e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

# A Systematic Review Of Educational Facial Emotion Recognition: Datasets, Methods, Modality, And Potential Transfer To Vocational Teaching Contexts

**Pipit Utami [1)]; Masduki Zakarijah [2)]; Satrio Wiroyudho Pratomo [3)]; Mozan Osman Gebalr [4)]; Dwi Indriyani [5)]; Bismaka Sahasika [6)]; Hanif Nurkhalis [7)]; Tomy Herlambang [8)]**

[1,2,3,4,5,6,7,8)] ***Electronics Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta***

**Email:** [1)]pipitutami@uny.ac.id

***How to Cite :***

Utami, P., Zakaria, M., Pratomo, S.W., Gebalr, M.O., Indriyani, D., Sahasika, B., Nurkhalis, H., Herlambang, T. (2025). A Systematic Review of Educational Facial Emotion Recognition: Datasets, Methods, Modality, and Potential Transfer to Vocational Teaching Contexts. Jurnal Media Computer Science, 4(2)

**ARTICLE HISTORY**

*Received [25 Juni 2025]*
*Revised [28 Juli 2025]*
Accepted [31 Juli 2025]

***KEYWORDS***

*Facial Emotion Recognition (FER), Engagement Detection, Deep Learning, Affective Computing, Vocational Education, Affective EdTech.*



## ABSTRAK

Pengenalan ekspresi wajah (Facial Emotion Recognition, FER) berkembang sebagai komponen penting dalam teknologi pendidikan, namun penerapannya dalam praktik pedagogis masih belum merata, khususnya pada pembelajaran vokasi. Tinjauan sistematis ini menganalisis 38 studi empiris yang dipublikasikan antara tahun 2015 hingga 2025 melalui penelusuran berbasis kerangka PRISMA pada basis data akademik utama, termasuk Scopus. Hasil sintesis menunjukkan bahwa perkembangan FER pendidikan masih didominasi oleh penggunaan dataset benchmark dan penyempurnaan arsitektur model dalam kondisi terkontrol, sehingga membatasi keterterapan pada lingkungan pembelajaran berbasis praktik. Meskipun model ringan berbasis perhatian dan pendekatan multimodal meningkatkan kelayakan implementasi serta pemaknaan afektif, sebagian besar sistem masih bersifat open-loop dan belum mendorong adaptasi pedagogis yang berkelanjutan. Secara keseluruhan, tinjauan ini menegaskan bahwa kemajuan FER dalam pendidikan menuntut keselarasan yang lebih kuat antara praktik pengelolaan data, desain model dan modalitas, serta realitas pedagogis pendidikan vokasi.

## *ABSTRACT*

*Facial emotion recognition (FER) has emerged as a promising component of educational technology, yet its integration into pedagogical practice remains uneven, particularly in vocational learning contexts. This systematic review examines 38 empirical studies published between 2015 and 2025, identified through a PRISMA-guided search of major academic databases, including Scopus. The synthesis explores how datasets, model architectures, multimodal learning signals, and system design shape the applicability of affect analytics in authentic instructional settings. The findings indicate that progress in educational FER is largely driven by benchmark datasets and incremental architectural refinement evaluated under controlled conditions, which limits transferability to hands-on learning environments. While lightweight, attention-enhanced models and multimodal approaches improve deployment feasibility and affective interpretation, most systems remain open-loop and rarely support sustained pedagogical adaptation. Overall, the review highlights that advancing educational FER requires closer alignment between data practices, model and modality design, and the pedagogical realities of vocational education.*

## INTRODUCTION

In the evolving landscape of digital education, the ability to understand and respond to learners' emotional states has become an essential component of effective pedagogy. The integration of Facial Emotion Recognition (FER) and engagement detection systems has transformed how educators interpret participation, motivation, and attention in both virtual and physical classrooms. Emotion-aware technologies powered by deep learning enable personalized feedback, dynamic content adaptation, and improved emotional support for learners (H. S. Kim & Cho, 2024; C. Wang et al., 2023).

While conventional learning analytics focus primarily on cognitive and behavioral metrics, affective data—such as facial expressions, gaze, and posture—provide a richer reflection of learner experience. Emotions shape concentration, persistence, and performance, making their analysis vital for creating meaningful learning environments (Das et al., 2024). The emergence of FER integrated with Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformers has increased the precision of emotion detection in educational contexts, achieving accuracies exceeding 90% on standard benchmarks (Yuvaraj et al., 2025).

However, most FER datasets (e.g., FER2013, AffectNet, DAiSEE) originate from general-purpose or Western-centric settings, which limits their transferability to diverse and vocational learning environments (W. Zhao & Qiu, 2025). Emotion dynamics in hands-on learning—such as confidence, frustration, and curiosity—require contextual understanding that current datasets rarely capture. This motivates further investigation into FER applications that reflect the authentic emotional ecology of vocational education.

In Indonesian vocational schools, where learning is practical and skill-oriented, affective engagement strongly correlates with psychomotor performance and reflective behavior (Utami et al., 2019). Recognizing and responding to these emotions helps educators foster empathy, maintain motivation, and scaffold learning more effectively. Thus, FER and engagement detection can serve not only as monitoring tools but as pedagogical instruments that strengthen teacher–student relationships.

Despite rapid technical progress, research on educational FER remains fragmented. Few studies focus on integrating FER into authentic classrooms or vocational training contexts, and ethical governance of affective data remains underdeveloped. Therefore, this systematic review aims to synthesize existing literature on educational FER and engagement detection, identify methodological and theoretical patterns, and explore their transferability to vocational teaching contexts. The objectives of this study are to: (1) Datasets and Benchmarking – mapping available educational FER datasets and their applicability; (2) Algorithmic Developments and Model Performance – comparing CNN, Transformer, and ensemble architectures; (3) Multimodal Integration and Real-Time Applications – analyzing fusion strategies and deployment feasibility; (4) Transferability to Vocational Teaching Contexts – exploring adaptation strategies and pedagogical implications. This review seeks to identify current limitations, theoretical insights, and methodological progress in FER and engagement analytics, ultimately outlining a framework for transferring emotion-aware AI systems to vocational education, where emotional and cognitive engagement directly impact skill mastery and professional readiness.

## Theoretical Background

### Cognitive–Affective Foundations

Learning involves intertwined cognitive and emotional processes. The Cognitive-Affective Theory of Learning with Media (CATLM) offers a framework for understanding how emotions regulate learning performance through motivational and attentional mechanisms. According to CATLM, positive affect expands cognitive capacity and promotes sustained engagement, whereas

negative affect increases cognitive load and may hinder knowledge integration (Das et al., 2024). Emotional regulation functions as a mediator between mental effort and learning outcomes, highlighting the importance of affective states in shaping meaningful learning experiences.

Affective Computing operationalizes these theoretical assumptions into computational models that can detect and interpret emotional cues. Through deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Transformers, emotional signals can be extracted from multimodal data streams—facial, vocal, and physiological (Lampropoulos et al., 2024). When embedded into educational platforms, these models enable adaptive instruction responsive to learners' cognitive and emotional conditions.

**Facial Emotion Recognition in Education**

Facial Emotion Recognition (FER) connects psychology and artificial intelligence by converting facial cues such as micro-expressions, gaze, and muscle activation into emotional categories. Based on Ekman's taxonomy of basic emotions (happiness, sadness, anger, fear, disgust, and surprise), FER models are now being extended to identify complex academic emotions such as curiosity, anxiety, and satisfaction (Banzon et al., 2024). In education, FER supports emotion-aware feedback that helps teachers monitor engagement and respond empathetically in real time.

Despite progress in FER accuracy, most datasets including FER2013, AffectNet, and DAiSEE are collected in controlled or Western-centered environments, limiting generalizability to diverse or vocational settings (Ganepola, 2022). Context-sensitive datasets are essential for authentic affective modeling since emotional patterns vary by domain, environment, and role. Teacher emotions, for instance, often involve subtle and regulated expressions that require specialized feature representations.

**Ethical and Fairness Considerations**

As affective computing enters classrooms, concerns arise regarding privacy, bias, and data governance. Dataset imbalance can lead to biased predictions related to ethnicity, gender, or lighting conditions, reducing reliability (Cioruţa et al., 2024). Ethical FER frameworks emphasize transparency, consent, and human oversight. The new paradigm of Explainable FER (XFER) promotes interpretability and user trust by clarifying how models infer emotional states (Torres-Hernández, 2025).

Responsible AI in education requires anonymization, secure data storage, and informed consent to protect psychological safety, especially for minors (Takyi Mensah et al., 2023). These principles align FER development with educational values of empathy, inclusivity, and well-being, ensuring ethical and human-centered application rather than surveillance.

**Vocational Context and Affective Dimensions**

Vocational education is experiential and practice-oriented, combining cognitive precision with emotional resilience. Emotions such as confidence, frustration, and curiosity influence psychomotor performance and reflective learning. Emotion-aware analytics help instructors identify these affective signals to guide mentoring, assessment, and skill acquisition. Research in Indonesian vocational contexts highlights the importance of FER for analyzing teaching affectivity and learner engagement in authentic practice environments (Utami et al., 2022).

## Research Methodology

**Systematic Review Framework**

This study adopted a Systematic Literature Review (SLR) approach to synthesize trends and evidence in facial emotion and engagement recognition within educational contexts. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) framework to ensure transparency, replicability, and methodological rigor. PRISMA was selected

because it provides a structured process for identifying, screening, evaluating, and synthesizing empirical research, reducing selection bias and ensuring reproducibility (C. Wang et al., 2023). The protocol included three primary stages: (1) identification of relevant publications, (2) screening and eligibility assessment, and (3) inclusion of qualified studies. Inclusion criteria required studies to: (a) focus on FER or engagement detection in educational or learning contexts, (b) employ deep learning or computer vision methods, (c) be published between 2015–2025, and (d) report measurable outcomes such as accuracy, F1-score, or AUROC. Exclusion criteria omitted non-empirical papers, reviews without datasets, and studies outside educational applications.

**Search Strategy and Databases**

The literature search was conducted across major scientific databases including Scopus, IEEE Xplore, ScienceDirect, and SpringerLink. Boolean operators were used to refine search queries, combining keywords such as: ("facial expression recognition" OR "emotion recognition" OR "engagement detection") AND (education OR classroom OR "vocational") AND (dataset OR benchmark OR corpus) AND (CNN OR Transformer OR YOLO OR EfficientNet). The query produced 157 initial results, which were then filtered through title, abstract, and keyword screening. After full-text evaluation, 38 studies met all inclusion criteria and were retained for synthesis. This process is visualized in Figure 1, representing the systematic progression from identification to inclusion.

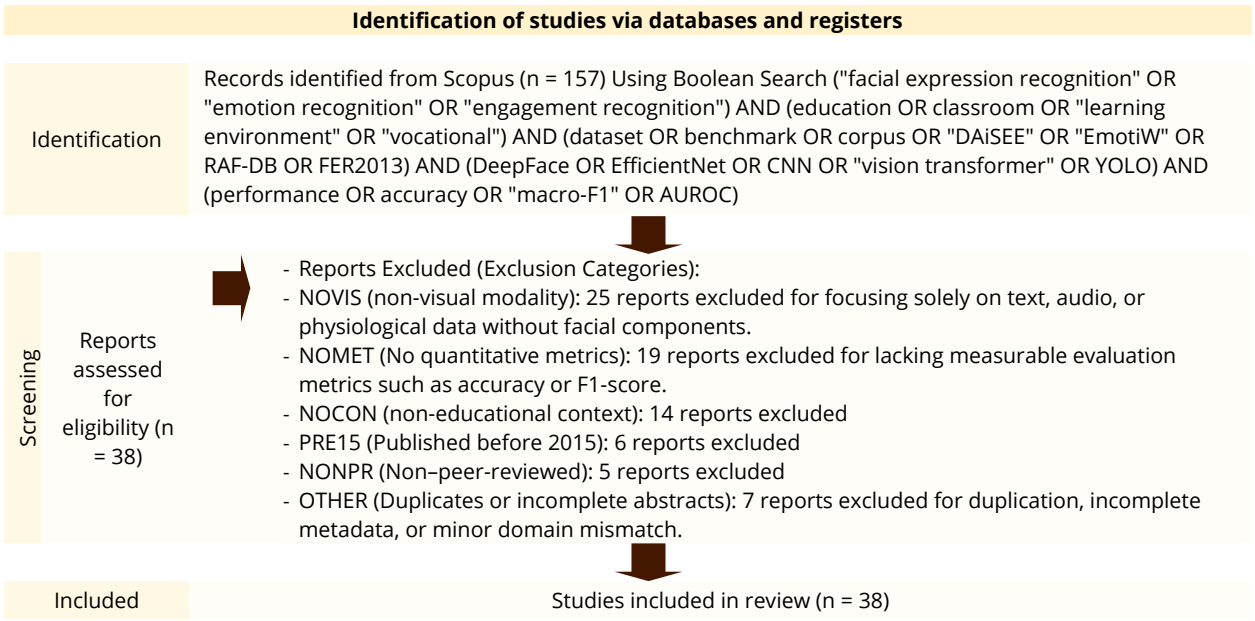



**Figure 1. PRISMA Flow Diagram**

**Quality Assessment**

To evaluate methodological rigor, the selected studies were assessed using the Mixed Methods Appraisal Tool (MMAT), which examines quality indicators across quantitative, qualitative, and mixed-methods designs (Gawande, 2025). Each study was scored on transparency of dataset use, validation procedures, and reproducibility of model evaluation. Studies with ambiguous datasets or unverified metrics were documented but weighted lower in the synthesis matrix. The MMAT results guided the identification of high-confidence findings used in the analysis sections.

**Data Extraction and Synthesis**

Key study information was extracted and organized into analytical categories aligned with the research objectives: (1) datasets and benchmarks, (2) deep learning models and algorithmic performance, (3) multimodal fusion techniques, and (4) transferability to vocational contexts. Each

study was coded for dataset name, sample size, emotion type, model architecture, and reported accuracy. The extracted data formed the basis of Tables 1–4, which summarize comparative insights across studies. A narrative synthesis approach was applied to integrate quantitative results with qualitative interpretations. Statistical patterns such as accuracy distribution and model type frequency were combined with theoretical insights regarding emotion modeling and pedagogical implications. This integrative synthesis highlights methodological tendencies, empirical strengths, and contextual research gaps.

**Research Ethics and Relevance**

All reviewed studies were evaluated for ethical considerations, particularly regarding facial data privacy, participant consent, and cultural representation. Ethical compliance was verified through dataset documentation and institutional review statements when available (Cioruţa et al., 2024). In addition, the review emphasizes alignment with educational ethics by considering emotional sensitivity and contextual respect in vocational learning settings (Utami et al., 2022). Through this structured and ethically grounded methodology, the review ensures a comprehensive, credible, and contextually relevant synthesis of how deep learning-based FER advances emotion-aware education and its potential transfer to vocational learning environments.

## RESULTS AND DISCUSSION

**Results**

The results of this review synthesize evidence from 38 empirical studies to provide a structured understanding of how affect analytics, including facial emotion recognition and engagement detection, has evolved within educational settings. Rather than reporting individual study outcomes, the findings are organized to reveal broader patterns across datasets, model choices, sensing modalities, and design orientations, with particular attention to their relevance for vocational learning contexts. This section emphasizes how methodological advances intersect with practice-oriented and context-sensitive instructional environments. By progressively moving from data landscapes to model trends, modality usage, and design implications, the results aim to highlight not only what has been achieved, but also where structural gaps persist. This layered presentation establishes a coherent foundation for interpreting transferability challenges and opportunities before engaging with the detailed analyses presented in the following subsections.

**Dataset and Contextual Landscape of Educational Affect Analytics**

The synthesis of the 38 studies included through the PRISMA-based screening process reveals a research landscape that is methodologically mature yet contextually uneven. A substantial portion of the literature relies on general-purpose benchmark datasets developed under controlled laboratory conditions. These datasets have played a decisive role in stabilizing model architectures and enabling reproducible experimentation. However, the educational contexts they represent are often abstracted from the realities of everyday teaching and learning, particularly those found in skill-oriented and practice-intensive environments. As a result, methodological robustness has advanced more rapidly than pedagogical alignment.

**Table 1. Dataset–Model Validation Summary with Transferability Aspect**

| Dataset (n) | Dominant Model | Educational Context, Vocational Relevance, and Representative Studies |
|---|---|---|
| FER2013 (9) | CNN, ResNet, EfficientNet, Ensemble CNN | Benchmark for classroom emotion analysis; easily deployable for facial expression monitoring in vocational labs (Deepa et al., 2025; Lawpanom et al., 2024; Tang et al., 2019) |
| DAiSEE (8) | CNN, , ViT, Hybrid CNN–LSTM, | Engagement and attention detection in online learning; transferable to blended and distance vocational classrooms |

| Dataset (n) | Dominant Model | Educational Context, Vocational Relevance, and Representative Studies |
|---|---|---|
| | EfficientNetB0/B7 | (Athauda et al., 2023; El Maazouzi & Retbi, 2025; Thiruthuvanathan & Krishnan, 2024) |
| AffectNet (5) | CNN–LSTM, Transformer, ViT-YOLO Hybrid | Multimodal emotion recognition enabling adaptive affect-aware support in vocational learning environments (Yang, 2025; N. Zhang & Leong, 2025) |
| KDEF (3) | CNN, ResNet, EfficientNet-B7 | Controlled expression recognition; suitable for vocational training spaces with variable lighting (Lasri et al., 2023; Liu et al., 2025a) |
| CK+ (3) | CNN, Conv3D–ConvLSTM–SEnet, Hybrid Attention | Spatio-temporal expression analysis supporting gesture and micro-expression assessment in skill-based training (Fu & Tian, 2024; Tian et al., 2021) |
| JAFFE (3) | VGG-16/19 Transfer Learning | Cross-cultural emotion recognition supporting inclusive and multicultural vocational classrooms ((Lasri et al., 2023; Rathod et al., 2025) |
| RAVDESS (2) | CNN–LSTM, Transformer | Audio-visual emotion and frustration detection for multimodal teacher–student interaction analysis (Thomas et al., 2025; Tokhtarov et al., 2025) |
| VGGFace (1) | YOLOv9 + DeepFace Fusion | Real-time behavior and emotion monitoring in digital vocational classrooms (H. Zhang et al., 2025) |
| FESR (1) | Multi-scale CNN + Wavelet Attention | Authentic classroom emotion dataset directly applicable to vocational learning analytics (Liu et al., 2025b) |
| EIDB-13 (1) | InceptionResNetV2 + CBAM Attention | Teacher expression intensity analysis for objective evaluation of vocational instruction (Zheng et al., 2020) |
| MELD (1) | Federated 3D-CNN + LSTM | Emotion–sentiment fusion supporting affective language and communication training (Orosoo et al., 2024) |
| OLSFED (1) | APViT (Vision Transformer) | Emotional feedback system extendable to remote practicum and virtual mentoring (K. Wang & Cheng, 2024) |
| RAF-DB (1) | InceptionResNetV2 Transfer Learning | Macro-expression benchmark informing educator affect modeling in vocational contexts (Zheng et al., 2020) |
| Emotic (1) | Dual-Stream 3D-CNN | Scene-aware emotion recognition for simulation-based and AR/VR vocational training (W. Zhao & Qiu, 2025) |
| SCB Behavior + Custom (1) | YOLOv9 + DeepFace Fusion | Joint facial–behavioral monitoring for engagement tracking in hands-on vocational classes (H. Zhang et al., 2025) |
| Other / In-house (4) | Hybrid CNN, ViT, Fed 3D-CNN, RViT | Locally developed datasets enabling replicable affective analytics in vocational institutions (Chan et al., 2023; Evangeline & Parkavi, 2024; Gupta et al., 2023) |

As summarized in Table 1, the distribution of datasets shows a strong reliance on widely used benchmarks, while only a limited number of datasets explicitly capture authentic classroom or vocational conditions. Although several datasets support emotion or engagement analysis in online or generic educational scenarios, comparatively few reflect hands-on training, procedural activities, or workshop-based instruction. This imbalance highlights a persistent domain gap between publicly available datasets and genuine instructional contexts. In vocational education, where affective expressions are closely intertwined with physical action, tool use, and collaborative problem solving, such abstraction constrains the direct transferability of research outcomes.

e-ISSN 2828-0490
JURNAL
MEDIA COMPUTER SCIENCE

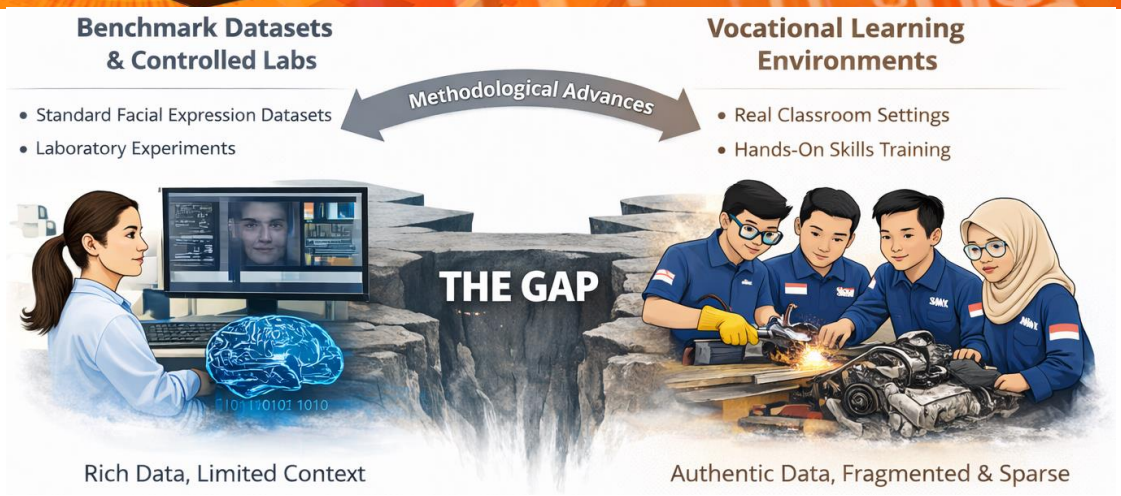



**Figure 2. Conceptual Distribution Diagram**

This structural gap is further illustrated in Figure 2, which conceptually maps how methodological advances have clustered around benchmark-centered data ecosystems, while authentic vocational datasets remain sparse and fragmented. The diagram does not present statistical proportions, but rather emphasizes the contrast between data-rich yet context-light benchmarks and data-poor but pedagogically rich vocational environments. At the same time, recent studies indicate a gradual methodological shift toward multimodal and context-aware datasets, reflecting growing concern for ecological validity and pedagogical realism. Taken together, Table 1 and Figure 1 suggest that advancing affect analytics for vocational education requires prioritizing contextual authenticity, pedagogical alignment, and selective multimodal integration to better capture learners' affective and cognitive dynamics within real instructional settings.

**Model Trends and Deployment-Oriented Performance**

Across the reviewed studies, the development of model architectures reflects a growing effort to balance performance with practical considerations of educational deployment, especially in vocational learning environments. Rather than being driven solely by technical optimization, architectural choices increasingly respond to demands for robustness, computational efficiency, and sensitivity to instructional context.

**Table 2. Model Architectures and Vocational Transferability**

| Model / Architecture | Accuracy Range | Key Characteristics, Vocational Transferability, and Evidence |
|---|---|---|
| CNN (Baseline) | ~70%–85% | Lightweight FER baseline for low-resource vocational labs (T. S. & Guddeti, 2020; Tang et al., 2019) |
| ResNet | ~82%–92% | Robust to pose and illumination variations in classroom settings (Lawpanom et al., 2024; Singh et al., 2022) |
| EfficientNet | ~85%–95% | High accuracy–efficiency trade-off for edge-based vocational monitoring (Athauda et al., 2023; Deepa et al., 2025) |
| CNN–LSTM (Hybrid) | ~80%–93% | Spatial–temporal modeling for engagement and fatigue detection (Athauda et al., 2023; Tokhtarov et al., 2025) |
| Conv3D / ConvLSTM | ~78%–90% | Motion-aware modeling for gesture-based vocational assessment (Fu & Tian, 2024; Tian et al., 2021) |
| Vision Transformer (ViT) | ~88%–96% | Global attention modeling for complex and multimodal learning data (K. Wang & Cheng, 2024; N. Zhang & Leong, 2025) |
| Transformer-based Multimodal | ~90%–97% | Attention-based fusion for affect-aware vocational learning systems (Yang, 2025; W. Zhao & Qiu, 2025) |

| Model / Architecture | Accuracy Range | Key Characteristics, Vocational Transferability, and Evidence |
|---|---|---|
| YOLO + DeepFace Fusion | ~85%–93% | Real-time facial behavior monitoring in hands-on vocational classes (H. Zhang et al., 2025) |
| Wavelet Attention CNN | ~88%–95% | Noise-robust emotion recognition in authentic classroom environments (Liu et al., 2025b) |
| CBAM-Attention CNN | ~83%–91% | Enhanced facial region focus for instructor affect analysis (Zheng et al., 2020) |
| Federated 3D-CNN + LSTM | ~80%–90% | Privacy-preserving learning for institution-level deployment (Orosoo et al., 2024) |
| Dual-Stream 3D-CNN | ~86%–94% | Scene-aware emotion modeling for simulation-based training ((W. Zhao & Qiu, 2025) |
| Hybrid / In-house Models | ~75%–90% | Customizable architectures for local vocational datasets (Gupta et al., 2023) |

Table 2 consolidates the principal categories of deep learning architectures employed in the reviewed studies and summarizes their typical accuracy ranges alongside key characteristics relevant to vocational learning contexts. A clear pattern emerges from this synthesis. Conventional Convolutional Neural Networks continue to serve as a foundational baseline due to their lightweight structure and reliable performance under constrained computational conditions. However, subsequent architectural developments, including residual networks, efficiency-oriented scaling, and attention-enhanced variants, demonstrate that performance improvements are increasingly achieved through design refinement rather than sheer model complexity. This progression suggests that accuracy gains are closely coupled with considerations of deployability, particularly in learning environments where real-time processing and limited infrastructure are common.

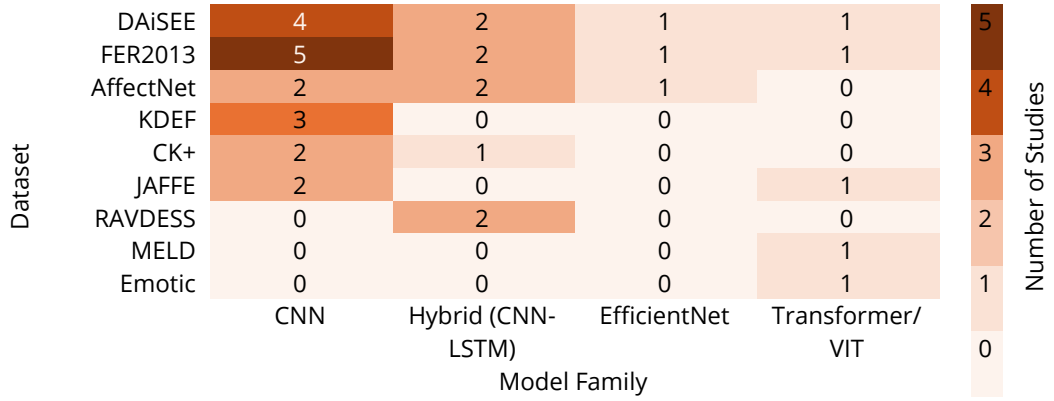



**Figure 3. Dataset–Model Co-Occurrence Heatmap**

The distribution of model architectures across datasets is further illustrated in Figure 3 (Dataset–Model Co-Occurrence Heatmap). This figure provides an indicative overview of recurring pairings between widely used datasets and dominant model architectures. CNN-based and hybrid CNN–LSTM approaches appear most frequently in conjunction with benchmark datasets such as FER2013 and DAiSEE, reflecting their stability and ease of integration. In contrast, transformer-based and multimodal architectures are observed less frequently and tend to be associated with more recent studies, signaling a gradual methodological shift rather than an abrupt replacement of existing approaches. Importantly, the heatmap is intended to visualize usage patterns rather than to imply comparative effectiveness.

e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

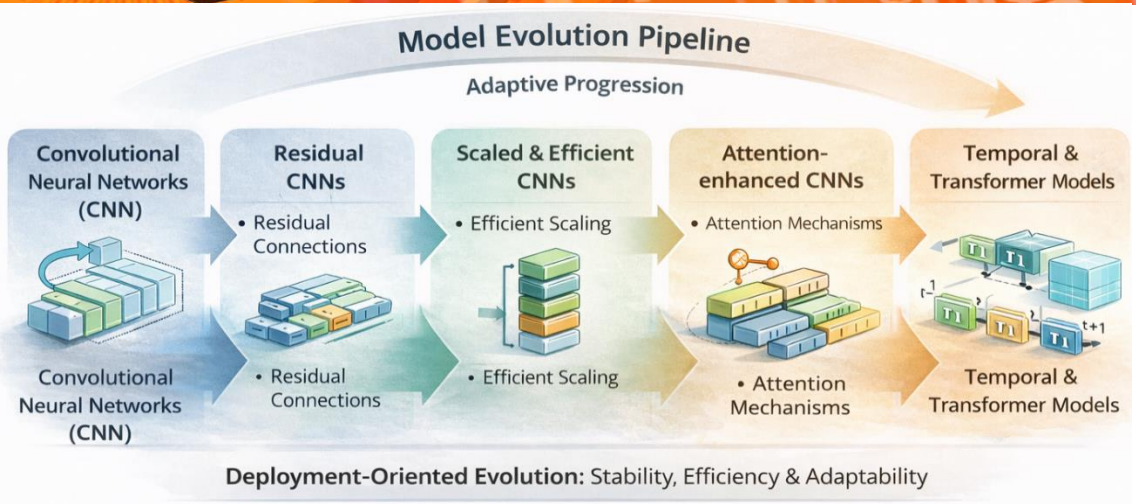



**Figure 4. Model Evolution Pipeline**

Building on this descriptive overview, Figure 4 (Model Evolution Pipeline) conceptualizes the adaptive progression of model architectures observed across the literature. The pipeline illustrates how CNN-based foundations have been incrementally extended through residual connections, efficient scaling strategies, attention mechanisms, and temporal or transformer-based modeling. This progression clarifies why CNNs remain central within educational FER research: they are not superseded, but continually adapted to meet emerging requirements for contextual sensitivity and temporal awareness. The pipeline thus reinforces a deployment-oriented perspective, in which architectural evolution is guided by the need for stability, efficiency, and adaptability within real instructional settings.

Taken together, Table 2 and Figures 3 and 4 indicate that advances in educational FER are driven less by wholesale architectural shifts than by incremental adaptation aligned with deployment constraints. For vocational education in particular, these findings underscore the importance of selecting model architectures that balance representational power with practical feasibility. Such balance enhances the likelihood that affect-aware systems can be integrated into everyday instructional practice, supporting emotion-sensitive and data-informed pedagogical decision making.

**Modality Integration in Educational Affect Analytics**

Across the reviewed studies, the use of learning modalities reflects a gradual expansion from visually grounded affect sensing toward more integrated multimodal configurations, shaped largely by educational context rather than technological novelty. Table 3 synthesizes how different learning signals are operationalized across datasets and mapped to specific vocational applications. Visual modalities remain central, particularly facial expressions and observable behaviors, as they provide a direct and non-intrusive means of interpreting learner engagement during hands-on laboratory and workshop activities. Datasets derived from real classroom settings further emphasize the role of visual observation in capturing authentic instructional dynamics without disrupting learning processes.

**Table 3. Learning Modalities and Vocational Application Contexts**

| Learning Modality / Signal Type | Representative Dataset(s) | Educational Context and Vocational Application |
|---|---|---|
| Visual – Facial Expression | FER2013; CK+; JAFFE; KDEF | Emotion and affect analysis for monitoring learner attention and motivation during hands-on vocational laboratory activities |
| Visual – Real | FESR; VGGFace | Real-time observation of learner behavior and |

| Learning Modality / Signal Type | Representative Dataset(s) | Educational Context and Vocational Application |
|---|---|---|
| Classroom | | engagement in authentic vocational classrooms and workshops |
| Engagement-Oriented Visual Analytics | DAiSEE | Attention and engagement classification supporting learning analytics in blended and distance vocational education |
| Audio–Visual Emotion | RAVDESS; MELD | Multimodal emotion and frustration detection for communication skills training and stress monitoring in vocational learners |
| Multimodal (Visual + Contextual) | AffectNet; Emotic | Context-aware emotion recognition enabling simulation-based and AR/VR-enhanced vocational training environments |
| Teacher Affect Recognition | EIDB-13; RAF-DB | Analysis of instructor emotional intensity to support objective evaluation and improvement of vocational teaching practices |
| Scene-Aware Behavioral Analysis | Emotic; SCB Behavior + Custom | Integrated facial and behavioral monitoring for safety-aware engagement tracking in workshop-based vocational practice |
| Privacy-Aware Multimodal Signals | MELD (Federated setting) | Distributed emotion and sentiment analytics supporting vocational learning under strict data governance constraints |
| Locally Developed / In-house Signals | Custom / In-house Datasets | Context-specific affective and engagement analytics replicable for local vocational institutions without licensing barriers |

Beyond visual signals, the table indicates selective incorporation of audio and contextual cues in scenarios where communication, collaboration, or situational awareness is pedagogically relevant. Audio–visual configurations are primarily associated with emotion and frustration analysis in communication-oriented training, while contextual and scene-aware signals support simulation-based, safety-critical, or environment-sensitive vocational practices. Importantly, these modality choices are not presented as hierarchical improvements, but as context-driven adaptations aligned with instructional goals and institutional constraints.

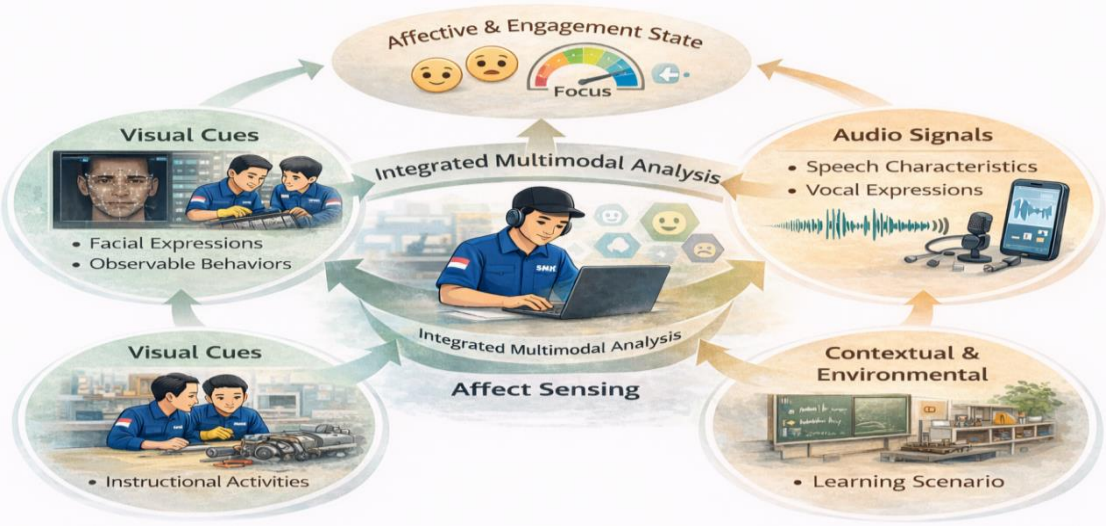



**Figure 5. Multimodal Learning Signal Diagram**

This relationship between modalities is conceptually illustrated in Figure 5 (Multimodal Learning Signal Diagram), which frames affect sensing as an integrative process rather than a collection of isolated inputs. The diagram highlights how visual, audio, and contextual signals converge to inform affective and engagement states within learning activities. Taken as a whole, the evidence synthesized in this subsection points to the importance of aligning modality integration with pedagogical relevance when designing affect analytics for vocational education. This perspective provides a natural transition toward discussing design implications, where the balance between informational richness, ethical considerations, and deployment feasibility becomes central to human-centered vocational learning systems.

**Design Implications and Closed-Loop Transfer to Vocational Education**

The synthesis of the reviewed studies indicates that transferring affect-aware systems to vocational education requires more than technical accuracy; it demands alignment between sensing mechanisms, pedagogical intent, and institutional constraints. Rather than viewing Facial Emotion Recognition (FER) as a standalone analytic component, the literature increasingly positions it within a broader instructional workflow where affective signals inform teaching and learning processes. Table 4 distills this cross-study synthesis into a set of design principles that reflect recurring patterns across datasets, modalities, and application contexts relevant to vocational education.

**Table 4. Design Principles for Transferring Affect Analytics to Vocational Education**

| Design Principle | Supporting Evidence (Representative Studies) | Implication for Vocational Education |
|---|---|---|
| Real-time affect and engagement sensing | Studies using FER2013, DAiSEE, FESR, and VGGFace consistently demonstrate the feasibility of continuous facial-based engagement and emotion monitoring | Enables instructors to monitor learner attention, motivation, and fatigue during hands-on vocational practice in real time |
| Multimodal fusion over single-modality sensing | AffectNet, Emotic, RAVDESS, and MELD-based studies highlight improved robustness when visual, audio, and contextual signals are combined | Supports holistic learner-state interpretation in complex vocational tasks involving communication, collaboration, and physical activity |
| Temporal modeling for sustained activities | CNN–LSTM, ConvLSTM, and 3D-CNN approaches in DAiSEE, CK+, and RAVDESS capture engagement dynamics over time | Essential for long-duration vocational training sessions, workshops, and project-based learning environments |
| Context-aware emotion interpretation | Scene-aware datasets (Emotic, AffectNet) demonstrate that emotional meaning depends on surrounding context | Improves accuracy of affect interpretation in simulated, AR/VR-based, and industry-like vocational learning settings |
| Lightweight and edge-capable architectures | EfficientNet, YOLO-based fusion, and compact CNNs are frequently adopted in applied classroom studies | Facilitates deployment in vocational institutions with limited computational infrastructure and real-time constraints |
| Authentic classroom data integration | FESR and real-classroom studies reveal gaps between benchmark datasets and real learning environments | Highlights the need for locally collected vocational datasets to improve ecological validity |
| Instructor affect monitoring | EIDB-13 and RAF-DB studies show the role of teacher emotional expression in instructional effectiveness | Enables reflective teaching practices and professional development in vocational education |
| Privacy-aware and distributed analytics | Federated learning approaches (e.g., MELD-based studies) address data-sharing and governance concerns | Supports scalable vocational analytics while complying with institutional and ethical data |

| Design Principle | Supporting Evidence (Representative Studies) | Implication for Vocational Education |
|---|---|---|
| | | constraints |
| Adaptation still largely open-loop | Across studies, adaptive feedback is often partial rather than fully closed-loop | Indicates a research gap toward end-to-end adaptive vocational learning systems integrating sensing, analytics, and feedback |

As summarized in Table 4, effective transferability is grounded in real-time affect sensing, contextual awareness, and ethical feasibility. Studies consistently emphasize lightweight and edge-capable architectures as a practical foundation, enabling deployment within resource-constrained vocational institutions. At the same time, the integration of temporal modeling and selective multimodal fusion supports sustained engagement analysis during extended training sessions and collaborative tasks. Importantly, several studies highlight the pedagogical value of monitoring not only learner affect but also instructor emotional expression, reinforcing reflective teaching practices as part of vocational pedagogy.

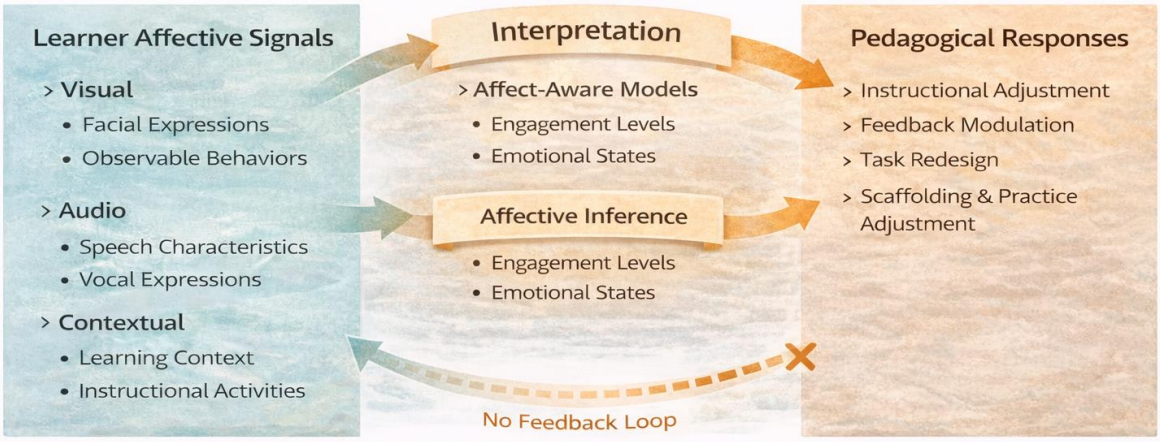
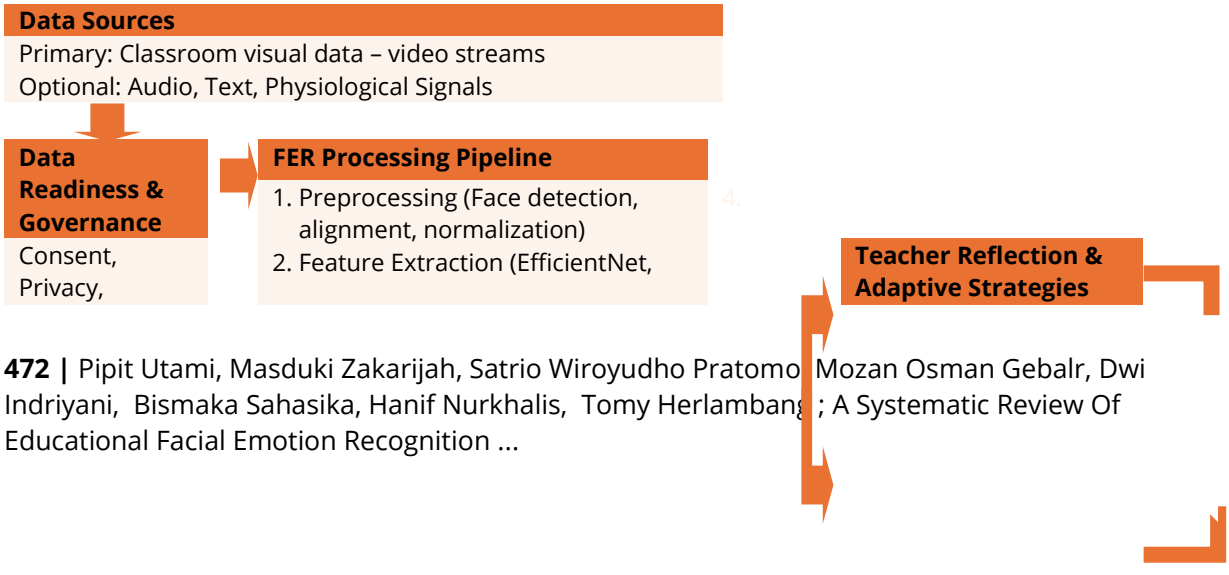



**Figure 6. Closed-Loop Pedagogical System Diagram**

The closed-loop perspective articulated in Figure 6 synthesizes these design principles into a coherent pedagogical system. The diagram visualizes how affective signals are sensed, interpreted, and translated into pedagogical responses, which in turn shape subsequent learner behavior and emotional states. Unlike open-loop approaches that terminate at detection or reporting, the closed-loop model emphasizes continuous feedback, enabling instructional adaptation and learner self-regulation. This perspective reflects a conceptual shift observed across the reviewed studies, from affect recognition as post-hoc analysis toward affect analytics as an integral component of adaptive instruction.



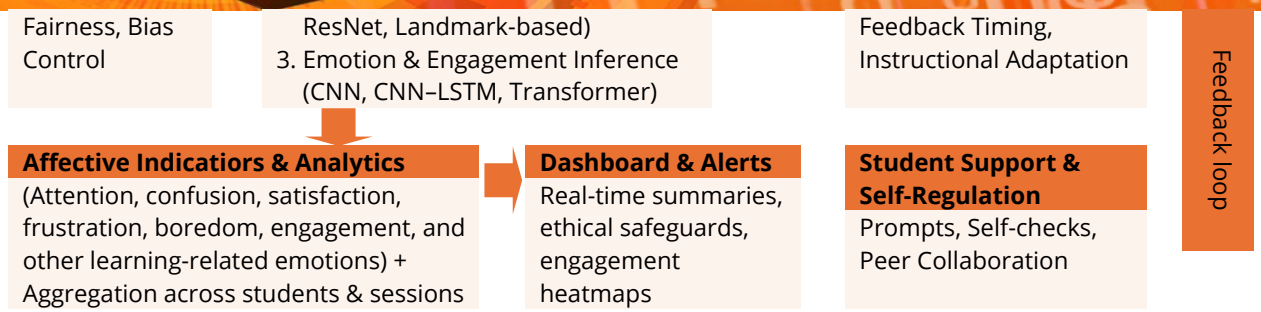



**Figure 7. Framework for FER Transfer to Vocational Teaching Contexts.**

Complementing this conceptual view, Figure 7 operationalizes closed-loop principles within vocational teaching contexts by mapping data sources, governance considerations, analytic pipelines, and instructional actions. The framework highlights that successful deployment depends on balancing informational richness with privacy, fairness, and feasibility. Viewed collectively, the materials presented in this section suggest that current FER-based systems in vocational education are still largely open-ended, with adaptation occurring in fragmented ways. The progression toward genuinely adaptive learning environments therefore depends on closing this gap by explicitly linking affect sensing, analytic interpretation, and pedagogical action within an ethically grounded and context-sensitive instructional cycle.

## Discussion

### Dataset and Contextual Landscape of Educational Affect Analytics

The synthesis of studies reviewed in this section indicates that educational affect analytics remains strongly anchored in a limited ecosystem of benchmark facial emotion recognition datasets, most notably FER2013, AffectNet, and CK+. While these datasets have played a pivotal role in advancing methodological consistency and enabling comparative evaluation across studies, their origins largely outside authentic classroom environments present a structural limitation. As documented in prior work, the majority of these datasets were collected under controlled or semi-controlled conditions, relying on posed expressions or generalized public imagery rather than naturally occurring classroom interactions (S. Li & Deng, 2022; Pan et al., 2022; Soeparno, 2023). This disconnect constrains the ecological validity of trained models, as emotional expressions in real educational settings are shaped by situational demands, social interaction, and task engagement that are rarely represented in benchmark data.

The limitations of this dataset landscape become more pronounced when considering vocational and hands-on learning environments. Controlled laboratory datasets often fail to capture spontaneous affective responses that emerge during applied, skill-oriented activities, where learners navigate frustration, confidence, uncertainty, and collaboration in dynamic ways (Anwar et al., 2023; Costa et al., 2023; Gong et al., 2022). Reviews consistently report that models trained on such benchmarks struggle under conditions involving occlusion, variable lighting, and contextual complexity, leading to reduced generalization and overfitting when deployed in real classrooms (D'Incà et al., 2023; Duan, 2023; Kollias & Zafeiriou, 2021). Although some studies argue for the transferability of general-purpose datasets as a pragmatic starting point, they also acknowledge persistent representation bias and contextual gaps that limit reliability in vocational settings (Mukherjee et al., 2024; S. Zhang et al., 2020). Collectively, this body of evidence underscores the need to shift from dataset-centric validation toward context-aware data development, positioning authentic educational and vocational datasets as a critical foundation for advancing affect analytics that is pedagogically meaningful and practically transferable. This pattern suggests that current progress in educational FER is shaped more by dataset availability than by pedagogical authenticity.

**Model Trends and Deployment-Oriented Performance**

The synthesis of recent studies reveals a clear evolution in facial emotion recognition architectures used in educational research, marked by a gradual shift from conventional CNN-based models toward attention-enhanced and transformer-based designs. Rather than indicating a complete architectural replacement, this trend reflects an expanding design space in which transformers and attention mechanisms are introduced to address the growing complexity of visual and behavioral cues in classroom environments. Empirical evidence shows that hybrid vision transformer models and transformer-integrated feature pyramid networks achieve improved emotion and behavior recognition by capturing long-range dependencies and salient facial regions more effectively than standard CNNs (N. Li et al., 2024; Z. Wang et al., 2023). These developments suggest that attention-driven architectures are increasingly valued for their representational flexibility, particularly in learning contexts characterized by heterogeneous lighting, occlusion, and dynamic learner behavior.

At the same time, the reviewed literature consistently emphasizes that performance gains in educational FER are constrained by deployment realities, especially in vocational and resource-limited settings. Transformer-based models, while powerful, incur higher computational costs that may hinder real-time classroom use, prompting continued reliance on EfficientNet and lightweight CNN variants for practical deployment (Riaz et al., 2020; X. Zhang et al., 2022). Studies demonstrate that architectural refinement, such as embedding attention modules within compact networks or employing hybrid CNN–transformer configurations, often yields substantial accuracy improvements without proportional increases in complexity (Deramgozin et al., 2023; Tanweer et al., 2022). Comparative evaluations further indicate that streamlined and hybrid transformer architectures can achieve a workable balance between accuracy and efficiency when carefully optimized for edge devices (S. Lin et al., 2020; Wu et al., 2024; Y. Zhao et al., 2023). Collectively, this body of evidence supports a deployment-oriented perspective, in which model selection is guided less by maximal architectural sophistication than by the alignment between representational capacity, computational feasibility, and the operational demands of real educational environments (Gera, 2020; Laraib et al., 2023). In this sense, architectural evolution in educational FER reflects not a pursuit of novelty, but a pragmatic negotiation between performance ambition and instructional feasibility.

**Modality Integration in Educational Affect Analytics**

The synthesis of recent studies indicates that educational affect analytics increasingly relies on multimodal configurations to capture the complexity of learner engagement and emotional experience. Visual signals, particularly facial expressions, remain the most widely adopted modality due to their non-intrusive nature and direct relevance to observable learning behaviors. However, evidence shows that combining visual cues with audio and contextual information enables a more comprehensive interpretation of affective states, especially in interactive and collaborative learning settings (Donate-Beby et al., 2023; Liao, 2023; Ochôa, 2022). Audio signals contribute insights into vocal tone, participation patterns, and ambient classroom dynamics, while contextual cues derived from environmental conditions and behavioral traces situate emotional expressions within meaningful instructional activities. This integrative approach supports a more nuanced understanding of learners' cognitive and emotional processes, particularly when learning unfolds through dialogue, teamwork, and situated practice.

At the same time, the reviewed literature emphasizes that the effectiveness of multimodal fusion is highly scenario-dependent rather than universally advantageous. Studies demonstrate that audiovisual and contextual integration yields notable robustness in instructional contexts centered on communication, collaboration, and simulation-based learning, where emotional expression is distributed across multiple channels (Xie et al., 2021; Yan et al., 2024). In contrast, hands-on and workshop-based environments introduce practical constraints related to data synchronization,

sensor reliability, and real-time processing demands, which can limit the feasibility of complex multimodal systems (Zhan, 2024; Zhu et al., 2022). These findings underscore the importance of aligning modality selection with instructional goals and contextual conditions, suggesting that multimodal analytics should be designed as a pedagogically informed strategy rather than a default technical enhancement. In this sense, the value of multimodal integration lies not in the accumulation of signals, but in their selective orchestration to support meaningful and context-sensitive learning processes

**Design Implications and Closed-Loop Transfer to Vocational Education**

The reviewed literature indicates that educational affect analytics most commonly translates affective inference into pedagogical action through a linear process that prioritizes immediate interpretation over sustained instructional adaptation. In many studies, emotional cues are collected and analyzed to infer learner engagement or distress, yet the process frequently terminates at this analytic stage without feeding learner responses back into subsequent instructional decisions. This open-ended configuration constrains the capacity of affect-aware systems to address the evolving nature of learners' emotional states over time, thereby limiting their long-term pedagogical impact. Although a small number of studies report closed-loop approaches in which real-time feedback enables affective signals to dynamically inform instructional strategies, such implementations remain relatively rare and are often confined to experimental or highly controlled settings. As a result, affect analytics in education is still predominantly positioned as a decision-support mechanism rather than an adaptive instructional system.

Transferring closed-loop affect-aware systems into vocational and hands-on learning environments introduces additional layers of complexity that extend beyond technical model performance. Workshop-based settings demand rapid interpretation and timely pedagogical response, yet multimodal systems in these contexts are challenged by infrastructural constraints, data integration overload, and the practical realities of instructor workload. Common pedagogical responses following affect detection, including feedback modulation, task redesign, and scaffolding, offer promising pathways for enhancing learner engagement, but their effectiveness depends on educators' readiness to interpret affective data and act upon it in situ (Mangaroska et al., 2020; Ruan et al., 2023). Moreover, ethical considerations related to privacy, consent, and institutional governance further shape the feasibility of closed-loop implementation in authentic educational settings (Chango et al., 2022; Kwon et al., 2022; Z. Zhang et al., 2024). Taken together, these findings suggest that advancing closed-loop affect analytics for vocational education requires not only technical integration, but also pedagogically grounded design, instructor capacity building, and institutional alignment to support responsive and human-centered instructional practice.

**Implications**

Synthesizing the findings across the four discussion subsections, this review demonstrates that current educational facial emotion recognition research has achieved notable methodological maturity while remaining unevenly aligned with authentic instructional practice, particularly in vocational contexts. The results show that progress is largely shaped by benchmark-centered datasets that offer methodological stability but limited ecological validity, constraining transferability to hands-on and practice-oriented learning environments. At the model level, architectural evolution reflects a deployment-oriented trajectory in which performance gains are driven by incremental refinement and efficiency considerations rather than wholesale replacement, underscoring the tension between representational ambition and real-world feasibility. The integration of multimodal signals further reveals that affect analytics is most effective when modality selection is guided by instructional context rather than technical accumulation, with robustness emerging in scenario-specific applications rather than universally. Finally, design-oriented evidence indicates that most FER systems remain open-loop, where affective inference informs analysis but rarely translates into sustained pedagogical adaptation due to infrastructural, pedagogical, and ethical constraints. Taken

together, these results suggest that the central challenge in advancing educational FER lies not in improving detection accuracy, but in aligning datasets, models, modalities, and system design with the contextual and pedagogical realities of vocational learning.

The synthesis of the Results and Discussion highlights that the primary limitations of current educational facial emotion recognition research stem from structural mismatches between methodological advancement and authentic instructional contexts. A central constraint lies in the continued reliance on benchmark FER datasets collected under controlled conditions, which limits ecological validity when models are deployed in vocational and hands-on learning environments characterized by physical activity, interaction, occlusion, and spontaneous affective expression. This dataset bias constrains generalization and reduces pedagogical reliability in real classrooms (Akputu et al., 2022; S. Li & Deng, 2022; Ullah et al., 2022). These data limitations are compounded by deployment challenges associated with high-performing models, particularly transformer-based architectures, whose computational demands and real-time processing requirements are often misaligned with the resource constraints of vocational institutions (Agung et al., 2024; Saurav et al., 2023; Shen & Jin, 2023). While multimodal affect analytics offers richer interpretive potential, its implementation in workshop-based settings remains constrained by integration complexity, processing overhead, and limited instructor readiness to translate complex outputs into timely pedagogical action (Pan et al., 2022; Soeparno, 2023; Yan et al., 2024; Zhu et al., 2022). Collectively, these factors contribute to the predominance of open-loop FER systems, where affective inference rarely translates into sustained instructional adaptation due to infrastructural, ethical, and pedagogical barriers (Dong et al., 2022; Keshtkaran et al., 2021; Mwangi et al., 2021).

Grounded in the synthesized results of this review, future research in educational affect analytics should focus on addressing the structural gaps identified between methodological advancement and authentic pedagogical practice, particularly in vocational learning contexts. Priority should be given to the development of context-specific and classroom-authentic datasets that capture spontaneous, task-embedded, and interaction-driven affective expressions, thereby improving ecological validity beyond what benchmark datasets can offer (Giảng et al., 2023; L. Lin et al., 2024). In parallel, research should advance from predominantly open-loop implementations toward closed-loop affect-aware systems that more explicitly integrate affect sensing, interpretation, and pedagogical action, enabling instructional adjustment based on evolving learner states rather than static inference (Seok et al., 2024). Progress in this direction must remain sensitive to deployment realities highlighted in the results, including resource constraints, instructor readiness, and the risk of analytic overload, while also foregrounding human-centered and ethical considerations related to bias, privacy, and consent (Wongvorachan et al., 2024). Collectively, these directions point toward a measured shift from accuracy-driven innovation toward context-grounded, adaptive, and ethically informed affect analytics that align more closely with the instructional and experiential demands of vocational education (Demszky & Hill, 2023).

Synthesizing the results of this review reveals that current theoretical foundations of educational facial emotion recognition remain partially misaligned with the realities of authentic learning contexts, particularly in vocational education. The heavy reliance on benchmark datasets centered on posed and controlled expressions underpins theoretical models that treat emotion as a stable, context-independent construct, despite evidence that affect in real classrooms is situational, interactional, and dynamically shaped by task demands, physical activity, and social engagement. At the model level, the findings challenge scale-driven assumptions of intelligence by showing that architectural refinement, inductive bias, and attention-guided design contribute more meaningfully to generalization than increased complexity alone, suggesting that theoretical accounts of model capability must prioritize structural alignment with affective phenomena. Similarly, evidence from multimodal analytics contests hierarchical and universal assumptions of modality integration, indicating instead that affective meaning emerges through selective, context-sensitive orchestration of signals shaped by instructional goals and learning scenarios. Finally, the predominance of open-

loop systems exposes a theoretical gap in which emotion is conceptualized as a static analytic output rather than a dynamic element within pedagogical interaction, underscoring the need for closed-loop frameworks that theorize affect as both influencing and being influenced by instructional action. Collectively, these results point toward a reorientation of educational FER theory from abstract detection paradigms toward context-grounded, adaptive, and pedagogically embedded conceptions of emotion in learning.

The findings of this review suggest that effective implementation of educational facial emotion recognition in vocational settings depends on context-sensitive and pedagogically grounded decisions rather than generic, accuracy-driven adoption. In practice, institutions should prioritize the development of FER datasets derived from authentic classroom and workshop activities, with careful attention to ethical governance, consent, and sociocultural variation in emotional expression, in order to improve ecological validity and instructional relevance (Banzon et al., 2024; Jia et al., 2021; J. H. Kim et al., 2021). Model deployment should favor lightweight and attention-enhanced architectures that balance recognition performance with computational efficiency, enabling real-time responsiveness within the infrastructural constraints typical of vocational environments (Aly et al., 2023; Huang et al., 2023; S. Li & Deng, 2022). Where multimodal analytics are employed, modality selection should be guided by instructional goals, learner interaction patterns, and technical feasibility, rather than maximal integration of available signals (Moon et al., 2024; Vistorte et al., 2024). Finally, translating affective inference into meaningful instructional action requires organizational readiness, including teacher capacity to interpret emotional data, supportive institutional policies, and gradual integration of feedback mechanisms that can inform instructional adjustment without overwhelming classroom practice (Siddiqui et al., 2022; Banerjee et al., 2021; García-Hernández et al., 2024).

## CONCLUSION AND SUGGESTIONS

### Conclusion

This systematic review synthesizes evidence from recent educational facial emotion recognition research and demonstrates that the field has achieved substantial methodological maturity while remaining unevenly aligned with authentic instructional practice, particularly in vocational learning contexts. The findings show that progress has been largely driven by benchmark-centered datasets, incremental architectural refinement, and selective multimodal integration, all of which have contributed to improved detection accuracy and analytic capability. However, these advances also reveal a persistent gap between technical performance and pedagogical relevance, as many systems are developed and evaluated outside the dynamic, hands-on environments where emotional engagement plays a central role in learning. Overall, the review highlights that the primary challenge facing educational FER is no longer the feasibility of emotion detection, but the meaningful integration of affect analytics into real instructional processes. In vocational education, where learning is experiential, socially situated, and emotionally demanding, FER holds promise not merely as a monitoring technology but as a pedagogical instrument that can support reflective teaching, learner engagement, and adaptive instruction when grounded in contextual authenticity, ethical awareness, and deployment realism.

### Suggestions

Based on the synthesized findings, future efforts in educational FER should emphasize alignment between technological design and instructional context. Researchers are encouraged to focus on developing and validating datasets that reflect authentic classroom and workshop conditions, while practitioners should prioritize model choices and modality configurations that balance analytic capability with practical feasibility. Strengthening the connection between affective inference and pedagogical action through gradual adoption of feedback-driven instructional

strategies is also essential for enhancing the educational value of FER systems. In addition, sustained progress will depend on institutional readiness and human capacity. This includes supporting educators in interpreting affective insights, establishing ethical and governance frameworks that protect learner well-being, and fostering collaboration between researchers, educators, and system developers. Through these measures, educational FER can evolve toward context-sensitive, adaptive, and human-centered applications that genuinely support vocational learning and professional skill development.

## REFERENCES


Agung, E. S., Rifai, A. P., & Wijayanto, T. (2024). Image-Based Facial Emotion Recognition Using Convolutional Neural Network on Emognition Dataset. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-65276-x

Akputu, O. K., Inyang, U. G., Msugh, O., Mughal, F., & Usoro, A. (2022). Recognizing Facial Emotions for Educational Learning Settings. Iaes International Journal of Robotics and Automation (Ijra), 11(1), 21. https://doi.org/10.11591/ijra.v11i1.pp21-32

Aly, M., Ghallab, A., & Fathi, I. S. (2023). Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model. Ieee Access, 11, 121419–121433. https://doi.org/10.1109/access.2023.3325407

Anwar, A., Rehman, I. U., Nasralla, M. M., Khattak, S. B. A., & Khilji, N. (2023). Emotions Matter: A Systematic Review and Meta-Analysis of the Detection and Classification of Students' Emotions in STEM During Online Learning. Education Sciences, 13(9), 914. https://doi.org/10.3390/educsci13090914

Athauda, H., Jayasinghe, U., & Ragel, R. (2023). CNN for Facial Emotion Recognition in Online Learning Platforms to Identify Learner Engagement. 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), 128–133. https://doi.org/10.1109/ICIIS58898.2023.10253525

Banzon, A. M., Beever, J., & Taub, M. (2024). Facial Expression Recognition in Classrooms: Ethical Considerations and Proposed Guidelines for Affect Detection in Educational Settings. IEEE Transactions on Affective Computing, 15(1), 93–104. https://doi.org/10.1109/TAFFC.2023.3275624

Chan, X. Y., Connie, T., & Goh, M. K. O. (2023). Facial and Body Gesture Recognition for Determining Student Concentration Level. International Journal on Advanced Science, Engineering and Information Technology, 13(5), 1693–1702. https://doi.org/10.18517/ijaseit.13.5.19035

Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A Review on Data Fusion in Multimodal Learning Analytics and Educational Data Mining. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, 12(4). https://doi.org/10.1002/widm.1458

Cioruţa, B. V., Sabou Cioruţa, I. E., & Pop, A. L. (2024). Digitalization and Computer-assisted training - from Methodological Scenarios to Practical Values in the Instructional-Educational Process. BULETIN ŞTIINŢIFIC SERIA A Fascicula Pedagogie-Psihologie-Metodică, 24, 43–52. https://doi.org/10.37193/BS-PPM.24.05

Costa, W., Talavera, E., Oliveira, R., Figueiredo, L., Teixeira, J. M., Lima, J. P., & Teichrieb, V. (2023). A Survey on Datasets for Emotion Recognition From Vision: Limitations and in-the-Wild Applicability. Applied Sciences, 13(9), 5697. https://doi.org/10.3390/app13095697

Das, D. K., Patnaik, P., Nayak, N., Das, S. K., & Baral, M. (2024). Affective Computing in Education (pp. 65–82). https://doi.org/10.4018/979-8-3693-7011-7.ch005

Deepa, D., Bhimaavarapu, K., Sholapurapu, P. K., & Sarupriya, S. (2025). Real-Time Classroom Emotion Analysis Using Machine and Deep Learning for Enhanced Student Learning. Journal of Intelligent Systems and Internet of Things, 16(2), 82–101. https://doi.org/10.54216/JISIoT.160207

Demszky, D., & Hill, H. C. (2023). The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. https://doi.org/10.18653/v1/2023.bea-1.44

Deramgozin, M. M., Jovanović, S., Arevalillo-Herráez, M., Ramzan, N., & Rabah, H. (2023). Attention-Enabled Lightweight Neural Network Architecture for Detection of Action Unit Activation. Ieee Access, 11, 117954–117970. https://doi.org/10.1109/access.2023.3325034

D'Incà, M., Beyan, C., Niewiadomski, R., Barattin, S., & Sebe, N. (2023). Unleashing the Transferability Power of Unsupervised Pre-Training for Emotion Recognition in Masked and Unmasked Facial Images. Ieee Access, 11, 90876–90890. https://doi.org/10.1109/access.2023.3308047

Donate-Beby, B., García-Peñalvo, F. J., & Amo, D. (2023). Analíticas De Aprendizaje en La Educación Primaria Y Secundaria en España: Una Revisión Sistemática De La Literatura. Universitas Tarraconensis Revista De Ciències De L Educació, 2, 63–86. https://doi.org/10.17345/ute.2023.3685

Dong, H., Hou, J., Song, Z., Xu, R., Meng, L., & Ming, D. (2022). An Adaptive Reflexive Control Strategy for Walking Assistance System Based on Functional Electrical Stimulation. Frontiers in Neuroscience, 16. https://doi.org/10.3389/fnins.2022.944291

Duan, C. (2023). A Survey of Facial Expression Recognition in the Wild. Applied and Computational Engineering, 6(1), 98–106. https://doi.org/10.54254/2755-2721/6/20230760

El Maazouzi, Q., & Retbi, A. (2025). Multimodal Detection of Emotional and Cognitive States in E-Learning Through Deep Fusion of Visual and Textual Data with NLP. Computers, 14(8), 314. https://doi.org/10.3390/computers14080314

Evangeline, D., & Parkavi, A. (2024). Facial Emotion Recognition of Online Learners Using a Hybrid Deep Learning Model. International Journal of Intelligent Engineering and Systems, 17(6), 735–751. https://doi.org/10.22266/ijies2024.1231.56

Fu, R., & Tian, M. (2024). Classroom Facial Expression Recognition Method Based on Conv3D-ConvLSTM-SEnet in Online Education Environment. Journal of Circuits, Systems and Computers, 33(07). https://doi.org/10.1142/S0218126624501317

Ganepola, D. (2022). Affective Computing for Facilitation of Inclusive Education. https://doi.org/10.31219/osf.io/69hms

Gawande, R. (2025). MoodMapper: Facial Emotion Recognition. International Scientific Journal of Engineering and Management, 04(06), 1–9. https://doi.org/10.55041/ISJEM04087

Gera, D. (2020). Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information Based Facial Expression Recognition. https://doi.org/10.48550/arxiv.2007.10298

Giảng, N. V, Chan, N., Nguyen, H. T., Dich, Q., & Dong, T. (2023). BK-SAD: A Large Scale Dataset for Student Activity Recognition. Ssad, 33(3), 16–23. https://doi.org/10.51316/jst.168.ssad.2023.33.3.3

Gong, W., Wang, C., Jia, J., Qian, Y., & Fan, Y. (2022). Multi-Feature Fusion Network for Facial Expression Recognition in the Wild. Journal of Intelligent & Fuzzy Systems, 42(6), 4999–5011. https://doi.org/10.3233/jifs-211021

Gupta, S., Kumar, P., & Tekchandani, R. (2023). EDFA: Ensemble deep CNN for assessing student's cognitive state in adaptive online learning environments. International Journal of Cognitive Computing in Engineering, 4, 373–387. https://doi.org/10.1016/j.ijcce.2023.11.001

Huang, J.-J., Yang, C., Lin, Y., Shen, V. R., Lin, C.-T., & Shen, F. H. (2023). Novel Emotion Recognition System Using Edge Computing Platform With Deep Convolutional Networks. Journal of Intelligent & Fuzzy Systems, 45(2), 2669–2683. https://doi.org/10.3233/jifs-223801

Jia, S., Wang, S., Hu, C., Webster, P., & Li, X. (2021). Detection of Genuine and Posed Facial Expressions of Emotion: Databases and Methods. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.580287

Keshtkaran, M. R., Sedler, A. R., Chowdhury, R. H., Tandon, R., Basrai, D., Nguyen, S. L., Sohn, H., Jazayeri, M., Miller, L. E., & Pandarinath, C. (2021). A Large-Scale Neural Network Training Framework for Generalized Estimation of Single-Trial Population Dynamics. https://doi.org/10.1101/2021.01.13.426570

Kim, H. S., & Cho, S. H. (2024). Advancements in AI Educational Tools and Learners' Emotional Well-Being: A Trend Analysis. Journal of Next-Generation Convergence Information Services Technology, 13(5), 581–591. https://doi.org/10.29056/jncist.2024.10.01

Kim, J. H., Poulose, A., & Han, D. S. (2021). The Extensive Usage of the Facial Image Threshing Machine for Facial Emotion Recognition Performance. Sensors, 21(6), 2026. https://doi.org/10.3390/s21062026

Kollias, D., & Zafeiriou, S. (2021). Affect Analysis in-the-Wild: Valence-Arousal, Expressions, Action Units and a Unified Framework. https://doi.org/10.48550/arxiv.2103.15792

Kwon, S., Ahn, J.-H., Choi, H., Jeon, J., Kim, D., Kim, H., & Kang, S.-J. (2022). Analytical Framework for Facial Expression on Game Experience Test. Ieee Access, 10, 104486–104497. https://doi.org/10.1109/access.2022.3210712

Lampropoulos, G., Fernández-Arias, P., Antón-Sancho, Á., & Vergara, D. (2024). Affective Computing in Augmented Reality, Virtual Reality, and Immersive Learning Environments. Electronics, 13(15), 2917. https://doi.org/10.3390/electronics13152917

Laraib, U., Shaukat, A., Khan, R. A., Mustansar, Z., Akram, M. U., & Asgher, U. (2023). Recognition of Children's Facial Expressions Using Deep Learned Features. Electronics, 12(11), 2416. https://doi.org/10.3390/electronics12112416

Lasri, I., Riadsolh, A., & Elbelkacemi, M. (2023). Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. Education and Information Technologies, 28(4), 4069–4092. https://doi.org/10.1007/s10639-022-11370-4

Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024). Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach. Applied Sciences, 14(3), 1156. https://doi.org/10.3390/app14031156

Li, N., Huang, Y., Wang, Z., Fan, Z., Li, X., & Xiao, Z. (2024). Enhanced Hybrid Vision Transformer With Multi-Scale Feature Integration and Patch Dropping for Facial Expression Recognition. Sensors, 24(13), 4153. https://doi.org/10.3390/s24134153

Li, S., & Deng, W. (2022). Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Liao, Z. (2023). CH-CC: A Chinese Multimodal Classroom Atmosphere Analysis Dataset Based on Teachers' Behavior and Voice. https://doi.org/10.20944/preprints202301.0551.v1

Lin, L., Yang, H., Xu, Q., Xue, Y., & Li, D. (2024). Research on Student Classroom Behavior Detection Based on the Real-Time Detection Transformer Algorithm. Applied Sciences, 14(14), 6153. https://doi.org/10.3390/app14146153

Lin, S., Wu, C., Chen, S., Lin, T.-L., & Tseng, Y.-W. (2020). Continuous Facial Emotion Recognition Method Based on Deep Learning of Academic Emotions. Sensors and Materials, 32(10), 3243. https://doi.org/10.18494/sam.2020.2863

Liu, J.-W., Lin, X.-Y., Ji, P.-F., Chen, J.-M., & Zhang, J. (2025a). Multiscale wavelet attention convolutional network for facial expression recognition. Scientific Reports, 15(1), 22219. https://doi.org/10.1038/s41598-025-07416-5

Liu, J.-W., Lin, X.-Y., Ji, P.-F., Chen, J.-M., & Zhang, J. (2025b). Multiscale wavelet attention convolutional network for facial expression recognition. Scientific Reports, 15(1), 22219. https://doi.org/10.1038/s41598-025-07416-5

Mangaroska, K., Sharma, K., Gašević, D., & Giannakos, M. N. (2020). Multimodal Learning Analytics to Inform Learning Design: Lessons Learned From Computing Education. Journal of Learning Analytics, 7(3), 79–97. https://doi.org/10.18608/jla.2020.73.7

Moon, J., Yeo, S., Banihashem, S. K., & Noroozi, O. (2024). Using Multimodal Learning Analytics as a Formative Assessment Tool: Exploring Collaborative Dynamics in Mathematics Teacher Education. Journal of Computer Assisted Learning, 40(6), 2753–2771. https://doi.org/10.1111/jcal.13028

Mukherjee, D., Hong, J., Vats, H., Bae, S., & Najjaran, H. (2024). Personalization of Industrial Human–robot Communication Through Domain Adaptation Based on User Feedback. User Modeling and User-Adapted Interaction, 34(4), 1327–1367. https://doi.org/10.1007/s11257-024-09394-1

Mwangi, K. W., Mainye, N., Ouso, D. O., Esoh, K., Muraya, A., Mwangi, C. K., Naitore, C., Karega, P., Rono, G. K., Musundi, S., Mutisya, J., Mwangi, E., Mgawe, C., Miruka, S. A., & Kibet, C. K. (2021). Open Science in Kenya: Where Are We? Frontiers in Research Metrics and Analytics, 6. https://doi.org/10.3389/frma.2021.669675

Ochôa, X. (2022). Multimodal Learning Analytics: Rationale, Process, Examples, and Direction. 54–65. https://doi.org/10.18608/hla22.006

Orosoo, M., Rajkumari, Y., Ramesh, K., Fatma, G., Nagabhaskar, M., Gopi, A., & Rengarajan, M. (2024). Enhancing English Learning Environments Through Real-Time Emotion Detection and Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 15(7). https://doi.org/10.14569/IJACSA.2024.0150787

Pan, X., Liu, W., Wang, Y., Lü, X., & Liu, B. (2022). MSL-FER: Mirrored Self-Supervised Learning for Facial Expression Recognition. 1601–1605. https://doi.org/10.1109/icip46576.2022.9898036

Rathod, T., Patil, S., Shahade, A. K., Kadam, P., & Kulkarni, A. (2025). Facial emotion recognition using deep Siamese neural networks: multi-classifier fusion for single-emotion and multi-emotion models across age groups. Journal of Big Data, 12(1), 222. https://doi.org/10.1186/s40537-025-01287-3

Riaz, M. N., Shen, Y., Sohail, M., & Guo, M. (2020). eXnet: An Efficient Approach for Emotion Recognition in the Wild. Sensors, 20(4), 1087. https://doi.org/10.3390/s20041087

Ruan, X., Palansuriya, C., & Constantin, A. (2023). Affective Dynamic Based Technique For Facial Emotion Recognition (FER) To Support Intelligent Tutors In Education. 774–779. https://doi.org/10.1007/978-3-031-36272-9_70

Saurav, S., Saini, R., & Singh, S. K. (2023). A Dual-channel Ensembled Deep Convolutional Neural Network for Facial Expression Recognition in the Wild. Computational Intelligence, 39(5), 666–706. https://doi.org/10.1111/coin.12586

Seok, C. L., Park, Y., Baek, J., Lim, H., Roh, J., Kim, Y., Kim, S., & Lee, E. C. (2024). AffectiVR: A Database for Periocular Identification and Valence and Arousal Evaluation in Virtual Reality. Electronics, 13(20), 4112. https://doi.org/10.3390/electronics13204112

Shen, L., & Jin, X. (2023). VaBTFER: An Effective Variant Binary Transformer for Facial Expression Recognition. Sensors, 24(1), 147. https://doi.org/10.3390/s24010147

Singh, S., Gupta, A., & Pavithr, R. S. (2022). Automatic Classroom Monitoring System Using Facial Expression Recognition (pp. 151–165). https://doi.org/10.1007/978-981-16-8542-2_12

Soeparno, H. (2023). Facial Emotion Recognition Using Convolutional Neural Network Based on the Visual Geometry Group-19. Jurnal Tam (Technology Acceptance Model), 14(1), 48. https://doi.org/10.56327/jurnaltam.v14i1.1475

T. S., A., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Education and Information Technologies, 25(2), 1387–1415. https://doi.org/10.1007/s10639-019-10004-6

Takyi Mensah, E., Chen, M., Ntim, S. Y., & Gabrah, A. (2023). Analysing Dewey's vocational aspects of education and Maslow's theory of motivation in support of vocational education and training. Discover Education, 2(1), 18. https://doi.org/10.1007/s44217-023-00042-1

Tang, J., Zhou, X., & Zheng, J. (2019). Design of Intelligent classroom facial recognition based on Deep Learning. Journal of Physics: Conference Series, 1168, 022043. https://doi.org/10.1088/1742-6596/1168/2/022043

Tanweer, R., Tanveer, H., Mayo, A. A., Ain, Q. U., & Ahmad, J. (2022). Real-Time Intelligent Facial Expression Recognition System. Jisr-C, 20(2). https://doi.org/10.31645/jisrc.22.20.2.4

Thiruthuvanathan, M. M., & Krishnan, B. (2024). Multitask EfficientNet affective computing for student engagement detection. Multimedia Tools and Applications, 84(18), 19039–19063. https://doi.org/10.1007/s11042-024-19815-3

Thomas, L., Shilpa, K. C., Shilpa, M. I., & Sandeep Telkar, I. (2025). Automated User Engagement Analysis Through Multi-Modal Facial Expression Recognition: A Deep Learning Approach. 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), 1–6. https://doi.org/10.1109/AMATHE65477.2025.11080897

Tian, Y., Han, T., & Wu, L. (2021). Teacher Facial Expression Recognition Based on GoogLeNet-InceptionV3 CNN Model. In Artificial Intelligence in Education and Teaching Assessment (pp. 69–78). Springer Singapore. https://doi.org/10.1007/978-981-16-6502-8_8

Tokhtarov, A., Toxanbayeva, N., Seisekenova, M., Baidildinov, T., Baigozhanova, D., & Abden, A. (2025). AI-Based Analysis of Student Frustration: Speech and Facial Expression Recognition. Electronic Journal of E-Learning, 23(2), 143–157. https://doi.org/10.34190/ejel.23.2.4043

Torres-Hernández, E. F. (2025). Vocación a la Docencia: Su Relación con la Regulación Emocional y la Resiliencia. Revista de Estudios y Experiencias En Educación, 24(54), 233–249. https://doi.org/10.21703/rexe.v24i54.2953

Ullah, Z., Mohmand, M. I., Rehman, S. U., Zubair, M., Driss, M., Boulila, W., Sheikh, R., & Alwawi, I. (2022). Emotion Recognition From Occluded Facial Images Using Deep Ensemble嫫odel. Computers Materials & Continua, 73(3), 4465–4487. https://doi.org/10.32604/cmc.2022.029101

Utami, P., Hartanto, R., & Soesanti, I. (2019). A Study on Facial Expression Recognition in Assessing Teaching Skills: Datasets and Methods. Procedia Computer Science, 161, 544–552. https://doi.org/10.1016/j.procs.2019.11.154

Utami, P., Hartanto, R., & Soesanti, I. (2022). A Brief Study of The Use of Pattern Recognition in Online Learning: Recommendation for Assessing Teaching Skills Automatically Online Based. Elinvo (Electronics, Informatics, and Vocational Education), 7(1), 48–62. https://doi.org/10.21831/elinvo.v7i1.51354

Vistorte, A. O. R., Deroncele-Acosta, A., Ayala, J. L. M., Barrasa, A., López-Granero, C., & Martí-González, M. (2024). Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1387089

Wang, C., Dai, J., Chen, Y., Zhang, X., & Xu, L. (2023). A Learning Analytics Model Based on Expression Recognition and Affective Computing: Review of Techniques and Survey of Acceptance. In Proceedings of the 2022 International Conference on Educational Innovation and Multimedia Technology (EIMT 2022) (pp. 169–178). Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-012-1_19

Wang, K., & Cheng, M. (2024). Teaching Feedback System Based on ViT Expression Recognition in Distance Education. 2024 13th International Conference on Educational and Information Technology (ICEIT), 93–97. https://doi.org/10.1109/ICEIT61397.2024.10540794

Wang, Z., Yao, J., Zeng, C., Li, L., & Tan, C. (2023). Students' Classroom Behavior Detection System Incorporating Deformable DETR With Swin Transformer and Light-Weight Feature Pyramid Network. Systems, 11(7), 372. https://doi.org/10.3390/systems11070372

Wongvorachan, T., Bulut, O., Liu, J. X., & Mazzullo, E. (2024). A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning. Information, 15(6), 326. https://doi.org/10.3390/info15060326

Wu, Y., Zhang, S., & Li, P. (2024). Improvement of Multimodal Emotion Recognition Based on Temporal-Aware Bi-Direction Multi-Scale Network and Multi-Head Attention Mechanisms. Applied Sciences, 14(8), 3276. https://doi.org/10.3390/app14083276

Xie, B., Sidulova, M., & Park, C. H. (2021). Robust Multimodal Emotion Recognition From Conversation With Transformer-Based Crossmodality Fusion. Sensors, 21(14), 4913. https://doi.org/10.3390/s21144913

Yan, L., Gašević, D., Echeverría, V., Zhao, L., Jin, Y., Li, X., & Martínez-Maldonado, R. (2024). In Sync or Out of Sync? Understanding Stress and Learning Performance in Collaborative Healthcare Simulations Through Physiological Synchrony and Arousal. https://doi.org/10.21203/rs.3.rs-4855446/v1

Yang, L. (2025). Monitoring of Student Classroom Learning Status based on Hybrid ViT-YOLOv5s. 2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 1–6. https://doi.org/10.1109/ICDCECE65353.2025.11035485

Yuvaraj, R., Mittal, R., Prince, A. A., & Huang, J. S. (2025). Affective Computing for Learning in Education: A Systematic Review and Bibliometric Analysis. Education Sciences, 15(1), 65. https://doi.org/10.3390/educsci15010065

Zhan, X. (2024). Research on the Application of English Short Essay Reading Emotional Analysis in Online English Teaching Under IoT Scenario. Internet Technology Letters, 8(2). https://doi.org/10.1002/itl2.535

Zhang, H., Peng, Y., & Liu, Y. (2025). Multimodal fusion for real-time classroom engagement assessment using YOLOv9 and DeepFace. The Visual Computer, 41(14), 12325–12337. https://doi.org/10.1007/s00371-025-04159-2

Zhang, N., & Leong, W. Y. (2025). Intelligent emotional computing with deep convolutional neural networks: Multimodal feature analysis and application in smart learning environments. Eurasia Journal of Mathematics, Science and Technology Education, 21(8), em2680. https://doi.org/10.29333/ejmste/16661

Zhang, S., Huang, Z., Paudel, D. P., & Gool, L. V. (2020). Facial Emotion Recognition With Noisy Multi-Task Annotations. https://doi.org/10.48550/arxiv.2010.09849

Zhang, X., Yan, C., & Miao-miao, S. (2022). Facial Expression Recognition Based on Improved Mobile-Former. https://doi.org/10.21203/rs.3.rs-2195625/v1

Zhang, Z., Zhang, S., Ni, D., Wei, Z., Yang, K., Jin, S., Huang, G., Liang, Z., Zhang, L., Li, L., Ding, H., Zhang, Z., & Wang, J. (2024). Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data. Sensors, 24(12), 3714. https://doi.org/10.3390/s24123714

Zhao, W., & Qiu, L. (2025). Emotion recognition and interaction of smart education environment screen based on deep learning networks. Journal of Intelligent Systems, 34(1). https://doi.org/10.1515/jisys-2024-0082

Zhao, Y., Guo, M., Sun, X., Chen, X., & Zhao, F. (2023). Attention-based Sensor Fusion for Emotion Recognition From Human Motion by Combining Convolutional Neural Network and Weighted Kernel Support Vector Machine and Using Inertial Measurement Unit Signals. Iet Signal Processing, 17(4). https://doi.org/10.1049/sil2.12201

Zheng, K., Yang, D., Liu, J., & Cui, J. (2020). Recognition of Teachers' Facial Expression Intensity Based on Convolutional Neural Network and Attention Mechanism. IEEE Access, 8, 226437–226444. https://doi.org/10.1109/ACCESS.2020.3046225

Zhu, J., Zhang, X., Wang, R., Wang, M., Chen, P., Cheng, L., Wu, Z., Wang, Y., Liu, Q., & Liu, M. (2022). A Heterogeneously Integrated Spiking Neuron Array for Multimode-Fused Perception and Object Classification. Advanced Materials, 34(24). https://doi.org/10.1002/adma.202200481