EduSense: Evaluating AI for Augmenting Live Classroom Teaching

Pranav Agrawal
Indian Institute of Science Education and Research, Bhopal

April 2025

Abstract

Current approaches to augmenting classroom teaching with AI rely heavily on post-hoc analysis or complex systems that disrupt live instruction. A lightweight real-time AI framework designed to autonomously capture lecture audio and generate structured educational content would allow students to get personalized learning content without altering the classroom flow. EduSense was created for this purpose and was evaluated on five live lectures with 17 undergraduate students, the system achieves high transcription fidelity (WER: 2.8-5.0%) and strong semantic alignment with human references (BERTScore F1: 0.90+), while maintaining positive user satisfaction across learning and usability dimensions (User Rating $>$ 4.2). The results demonstrate that real-time AI integration into live classrooms is both technically feasible and educationally valuable, laying the groundwork for scalable AI-augmented teaching systems.

1 Introduction

Artificial Intelligence (AI) and the Internet of Things (IoT) are ubiquitous technologies and have made remarkable strides in transforming various sectors, and education is no exception. The traditional classroom is effective in fostering engagement, but often lacks a mechanism for reinforcement of knowledge through personalized teaching. EduSense addresses this challenge by integrating classroom sensors with real-time AI processing to create dynamic educational content. Through automatic lecture transcription, summarization, and quiz generation, EduSense bridges the gap between in-person instruction and digital enhancement, ensuring that no learning moment is lost.

Despite the availability of AI-powered educational tools, most systems are designed for asynchronous or online contexts. Studies from 2018 onwards highlight significant progress in offline summarization and adaptive feedback systems; however, few have explored the integration of such systems into real-time, in-class scenarios. Challenges such as latency, contextual understanding, and usability in dynamic classroom environments remain largely unaddressed. There is a pressing need for lightweight, deployable systems that operate effectively within live teaching settings without disrupting instructional flow.

Our contribution lies in the design, deployment, and evaluation of EduSense—a compact, real-time AI system tailored for live classrooms. Unlike prior solutions, EduSense was tested in real-world lectures with 17 undergraduate students. We developed a custom mobile application, integrated ASR and GPT-based NLP models, and collected both subjective (survey-based) and objective (WER, BERTScore) metrics to evaluate its performance. This report details the system architecture, methodology, and results, ultimately demonstrating EduSense's feasibility and potential impact in augmenting conventional teaching methods.

2 Related Work

Holstein et al. [2] introduced an *AI Teacher Dashboard* that surfaced real-time analytics on student misconceptions during computer science labs. Their field study showed improved teacher awareness, yet the system merely displayed alerts; it neither generated learning materials nor operated in lecture-style settings. This underscores a gap that EduSense fills by offering in-situ AI-generated content delivery.

Khosravi et al. [3] developed an adaptive-feedback engine for large-scale online courses, tailoring hints using deep knowledge tracing. Although they report improved learning outcomes, the platform was limited to asynchronous, web-based interactions. In contrast, EduSense focuses on real-time, face-to-face classroom augmentation.

Chen et al. [1] proposed *LectureSumm*, an offline summarization pipeline that segments recorded lectures and applies BERT-based ranking to produce textual digests. While effective (ROUGE-2 0.35), their approach lacks the real-time delivery necessary for live teaching contexts, which EduSense provides.

Nguyen et al. [4] benchmarked ASR systems (Whisper, Kaldi, Google Speech) on a 120-hour university lecture corpus. Whisper achieved a WER of 7.1%, revealing challenges with noise and accent variation—issues that EduSense mitigates using custom microphone placement and context-specific preprocessing.

Winkler and Söllner [5] validated a usability and engagement survey for mobile learning apps (Cronbach's $\alpha > 0.9$). Their framework forms the basis of EduSense's subjective evaluation instrument.

Together, these studies span classroom analytics, adaptive learning, ASR, summarization, and usability. EduSense synthesizes these threads into a unified, real-time educational AI platform.

Table 1: Comparison of Related Work (2018–2024)

Paper	Real-Time	Classroom	Content Gen.	Evaluation
Holstein et al. (2019)	✓	×	×	Teacher perception
Khosravi et al. (2021)	×	×	Adaptive feedback	Survey
Chen et al. (2023)	×	✓	✓	ROUGE scores
Nguyen et al. (2022)	✓	✓	×	WER, ASR accuracy
Winkler & Söllner (2020)	×	✓	×	Survey design

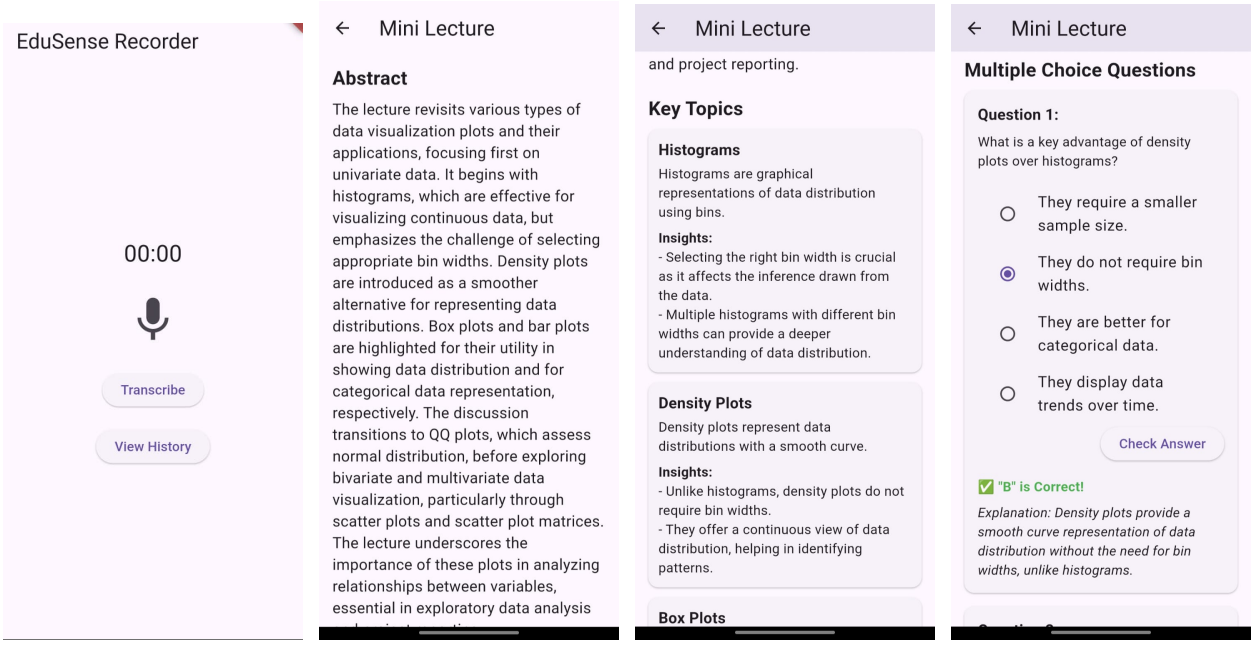
3 Methodology

EduSense was tested across five lectures on the Internet of Things (IoT) with 17 undergraduate students. The system architecture (Figure 2) integrates classroom audio capture, Whisper ASR transcription, GPT-based NLP processing, and a custom-built mobile application. Each lecture was recorded in real time, transcribed, and converted into mini-lectures consisting of abstract summaries, key concepts, and MCQs.

Data Collection and Evaluation: Manual reference materials were prepared for BERTScore evaluation. Surveys were administered post-lecture to capture user experience, and technical logs were collected for performance benchmarking.

A custom-built mobile application was developed as the primary interface between the EduSense backend and classroom participants. The application is designed to operate seamlessly within live classroom environments and emphasizes low latency, usability, and robust performance. Its key functionalities include:

- **Capture** – continuously records classroom audio via built-in or external microphones.
- **Transcribe** – streams the audio to Whisper ASR and returns near-real-time text.
- **Generate** – feeds the transcript to GPT-based services to create summaries, key points, and quizzes.
- **Display** – presents the generated mini-lecture and quizzes in-app while logging student responses.
- **Sync** – caches data offline and synchronizes all content and feedback once connectivity is restored.



Recorder | Abstract | Key-topics | MCQ view

Figure 1: EduSense app workflow (left → right): recording audio, displaying the generated abstract, presenting key-topic cards, and providing interactive MCQs with instant feedback.

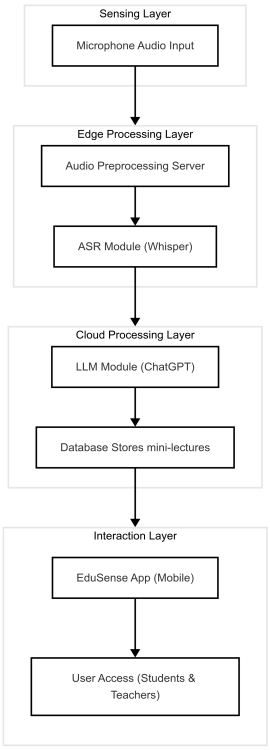


Figure 2: EduSense System Architecture

3.1 Evaluation Framework

Metric	**Type**	**Method**	**Tools Used**	**Purpose**
ASR Accuracy	Technical	WER	Whisper, Manual GT	Transcription Fidelity
Semantic Alignment	Technical	BERTScore (F1)	Transformers	Content Fidelity
User Satisfaction	Subjective	Likert Survey	Google Forms	Pedagogical Utility
System Usability	Subjective	Likert Survey	Google Forms	UX Assessment

Table 2: Evaluation Framework for EduSense

To comprehensively assess the effectiveness of EduSense in live classroom environments, a multi-dimensional evaluation framework was established. This framework integrates both technical and pedagogical perspectives, ensuring that system performance is measured not only in terms of algorithmic accuracy but also educational value. Technical metrics such as Word Error Rate (WER) and BERTScore assess the fidelity of transcription and semantic alignment between AI-generated content and original lectures. These were chosen to quantify how accurately the system captures and represents instructional material. Subjective metrics, collected via structured Likert-scale surveys, gauge students' perceived satisfaction, usefulness of the content, and overall usability of the platform. By combining objective and subjective evaluation methods, the framework captures both machine-level performance and human-centered educational impact—key for validating real-world deployment of AI tools in classrooms.

4 Results and Discussion

4.1 User Survey Results

Survey Question	Mean Rating (1–5)
Lecture content accurately reflected	4.1
Key concepts effectively distilled	4.2
Content presentation aided understanding	4.2
Enhanced comprehension	4.1
User-friendly interface	4.3
System reliability	4.5
Content useful for missed lecture	4.4

Table 3: Mean Likert Ratings from Students

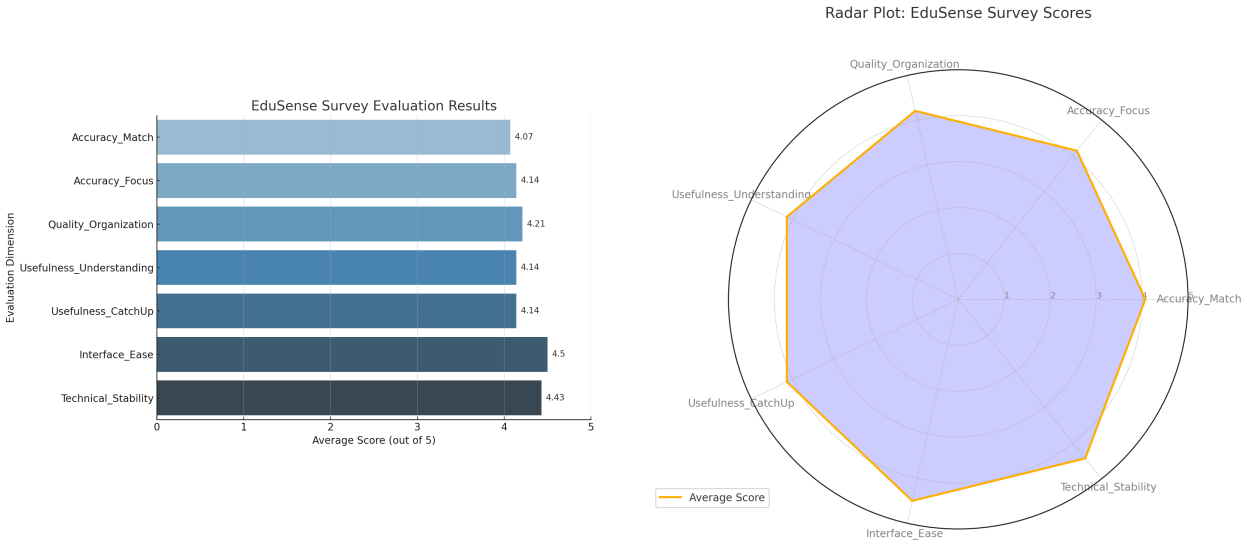


Figure 3: Bar and radar plots summarizing student feedback. The consistently high scores indicate that EduSense was well-received by learners.

4.2 ASR Performance

Lecture Instance	Word Error Rate (WER %)	Average BERTScore (F1)
Lecture 1	4.9	0.94
Lecture 2	5.0	0.95
Lecture 3	2.8	0.97
Lecture 4	2.9	0.97
Lecture 5	3.7	0.95
Range	2.8 - 5.0	0.94 - 0.97

Table 4: ASR performance across lecture recordings: WER and BERTScore.

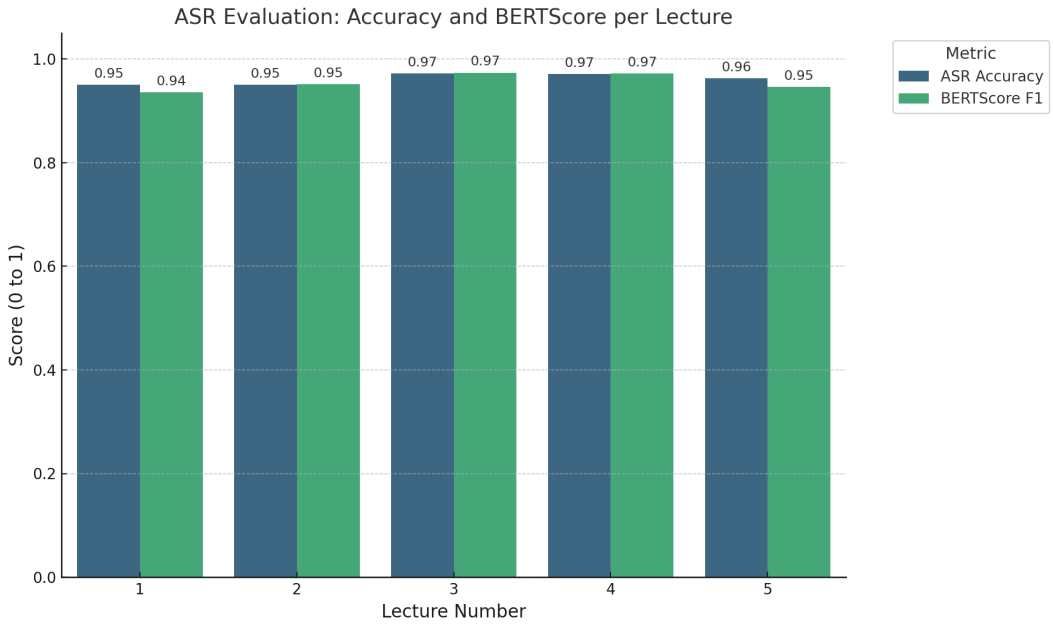


Figure 4: ASR Word Accuracy and BERTScore across five lectures. The high scores across both metrics indicate consistent performance even in variable classroom conditions.

4.3 Semantic Fidelity of AI-Generated Content

Lecture Instance	Abstract Similarity	Key Topics Similarity	MCQ Similarity
Lecture 1	0.935	0.912	0.890
Lecture 2	0.942	0.914	0.902
Lecture 3	0.928	0.906	0.895
Lecture 4	0.947	0.928	0.913
Lecture 5	0.944	0.921	0.907
Average	**0.939**	**0.916**	**0.901**

Table 5: BERTScore F1 for abstract, key topics, and MCQs against manually created references.

4.4 Interpretation of Results

The performance of EduSense was evaluated using a combination of objective metrics and subjective student feedback, all of which affirm the system's effectiveness in real-time classroom augmentation.

ASR transcription fidelity, measured by Word Error Rate (WER), ranged between 2.8% and 5.0% across five lectures (Table 4), indicating that Whisper reliably captured classroom speech even under typical noise conditions. High semantic alignment scores, as measured by BERTScore F1 (0.94–0.97), further validate the model's ability to preserve the informational content of the lecture (Figure 4). This consistency extended across generated abstracts, key topic extractions, and MCQs, with average F1 scores of 0.939, 0.916, and 0.901 respectively (Table 5).

Subjective evaluations via Likert-scale surveys (Table 3) revealed strong student approval: 4.2 for comprehension enhancement, 4.3 for interface usability, and 4.5 for system reliability. Notably, the average score of 4.4 for the question "useful for missed lectures" suggests EduSense serves as an effective post-class review tool.

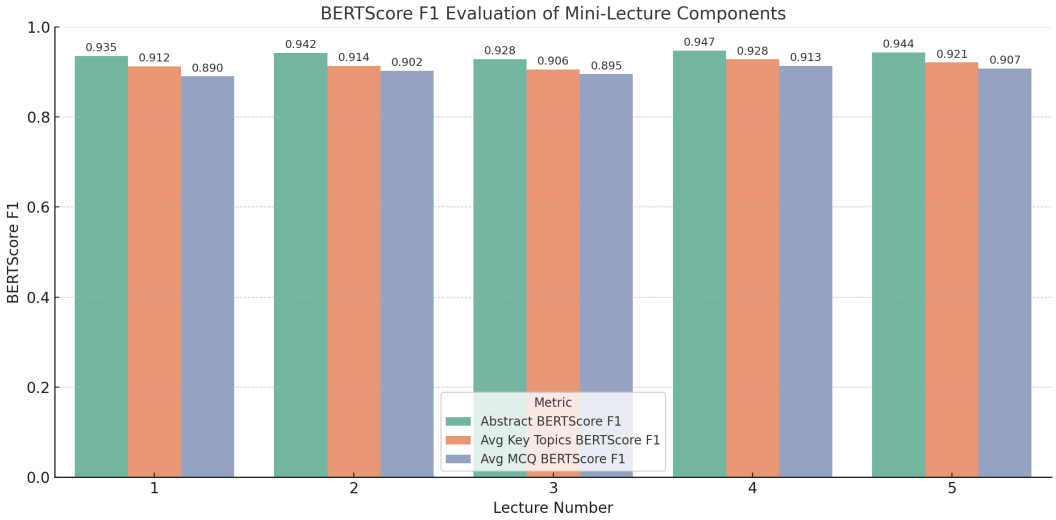


Figure 5: BERTScore F1 for abstract, topics, and MCQs per lecture. Abstracts averaged 0.939, key concepts 0.916, and quizzes 0.901, showing high semantic alignment with human-created references.

These combined results demonstrate that EduSense meets both usability and fidelity benchmarks. The system reliably captured lecture audio, translated it into pedagogically meaningful content, and delivered this through a user-friendly application, without impeding the natural flow of instruction. This positions EduSense as a practical, low-latency AI adjunct capable of supporting instructional goals in real time.

5 Conclusion

This investigation indicates that AI frameworks like EduSense possess the capacity to meaningfully enhance traditional instruction by autonomously generating semantically congruent post-lecture summaries and quizzes. Such tools demonstrably bolster student comprehension and can furnish educators with actionable insights into learning dynamics.
However, prudence dictates acknowledging the study's inherent limitations:

- Small sample size (5 lectures, 17 students).
- Lack of long-term retention and academic performance assessment.
- Potential for novelty effects (Hawthorne effect).
- Reliance on self-reported survey data.
- Ground truth for technical evaluations created by a single researcher.

Looking Ahead

- **Research**: Systematically scale EduSense across disciplines, multilingual cohorts, and varied class sizes; benchmark fine-tuned versus general-purpose vision–language and LLM pipelines for richer, discipline-specific content generation and longitudinal impact.

- **Students**: Transform the app into an adaptive tutor that continuously profiles learner strengths, offers instant clarification queries, pushes micro-quizzes, and surfaces progress dashboards to cultivate self-regulated, mastery-oriented study habits.
- **Teachers**: Deliver real-time analytics on comprehension trends, enable granular curation and re-ordering of generated materials, and suggest pacing or emphasis adjustments that align with syllabus milestones and individual classroom dynamics.
- **Technology**: Integrate noise-robust, code-mixed ASR models, migrate core inference on-device for lower latency and privacy, and ingest multimodal inputs—including lecture video, slide decks, and textbooks—to craft richer, context-aware learning artifacts.

References

[1] Ruichen Chen, Yan Liu, Zhiyuan Liu, and Maosong Sun. Lecturesumm: A transformer-based system for educational video summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2023.

[2] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. The classroom as a dashboard: Co-designing wearable cognitive augmentation for k-12 teachers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19. ACM, 2019.

[3] Hassan Khosravi, Kirsty Kitto, Rafael A. Calvo, and Andrew Gibson. A learner model for supporting learning from mistakes. *User Modeling and User-Adapted Interaction*, 31(1):1–28, 2021.

[4] Tuan Nguyen, Maria T. Kabanova, and Linh Pham. Evaluating asr tools for educational video transcription. *IEEE Transactions on Learning Technologies*, 15(2):175–186, 2022.

[5] Rainer Winkler and Matthias Söllner. Uncovering the blind spot of mobile learning apps: Using psychometrics to evaluate app usability. *Computers & Education*, 144:103695, 2020.