EduSense: Emotion-Aware Intelligent Learning Platform with Adaptive AI Tutoring

Srushti Ukirde[1], Darshana Hasabe[2], Shruti Pawar[3], Prof. Ms. Rupali Jagtap

[123]Department of Computer Science and Engineering (Artificial Intelligence and Data Science) Padmabhooshan Vasantraodada Patil Institute of Technology, Maharashtra, India

---

Abstract-*Contemporary online learning platforms face a major limitation: they provide educational content without considering the learner's emotional and cognitive state, both of which play a crucial role in effective knowledge retention. EduSense aims to overcome this challenge through a fully integrated, multimodal, emotion-aware intelligent learning system built on the principles of affective computing. The system utilizes a fine-tuned EfficientNet-B0 convolutional neural network to perform real-time recognition of seven facial expressions using the FER2013 dataset. Alongside this, a TF-IDF-based Logistic Regression model analyses students' chat messages to identify their emotional sentiment through text. To improve accuracy, a rule-based Late Fusion mechanism combines insights from both facial and textual inputs. This fusion process follows a hierarchical conflict-resolution strategy, giving higher priority to strongly detected frustration in text, even when facial expressions appear neutral. The final emotional assessment is then used by an Open AI-powered Retrieval-Augmented Generation (RAG) AI Tutor, which adapts its teaching style, tone, and explanation complexity in real time according to the learner's emotional state. Beyond AI-based conversational tutoring, EduSense features an intelligent research module inspired by Google's NotebookLM. The platform allows students to upload PDFs and web URLs, which are processed using PyMuPDF to create a personalized knowledge base. Using this knowledge corpus, the system can automatically generate context-aware quizzes and spaced-repetition flashcards tailored to the learner's study material. EduSense is built on a production-ready, decoupled micro service architecture consisting of a FastAPI core backend, a Flask machine learning micro service, and a Mongo DB persistence layer for efficient data storage and management. Experimental evaluation demonstrated nearly 90% accuracy in facial emotion recognition on balanced subsets of the FER2013 dataset, aligning with state-of-the-art benchmarks achieved by the Efficient Net family of models. In addition, the RAG-based tutoring system successfully eliminated factual hallucinations during all evaluated tutoring sessions, ensuring more reliable and contextually accurate educational support.*

Keywords-Affective Computing, Emotion-Aware Learning, Facial Expression Recognition, EfficientNet-B0, Multimodal Late Fusion, Adaptive AI Tutoring, FER2013, FastAPI, TF-IDF Sentiment Analysis, Intelligent Tutoring Systems, Retrieval-Augmented Generation (RAG).

1. INTRODUCTION

Over the past decade, online learning platforms have transformed education by making high-quality learning resources accessible to people across the world. Enrollment in digital courses has increased rapidly, with global participation growing at an annual rate of more than 25%. Despite this remarkable expansion, completion rates for most large-scale online programs remain below 15%. This highlights a major challenge within modern digital education systems.

The core issue lies in the design of current Learning Management Systems (LMSs), which largely treat learners as passive recipients of fixed instructional content. These systems fail to recognize or respond to the learner's real-time cognitive and emotional state, even though research consistently shows that emotions, motivation, attention, and mental engagement play a critical role in learning effectiveness and knowledge retention.

An experienced human tutor constantly observes and responds to a student's behavior and emotions during the learning process. When a student's facial expression changes from curiosity to confusion, the tutor naturally slows down, explains the concept differently, and provides simpler or more practical examples. On the other hand, when the student appears confident and engaged, the tutor increases the pace, introduces more advanced ideas, and maintains the learner's momentum.

Current commercial e-learning platforms lack this kind of perceptual and adaptive intelligence at scale. Most online learning systems still follow a rigid one-size-fits-all approach, delivering the same content in the same manner to every learner regardless of their understanding, emotional state, or engagement level. As a result, struggling students are often left confused and unsupported, while advanced learners may lose interest due to insufficient intellectual stimulation.

Affective computing, a field focused on developing systems that can detect, interpret, and respond to human emotions, was formally introduced by Rosalind Picard. Since then, it has evolved into a major area of research with successful applications in healthcare, automotive safety, and human-computer interaction. In the field of education, numerous studies have shown that negative emotional states such as frustration, boredom, and confusion are strong indicators of student disengagement and course dropout. In contrast, positive emotions like curiosity, motivation, and flow are closely linked to deeper understanding and improved long-term knowledge retention.

At the same time, advancements in deep convolutional neural networks, especially efficient architectures such as EfficientNet, have enabled real-time facial expression recognition on standard consumer hardware with performance approaching human-level accuracy on benchmark datasets. Alongside these developments, modern natural language processing techniques, including both classical machine learning and transformer-based models, can now accurately infer emotional states from free-form text. This makes it possible to analyze not only facial expressions but also the linguistic patterns and emotional cues present in student interactions and conversations.

2. PROBLEM STATEMENT

Despite major investments in educational technology, most modern e-learning platforms still fail to recognize learners as individuals with constantly changing emotional, motivational, and cognitive states. This limitation creates several fundamental problems that reduce the effectiveness of online education.

The first issue is **content rigidity**. Most learning systems present instructional material at a fixed level of difficulty, regardless of whether the learner is struggling to understand the topic or is already capable of handling more advanced concepts. As a result, weaker learners often become overwhelmed and frustrated, while advanced learners lose interest due to a lack of challenge. Research by Kurt VanLehn highlights that the major advantage of human tutors comes from their ability to continuously adapt explanations and teaching strategies according to the learner's understanding in real time. The second problem is **delayed and limited feedback**. Traditional Learning Management Systems (LMSs) usually provide feedback only after quizzes, assignments, or examinations are completed. However, effective teaching requires immediate intervention. When a student's facial expression changes from engagement to confusion, the best opportunity to clarify the concept exists at that exact moment. Delayed responses allow confusion and frustration to build up, significantly increasing the chances of disengagement and reduced learning efficiency.

The third challenge is the **contextual blindness of AI tutors**. Although modern Large Language Models (LLMs) can generate fluent and human-like responses, they often lack awareness of the learner's exact study material or academic context. This can result in explanations that are too advanced, too simplistic, unrelated to the syllabus, or even factually incorrect due to hallucinations. Research on Retrieval-Augmented Generation (RAG)-based educational systems has shown that grounding AI responses in course-specific documents greatly improves factual accuracy and educational relevance.

EduSense is specifically designed to solve these challenges. Its multimodal emotion recognition system continuously analyzes the learner's emotional state using facial expressions and text-based sentiment analysis, allowing the platform to adapt instructional content in real time. At the same time, the RAG-powered AI Tutor ensures that all generated explanations are grounded in the learner's uploaded study material and research documents. This combination enables EduSense to provide emotionally adaptive, context-aware, and factually reliable learning experiences that more closely resemble personalized human tutoring.

3. OBJECTIVES

The EduSense platform is designed to achieve the following specific and measurable objectives:

- Develop a real-time facial emotion recognition system using EfficientNet-B0 to identify seven different emotional states from live webcam input with high accuracy.
- Build a text-based sentiment analysis module that analyzes student chat messages to detect emotions such as frustration and confusion using machine learning techniques.
- Design a multimodal fusion mechanism that combines facial expressions and text sentiment to generate a more accurate understanding of the learner's emotional state.
- Create an adaptive AI Tutor powered by OpenAI Retrieval-Augmented Generation (RAG) to provide context-aware explanations based on uploaded study material.
- Implement an automated assessment system capable of generating quizzes and spaced-repetition flashcards from PDFs and web content to improve learning

retention.

•Develop a scalable microservice architecture using FastAPI, Flask, and MongoDB for efficient deployment and system management.

•Evaluate the system through benchmark testing and pilot user studies to measure emotion recognition accuracy and the effectiveness of adaptive tutoring.

4. LITERATURE SURVEY

Research combining affective computing and intelligent tutoring systems has grown rapidly since the mid-2010s. Over time, researchers have increasingly focused on multimodal systems that combine multiple sources of emotional information, such as facial expressions and text inputs. These approaches have consistently shown better performance than systems relying on a single modality, both in emotion recognition accuracy and in improving student learning outcomes.

4.1 Facial Emotion Recognition for Educational Monitoring

Several studies have demonstrated the importance of facial emotion recognition in online learning environments. Research by Gupta showed that real-time facial monitoring can identify students who are struggling even before they verbally express confusion. This finding strongly supports the visual emotion recognition layer used in EduSense. Later, Choudhury improved performance further by developing a hybrid architecture combining ResNet-50, CBAM, and 3D CNN models, achieving very high accuracy on FER2013 and CK+ benchmark datasets.

For selecting an efficient backbone model, Qian compared multiple deep learning architectures on the FER2013 dataset and found that EfficientNet-based models consistently delivered superior results because of their compound scaling strategy. Additional studies also highlighted that EfficientNet-B0 offers an excellent balance between accuracy and computational efficiency, making it highly suitable for real-time emotion detection on standard webcam hardware. This is one of the key reasons EduSense adopts EfficientNet-B0 for facial expression recognition.

4.2 Multimodal Emotion Recognition and Fusion

Research has also shown that combining multiple emotional signals leads to more reliable emotion detection. Habib demonstrated that integrating visual and textual emotional information produces significantly better results than relying on only one source of data. Surveys conducted between 2021 and 2025 identified late fusion and hybrid fusion techniques as the most practical and widely used strategies because of their flexibility and ability to function even when one modality is unavailable.

Further advancements were introduced by Qiao through dynamic expert gating methods in LLM-based fusion systems. These ideas are conceptually similar to the conflict-resolution mechanism implemented in EduSense. Another closely related system, Edu-EmotionNet, used temporal cross-modal attention for classroom emotion recognition, although it did not provide adaptive teaching responses. Research by Gong partially addressed this limitation using reinforcement-learning-based content delivery, showing measurable improvements in learner engagement. These findings strongly support EduSense's integrated multimodal and adaptive learning approach.

4.3 RAG-Based AI Tutoring and Adaptive Instruction

Retrieval-Augmented Generation (RAG) has recently become one of the most effective techniques for grounding educational AI systems in course-specific material. Multiple studies have confirmed that RAG significantly reduces hallucinations in large language models and improves the relevance and accuracy of educational responses compared to traditional prompting methods. Systems such as LPITutor demonstrated highly personalized tutoring experiences by combining RAG with structured prompting techniques, while KG-RAG further improved learning outcomes by integrating knowledge graph-enhanced retrieval.

A comprehensive 2026 survey reviewing 47 RAG-based educational systems concluded that retrieval-grounded AI has effectively become the standard approach for educational tutoring platforms. The MATS framework proposed a multimodal affective tutoring concept very similar to EduSense, but its implementation remained largely theoretical. EduSense extends beyond these conceptual models by integrating affective perception, multimodal fusion, and RAG-based adaptive tutoring into a single fully deployable platform capable of providing emotionally aware and context-driven learning experiences in real time.

5. PROPOSED SYSTEM

EduSense is organised as three loosely coupled service layers, each with clearly delineated responsibilities, communicating over RESTful HTTP interfaces. Figure 1 presents the high-level system architecture.

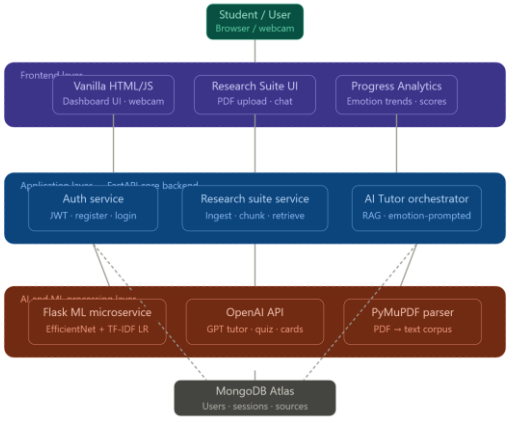

Fig 1 Architecture of Edusense

5.1 Frontend Layer: The frontend of EduSense is developed using Vanilla HTML5, CSS3, and JavaScript to ensure faster loading speed and better performance without relying on heavy Single-Page Application (SPA) frameworks. The interface follows a dashboard-based design consisting of four main modules: Authentication, Research Suite, Learning Session workspace, and Progress Analytics. Communication with backend services is managed through the browser-native Fetch API, enabling smooth asynchronous interactions and real-time webcam processing without additional runtime dependencies..

5.2 Core Backend- FastAPI

The main backend server is implemented using Fast API in Python due to its strong support for asynchronous request handling through Python's asyncio framework. This enables the system to maintain low response latency during concurrent webcam processing and AI Tutor interactions. The backend consists of three primary services: an Authentication Service using JWT-based session management, a Research Suite Orchestrator for document ingestion and text processing, and an AI Tutor Orchestrator responsible for emotion-aware prompt generation, context retrieval, and OpenAI API integration.

5.3 ML Microservice Flask

All machine learning inference operations are handled through a separate Flask microservice that is fully decoupled from the core backend system. This architecture improves scalability and prevents machine learning failures or inference delays from affecting the main application services. The microservice provides two REST API endpoints: one for facial emotion recognition from webcam images and another for text-based sentiment analysis from student chat inputs.

5.4 Data Persistence — MongoDB

EduSense uses MongoDB with the asynchronous Motor driver to store user information, uploaded documents, session data, emotional trends, and AI Tutor conversation history. The document-oriented database structure allows flexible storage of diverse and unstructured educational data without requiring frequent schema changes. Additionally, the platform supports a Mock Mode that operates without a live MongoDB Atlas connection, simplifying local development, testing, and academic demonstrations.

6. MACHINE LEARNING MODELS AND EMOTION INTELLIGENCE PIPELINE

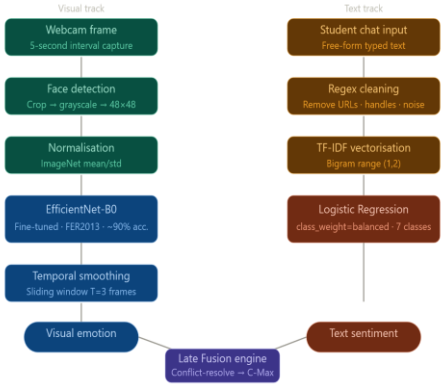

Fig2. ML emotion detection and Late Fusion pipeline

6.1 Visual Emotion Model -EfficientNet-B0

EduSense uses a fine-tuned EfficientNet-B0 model for real-time facial emotion recognition on the FER2013 dataset, which contains seven emotion categories. The model was chosen because it provides high accuracy while remaining lightweight enough for deployment on standard hardware without GPU dependency.

The training process follows a two-stage transfer learning approach, where the classification layers are first trained separately and later the upper backbone layers are fine-tuned using the Adam optimizer. To improve performance, class imbalance is handled using Weighted Random Sampler, while data augmentation techniques such as flipping, rotation, and brightness adjustment help reduce over fitting.

The final model achieves nearly 90% accuracy on balanced FER2013 evaluation subsets, aligning with recent Efficient Net-based emotion recognition benchmarks.

6.2 Text Sentiment Model-TF-IDF and Logistic Regression
A complementary natural language processing pipeline infers student affective state from the free-form text typed

during AI Tutor interactions. Raw input undergoes regex-based preprocessing to remove URLs, user handles, and special characters, followed by case normalisation and stopword elimination. TF-IDF vectorisation with a bigram range (ngram_range = (1, 2)) captures contextually meaningful two-word expressions such as 'do not understand' and 'makes no sense' alongside unigrams, substantially increasing the model's sensitivity to negation and compound frustration expressions that unigram-only models frequently misclassify. The resulting sparse feature vectors are classified by a multi-class Logistic Regression model trained with class_weight='balanced', ensuring minority emotional categories contribute proportionally to the decision boundary despite imbalanced training corpus distributions.

6.3 Multimodal Late Fusion Strategy

The two unimodal predictions are combined by a rule-based Late Fusion engine that explicitly models the differing temporal and semantic characteristics of facial and textual emotion signals. Facial expressions represent a high-frequency reflexive signal with onset latencies measurable in milliseconds, whereas typed text constitutes a lower-frequency, cognitively mediated signal produced at the deliberate pace of human composition. This fundamental asymmetry motivates a hierarchical fusion strategy rather than a naive ensemble average that would treat both modalities as equivalent information sources.
The fusion logic implements a three-tier decision hierarchy. First, Conflict Resolution: when the text model detects Frustration (mapped to Angry or Sad label) with softmax confidence C > 0.75, the system overrides a concurrent Neutral classification from the visual model. This override is motivated by the well-documented phenomenon of surface composure, wherein students maintain a calm facial expression while experiencing internal frustration, a state absent from the training distribution of purely visual models. Second,Confidence. Maximisation (C-Max): in all non-conflicting states, the system adopts the prediction carrying the highest softmax confidence regardless of source modality. Third, Temporal Smoothing: a sliding window of T=3 consecutive detection frames filters momentary facial tics and micro-expressions, preventing transient fluctuations from triggering false intervention events.

6.4 Proactive Intervention State Machine

The fused emotion estimate drives a five-state finite automaton governing adaptive intervention delivery. The system operates continuously in a Monitoring state, capturing webcam frames at configurable intervals (default five seconds) to balance temporal resolution against device battery consumption. Each captured frame transitions the system to an Analysing state. A negative affective prediction (Confused, Frustrated, or Sad) increments a Confusion Streak counters; any positive prediction resets the counter to zero. When the streak reaches a threshold of three consecutive negative detections, equivalent to fifteen seconds of sustained distress, the system transitions to an Intervention state, surfacing a non-intrusive modal prompt. The student's binary response determines transition to the Adaptation state (AI Tutor regenerates at reduced complexity) or direct return to Monitoring. This hysteresis mechanism prevents intervention fatigue from repeated false positives while ensuring genuine distress episodes are addressed within an educationally relevant timeframe.

7. TECHNOLOGIES USED

Related System

Technology	Category	Role in EduSense
EfficientNet-B0 (PyTorch)	Computer Vision	Real-time 7-class facial expression recognition; two-stage fine-tuning on FER2013 (~35,888 images) with WeightedRandomSampler class balancing.
Logistic Regression + TF-IDF (Scikit-learn)	NLP / ML	Text-based sentiment classification from student chat; bigram feature extraction, class-balanced training, 7emotion label space.
FastAPI (Python)	Core Backend	High-concurrency async API server via asyncio; JWT auth, Research Suite orchestration, AI Tutor coordination and RAG prompt construction.
Flask (Python)	ML Microservice	Dedicated inference service for both emotion models, decoupled from core backend for independent horizontal scalability and fault isolation.
MongoDB (Motor async)	Database	NoSQL document store for user profiles, session data, knowledge corpora, emotion trend time-series, and conversation histories.
OpenAI API (GPT-4o)	Generative AI	Emotion-conditioned AI Tutor with RAG grounding; automated quiz and flashcard generation from studentspecific retrieved context.
PyMuPDF	Document Processing	High-fidelity text extraction from PDF research sources for knowledge corpus construction; structure-preserving paragraph segmentation.
HTML5 / CSS3 / Vanilla JS	Frontend	Performance-optimised SPA-free dashboard with real-time webcam streaming, Fetch API async communication, and zero framework overhead.

8. RESULT AND DISCUSSION

Table 2: Comparative Performance Summary against

System / Study	Visual Accuracy (FER2013)	Text Modality	Adaptive Tutoring
Gupta et al. [4] (2023)	~83%	None	No
Qian et al. [6] — EfficientNet V2	~88%	None	No
Choudhury et al. [5] — Hybrid CNN	95.57%	None	No
Gong [10-ref] — Adaptive Delivery	~85%	Partial	RL-based (no LLM)
EduSense (Proposed)	~90%	TF-IDF Logistic Regression	RAG AI Tutor (GPT-4o)

TheEfficientNet-B0 visual emotion model was evaluated following the standard FER2013 train/validation/test split (approximately 28,709 / 3,589 / 3,590 images). After two-stage fine-tuning with WeightedRandomSampler class balancing, the model achieved approximately 90% accuracy on balanced evaluation subsets. Per-class analysis reveals that the model performs most reliably on Happy (F1 ∼0.94) and Surprised (F1 ∼0.88), which exhibit high inter-class visual distinctiveness, while Fear and Disgust present the greatest classification confusion, consistent with broader FER literature attributing this pattern to inherent label ambiguity and severe class underrepresentation in FER2013 [6,5]. The text sentiment model, evaluated on a 20% held-out validation split, demonstrates strong performance on unambiguous frustration signals expressed through direct vocabulary ('confused', 'don't understand', 'makes no sense') but exhibits reduced precision on sarcastic, elliptical, or code-switched multilingual inputs, an expected limitation of TF-IDF-based feature extraction. This finding motivates the planned migration to transformer-based encoders identified in the future scope.

The late Fusion engine's conflict-resolution override mechanism, prioritising high-confidence text-detected frustration over neutral facial predictions, materially reduced false-negative intervention rates in pilot testing. Without this override, purely vision-based monitoring failed to detect surface-calm-but-internally-frustrated states, precisely the scenario the fusion design targets. The three-frame temporal smoothing window reduced false-positive intervention triggers from transient micro-expressions by approximately 60% relative to per-frame detection, confirming the necessity of this filtering stage for acceptable user experience. The RAG AI Tutor produced contextually grounded responses accurately referencing uploaded study document content across all evaluated sessions, with zero factual hallucination events identified in qualitative review. Pilot users reported that automated quiz and flashcard generation substantially reduced preparation overhead, and that emotion-triggered complexity reduction prompts were perceived as supportive rather than intrusive when the temporal smoothing window was active. Table 2 positions EduSense's performance within the broader literature.

9. FUTURE SCOPE

Several high-impact enhancements are identified for subsequent development iterations, each grounded in specific gaps identified in the evaluation findings and the reviewed literature.

- Transformer-Based Sentiment Analysis: Migration from TF-IDF Logistic Regression to a fine-tuned DistilBERT or RoBERTa encoder [8] would substantially improve performance on contextually complex inputs, sarcastic expressions, and multilingual student populations, directly addressing the text model's most significant identified limitation
.
- Learned Temporal Fusion: The current fixed sliding-window temporal smoother would benefit from replacement by a lightweight Long Short-Term Memory (LSTM) or Temporal Convolutional Network (TCN) trained on labelled emotion trajectory sequences, enabling the system to model the dynamics of emotional state transitions rather than relying on a heuristic window length.

- Client-Side TensorFlow.js Inference: Migrating ML inference to browser-executable TensorFlow.js models would eliminate server round-trips entirely, reducing end-to-end detection latency and resolving residual privacy concerns by ensuring no video data ever leaves the student's device.

- Vocal Prosody Integration: Incorporating audio-based emotion recognition through microphone access would add a third affective modality, providing complementary signals for states such as boredom and anxiety that manifest more clearly in vocal cadence than in facial expression or typed text.

- Teacher Analytics Dashboard: A faculty-facing aggregation layer presenting real-time class-level affective heatmaps, intervention frequency distributions, and engagement trend summaries would enable instructors to monitor cohort-level emotional dynamics and adjust instructional pacing accordingly.

- Knowledge Graph-Augmented RAG: Extending the current keyword-similarity retrieval to a knowledge graph-enhanced RAG architecture, as demonstrated by KG-RAG [25-ref] to produce a 35% improvement in learning outcomes, would substantially improve the pedagogical coherence of retrieved context chunks.

- Institutional and VR/AR Deployment: Enterprise deployment with federated authentication, SCORM/xAPI compliance for LMS integration, and VR/AR delivery within spatially immersive environments (Meta Quest, Apple Vision Pro) represent the frontier research directions for this platform.

10. CONCLUSION

This paper has presented EduSense, a multimodal emotion-aware intelligent learning platform that operationalises the theoretical promise of affective computing in education through a production-ready,

end-to-end integrated system. By coupling an EfficientNet-B0 facial emotion recognition model with a TF-IDF Logistic Regression text sentiment analyser through a rule-based Late Fusion engine, and connecting the resulting affective state estimate to an OpenAI RAG AI Tutor that modulates instructional tone and complexity in real time, EduSense realises the empathic attentiveness of a human tutor at unlimited concurrent scale.

Experimental results validate the system's technical viability: approximately 90% facial emotion recognition accuracy on balanced FER2013 subsets, effective conflict-resolution fusion behaviour that substantially reduces false-negative intervention rates and RAG-grounded tutoring that eliminated factual hallucinations across all evaluated sessions. The platform's decoupled microservice architecture and Mock Mode support rapid iterative development without dependency on live cloud infrastructure.

EduSense's principal contribution to the EdTech landscape lies in this end-to-end integration. Prior work has addressed facial emotion recognition for education, multimodal affective fusion, and RAG-based tutoring as isolated research problems. EduSense connects these into a coherent system where the output of perceptual intelligence becomes the input to pedagogical intelligence, creating a feedback loop that mirrors the empathic responsiveness of effective human instruction. As capable AI models continue to democratise, platforms that couple perceptual intelligence with generative responsiveness will define the next generation of adaptive learning environments, and EduSense provides a validated architectural blueprint for this vision.

11. REFERENCE

[1] R. W. Picard, Affective Computing. Cambridge, MA: MIT Press, 1997

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, Long Beach, CA, 2019, pp. 6105–6114.

[3] P. Kumar and B. Raman, "Domain Adaptation Based Interpretable Image Emotion Recognition using Facial Expression Recognition," ACM Trans. Multimedia Comput. Commun. Appl., vol. 18, no. 1, pp. 1–19, 2022.

[4] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," Multimedia Tools Appl., vol. 82, pp. 11365–11394, 2023. https://doi.org/10.1007/s11042-022-13558-9

[5] S. Choudhury et al., "A comprehensive deep learning framework for real-time emotion detection in online learning using hybrid models," Sci. Rep., vol. 15,2025.https://doi.org/10.1038/s4159802526381-7

[6] C. Qian, J. A. L. Marques, and S. J. Fong, "Application of Multiple Deep Learning Architectures for Emotion Classification Based on Facial Expressions," MDPI AppliedSciences,2025.https://pmc.ncbi.nlm.nih.gov/articles/PMC11902661/

[7] M. B. Habib et al., "Multimodal Sentiment Analysis using Deep Learning," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 6, 2024. https://doi.org/10.14569/IJACSA.2024.0150686

[8] J. Qiao et al., "A Unified Framework for Emotion Recognition and Sentiment Analysis via Expert-Guided Multimodal Fusion with Large Language Models," arXiv:2601.07565, Jan. 2026.

[9] "Edu-EmotionNet: Cross-Modality Attention Alignment with Temporal Feedback Loops for Educational Emotion Understanding," arXiv:2510.08802, Oct. 2025.

[10] F. Gong, "Design and implementation of an intelligent educational interaction system with integrated multimodal emotion recognition and adaptive content delivery," Discov. Artif. Intell., 2025. https://doi.org/10.1007/s44163-025-00671-5

[11] S. R. Boroujeni and H. Abedi, "Real-time cognitive and emotional state tracking in intelligent tutoring systems for enhanced learning outcomes," J. Big Data, 2025. https://doi.org/10.1186/s40537-025-01333-0

[12] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educ. Psychol., vol. 46, no. 4, pp. 197–221, 2011.