstrates the instrument's feasibility and provides empirical recommendations for multi-camera configurations in future implementations. Practically, this instrument enables school leaders to conduct data-driven clinical pedagogical supervision and targeted instructional coaching.

Keywords: Computer Vision, Deep Learning, Pedagogical Learning Assessment

How to cite: Karunia, T. B. P., Hanif, M., Asmarajati, D., Ihsannuddin, I. A., Setiawan, B., & Khosihan, A. (2026). Development of a computer vision-based instrument for assessing mindful, meaningful, and joyful learning in elementary schools. *Didaktika, 6*(3), 300-314.

Info Artikel

Riwayat Artikel
Dikirim:
09 Agustus 2026
Diterima:
10 September 2026
Diterbitkan:
20 September 2026

Abstrak

Pendekatan Deep Learning (yang mencakup aspek Mindful, Meaningful, dan Joyful) merupakan kerangka pedagogis yang berpusat pada siswa, sebagaimana diadopsi dalam kurikulum Merdeka Belajar di Indonesia. Namun, ketiadaan instrumen penilaian yang objektif dan dapat diskalakan masih menjadi tantangan yang signifikan. Studi ini bertujuan untuk mengembangkan dan memvalidasi EduClassify, sebuah instrumen penilaian berbasis Computer Vision untuk pembelajaran Deep Learning di sekolah dasar. Dengan menggunakan model penelitian dan pengembangan (R&D) ADDIE, sistem ini dievaluasi melalui 28 sesi observasi kelas. EduClassify mengintegrasikan empat model Computer Vision (YOLOv8, MediaPipe, DeepFace, dan pelacakan tatapan mata/gaze tracking) untuk mendeteksi indikator-indikator dalam ketiga dimensi tersebut secara otomatis. Hasil penelitian menunjukkan bahwa 66,7% kelas memenuhi kategori 'Cukup' (skor $\geq$ 50). Pembelajaran bermakna (Meaningful Learning) menunjukkan korelasi terkuat dengan skor Indeks Deep Learning secara keseluruhan ($r = 0,847$). Korelasi positif yang kuat juga ditemukan antara durasi siswa berbicara dengan skor Meaningful ($r = 0,910$), yang mengonfirmasi bahwa interaksi yang berpusat pada siswa merupakan pendorong utama pembelajaran yang bermakna. Pengujian reliabilitas menggunakan Cronbach's Alpha menghasilkan nilai $\alpha = 0,722$ untuk dimensi Mindful dan $\alpha = 0,748$ untuk dimensi Meaningful, sementara dimensi Joyful memperoleh skor $\alpha = 0,031$. Kelemahan pada aspek Joyful ini disebabkan oleh keterbatasan sudut pandang kamera yang ditempatkan di bagian belakang. Studi ini menunjukkan kelayakan instrumen tersebut dalam menilai implementasi Deep Learning serta memberikan rekomendasi empiris terkait konfigurasi multi-kamera untuk implementasi di masa mendatang. Secara praktis, instrumen ini memungkinkan kepala sekolah untuk melaksanakan supervisi pedagogis klinis berbasis data serta memberikan pembinaan instruksional yang lebih terarah.

Kata Kunci: Computer Vision, Pembelajaran Mendalam, Penilaian Pedagogis Pembelajaran

Cara mengutip: Karunia, T. B. P., Hanif, M., Asmarajati, D., Ihsannuddin, I. A., Setiawan, B., & Khosihan, A. (2026). Development of a computer vision-based instrument for assessing mindful, meaningful, and joyful learning in elementary schools. *Didaktika, 6*(3), 300-314.

©2026 Universitas Pendidikan Indonesia
e-ISSN: 2775-9024, p-ISSN: 2987-9388

INTRODUCTION

Quality education hinges on pedagogical designs that advance beyond rote memorization toward authentic deep learning, characterized by cognitive reasoning, conscious inquiry, and emotional resonance within collaborative classroom groups (Bentz, 1992; Kovač et al., 2025). Cultivating such student-centered environments empowers learners as active agents rather than passive recipients (Tal & Tsaushu, 2018). In Indonesia, this imperative is operationalized under the Kurikulum Merdeka, which embeds Ki Hadjar Dewantara's holistic philosophy through the "3M" deep learning framework: Mindful (conscious attentiveness), Meaningful (contextual conceptual understanding), and Joyful (positive affective engagement) (Suyanto, 2025). Grounding elementary education in these core pillars is critical for fostering foundational self-regulation, motivation, and inquiry skills.

Evaluating the realization of these instructional qualities in authentic primary classrooms requires operationalizing the 3M dimensions into observable behavioral markers (Bentz, 1992). However, conventional supervisory audits remain predominantly manual, episodic, and vulnerable to observer subjectivity and Hawthorne distortion, which obscures actual classroom dynamics (Lin et al., 2024; Trabelsi et al., 2023). This measurement bottleneck underscores the urgent need for scalable, automated assessment instruments that can capture multimodal classroom dynamics continuously and objectively.

Despite its policy emphasis, the classroom implementation of Deep Learning faces systemic assessment bottlenecks. Conventional classroom supervision in Indonesian elementary schools relies almost exclusively on manual, paper-based observational rubrics conducted intermittently by school principals or external supervisors. Recent empirical evidence demonstrates that manual pedagogical audits suffer from severe inherent biases: high evaluative subjectivity conditioned by personal observer preferences (Trabelsi et al., 2023), poor scalability due to prohibitive human-resource costs across vast school districts, and the Hawthorne effect, wherein both teachers and students artificially modify their authentic classroom behaviors in the presence of an in-person evaluator (Parsons, 1974). Consequently, manual supervisory visits yield cross-sectional, anecdotal data rather than continuous, reproducible evidence, leaving educators without granular diagnostic insights into their pedagogical interactions. In response, modern educational paradigms increasingly demand continuous, big-data-driven evaluation frameworks that can objectively synthesize complex classroom behavioral indicators at scale (Jin et al., 2025; Lin et al., 2024).

To overcome the subjectivity and latency of manual audits, computer vision (CV) and multimodal sensing have emerged as a transformative paradigm for continuous classroom monitoring (Liu et al., 2025; Szeliski, 2022). Recent comprehensive reviews highlight significant progress in automated student action recognition and behavioral monitoring (Liu et al., 2025); however, existing implementations largely classify generic behavioral markers without grounding these visual cues in a coherent, validated pedagogical framework such as the 3M learning paradigm. A synthesis of the literature shows substantial advances in automated behavioral sensing: Pillai (2022) used convolutional neural networks (CNNs) to detect student engagement via facial affect with 60.7% accuracy; Xu et al. (2023) coupled facial expression tracking with photoplethysmography-based video heart-rate estimation to capture cognitive load.

Recent studies have explored machine learning architectures to capture learner attention: Alruwais and Zakariah (2023) implemented classroom engagement detection algorithms, while Gupta et al. (2023) integrated multimodal facial cues using deep learning within virtual settings. However, critical gaps persist across existing studies, which largely rely on fragmented modalities

detached from grounded pedagogical frameworks, and predominantly target controlled e-learning environments rather than dynamic, face-to-face primary classrooms (Gupta et al., 2023; Liu et al., 2025). Complementarily, in the auditory domain, recent investigations leveraging Natural Language Processing and voice activity analysis have demonstrated that teacher-student vocalization ratios and acoustic dynamics serve as reliable proxies for interactive instruction (Nursehag & Amalia, 2024; Zhang et al., 2024).

Nevertheless, critical gaps persist across existing studies, which largely rely on fragmented, single-stream modalities rather than integrated visual and acoustic dynamics. Most models detect isolated, generic behaviors detached from grounded pedagogical frameworks like 3M learning. Additionally, these systems are typically designed for high-resource environments and lack ecological validity in dense, hardware-constrained elementary classrooms typical of Global South contexts such as Indonesia. These compounding limitations underscore an urgent, unaddressed need for a lightweight, non-intrusive, and computationally efficient multimodal assessment platform engineered specifically for naturalistic elementary classrooms. To bridge these technological and pedagogical chasms, this study introduces a system designed to automatically operationalize, extract, and quantify the behavioral correlates of Mindful, Meaningful, and Joyful learning.

METHODS

This study employed a Research and Development (R&D) approach using the ADDIE (Analysis, Design, Development, Implementation, Evaluation) framework to engineer and validate EduClassify. The ADDIE paradigm was selected due to its proven efficacy in systematically translating complex pedagogical requirements into iterative, scalable technological solutions (Adeoye et al., 2024; Spatioti et al., 2022). The procedures of each phase are below:

Phase 1: Analysis

The analysis phase comprised a literature review and an empirical field needs assessment. The theoretical synthesis operationalized Ki Hadjar Dewantara's Deep Learning pedagogical philosophy and the Kurikulum Merdeka guidelines into observable classroom constructs. The field needs assessment was conducted via Focus Group Discussions (FGDs) and structured questionnaires involving elementary school teachers and instructional supervisors. The assessment gathered baseline data on the limitations of conventional manual supervision with high observer subjectivity, latency in providing feedback, scheduling bottlenecks, and the Hawthorne effect as key challenges. These empirical insights established the system's functional requirements, underscoring the necessity for a continuous, non-intrusive, and objective multimodal computational instrument.

Phase 2: Design

The design phase translated the abstract 3M constructs into machine-extractable behavioral indicators and established the technical acquisition protocol. Each pedagogical dimension was mapped to a targeted computer vision or acoustic model (summarized in Table 1). Data acquisition parameters were standardized to ensure environmental replicability: video recordings were captured at 1080p resolution at 30 frames per second (FPS), with wide-angle cameras positioned at an elevation of 2.2 meters at the rear and lateral corners of the classroom. Multimodal synchronization utilized sliding temporal windows (5-second intervals) to align continuous audio waveforms with extracted visual frames.

Table 1. Multimodal AI Mapping Matrix and Behavioral Indicators for Deep Learning Assessment

Pedagogical Dimension	Aspects	Behavioral Indicator	Model AI used
Mindful (33%)	Gaze Attention (40%), Skeletal Posture (35%), and Productive Silence (25%)	Students look at the board/teacher; maintain an upright posture; do not engage in activities unrelated to the lesson; and classroom silence ratio while the teacher is explaining	MediaPipe, Gaze Tracking Voice Activity Detection
Meaningful (34%)	Student-Teacher Talk Ratio (40%), Group Formation (30%), and Teacher Spatial Mobility (30%)	Teacher-to-student talk duration and formation for active discussion groups detected during the session	Pyannote Audio + Whisper YOLOv8
Joyful (33%)	Positive Facial Affect (35%), Acoustic Engagement (35%), and Collaborative Activity (30%).	Students showed happy/enthusiastic expressions; moments of laughter, applause, or active movement were detected times per session	DeepFace Acoustic Event Detection

To quantify overall learning quality, a hierarchical scoring matrix was developed on a 0–100 scale, with global dimension weights assigned based on pedagogical consensus in Table 1. Mindful was calculated from Gaze Attention (40%), Skeletal Posture (35%), and Productive Silence (25%); Meaningful was derived from Student-Teacher Talk Ratio (40%), Group Formation (30%), and Teacher Spatial Mobility (30%); Joyful was measured via Positive Facial Affect (35%), Acoustic Engagement (35%), and Collaborative Activity (30%). Behavioral thresholds (such as the target 60:40 teacher-student talk ratio) were aligned with active learning literature benchmarks. Each high-level indicator was mapped to an underlying algorithmic pipeline; for instance, the "Collaborative Group Activity" construct was computationally computed via Density-Based Spatial Clustering of Applications with Noise (DBSCAN) applied to coordinate bounding boxes generated by YOLOv8.

Phase 3: Development

The development phase encompassed two primary undertakings: (a) software engineering of the EduClassify platform and (b) instrument content validation by subject matter experts. EduClassify was developed with a decoupled architecture utilizing FastAPI for high-performance backend AI inference and React.js for the client-facing dashboard. The backend integrated a multimodal processing pipeline combining visual models (YOLOv8 for object and spatial formation detection, MediaPipe Pose for skeletal tracking, DeepFace for facial affect classification, and MediaPipe Face Mesh for iris/gaze tracking) and acoustic subsystems (Pyannote Audio for speaker diarization and Whisper for automated speech-to-text). Instrument content validation was conducted by a panel of three expert evaluators specializing in Artificial Intelligence/Informatics and Educational Technology. The team iteratively incorporated feedback across three engineering cycles to align computational metrics with the theoretical constructs of the 3M framework under the Kurikulum Merdeka.

Phase 4: Implementation

The field implementation evaluated a video corpus of 28 authentic classroom sessions (totaling ±18.6 hours of footage; 35–40 minutes per session) across upper-grade elementary classrooms (Grades IV–VI) in three elementary schools. Prior to data collection, institutional ethics approvals and parental informed consent were secured. Recorded sessions were processed

through a dedicated GPU-accelerated computing environment (NVIDIA RTX) to extract continuous time-series metrics, including skeletal posture vectors, gaze directions, affective states, and teacher-student vocal turn-taking ratios. Extracted streams were temporally aligned and aggregated into composite index scores (0–100) representing the Mindful, Meaningful, and Joyful dimensions.

Phase 5: Evaluation

The evaluation phase assessed the instrument's technical and empirical validity. First, internal consistency was examined using Cronbach's Alpha calculated independently across each Deep Learning dimension against an acceptable benchmark threshold of $\alpha \geq 0.70$. Second, bivariate Pearson correlation analyses (tested at $p < 0.05$ and $p < 0.001$) were computed across behavioral metrics and composite scores to evaluate internal validity and identify the primary predictive indicators of overall deep learning quality.

RESULTS AND DISCUSSION

Phase 1: Analysis

The empirical field needs assessment confirmed that elementary schools currently lack objective and scalable instruments to evaluate the 3M learning standards under the Kurikulum Merdeka. Field findings revealed that conventional supervisory audits—which rely heavily on manual, paper-based checklists—remain highly vulnerable to evaluator subjectivity, reporting delays, and the Hawthorne effect, where classroom dynamics shift because of an observer's presence. These findings established key functional requirements for the proposed system: an unobtrusive recording mechanism, automated feature extraction, and an interpretable visual dashboard for longitudinal tracking. This automated approach ensures that the captured instructional data are both objective and reproducible across different classroom sessions. The resulting system functional specifications and user interactions are modeled in the EduClassify Use Case Diagram shown in Figure 1.

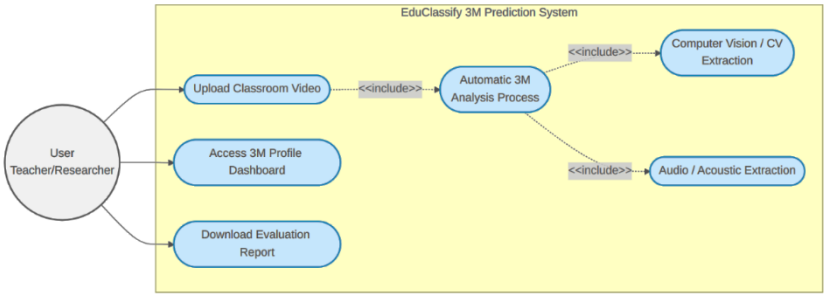


Figure 1. User Interaction Diagram (Use Case Diagram) for the EduClassify system

Phase 2: Design

The EduClassify system was designed using an integrated multimodal architecture that executes visual and acoustic inference in parallel. Within the visual subsystem, YOLOv8 (Jocher et al., 2022) is deployed for person detection and DBSCAN-based spatial cluster tracking; MediaPipe Pose (Lugaresi et al., 2019) tracks 33 skeletal landmarks to evaluate bodily posture and attentional gaze proxies; and DeepFace (Serengil & Ozpinar, 2020) classifies facial affective states. Concurrently, the acoustic subsystem integrates Pyannote Audio (Bredin et al., 2020) and

Whisper (Radford et al., 2023) to perform voice activity detection (VAD) and speaker diarization, accurately quantifying teacher-to-student talk time ratios. Heterogeneous data streams are temporally aligned using a 5-second sliding window before aggregating composite indicator metrics into an SQLite database. The platform architecture couples a FastAPI backend for computational inference with a React.js client interface for interactive reporting. Table 1 details the operational mapping matrix linking 3M constructs to computational models and Figure 2 illustrates the complete end-to-end data pipeline.

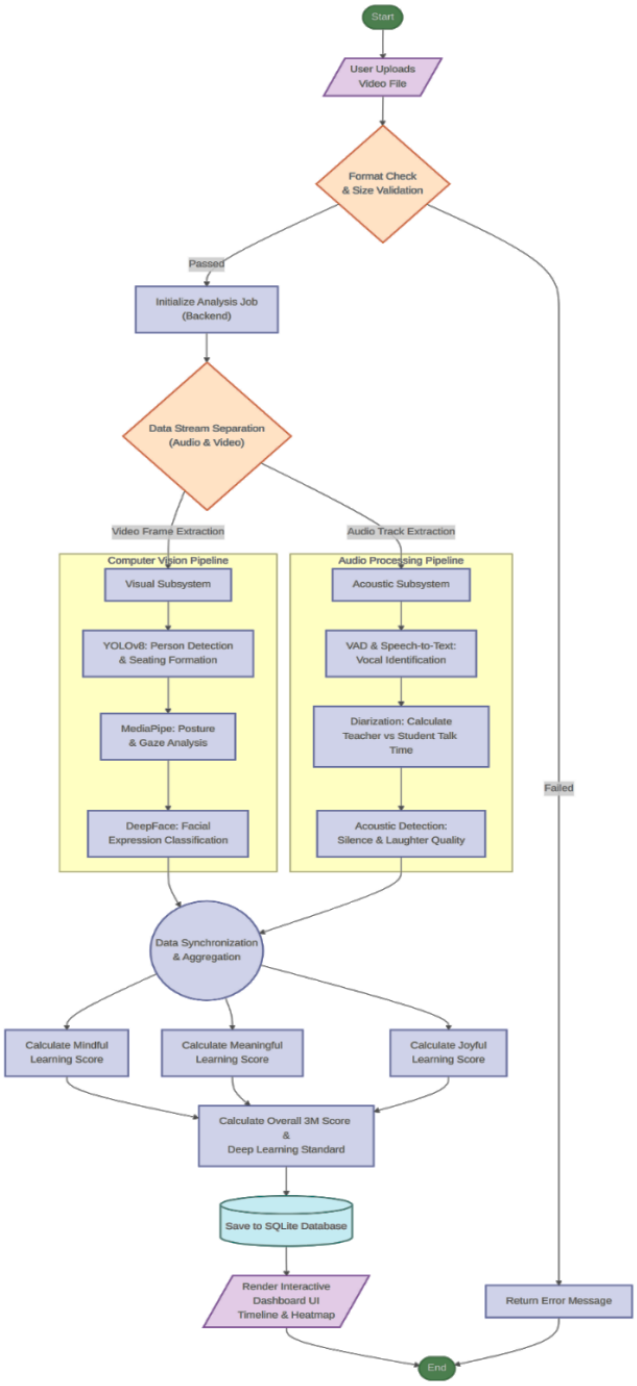


Figure 2. AI Prediction Workflow (System Pipeline Flowchart) for EduClassify

Phase 3: Development

The development of the EduClassify instrument and platform followed an iterative process through three engineering and expert validation cycles. Content and construct validation involved a multidisciplinary panel of three subject matter experts: two specialists in Informatics Engineering (focusing on computer vision pipelines and system integration) and one specialist in Educational Technology. The informatics experts evaluated the computational feasibility of the parallel inference architecture, model latency, and multimodal synchronization accuracy. Concurrently, the educational technology expert assessed the theoretical alignment between the automated behavioral indicators and the 3M learning constructs within the Kurikulum Merdeka framework. Iterative revisions confirmed the instrument's structural and content validity for field deployment. The final software artifact culminated in an interactive web-based dashboard that delivers comprehensive diagnostic visualizations, including a composite 3M radar chart, talk-time vocal distribution rings, and granular sub-indicator progress bars, as illustrated in Figure 3.

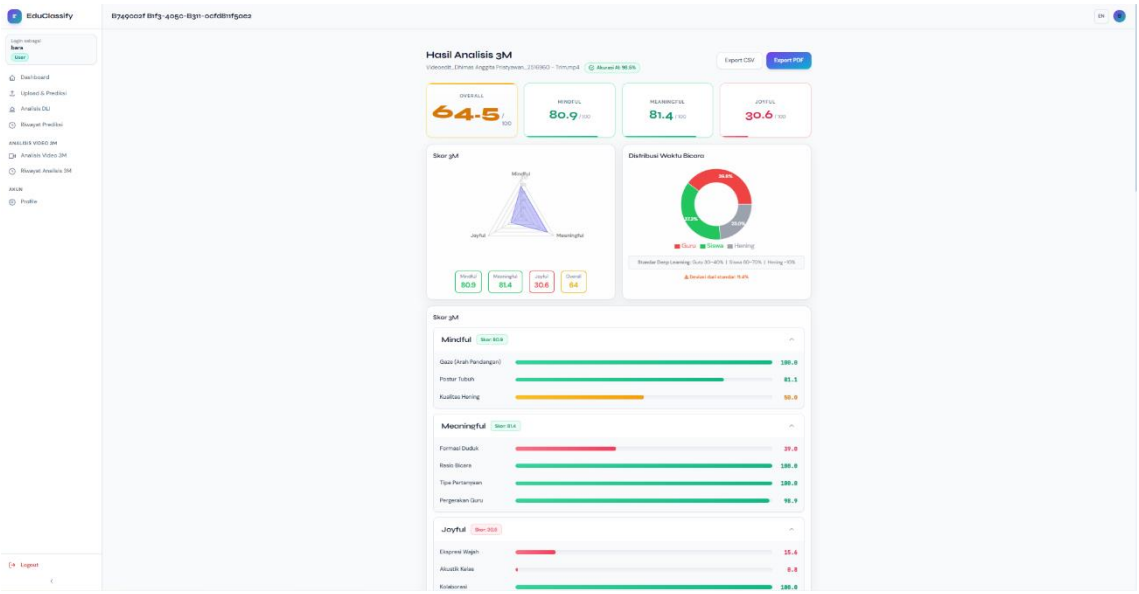


Figure 3. Dashboard View of the 3M Learning Analysis Results in the EduClassify System

Phase 4: Implementation

Field implementation was conducted by evaluating video recordings of 28 authentic classroom sessions covering thematic subjects across grades IV–VI in three public elementary schools in Semarang. All recorded sessions were processed through the automated inference pipeline of EduClassify to extract behavioral indicators, generating the quantitative 3M learning metric profile summarized in Table 2.

Based on the predefined system benchmarks, 66.7% of the observed sessions (n = 19) met the 'Fairly Good' threshold (composite score ≥ 50), while 33.3% (n = 9) fell into the 'Needs Improvement' category (score < 50). None of the evaluated classrooms achieved the ideal benchmark for Deep Learning (score ≥ 70). Across the individual dimensions, Meaningful Learning had the highest overall mean score (71.41) but also the widest dispersion (standard deviation = 24.16). This high variance highlights substantial instructional disparity among observed classrooms: some classes demonstrated highly active, student-centered engagement

(peaking at 93.39), while others remained predominantly teacher-centered, yielding a low score of 30.00.

Table 2. Descriptive Statistics for 3M Scores from 28 Observation Sessions

Dimension	Mean	Std. Dev.	Min.	Max.	Category
Mindful Learning	53.94	10.50	32.50	72.10	Fairly Good
Meaningful Learning	71.41	24.16	30.00	93.39	Fairly Good
Joyful Learning*	28.95	8.74	8.75	40.09	Need Improvement
Overall Deep Learning	**51.43**	**10.05**	**38.33**	**67.79**	**Fairly Good**

The acoustic diarization subsystem revealed that teachers accounted for an average of 53.46% of total classroom session duration, whereas student vocal participation comprised only 25.93%, establishing a teacher-to-student talk time ratio of 2.06:1. The remaining 20.61% of the instructional duration was classified as productive silence or instructional transition intervals. This imbalance in vocal distribution confirms that direct, teacher-led lecture patterns continue to dominate primary classroom environments. Concurrently, the visual processing pipeline operated robustly in real-world conditions, as demonstrated in Figure 4, which illustrates the simultaneous localization of students via YOLOv8 bounding boxes and skeletal keypoint tracking using MediaPipe Pose.



Figure 4. Evidence of Computer Vision-Based Visual Detection

A comparative evaluation between the 'Needs Improvement' (n = 9) and 'Fairly Good' (n = 19) cohorts is summarized in Table 3 and illustrated in Figure 5. The most pronounced divergence emerged in student vocal participation: classrooms in the Fairly Good category achieved an average student talk time of 35.2%, compared with just 7.4% in the Needs Improvement cohort, a substantial margin of 27.8 percentage points ($p < 0.001$). In low-

performing sessions, students spoke for approximately 3 minutes across a standard 40-minute period, evidencing a passive transmission model of instruction. This deficit corresponded directly with a 39.9-point disparity in Meaningful Learning scores, identifying student vocal activation as the primary leverage point for overall 3M optimization.

Table 3. Comparison of Learning Profiles Between Two Groups

Matrix	Need Improvement (<50)	Fairly Good (≥ 50)	Difference (Δ)	Interpretation
Meaningful Score	44.8	84.7	**+39.9**	Strongly significant
% Students Talk Time	7.4%	35.2%	**+27.8%**	Strongly significant
% Teacher Talk Time	77.6%	41.4%	**−36.1%**	Teacher is more interactive
Collaboration Score	42.3	62.0	**+19.7**	Significant

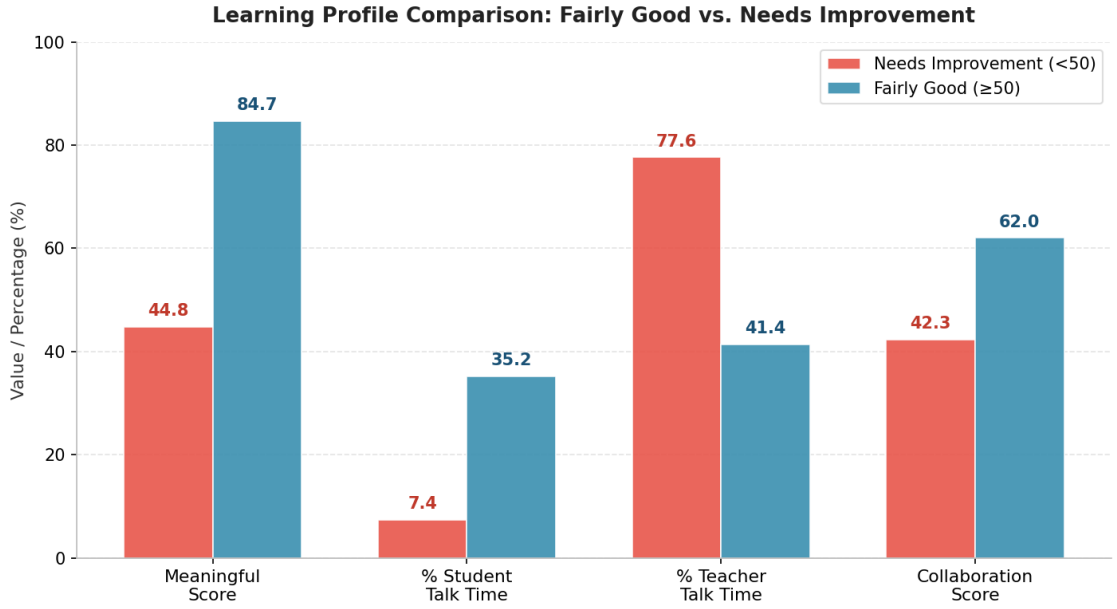


Figure 5. Bar Chart Comparing Learning Profiles Between the Two Groups

Bivariate Pearson correlation analysis corroborated these observations with high statistical significance: the association between student talk time and the Meaningful Learning dimension reached $r = 0.910$ ($p < 0.001$), confirming that collaborative verbal exchange underpins meaningful knowledge acquisition. Furthermore, Meaningful Learning served as the strongest predictor of overall Deep Learning composite scores ($r = 0.847$, $p < 0.001$), followed by the Mindful dimension ($r = 0.794$, $p < 0.001$), empirically delineating the internal hierarchy of 3M constructs. This finding provides a distinct contribution to educational computer vision literature: whereas existing automated engagement systems typically treat multimodal metrics uniformly (Pillai, 2022; Xu et al., 2023), our findings demonstrate an asymmetric structural weighting centered on student agency. Aligned with Vygotsky's (1978) Social Constructivist Theory, verbal dialogue functions as the foundational mechanism of internal cognitive formulation.

Consequently, instructional development should prioritize expanding student-led discursive spaces over traditional teacher-monologue formats.

Phase 5: Evaluation

The reliability assessment was conducted by computing Cronbach's Alpha independently for each dimension to evaluate the internal consistency of the multimodal constructs. The empirical findings confirmed that the Mindful ($\alpha = 0.722$) and Meaningful ($\alpha = 0.748$) dimensions satisfied the Acceptable threshold (George & Mallery, 2020), demonstrating high stability in automated feature extraction for postural tracking and acoustic diarization. Conversely, the Joyful dimension yielded an unacceptable reliability coefficient ($\alpha = 0.031$).

From an information systems and sensor processing perspective, this failure is attributable to a zero-variance anomaly induced by camera occlusion. Deploying a single camera at the rear of the classroom prevented DeepFace (which requires frontal facial landmarks) and the MediaPipe Face Mesh iris tracker (points 468–477) from capturing student faces. Consequently, the inference engine repeatedly outputted static neutral default scores, suppressing the statistical variance across sessions. While the Mindful dimension preserved adequate internal consistency due to the orientation-agnostic robustness of MediaPipe Pose keypoints, the Joyful construct suffered from this loss of facial affective variance.

This diagnostic finding provides a concrete engineering blueprint for sensory setup optimization: future iterations must adopt a dual-stream camera topology (a rear wide-angle sensor for spatial tracking and frontal/lateral sensors for facial affect and gaze vector estimation). Nevertheless, the resilience of the Meaningful dimension confirms that EduClassify delivers actionable, objective diagnostic data for instructional evaluation even under single-sensor constraints. The failure to capture valid facial landmarks from the rear and lateral margins of the room highlights the inherent geometric challenges of single wide-angle optical sensors in educational spaces. Deploying wide-angle lenses in classroom monitoring introduces significant perspective distortions and coordinate skewness, which hinder facial feature extraction unless paired with specialized perspective rectification (Lai et al., 2022; Ngo et al., 2020).

Discussion

The theoretical and pedagogical contributions of this study center on operationalizing indigenous educational philosophy into objective, computational metrics, transforming how classroom dynamics are evaluated in primary education. Historically, educational assessment platforms and computer vision applications have treated student engagement as a monolithic, surface-level construct, often restricting measurement to passive physical indicators or controlled e-learning behaviors (Alruwais & Zakariah, 2023; Gupta et al., 2023; Pillai, 2022). In contrast, this research mathematically grounds multimodal sensing in Ki Hadjar Dewantara's 3M pedagogical paradigm (Mindful, Meaningful, Joyful), bridging the long-standing disconnect between high-level curriculum mandates under Kurikulum Merdeka and verifiable classroom behaviors. By demonstrating that the Meaningful dimension serves as the primary empirical engine of overall instructional quality, our findings computationally validate Vygotskian social constructivist theory, confirming that cognitive internalization is fundamentally driven by verbal dialogue and student agency rather than unilateral instructional delivery (Suyanto, 2025; Vygotsky, 1978).

This shift from passive behavioral surveillance to discursive interaction tracking addresses critical limitations in conventional educational supervision. In typical elementary school settings,

instructional audits conducted through manual observation sheets are inherently vulnerable to observer subjectivity, administrative latency, and acute Hawthorne distortion, which frequently induces defensive instructional behaviors rather than authentic pedagogical growth (Parsons, 1974; Trabelsi et al., 2023). Implementing automated acoustic and visual analytics repositions supervision as a clinical, evidence-based dialogue. Rather than serving as a punitive evaluation mechanism, continuous metric streams provide teachers with an objective mirror for self-directed micro-reflection, enabling them to diagnose instructional transition points, measure the democratization of classroom talk time, and intentionally design collaborative learning spaces (Zhang et al., 2024).

The pronounced impact of student vocal participation on Meaningful Learning aligns with empirical demonstrations that authentic deep learning flourishes exclusively when instructional agency shifts toward the learners (Tal & Tsaushu, 2018). Furthermore, integrating automated multimodal sensing preserves the humanistic core of education against algorithmic reductionism, ensuring that technological evaluation scaffolds authentic intellectual and emotional engagement rather than mechanical compliance (Bentz, 1992; Peters, 2018). By providing granular, continuous diagnostic indicators, EduClassify repositions instructional supervision as an empirical, data-driven decision-making process, empowering educators and administrators to optimize instructional strategies and management efficiency based on verifiable classroom evidence (Jin et al., 2025). Furthermore, by demonstrating functional feasibility within dense, naturalistic Indonesian classrooms, this research establishes ecological validity in Global South learning environments that are often neglected by high-resource, e-learning-centric surveillance platforms developed in Western contexts (Szeliski, 2022).

Methodologically, the empirical breakdown observed in the Joyful dimension provides a vital engineering insight for educational information systems: multimodal classroom tracking cannot succeed under a homogeneous, single-camera sensor setup. The loss of statistical variance from rear-view occlusions underscores the need to move toward synchronized multi-sensor topologies, pairing lateral or frontal optical sensors with wide-angle spatial cameras to capture both collaborative physical clustering and facial affect without intrusive wearable hardware. Moving forward, the integration of automated 3M assessment instruments into longitudinal teacher-in-service programs presents a compelling avenue to examine how iterative, data-driven reflective cycles correlate with measurable gains in elementary students' higher-order thinking skills and conceptual understanding over time.

To mitigate spatial and facial occlusions in future iterations, the physical deployment of multi-camera arrays must be accompanied by computational perspective and distortion corrections. Incorporating homography mapping and planar perspective rectification allows the visual pipeline to project oblique camera views into standardized bird's-eye coordinates for spatial clustering, while dedicated wide-angle correction algorithms prevent metric degradation in facial affect models (Lai et al., 2022; Ngo et al., 2020; Wang et al., 2025).

CONCLUSION

This study successfully engineered and validated EduClassify, an automated multimodal instrument that operationalizes Ki Hadjar Dewantara's 3M pedagogical philosophy (Mindful, Meaningful, Joyful) into objective, machine-extractable behavioral metrics under the Kurikulum Merdeka. By synthesizing computer vision and acoustic diarization pipelines, the system demonstrates the practical feasibility of conducting non-intrusive, scalable, and continuous instructional diagnostics in naturalistic elementary classrooms. The empirical evaluation across

authentic classroom sessions confirms that active student vocal participation functions as the primary leverage point driving instructional depth, validating social constructivist tenets that authentic learning requires discursive student agency rather than passive visual compliance. Furthermore, while the postural and acoustic subsystems proved statistically robust and reliable, the empirical vulnerability observed in facial affect tracking highlights a critical architectural lesson regarding the limitations of single-sensor placements in dynamic classroom spaces. Consequently, EduClassify establishes a viable foundation for evidence-based clinical supervision and self-directed teacher reflection, bridging national curriculum mandates with data-driven classroom practices. Future investigations should deploy synchronized multi-camera topologies to resolve visual occlusion constraints, expand the multi-site evaluation corpus across broader geographical cohorts in Indonesia, and investigate the longitudinal predictive validity of machine-extracted 3M metric profiles on elementary students' higher-order thinking skills and academic outcomes.

ACKNOWLEDGMENTS

This research was funded and supported by Universitas Pendidikan Indonesia (UPI) under the institutional research grant scheme of Lecturer Affirmation Research grant with Contract No. 332/UN40.F7/PT.01.01/2026. The authors express sincere gratitude to the collaborator team from the Faculty of Engineering and Computer Science (FASTIKOM), Universitas Sains Al-Qur'an (UNSIQ) Jawa Tengah, for their valuable technical guidance and system engineering support throughout the development of the EduClassify platform. The authors also extend their deepest appreciation to the collaborative research team from Universitas Pendidikan Indonesia for their substantial pedagogical insights into 3M learning frameworks and their dedicated support in instrument validation, field implementation, and the facilitation of this publication.

REFERENCES

Adeoye, M. A., Wirawan, K. A. S. I., Pradnyani, M. S. S., & Septiarini, N. I. (2024). Revolutionizing education: Unleashing the power of the ADDIE model for effective teaching and learning. *Jurnal Pendidikan Indonesia, 13*(1), 202–209. https://doi.org/10.23887/jpiundiksha.v13i1.68624

Alruwais, N., & Zakariah, M. (2023). Student-engagement detection in classroom using machine learning algorithm. *Electronics, 12*(3), 731. https://doi.org/10.3390/electronics12030731

Bentz, V. M. (1992). Deep learning groups: Combining emotional and intellectual learning. *Clinical Sociology Review, 10*(1), 71–89. https://digitalcommons.wayne.edu/csr/vol10/iss1/9/

Bredin, H., Yin, R., Coria, J. M., Broux, G., Hadjimarkou, P. A., Bennaceur, M., ... & Gill, M. P. (2020). Pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7124-7128). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9052974

George, D., & Mallery, P. (2020). *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference (16th Ed.).* Routledge. https://doi.org/10.4324/9780429056765

Gupta, S., Kumar, P., & Tekchandani, R. (2023). A multimodal facial cues based engagement detection system in e-learning context using deep learning approach. *Multimedia Tools and Applications, 82*, 28589–28615. https://doi.org/10.1007/s11042-023-14392-3

Jin, R., Peng, Y., Wang, Z., Wang, J., Tang, J., & Zhang, M. (2025). Data-driven educational decision-making: How to enhance educational quality and management efficiency. *Journal of Higher Education Research, 5*(6), 550-553. https://doi.org/10.32629/jher.v5i6.3385

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Xie, T., Fang, J., Yifu, Z., … & Jain, M. (2022). Ultralytics YOLO: YOLOv5 SOTA Realtime Instance Segmentation (Version v7.0) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.7347926

Kovač, V. B., Nome, D. Ø., Jensen, A. R., & Skreland, L. L. (2025). The why, what and how of deep learning: Critical analysis and additional concerns. *Education Inquiry, 16*(2), 237–253. https://doi.org/10.1080/20004508.2023.2194502

Lai, W-S., Shih, Y-C., Liang, C-K., & Yang, M-H. (2022). Correcting face distortion in wide-angle videos. *IEEE Transactions on Image Processing, 31*, 366-378. https://doi.org/10.1109/TIP.2021.3131047

Lin, L., Zhou, D., Wang, J., & Wang, Y. (2024). A Systematic Review of Big Data Driven Education Evaluation. *Sage Open*, *14*(2). https://doi.org/10.1177/21582440241242180

Liu, Q., Jiang, X., & Jiang, R. (2025). Classroom behavior recognition using computer vision: A systematic review. *Sensors, 25*(2), 373. https://doi.org/10.3390/s25020373

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C-L., Yong, M. G., …. & Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. *arXiv, 1906,* 08172. https://doi.org/10.48550/arXiv.1906.08172

Ngo, H.-H., Lin, F.-C., Sehn, Y.-T., Tu, M., & Dow, C.-R. (2020). A Room Monitoring System Using Deep Learning and Perspective Correction Techniques. Applied Sciences, 10(13), 4423. https://doi.org/10.3390/app10134423

Nursehag, D. A., & Amalia, L. L. (2024). An investigation of teacher talk and student talk in Indonesian EFL classroom interaction. *Paedagogia: Jurnal Penelitian Pendidikan, 27*(2), 298–308. https://doi.org/10.20961/paedagogia.v27i1.84086

Parsons, H. M. (1974). What happened at Hawthorne? *Science, 183*(4128), 922–932. https://doi.org/10.1126/science.183.4128.922

Peters, M. A. (2018). Deep learning, education and the final stage of automation. *Educational Philosophy and Theory, 50*(6–7), 549–553. https://doi.org/10.1080/00131857.2017.1355799

Pillai, A. S. (2022). Student engagement detection in classrooms through computer vision and deep learning: A novel approach using YOLOv4. *Sage Science Review of Educational Technology, 5*(1), 87–97. https://doi.org/10.5281/zenodo.15061626

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of Machine Learning Research, 202,* 28492-28518. https://doi.org/10.5555/3618408.3619590

Serengil, S. I., & Ozpinar, A. (2020). LightFace: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 23–27). IEEE. https://doi.org/10.1109/ASYU50717.2020.9259802

Spatioti, A. G., Kazanidis, I., & Pange, J. (2022). A comparative study of the ADDIE instructional design model in distance education. *Information, 13*(9), 402. https://doi.org/10.3390/info13090402

Suyanto, S. (2025). *Naskah Akademik Pembelajaran Mendalam Menuju Pendidikan Bermutu untuk Semua*. Pusat Kurikulum dan Pembelajaran Badan Standar, Kurikulum, dan Asesmen Pendidikan Kementerian Pendidikan Dasar dan Menengah Republik Indonesia. https://repositori.kemendikdasmen.go.id/39385/

Szeliski, R. (2022). *Computer Vision: Algorithms and Applications (2nd ed.).* Springer. https://doi.org/10.1007/978-3-030-34372-9

Tal, T., & Tsaushu, M. (2018). Student-centered introductory biology course: evidence for deep learning. *Journal of Biological Education, 52*(4), 376–390. https://doi.org/10.1080/00219266.2017.1385508

Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. *Big Data and Cognitive Computing, 7*(1), 48. https://doi.org/10.3390/bdcc7010048

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Wang, C., Ding, Y., Cui, K., Li, J., Xu, Q., & Mei, J. (2025). A perspective distortion correction method for planar imaging based on homography mapping. *Sensors (Basel), 25*(6), 1891. https://doi.org/10.3390/s25061891

Xu, Q., Wei, Y., Gao, J., Yao, H., & Liu, Q. (2023). ICAPD framework and simAM-YOLOv8n for student cognitive engagement detection in classroom. *IEEE Access, 11,* 136063–136076. https://doi.org/10.1109/ACCESS.2023.3338721

Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications, 237*, 121692. https://doi.org/10.1016/j.eswa.2023.121692