An AI-Based Framework for Analyzing Classroom Audio to Characterize Teaching Practice

Federico Pardo[1], Óscar Cánovas[1] and Félix J. García Clemente[1]

[1]*Department of Computer Engineering and Technology, University of Murcia*

Abstract

Traditional classroom observation faces significant scalability limitations, hindering effective pedagogical feedback. This paper introduces a modular AI framework for the scalable and interpretable analysis of teaching practices via classroom audio. Our work directly addresses critical research gaps in interpretability, modality fragmentation, and feedback loops. Key contributions include the curation of over 200 meticulously labeled classroom audio recordings and the engineering of a robust, API-accessible processing pipeline. The framework leverages state-of-the-art techniques—including speaker diarization, transcription, multimodal fusion, and AI models—to classify teacher interventions and generate insights. Preliminary results demonstrate high accuracy in multimodal classification and positive utility feedback from participating educators. While promising, ongoing challenges in multimodal fusion complexity, generalization across diverse contexts, LLM implementation, and ensuring xAI accessibility for non-technical stakeholders are actively being addressed in our continuing research.

1. Introduction

Traditional classroom observation methods face critical scalability limitations, requiring trained professionals to provide meaningful feedback—a resource-intensive process that struggles to meet the demands of large-scale educational systems. Recent advances in artificial intelligence (AI) and machine learning (ML) offer new opportunities to automate the analysis of teaching practices through scalable processing of classroom audio recordings, addressing this fundamental constraint.

In previous work [1, 2, 3, 4], we developed and evaluated methods for speaker diarization, acoustic feature extraction, and discourse classification in real teaching environments, demonstrating that automated analysis can detect relevant interaction patterns and distinguish instructional formats. These technical foundations enable a paradigm shift from manual observation to AI-assisted reflection, where educators can systematically analyze aspects of their practice such as student participation dynamics, questioning strategies, and critical thinking facilitation -dimensions that directly impact learning outcomes.

Building on this foundation, we propose a modular framework that integrates multiple AI components to interpret classroom audio at scale. This framework leverages a sophisticated pipeline for extracting diverse audio features, including speaker diarization, low-level acoustics, and natural language processing (NLP) cues. It incorporates various machine learning models for robust data processing and analysis, with ongoing exploration into the integration of Generative AI for future enhancements. Another key focus of our current work is the exploration of model explainability, aiming to provide insights into how these complex models process data. Furthermore, the framework is designed to provide feedback to educators, with the usage of graphics and extracted teacher-students interaction metrics to support their professional reflection.

The rest of the paper is organized as follows: Section 2 identifies key challenges in existing audio analytics research, Section 3 outlines our research questions, Section 4 reviews related work, Section 5 details our methodology, Section 6 presents current results, and Section 7 concludes with future directions.

LASI Spain 2025 Doctoral Consortium, May 2025, Vitoria-Gasteiz

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-01-12

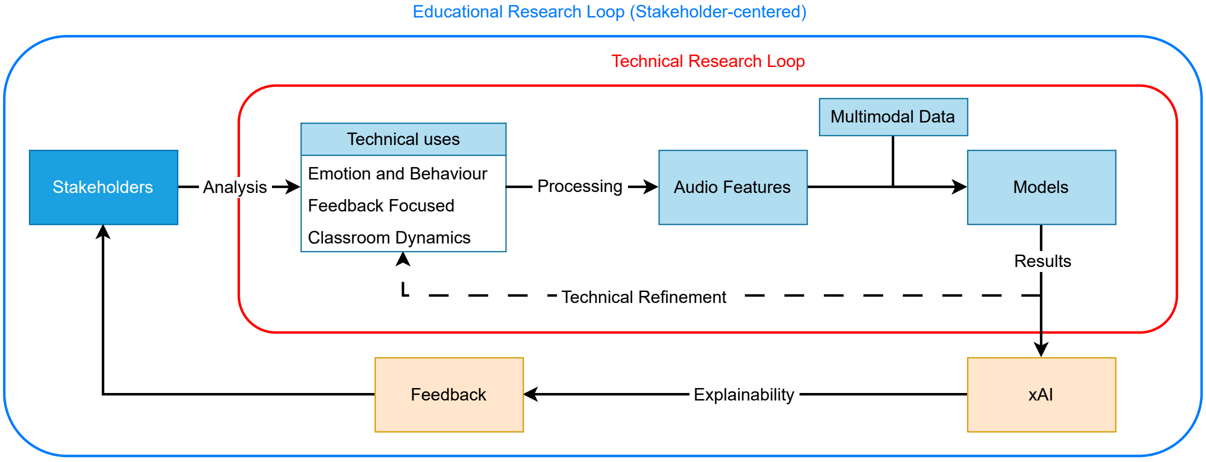

Figure 1: Conceptual diagram contrasting current and ideal research workflows using audio features in education. The red loop shows a technically focused pipeline lacking explainability and stakeholder input, while the blue loop emphasizes feedback and interpretability. The dashed arrow reflects a trend toward technical validation aimed at expanding use cases, often without pedagogical grounding.

2. Background and Problem Identification

As visualized in Figure 1, current audio analytics research predominantly follows the self-contained technical loop (red), where raw audio processing through machine learning models drives technical applications like emotion recognition or classroom dynamics analysis. Our systematic review of 82 studies [5] quantifies this imbalance: 87% of papers focused solely on technical validation metrics, while only 13% (n=11) involved real teacher participation, and none combined acoustic, diarization, and linguistic features simultaneously.

Three critical gaps emerge from this analysis:

- **Interpretability Deficit**: High-performing models remain black boxes, offering no insight into how features or patterns drive predictions.
- **Modality Fragmentation**: While 20% of studies combined two feature types, none integrated the triad of speaker diarization, acoustic features, and discourse analysis that our framework implements.
- **Feedback Disconnect**: The dashed "technical refinement" arrow dominates research trajectories, with only 11 papers establishing closed feedback loops between model outputs and teaching practice improvement.

Our framework addresses these limitations through three interlocked mechanisms. First, we aim to close the educational research loop (blue) by providing personalized PDF reports to teachers with:

- Speaker-diagrammed participation timelines.
- Turn-taking dynamics visualization.
- Annotated intervention transcripts.

Second, we implement multimodal fusion of diarization data (turn duration, overlap), acoustic features (pitch variance, speech tempo), and linguistic markers (lexical complexity, words per minute). Third, our ongoing integration of xAI techniques (SHAP, LIME) begins to address the interpretability gap by revealing feature contributions to classification decisions.

3. Research Goals and Questions

The primary objective of this research is to bridge the gap between technical audio analysis capabilities and actionable educational insights through AI methods that address the three critical limitations identified in Section 2: interpretability deficits, modality fragmentation, and disconnected feedback.

RQ1: Can information derived from classroom audio recordings be used to analyze teacher discourse and classroom dynamics?

RQ2: Can different audio-derived features (such as low-level acoustics, speaker diarization, and linguistic cues) be effectively combined to enhance the analysis and classification of teaching practices?

RQ3: Can we interpret the internal behavior of our models, in order to better understand how different aspects of classroom dynamics are being modeled and help stakeholders understand model decissions?

These research questions directly operationalize our commitment to developing an AI-based framework that transcends the limitations of current research. RQ1 addresses the critical deployment gap by validating the utility of classroom audio in real educational settings. Building upon this, RQ2 confronts the issue of modality fragmentation by exploring the effective fusion of diverse audio-derived features—acoustic, diarization, and linguistic cues—a tripartite integration conspicuously absent in prior work. Finally, RQ3 directly mitigates the interpretability deficit inherent in complex AI models by integrating techniques like SHAP and LIME, ensuring that our model's internal behaviors and classification decisions are transparent and comprehensible to non-technical stakeholders, thereby fostering trust and enabling actionable pedagogical insights.

4. State-of-the-Art and Existing Solutions

Recent work in educational analytics, as detailed in our systematic review [5], shows that combining audio features with advanced linguistic analyses can yield rich insights into classroom dynamics. These findings point to the lack of integrated models capable of integrating linguistic and acoustic features while maintaining interpretability. For example, VizChat [6] demonstrates how multimodal AI chatbots can deliver contextual explanations, while Lee et al. [7] show that techniques like SHAP can effectively clarify model decisions and mitigate bias. Similarly, Chejara et al. [8] underscore the benefits of multimodal learning analytics for building robust models. Building on these advances, our approach integrates multiple processing pipelines to provide feedback to educators.

5. Framework and Methodology

The proposed framework implements a **modular pipeline** (Figure 2) that addresses the three gaps identified in Section 2 through systematic audio processing. Developed over a two-year educational innovation project, this pipeline facilitated the collection of a substantial dataset of classroom audio recordings. This collaborative development also allowed for a continuous refinement of the process, informed by direct engagement and survey feedback from participating teachers.

Core technical components include:

- **Audio Preprocessing:** Raw recordings undergo noise reduction and amplitude normalization using LibROSA's spectral gating, followed by voice activity detection to isolate speech segments
- **Audio Diarization:** We use PyAnnote's embedding clustering, extracting turn-taking metrics (speaker switches, overlap duration) that address RQ1's dynamics analysis requirements.

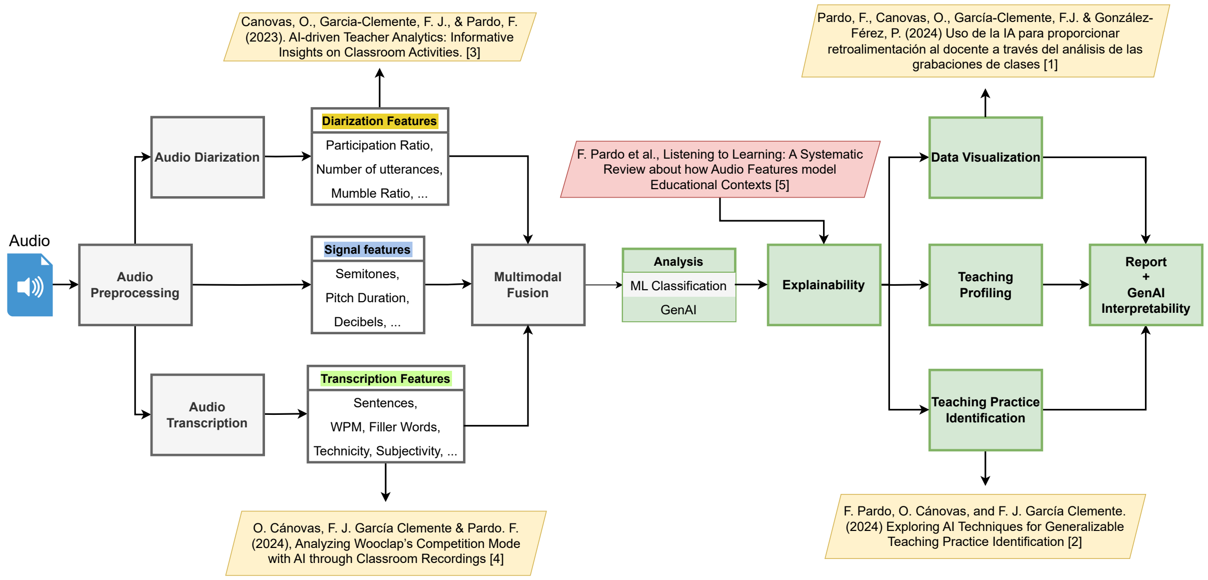
Canovas, O., Garcia-Clemente, F. J., & Pardo, F. (2023). AI-driven Teacher Analytics: Informative Insights on Classroom Activities. [3]
Pardo, F., Canovas, O., García-Clemente, F.J. & González-Férez, P. (2024) Uso de la IA para proporcionar retroalimentación al docente a través del análisis de las grabaciones de clases [1]
Audio
Audio Preprocessing
Audio Diarization
Diarization Features
Participation Ratio,
Number of utterances,
Mumble Ratio, ...
Signal features
Semitones,
Pitch Duration,
Decibels, ...
Audio Transcription
Transcription Features
Sentences,
WPM, Filler Words,
Technicity, Subjectivity, ...
Multimodal Fusion
F. Pardo et al., Listening to Learning: A Systematic Review about how Audio Features model Educational Contexts [5]
Analysis
ML Classification
GenAI
Explainability
Data Visualization
Teaching Profiling
Teaching Practice Identification
Report + GenAI Interpretability
O. Cánovas, F. J. García Clemente & Pardo. F. (2024), Analyzing Wooclap's Competition Mode with AI through Classroom Recordings [4]
F. Pardo, O. Cánovas, and F. J. García Clemente. (2024) Exploring AI Techniques for Generalizable Teaching Practice Identification [2]

Figure 2: Updated implementation pipeline of the proposed AI framework. Each module corresponds to a specific stage in the audio analysis process. Modules shown in gray were developed in previous iterations; newly added components are highlighted in color. This architecture reflects the current implementation status and highlights planned extensions of the system.

- **Audio transcription:** Whisper-large-v3 transcribes on educational content with a single microphone, temporally aligned to diarization output through dynamic time warping for accurate speaker attribution.
- **Multimodal Fusion:** Late fusion combines (1) diarization-derived participation ratios, (2) acoustic descriptors (pitch variance, MFCCs), and (3) SpaCy-processed linguistic markers (WPM, technicity index). This step is customized for every model we developed, as not every model need all the extracted information to work.
- **Analysis:** This new box includes all the models developed, from the ML classification models developed last year to the future Generative AI models (such as LLMs) for intervention classification.
- **Explainability:** This module integrates SHAP and LIME to provide insights into model decisions by analyzing feature contributions. Its inclusion directly addresses the interpretability deficits highlighted in our systematic review. [5]
- **Teaching Practice Identification:** Machine learning models identify practices such as lecturing, group work, or interactive activities based on multimodal inputs.
- **Reporting and Generative AI:** Periodic reports summarize indicators and include optional generated narratives to contextualize results.
- **Teacher Profiling:** Longitudinal analysis across sessions builds interaction profiles for each teacher, enabling personalized insights, focused on higher-education contexts.

The proposed methodology is primarily quantitative, as it relies on the extraction and modeling of structured features from classroom audio data using machine learning techniques. However, it also incorporates qualitative elements through the use of generative models for report generation and the interpretability tools that aim to support human understanding of model behavior. This mixed-methods approach aligns with applied learning analytics research by grounding technical outputs in teaching insights, directly addressing the stakeholder disconnect identified in Figure 1.

6. Current Work and Preliminary Results

Our ongoing research directly addresses the three research gaps identified in Section 2 through a series of tangible technical implementations and significant data engineering efforts. These developments, which form the core of our doctoral work, have been tested and refined using a substantial corpus of classroom recordings.

6.1. Key Developmental Achievements

Dataset Acquisition and Curation: Over two years, we have manually collected and curated a notable dataset comprising more than 200 classroom audio recordings. This collection spans diverse university courses and academic fields, featuring a variety of teaching methodologies. Recordings were systematically acquired using dedicated recorders positioned in the front row of classrooms, allowing for the comprehensive capture of all teacher-student interactions. The establishment of this extensive dataset itself represents a significant research achievement, providing a robust foundation for pedagogical analysis. Furthermore, a substantial portion of this dataset has been meticulously labeled for various analytical purposes, a time-intensive and valuable task that enhances its utility for diverse research applications.

End-to-End Pipeline Engineering: We have engineered an entire processing pipeline from scratch, integrating both publicly available, state-of-the-art technologies (such as PyAnnote for diarization and Whisper for transcription) with extensive custom-developed code. This bespoke development manages data routing, preprocessing, feature extraction, and output generation. Each module within this pipeline represents a distinct developmental effort, requiring meticulous design, programming, integration, and validation to ensure robust functionality.

System Modularity and Accessibility: The pipeline is designed with inherent modularity, ensuring that any component depicted in Figure 2 can be interchanged or updated, provided it adheres to specified input and output formats. This architectural flexibility promotes future scalability and adaptability. Moreover, the developed software exposes a straightforward Asynchronous API, enabling seamless requests for audio processing without requiring direct code access. This API facilitates integration with other ongoing projects within the university, demonstrating the system's practical applicability and collaborative potential.

Advanced Model Development and Generalization: Our efforts extend to the development of a multitude of machine learning models. This involved exploring various algorithms, testing different architectural variants, and employing extensive grid searches for hyperparameter tuning to optimize performance for diverse analytical tasks. Crucially, in developing these models, we emphasize generalization capabilities within our operational constraints. We adhere to established methodologies, such as those advocated by Chejara et al. [8], to validate the robustness and broad applicability of our models.

Reporting and Feedback Mechanism: To bridge the gap between technical outputs and practical pedagogical application, we have developed a system for generating periodic comprehensive PDF reports for participating teachers. These reports summarize key classroom interaction indicators, allowing educators to analyze their teaching behavior and student participation dynamics. We are also actively developing a digital format for these reports, aiming to enhance accessibility and user experience.

6.2. Preliminary Insights and Impact

Building on these foundational developments, our initial findings demonstrate promising capabilities:

- **Multimodal Feature Integration**: Through late fusion of PyAnnote-derived diarization features and Whisper transcriptions, our BERT models achieve a 75% accuracy in classifying teacher interventions, representing a 3% improvement over transcription-only baselines. We are currently expanding this classification by leveraging Large Language Models (LLMs) such as ChatGPT, exploring a 'Chain of Thought' approach to enhance interpretability and reasoning.

- **Explainability Foundations**: Our initial integration of SHAP analysis has proven effective in identifying key feature contributions to classification decisions, a crucial step towards making model outputs more transparent and comprehensible for non-technical stakeholders.
- **Teacher Feedback and Perceived Utility**: Post-deployment surveys from the 9 teachers involved in our educational innovation project reveal significant utility. A total of 5 out of 9 instructors specifically highlighted the value of metrics like Participation Speech Ratio (PSR) and timeline visualizations, with comments such as: *"The timeline of teacher-student speaking turns helped me identify participation patterns I hadn't noticed during class."* Furthermore, two instructors reported concrete adjustments to their teaching practices, exemplified by feedback like: *"Seeing low student participation rates motivated me to redesign activities – I now include more open-ended questions."* and *"The reports confirmed differences between student groups, guiding how I use interactive tools like Wooclap."* Overall, participants rated the system's utility at 4/5.

These qualitative insights strongly align with our technical focus on participation metrics (PSR, TTC) and temporal analysis during last year educational innovation project, providing crucial validation for the framework's practical relevance.

6.3. Limitations and Challenges

Despite the promising progress, our current framework faces several inherent limitations and challenges that guide our ongoing and future work. The multimodal fusion complexity presents a significant hurdle, as optimally combining disparate feature types (acoustic, diarization, and linguistic) requires continuous refinement to maximize analytical depth and predictive accuracy. Furthermore, ensuring generalization across diverse classroom contexts remains a key challenge, primarily due to the limited variety of audio data collected thus far in terms of pedagogical approaches, academic disciplines, and environmental conditions. The integration of Large Language Models (LLMs) and Generative AI also introduces considerable implementation challenges related to their computational size, associated operational costs, and the need to ensure some replicability of results. Finally, while explainability techniques like SHAP and LIME provide valuable insights for technical users, substantial work is still required to translate these complex explanations into formats that are genuinely accessible and actionable for non-technical educational stakeholders, fostering greater trust and practical utility.

7. Conclusion

Over the past two years, this research has successfully evolved into a robust, integrated framework for the AI-driven analysis of classroom audio, directly addressing identified gaps in the field. This journey encompasses the significant achievements of curating a substantial and meticulously labeled dataset, the from-scratch engineering of a modular and API-accessible processing pipeline, and the development of advanced, generalizable machine learning models for intervention classification. These tangible advancements provide the foundational technical infrastructure for detailed multimodal analysis of teaching practices, effectively bridging the gap between raw audio data and actionable pedagogical insights. Our current efforts include the strategic design of a multimodal feature fusion strategy and the initial, yet critical, development of explainability modules, all validated by positive feedback from participating teachers.

Looking ahead, our future work is strategically guided by identified complexities and limitations. We will focus on enhancing the teacher profiling module for longitudinal analysis and critically, on improving the generalization of our models across truly diverse classroom contexts. We are actively exploring and committed to refining the integration of Large Language Models and Generative AI, confronting challenges related to their computational demands and ensuring replicability of results. Finally, while explainability techniques are foundational, significant effort will be directed towards making these insights genuinely*accessible and actionable for non-technical educational stakeholders, thereby fostering greater trust and maximizing practical utility. Given the inherent scope and time constraints of doctoral research, certain advanced aspects of these developments may strategically transition into postdoctoral work to ensure comprehensive implementation and rigorous validation.

Acknowledgments

This work has been funded under grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033, NextGenerationEU/PRTR, UE, and grant PID2021-122466OB-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 2.5 in order to: Grammar and spelling check.

References

[1] F. P. García, Ó. Cánovas, F. J. G. Clemente, P. González-Férez, Uso de la IA para proporcionar retroalimentación al docente a través del análisis de las grabaciones de clases, Actas de las XXX Jornadas de la Eseñanza de la Informática 9 (2024) 241–249.

[2] F. Pardo, Óscar Cánovas, F. J. G. Clemente, Exploring AI Techniques for Generalizable Teaching Practice Identification, IEEE Access (2024).

[3] O. Canovas, F. J. Garcia-Clemente, F. Pardo, AI-driven teacher analytics: Informative insights on classroom activities, in: 2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), IEEE, 2023, pp. 1–8.

[4] Ó. C. Reverte, P. G. Férez, F. J. G. Clemente, F. P. García, Analyzing Wooclap's competition mode with AI through classroom recordings, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (2024).

[5] F. Pardo, et al., Audio features in education: A review of computational applications and research gaps, 2025. Systematic Review (in revision).

[6] L. Yan, L. Zhao, V. Echeverria, Y. Jin, R. Alfredo, X. Li, D. Gašević, R. Martinez-Maldonado, VizChat: Enhancing Learning Analytics Dashboards with Contextualised Explanations Using Multimodal Generative AI Chatbots, in: Proceedings of the International Conference on Artificial Intelligence in Education, 2024, pp. 180–193.

[7] H. Lee, C. Belitz, N. Nasiar, N. Bosch, XAI Reveals the Causes of Attention Deficit Hyperactivity Disorder (ADHD) Bias in Student Performance Prediction, in: Proceedings of LAK25: the 15th International Learning Analytics and Knowledge Conference, 2025, pp. 418–428.

[8] P. Chejara, L. P. Prieto, M. J. Rodriguez-Triana, R. Kasepalu, A. Ruiz-Calleja, S. K. Shankar, How to build more generalizable models for collaboration quality? lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics, in: Proceedings of LAK23: 13th International Learning Analytics and Knowledge Conference, 2023, pp. 111–121.