applied sciences

MDPI

Systematic Review

Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps

Federico Pardo *, Óscar Cánovas and Félix J. García Clemente

Department of Computer Engineering, Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain; ocanovas@um.es (Ó.C.); fgarcia@um.es (F.J.G.C.)

* Correspondence: federico.pardog@um.es

Abstract: This systematic review synthesizes 82 peer-reviewed studies published between 2014 and 2024 on the use of audio features in educational research. We define audio features as descriptors extracted from audio recordings of educational interactions, including low-level acoustic signals (e.g., pitch and MFCCs), speaker-based metrics (e.g., talk-time and participant ratios), and linguistic indicators derived from transcriptions. Our analysis contributes to the field in three key ways: (1) it offers targeted mapping of how audio features are extracted, processed, and functionally applied within educational contexts, covering a wide range of use cases from behavior analysis to instructional feedback; (2) it diagnoses recurrent limitations that restrict pedagogical impact, including the scarcity of actionable feedback, low model interpretability, fragmented datasets, and limited attention to privacy; (3) it proposes actionable directions for future research, including the release of standardized, anonymized feature-level datasets, the co-design of feedback systems involving pedagogical experts, and the integration of fine-tuned generative AI to translate complex analytics into accessible, contextualized recommendations for teachers and learners. While current research demonstrates significant technical progress, its educational potential is yet to be translated into real-world educational impact. We argue that unlocking this potential requires shifting from isolated technical achievements to ethically grounded pedagogical implementations.

Keywords: systematic literature review; audio features; artificial intelligence; educational context; explainability

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 22 May 2025

Revised: 6 June 2025

Accepted: 12 June 2025

Published: 19 June 2025

Citation: Pardo, F.; Cánovas, Ó.; Clemente, F.J.G. Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps. *Appl. Sci.* **2025**, *15*, 6911. https://doi.org/10.3390/app15126911

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Audio analysis in educational research can be traced back to the early 1960s when analog recording technologies first enabled the capture of classroom interactions [1]. Initially, researchers used tape recorders to document teacher–student communication, relying on manual transcription and qualitative observations to understand the dynamics of teaching and learning. The transition to digital recording in the 1990s brought a new era, as emerging digital signal processing techniques allowed for more systematic, quantitative analyses of speech patterns and interaction modalities. This evolution marked a shift from purely observational studies to data-driven investigations, enabling the systematic application of computational modeling in educational settings. By the early 2000s, as computational power increased, researchers began employing machine learning techniques to extract and analyze audio features more effectively. In the 2010s and beyond, deep learning and sophisticated artificial intelligence algorithms have further refined the process, enabling the detection of nuanced features such as emotional tone, speaker diarization, and linguistic patterns.

Audio features have emerged as a pivotal component in enhancing educational experiences, leveraging the rich information embedded within sound to facilitate learning and engagement. In educational contexts, audio features encompass a range of elements such as speech patterns, acoustic signals, and auditory cues that can be analyzed to assess student performance, provide feedback, and tailor instructional strategies. The integration of audio features into learning analytics and educational contexts has opened new opportunities for personalized learning, accessibility, and real-time assessment [2,3]. In this review, we define audio features as the measurable descriptors derived from raw audio recordings that are used to analyze educational interactions. These features may represent different levels of abstraction, with different tools for each task:

- Low-level acoustic features, such as pitch, intensity, or spectral representations (e.g., MFCCs or spectrograms), are directly extracted from the audio waveform using libraries such as Librosa [4].
- Diarization features, including speaker turn-taking, speaking time, and participant ratios, are obtained by segmenting and labeling who speaks when. For example, PyAnnote [5] is a very common tool to extract diarization information.
- Linguistic features derived from automatic transcriptions of speech, such as word usage, syntactic structure, or discourse-level indicators. Here, we must differentiate between extracting the transcription (for example, using Whisper) [6] and analyzing its content (e.g., spaCy) [7].

Despite their varying degrees of abstraction, all these features share a common origin: they are extracted from audio recordings of classroom or educational interactions. They serve as inputs for computational analyses aimed at modeling pedagogical behaviors, classroom climate, or learner engagement. Throughout this review, we use the term 'audio features' to refer collectively to this spectrum of descriptors, regardless of their proximity to the original waveform.

Although several reviews discuss multimodal analytics in education, none treat audio as a distinct and detailed data source. Most prior studies group audio under a generic "sensor" label or confine it to limited use cases, such as transcription or emotion detection, without examining the full range of acoustic, diarization, or linguistic metrics and their pedagogical implications. As a result, researchers lack a consolidated view of which audio feature types (e.g., prosodic versus syntactic analysis of transcripts) have proven effective for specific tasks, such as real-time feedback or modeling classroom interaction, and which remain challenging. Our review fills this gap by (1) comprehensively disaggregating the different categories of audio features used in educational research, (2) revealing how those metrics are extracted and combined, and (3) exposing methodological and ethical voids, such as limited interpretability and the scarcity of anonymized datasets, that have so far prevented findings from translating into practical solutions.

To synthesize the current landscape and expose the structural gaps that motivate this review, we present a conceptual diagram outlining two interconnected but misaligned workflows in the field (Figure 1). The current research loop (in red) captures the dominant, technology-driven approach, where audio features are extracted, processed, and modeled to produce analytical results. While this pipeline has yielded significant technical progress, it often operates in isolation from pedagogical practice. Crucial components, such as model explainability, feedback mechanisms, and, above all, the involvement of educational stakeholders, are frequently overlooked from this loop. In contrast, the educational impact loop (in blue) represents a broader, stakeholder-centered vision in which research findings are made explainable and translated into feedback that informs educational practice. The disconnect between these loops, particularly the absence of pedagogical validation and stakeholder engagement, underscores the need for a critical synthesis of how audio features are currently used in educational contexts and how they could be better aligned with real-world needs.

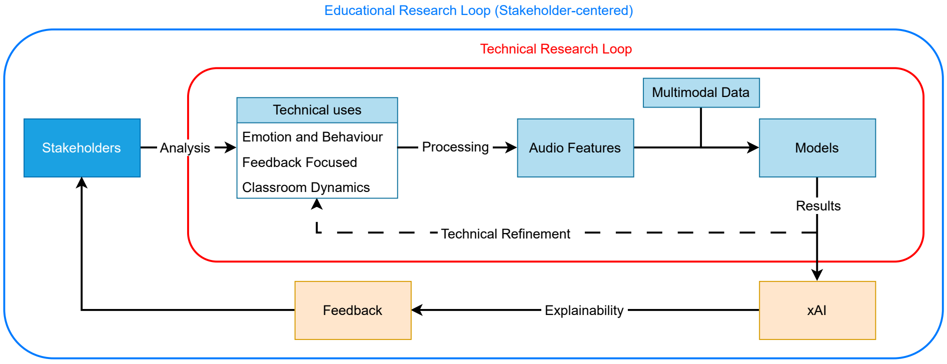

Figure 1. Conceptual overview of current and idealized research workflows involving audio features in educational contexts. The red loop represents the dominant technical pipeline, which focuses on extracting and modeling audio features but largely excludes explainability, feedback, and the participation of educational stakeholders. The blue loop reflects a more complete, stakeholder-centered perspective, where analysis leads to actionable outcomes through explainable models and feedback mechanisms. The dashed arrow highlights the lack of pedagogical validation connecting research outputs back to educational use cases.

Using audio features in education extends beyond signal processing or speech recognition, as it encompasses the broader challenge of extracting relevant information from audio streams to support computational analysis of educational settings. While previous work has explored trends in data-driven education more generally [8], no existing review has systematically examined how audio-derived data are functionally applied across studies. In this review, we focus on the technological uses of audio features, i.e., how they are applied within computational pipelines to detect, classify, or model phenomena observed in educational contexts. These uses reflect the current technical framing of audio features rather than their pedagogical purpose per se. Our goal is to uncover the functional roles that audio features play in existing research, establishing a typology of use cases that guides the subsequent, more technical questions. This leads to our first question: *What are the main technological uses of audio features in educational research? (RQ1).*

To comprehensively explore the technical landscape of audio analytics in education, we identify the most commonly used audio features and the methods employed for their extraction. Recent trends have shifted from traditional statistical approaches to AI-driven techniques, which enable the direct use of raw or minimally processed data while maintaining strong performance across diverse tasks [9,10]. Although these elements are central to audio-based analytics, no prior review has offered a systematic categorization of the features and extraction techniques used in educational contexts. With this in mind, we mapped the audio features found in the literature, classified them, and described them in our second research question: *What are the most common audio features used in educational studies and how are they extracted? (RQ2).*

Educational data are increasingly being collected from a variety of sources, reflecting the inherently multimodal nature of learning environments. Multimodal learning analytics (MMLAs) have emerged as a research area that seeks to integrate data from diverse modalities (e.g, audio, video, text, and physiological signals) to gain deeper insight into learning processes [11]. Within this context, audio features are rarely used in isolation; they are often combined with other forms of data to capture different dimensions of classroom interaction and learner behavior. These combinations allow researchers to explore richer representations of educational phenomena, though the methods and rationales for such integration vary widely. This leads us to our third research question: *How do researchers combine audio features with other data sources? (RQ3)*.

As audio features become more prevalent in educational research, a key concern is how they are computationally processed. In recent years, there has been a growing reliance on both traditional machine learning algorithms and deep learning models to handle tasks such as classification, clustering, and prediction [12]. These techniques are often applied to features derived from audio, such as speaker diarization outputs and automatic transcriptions, features that, despite being abstracted through additional processing, still originate from the raw audio stream. While these methods can yield strong performance, their complexity raises concerns about interpretability, which is a crucial factor when insights are intended to inform pedagogical decisions. Although explainable AI (xAI) has gained attention in other domains, its adoption within educational audio analytics remains limited. To examine this relationship between performance and transparency, we ask the following: *What techniques are employed to process audio features in educational studies, and to what extent are these solutions interpretable? (RQ4)*.

Finally, we turn to the practical implications of audio-based research in education by examining the extent to which these studies lead to actionable outcomes in real-world settings. Specifically, we explore whether researchers implement mechanisms to provide feedback to participants, such as teachers or students, based on insights derived from audio analysis. Feedback loops are essential for translating research into practice and fostering impact in educational environments [13]. This forms the basis of our final research question: *Which studies provide feedback for participants derived from obtained results? (RQ5)*.

This systematic review synthesizes the literature on the use of audio features in educational contexts, addressing how these features are applied, extracted, and interpreted across studies. The research questions are organized to reflect a gradual progression, from general trends and uses of audio data in education to the specific types of features most frequently extracted and the techniques used to process them.

This review not only synthesizes the current state of research on audio-based methods in educational contexts but also advances the field by identifying three core limitations and offering strategic directions to address them:

- Targeted analysis of audio features. Unlike earlier reviews that broadly categorize audio under "multimodal" or "sensor" headings, our research focuses specifically on audio as a standalone modality. It offers a focused examination of audio features within educational research. It covers a wide range of them, including low-level acoustic properties, linguistic indicators extracted via NLP, and speaker-based metrics obtained through diarization. As far as we are aware, no previous study has provided a systematic identification, categorization, and definition of the audio features employed in educational settings.
- Diagnosis of field-level limitations. While some existing papers describe individual case studies or isolated tools, few have highlighted systemic obstacles that impede pedagogical impact. Our synthesis reveals systemic barriers to pedagogical impact, including the scarcity of actionable feedback, low interpretability of AI models, fragmented and non-replicable datasets, and limited attention to privacy. These gaps highlight a misalignment between technical capability and practical utility.
- Actionable directions for future research. To advance the field, we propose three strategic directions: (1) the release of anonymized, standardized feature-level datasets; (2) the participatory design of feedback systems that actively involve educational practitioners and pedagogical experts; (3) the use of generative AI, particularly fine-tuned LLMs, to translate analytics into tailored, context-aware guidance for teachers and learners.

Summarizing, our research questions for this systematic review are as follows:

- RQ1: What are the main technological uses of audio features in educational research?
- RQ2: What are the most common audio features used in educational studies, and how are they extracted?
- RQ3: How do researchers combine audio features with other data sources?
- RQ4: What techniques are employed to process audio features in educational studies, and to what extent are these solutions interpretable?
- RQ5: Which studies provide feedback for participants derived from obtained results?

The rest of the paper is organized as follows: Section 2 describes the methodology, including some terminology clarifications, the research questions, databases and search terms, research selection, and the review process. Section 3 presents the analysis and synthesis of our results. Then, we end the paper with an analysis of our findings in Section 4 and conclusions in Section 5.

2. Methodology

This systematic review was performed in accordance with the PRISMA (Supplementary Materials) (preferred reporting items for systematic reviews and meta-analyses) guidelines [14], which are widely used for structuring evidence-based syntheses. Figure 2 shows the diagram representing the different stages of our systematic review. The methodology includes the following stages:

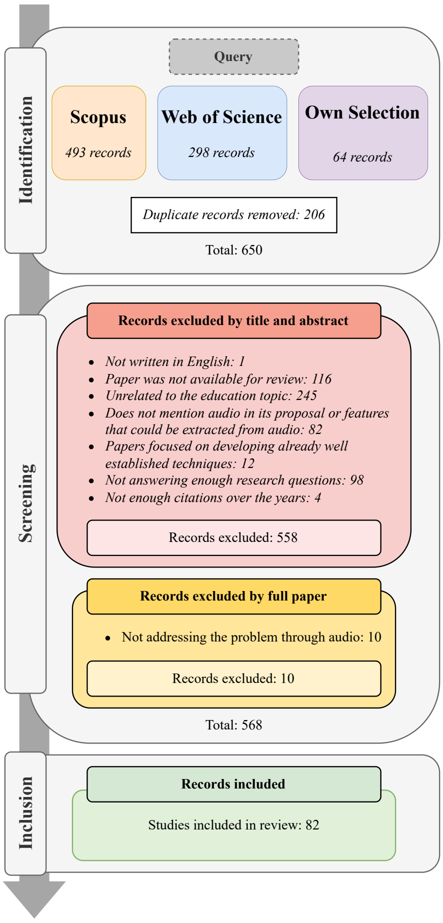
Identification
Query
Scopus
493 records
Web of Science
298 records
Own Selection
64 records
Duplicate records removed: 206
Total: 650
Screening
Records excluded by title and abstract
Not written in English: 1
Paper was not available for review: 116
Unrelated to the education topic: 245
Does not mention audio in its proposal or features that could be extracted from audio: 82
Papers focused on developing already well established techniques: 12
Not answering enough research questions: 98
Not enough citations over the years: 4
Records excluded: 558
Records excluded by full paper
Not addressing the problem through audio: 10
Records excluded: 10
Total: 568
Inclusion
Records included
Studies included in review: 82

Figure 2. Flow diagram of the PRISMA methodology followed.

2.1. Identification of Research Works

The literature data search was conducted on 27 November 2024. Scopus and Web of Science (WoS) were selected due to their broad coverage of peer-reviewed literature in education, technology, and computational sciences [15], making them appropriate for interdisciplinary reviews.

To perform the search on both databases, we restricted the query to the title and abstract to balance precision and recall, as full-text searches produced an unmanageable number of irrelevant results.

We used a structured query consisting of three conceptual blocks, along with a publication date restriction covering studies published after 2012.

- Data source: This component targets studies where audio serves as a primary or derived source of data. The query includes terms such as audio and sound to capture explicit references to raw audio. To encompass studies that use features extracted from audio rather than the waveform itself, it also includes terms like speech transcript, dialogue, and discourse features. This increased the likelihood of including work analyzing linguistic or prosodic features even when the term "audio" is not explicitly mentioned.
- Educational context: This block captures the environments in which learning interactions occur. Keywords such as collaborative learning, Group interaction, and teaching practice reflect classroom-based and peer-to-peer learning scenarios. Additionally, terms like meeting transcription are included to capture studies focusing on structured interactions in educational or academic contexts. This broader scope helps cover both formal classroom settings and informal learning environments such as workshops and seminars.
- Techniques: This component targets the computational methods used to process audio and audio-derived data. The query includes terms related to core technologies such as machine learning, deep learning, and artificial intelligence, along with domain-specific methods like speech recognition, voice activity detection, and natural language processing (NLP). These terms ensure the inclusion of studies applying advanced analytical frameworks. The inclusion of learning analytics further ensures alignment with educational objectives, emphasizing the intersection between computational processing and pedagogical insight.

The specific query terms used in this systematic review are presented in Figure 3. To account for terminological variation in a still-evolving research domain, we incorporated wildcard characters into search terms such as Speech Transcript$ and Speech Recogn*. The dollar sign ($) and asterisk (*) act as wildcard operators, allowing retrieval of multiple morphological variants of a base term. For instance, 'Transcript$' allows for both 'Transcript' and 'Transcription', while 'Recogn' retrieves terms like 'Recognize', 'Recognition', and similar word forms. This strategy increases recall by capturing terminological variation across studies, ensuring that semantically related works are not excluded due to minor wording differences. The initial search returned a total of 791 records: 493 from Scopus and 298 from Web of Science (WoS). To enhance completeness, we added 64 relevant papers that were not captured by the automated query but were identified through snowballing by reviewing the references and citations of initially selected articles, as well as through prior work by the authors. All additional papers were manually verified to meet the same inclusion criteria. After removing 206 duplicate entries from the combined dataset, the final corpus included 650 unique studies.

```
TITLE-ABS("Audio" OR "Speech Transcript$" OR "Dialogue" OR "Sound" OR
"Discourse Features")
AND TITLE-ABS("Teaching Practice" OR "Automated Feedback" OR
"Meeting Analysis" OR "Teaching Analytics" OR "Educational Technology"
OR "Collaborative Learning" OR "Group Interaction" OR  "Classroom" OR
"Meeting Transcription")
AND TITLE-ABS("Machine Learning" OR "Deep Learning" OR
"Artificial Intelligence" OR "AI" OR "NLP" OR
"Natural Language Processing" OR "Neural Network$" OR
"Learning Analytics" OR "Speech Recogn*" OR "Voice Activity Detect$")
AND PUBYEAR > 2012
```

Figure 3. Boolean search query used to identify relevant studies in Scopus and Web of Science. The query targets audio-related terms, educational contexts, and computational techniques in titles and abstracts, restricted to publications after 2012. Wildcards were included to account for morphological variation.

2.2. Screening of Articles

To ensure the relevance and quality of the included papers, we established a set of mandatory exclusion criteria designed to filter out studies misaligned with the review's objectives or lacking academic rigor. The screening process was conducted independently by two of the co-authors, using the available abstracts as the basis for screening. If reviewers were not sure whether any of the criteria applied, they used the full text to verify. Discrepancies were resolved through discussion until a consensus was reached. The exclusion criteria were applied sequentially, with any study not meeting a given condition being excluded immediately:

- The paper was not written in English due to language constraints in the review team.
- The full text of the paper was not accessible for review, either due to restricted institutional access, paywall limitations, or the unavailability of a preprint or author-supplied version upon request.
- The paper did not focus on educational settings or learning processes as a primary context or objective.
- The paper did not use audio in its proposal or features that could be extracted from audio (e.g., text transcriptions).
- The paper focused exclusively on technical improvements to well-established audio processing methods (e.g., diarization or speech transcription) without applying them to educational data.
- During title/abstract screening, we required each manuscript to address at least one of our five core research questions. If a paper's title or abstract showed no substantive discussion of audio-feature methodology or feedback-oriented use of audio, it was excluded under "Not answering enough research questions". A total of 98 papers fell into this category. For instance, some studies mentioned the presence of audio but did not analyze or extract features from it, nor did they integrate it into the research objectives in a meaningful way.
- The paper was published in 2022 or earlier and had no citations on Google Scholar as of 4 December 2024. Given the volume and the goal of identifying impactful contributions, we applied this criterion as a secondary filtering mechanism to exclude papers that had not generated scholarly attention over time. This step affected only four papers.

2.3. Inclusion of Papers for Review

After applying the exclusion criteria to the 650 unique records, a final sample of 82 studies was retained for review. The final set comprised 43 journal articles, 37 conference papers, 1 technical report, and 1 book chapter. A detailed overview of all 82 studies, including the author, title, year, educational level, and learning context, is provided in Appendix A. Regarding the educational level analyzed in the papers of this review, 42 were focused on K-12, 27 on higher education, 4 focused on toddlers, 3 included multiple educational levels, and 6 did not specify the educational level where the research took place. Figure 4 shows the annual distribution of selected studies by publication year. The number of studies using audio-related methodologies generally increased over the 2014–2024 period. This trend reflects the broader emergence of AI technologies capable of processing unstructured data, such as audio and images, which are increasingly adopted in various educational and analytical domains [16].

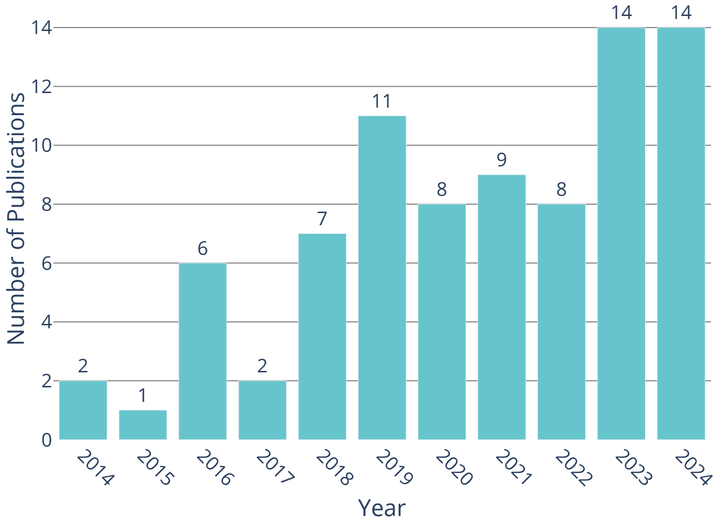

Figure 4. Number of selected studies per year of publication.

2.4. Data Analysis

Once all articles were selected, we prepared an Excel spreadsheet with columns corresponding to each element required to address our research questions. Two researchers independently coded each article by filling in all columns based on the full-text information. For each study, we extracted all available data relevant to our RQs, rather than selecting only subsets of results, ensuring that every feature, method, or feedback format mentioned was captured. After the initial coding, we held consensus meetings to compare both coders' entries, resolve any discrepancies, and harmonize the category labels (tags) so that the final structure accurately reflected the annotated data in each column. The agreed-upon version of the spreadsheet was then used to generate preliminary tables and figures, from which the narrative synthesis of results was developed.

Regarding the analysis of each research question, every paper was considered for every research question as long as the paper provided some information regarding that specific research question.

2.5. Synthesis of Results

Given the substantial heterogeneity across study designs, types of audio-based features, and educational outcomes, we did not perform a formal meta-analysis. Instead, we adopted a narrative synthesis approach to integrate findings. This choice was driven by three main factors: (1) variability in audio-feature extraction methods (e.g., raw waveform vs. transcript-based vs. prosodic features), (2) diverse machine learning and AI techniques applied, and (3) differing educational contexts and measured outcomes (e.g., classroom participation vs. automated feedback accuracy). By using a structured narrative framework, we grouped studies according to their primary research questions, types of audio data, and analytical methods. Within each group, we summarized key findings, methodological strengths, and limitations. This narrative strategy allowed us to highlight patterns and gaps in the literature.

Within this narrative framework, we also explored possible sources of heterogeneity by including studies comprising several factors such as educational level (e.g., K–12 vs. higher education), modality (in-person vs. online), and type of audio features used (e.g., prosodic features vs. transcription-based features).

2.6. Assessment of Bias and Certainty

We acknowledge the potential for publication bias qualitatively since this emerging area often provides greater visibility to studies reporting positive or innovative findings. Therefore, we conducted additional "snowball" searches (reviewing references and citations) to capture relevant papers that might not appear in standard databases, aiming to reduce bias toward highly cited or "headline" studies.

Additionally, we are aware of the risk that some null or less-"exciting" results may remain under-represented. This critical stance helps readers interpret our conclusions in light of possible reporting biases.

To gauge confidence in the collective findings, we applied a qualitative certainty assessment based on three key factors: (1) the presence of multiple independent studies reporting similar results, (2) transparency and reproducibility of methods, and (3) adequate sample sizes or dataset volumes.

3. Results

The following subsections present the findings of our systematic review, structured around five research questions. Each subsection synthesizes evidence from the selected studies to provide a focused analysis of the role of audio features in educational contexts.

3.1. What Are the Main Technological Uses of Audio Features in Educational Research? (RQ1)

Before presenting our classification, we clarify what we mean by the technological use of audio features. In this review, a use refers not to the pedagogical goal per se but to the function that audio-derived data fulfills within the study's analytical design. This includes tasks such as identifying question types, detecting emotional cues, inferring group behavior, or producing automated feedback. The key contribution lies in how audio features drive analytical understanding and facilitate automated processes.

This perspective allows us to frame audio usage as a continuum of increasing abstraction. At one end are studies that focus on low-level classification tasks (e.g., detecting specific interventions); at the other end are those that synthesize audio-derived insights into feedback intended for stakeholders.

Our classification scheme thus aims to make sense of the current research landscape by analyzing what problems audio features are being applied to solve. While many studies touch on multiple goals, we assigned each to a single primary category to support consistent comparison. The distribution of studies across these categories is shown in Figure 5.

The most common applications fall into two categories: feedback provision and behavioral/emotional analysis, which together account for more than half of the reviewed studies.

A smaller but still significant portion of the literature (18 out of 82 studies) concentrates on classroom dynamics, particularly in characterizing classroom climate and identifying teaching strategies. Finally, intervention classification emerges as the least frequent category in our review. These studies focus on assigning teacher or student interventions to predefined categories established during the research process.

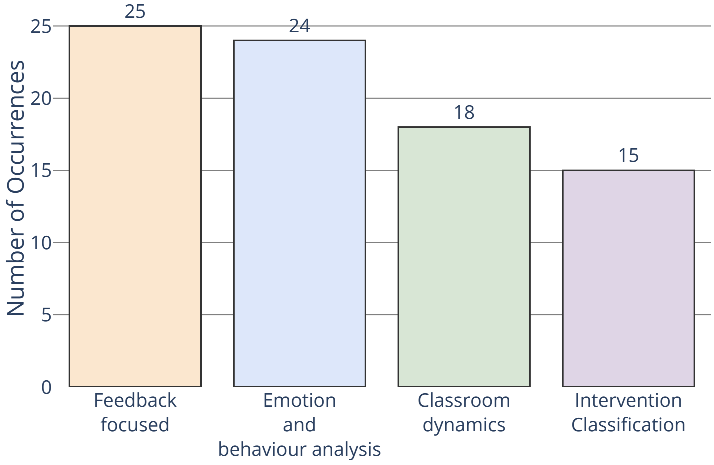

Figure 5. Main uses detected in the analyzed literature (RQ1).

The four categories follow a conceptual hierarchy based on the level of abstraction of the information extracted. At the lowest level, studies on intervention classification focus on isolated speech events; these feed into analyses of emotions and behaviors, which in turn contribute to understanding broader classroom dynamics. The highest level of abstraction is found in feedback-oriented studies, which synthesize insights from the previous categories to inform educational decision-making.

3.1.1. Intervention Classification

This category encompasses studies focused on the automated classification of individual interventions, typically utterances made by teachers or students. These approaches aim to identify pedagogically relevant patterns at the utterance level, such as the presence of questions, argument components, or instructional strategies.

A common objective is question detection, with early work such as [17] developing machine learning models to identify teacher questions that promote student participation. This direction has evolved to emphasize the detection of authentic questions, as explored by [18,19] and further discussed in [20,21].

Beyond questions, other studies apply natural language processing (NLP) techniques to classify rhetorical or argumentative structures. For instance, Lugini and Litman [22] investigates the identification of argument components within student interventions to better understand the structure of classroom discourse. Others, like [23], use low-level acoustic features for the classification.

Although these studies focus on isolated speech units, their findings often serve as the foundation for higher-level analyses of classroom interaction and engagement.

3.1.2. Emotion and Behavior Analysis

This category encompasses studies that analyze group behavior, student engagement, and emotional dynamics in classroom settings. Audio features are leveraged to extract information about interaction patterns, social regulation processes, and affective states, offering insights into how students collaborate and how emotions manifest during learning.

A subset of studies focuses on collaborative behavior and group dynamics. For instance, Dang et al. [24] investigates socially shared regulation in small groups, emphasizing the pedagogical role of silent pauses. Other studies, such as [25], aim to identify different interaction phases (cognitive, metacognitive, and social) during group work.

Diarization-based methods are employed by [26] to analyze participation while preserving student privacy.

Another line of work targets engagement and attention. Refs. [27,28] examine acoustic cues to infer students' interest and concentration levels throughout lessons, aiming to identify moments of disengagement or heightened attention.

Emotional analysis represents a parallel but related focus. Hou et al. [29] examines classroom expressions of encouragement and warmth, while Yuzhong [30] explores student emotions in politically oriented instruction. Teacher emotions are also considered in [31], which classifies emotional tones in teacher speech to understand how effective expression contributes to classroom climate.

Taken together, these studies illustrate how behavioral and emotional insights derived from audio features can illuminate aspects of classroom functioning that are not easily observable through traditional metrics. By capturing how students interact, regulate, and respond emotionally, this category provides critical inputs for the design of supportive learning environments. Moreover, these analyses often serve as a foundational layer for more complex educational interventions, such as those aimed at providing personalized feedback or informing pedagogical adjustments.

3.1.3. Classroom Dynamics

Classroom dynamics captures the evolving nature of classroom interactions by focusing on how teacher strategies, student responses, and the structure of classroom activities unfold over time. Unlike studies that examine static emotional states or isolated events, these works investigate how multiple signals (e.g., verbal, acoustic, or behavioral) interact across temporal sequences to shape the overall learning environment.

A significant focus within this category is the assessment of classroom climate. For example, James et al. [32] combines audio and video features to model the overall atmosphere of a class session. This line of work continues in [33,34], where the authors use the CLASS (classroom assessment scoring system) framework to automatically classify sessions as exhibiting either positive or negative climate characteristics.

Other studies emphasize the temporal structure of lessons and how teaching unfolds in real time. Uzelac et al. [35], for instance, uses student feedback to evaluate the perceived quality of lectures, implicitly assessing the classroom dynamic from the learner's perspective. In more technical approaches, Siddhartha et al. [36] explores the detection of classroom events under noisy conditions, while Cánovas and García [37] applies audio diarization techniques to classify different teaching methodologies based on the structure of interactions observed.

These studies reveal that classroom dynamics are not reducible to isolated actions or emotions but rather emerge from the ongoing interplay of multiple pedagogical and communicative signals. By capturing this temporal complexity, the works in this category provide a crucial lens for understanding teaching effectiveness and learning climate, factors that can guide the design of more adaptive and responsive educational practices.

3.1.4. Feedback Focused

This category includes studies that explicitly use audio-derived information to deliver feedback to educational stakeholders. Unlike works that remain at the descriptive or observational level, these studies aim to close the loop by translating insights into instructional guidance for teachers or learning support for students.

A prominent line of work centers on teacher feedback. Refs. [38,39] quantify teachers' uptake of student ideas, highlighting how conversational patterns can either reinforce or hinder student contributions. Similarly, Cánovas et al. [40] investigates how competitive versus non-competitive response systems influence teacher behavior, providing reflective feedback that helps educators adjust their instructional strategies.

Other studies focus more narrowly on improving questioning techniques. For example, Liu et al. [41] analyzes teachers' focus questions and provides targeted feedback to enhance student engagement. In the same vein, Hunkins et al. [42] examines how specific teacher interventions affect students' motivation, sense of identity, and classroom belonging. In line with this, Dale et al. [43] examine teacher speech to enable self-reflection and support instructional improvement.

On the student side, feedback mechanisms are used to guide learning progress. Gerard et al. [44] introduces an automated support system that helps students answer science questions, while Varatharaj et al. [45] proposes a predictive model to assess fluency and accuracy in language learning. These tools aim to replicate or augment teacher evaluations, providing learners with specific areas for improvement.

3.2. What Are the Most Common Audio Features Used in Educational Studies, and How Are They Extracted? (RQ2)

The studies examined employ a diverse array of audio features, which can be systematically grouped into three distinct groups: acoustic features, diarization-based features, and NLP-based features. These categories not only reflect the level of abstraction of the extracted information but also correspond to different analytical strategies employed across studies. To support interpretability, we provided a structured summary of representative features within each category, offering readers a reference point for the more technical descriptions that follow. To our knowledge, this synthesis constitutes the first structured mapping of audio features and their extraction methods within educational research contexts.

3.2.1. Acoustic Features

Acoustic features operate directly on the raw audio waveform, capturing attributes such as pitch, energy, spectral properties, and Mel-frequency cepstral coefficients (MFCCs). These features allow researchers to analyze classroom environments without relying on text transcriptions, making them particularly useful in noisy settings or when privacy concerns limit transcription. Time-frequency representations (e.g., Mel-spectrograms) are often computed using established signal processing libraries such as Praat, PyAudio, OpenSmile, or Librosa [31,36,46–48].

The main applications of acoustic features relate to classroom climate and emotion and behavior analysis. Variations in prosodic features, such as pitch, energy, or spectral envelopes, have been associated with teacher enthusiasm or student disengagement [49], as well as with shifts in instructional delivery [34,50]. In some cases, acoustic indicators are combined with NLP techniques to refine the detection of emotional states [51] or used as the basis for diarization techniques [52].

Extraction workflows for acoustic features typically rely on digital signal processing (DSP) techniques to compute metrics such as MFCCs, pitch, energy, and zero-crossing rate (ZCR). More recent approaches leverage deep learning models, such as wav2vec2 or OpenL3, to generate acoustic embeddings that capture complex prosodic and paralinguistic information [36,53]. These features have also proven effective for tasks such as teaching methodology classification [10,54] or as a basis for voice activity detection (VAD), which supports downstream processes like diarization [55].

3.2.2. Diarization-Based Features

Diarization-based features aim to determine who is speaking and when enabling detailed turn-taking analyses within classroom interactions. By segmenting audio into speaker-specific tracks, these features capture metrics such as speaking time, the number of interventions, and the distribution of talk among participants [56]. Voice activity detection (VAD) often serves as a preliminary step to isolate speech segments from silence or background noise [57]. These segments are then grouped into speaker clusters using acoustic similarity, typically via embedding-based techniques (e.g., x-vectors) and clustering algorithms. This turn-level information supports the analysis of engagement patterns, such as who initiates discussions or how often students respond to peers [58].

The main applications of diarization-based features fall within emotion and behavior analysis, particularly in the study of classroom participation and collaboration. Knowing who talks and how often provides insight into whether students engage equitably or whether discussions are dominated by specific individuals [59,60]. These same metrics are also used in feedback-focused studies where the ratio of teacher-to-student speaking time or the distribution of teacher attention across the class is quantified [61]. When combined with emotional or linguistic indicators, diarization features can contribute to the detection of classroom climate patterns, such as identifying which speakers tend to offer encouragement or whether teacher-led sequences reflect particular instructional strategies.

Diarization pipelines typically include a voice activity detection module, followed by segmentation and speaker clustering. These steps are commonly implemented using automated toolkits such as pyannote, Kaldi, or Whisper-based diarization, often complemented by pre-trained speaker embedding models. To improve accuracy, some studies incorporate manual corrections or use auxiliary inputs such as individual microphones or spatial data [9,62].

Notably, five studies combine diarization and acoustic features, as reported in Table 1. These combinations are often used to detect engagement patterns or to differentiate between teaching methodologies [9,63]. By aligning speaker turns with acoustic variations, these studies offer a more comprehensive understanding of classroom dynamics than using either feature type alone. For instance, detecting frequent short student utterances alongside shifts in pitch and energy may signal active but uneven participation, while extended teacher monologues with low vocal variability might suggest a more transmissive teaching style. Such multimodal insights provide a nuanced lens through which researchers can infer not just who is speaking, but how that speaking behavior relates to pedagogical effectiveness.

Table 1. Number of papers with every feature category and combination found (RQ2).

Features Combination	Count
Acoustic	21
Acoustic; Diarization	5
Acoustic; NLP	10
Diarization	13
Diarization; NLP	2
NLP	30

3.2.3. NLP-Based Features

Natural language processing (NLP)-based features are often derived from the textual representation of speech. Typically, the first step involves automatically transcribing classroom oral interactions using a speech recognition system (e.g., [64,65]), although in some cases, manual transcriptions are used [66–68]. The quality of transcription, whether automated or manual, critically influences the reliability of downstream NLP analyses. Based on these transcripts, linguistic analysis techniques are applied to extract various indicators, such as word frequency, syntactic complexity, question usage, or keyword identification [18,19,38,69].

Many studies focus primarily on examining how teachers formulate questions, provide feedback, or acknowledge student contributions [70,71]. This emphasis on linguistic dynamics allows for the analysis of teacher support (feedback-focused), as well as aspects of classroom climate and underlying emotions (emotion and behavior) when expressions of encouragement or positive comments are detected [42].

Beyond simple counts of specific words or phrases (e.g., "I" and "We" interrogative words) [19], some studies adopt more advanced NLP approaches, such as TF-IDF, transformer-based classification, or semantic embeddings, to classify intervention types, identify authentic questions, or map discourse flow [38,72]. These methodologies enable researchers to detect specific "pedagogical moves" (e.g., rephrasing, requests for justification, or corrective feedback).

The processing pipeline often begins with an automatic speech recognition system (e.g., Google Speech-to-Text, Otter.ai, IBM Watson, or Whisper) that generates the transcription [64,73]. Next, NLP algorithms are applied to extract linguistic features (e.g., keyword counts, sentiment analysis, grammatical tagging, or discourse segmentation). In more complex cases, deep learning models such as BERT are commonly employed for question classification and semantic encoding, while recurrent architectures (e.g., LSTM) are often used for emotion recognition or discourse modeling [58,74–76].

Finally, as with acoustic and diarization features, NLP features could also be combined with these previous features, as shown in Table 1. The combination with acoustic features is more common, as diarization information is typically integrated into ASR systems when they identify individual speakers. Examples of this integration include analyzing socially shared regulation in collaborative learning or understanding how teacher interventions affect student motivation [24,42].

This range of NLP-based strategies provides researchers with tools to quantify participation and interaction (emotion and behavior analysis), as well as to uncover patterns of teacher feedback and their potential effect on student performance (feedback-focused). Furthermore, NLP is the main group of features used in our defined intervention classification category. Given that verbal content is central to most educational exchanges, NLP-based features offer perhaps the most direct lens into instructional intent and pedagogical nuance, explaining their dominant role across multiple use categories.

To provide a comprehensive overview, we constructed a Sankey diagram (Figure 6) to visualize how the usage categories defined in RQ1 (e.g., feedback-focused and emotion and behavior analysis) relate to the audio features types employed across studies. The diagram illustrates how studies connect these categories to the primary audio feature groups (acoustic, diarization, and NLP-based) and, subsequently, to specific extracted features, such as MFCC, transcriptions, and participant ratios.

It is important to note that this visualization reflects every unique combination of usage and feature types; thus, the total number of connections exceeds the number of studies included. A disaggregated summary of feature counts by usage category is presented in Table 2. Notably, classroom dynamics and emotion and behavior analysis are more frequently associated with acoustic and diarization features, reflecting a focus on capturing environmental and interactional dynamics. In contrast, studies under feedback-focused and intervention classification categories predominantly rely on NLP features, emphasizing the analysis of linguistic content.

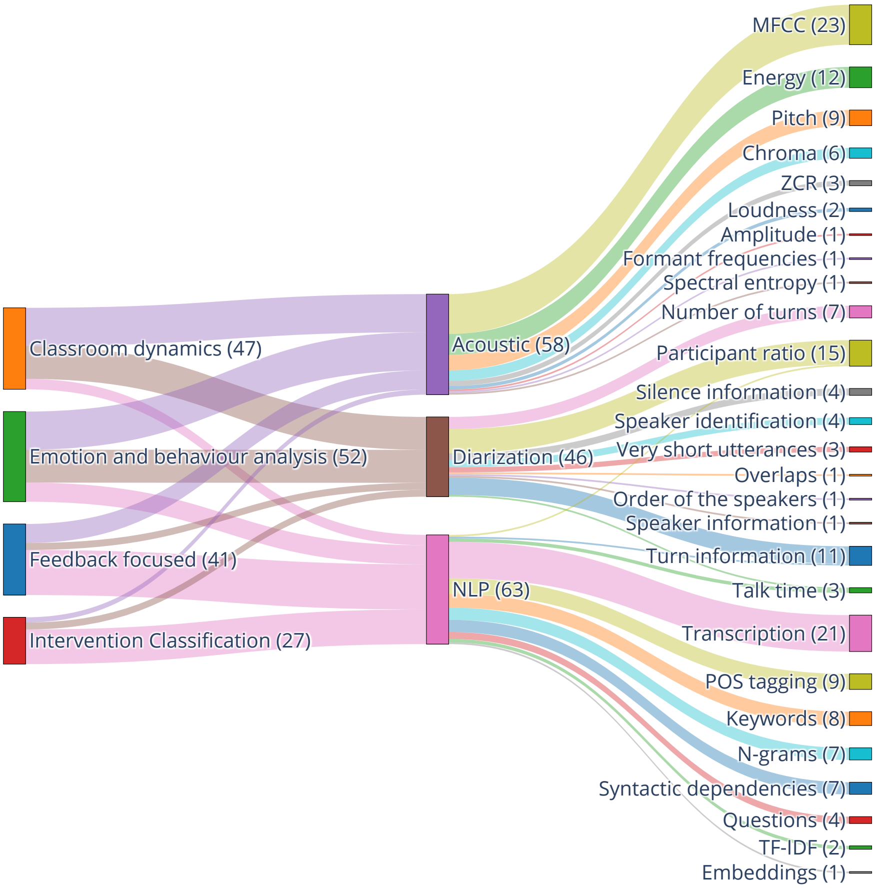

Figure 6. Relationships between types of usage categories (RQ1) and types of audio features identified (RQ2).

Table 2. Number of papers that use each kind of features by RQ1 category (RQ2).

Tag	Features	Count
Classroom dynamics	Acoustic	5
Classroom dynamics	Diarization	2
Emotion and behavior analysis	Acoustic	11
Emotion and behavior analysis	Diarization	10
Emotion and behavior analysis	NLP	8
Feedback focused	Acoustic	6
Feedback focused	Diarization	2
Feedback focused	NLP	18
Intervention Classification	Acoustic	9
Intervention Classification	Diarization	6
Intervention Classification	NLP	10
Technology Development	Acoustic	5
Technology Development	NLP	6

A summary of the most common acoustic features and their functions is presented in Table 3.

Table 3. Description of the main features identified in the literature for acoustic, diarization, and natural language processing (NLP) approaches (RQ2).

Feature Type	Description
Acoustic Features	These capture low-level acoustic properties of audio (e.g., timbre, pitch, intensity, and MFCC).
Spectral Features	Includes Mel-frequency cepstral coefficients (MFCC), filter bank energies (Fbank), formant frequencies, spectral entropy, and spectral centroid. These characterize the frequency content of the audio signal.
Prosodic Features	Encompasses pitch (fundamental frequency), energy/amplitude (volume), and intensity (decibels), which reveal emphasis, speaking style, or emotional cues.
Time-Domain/Statistical Features	Covers zero-crossing rate (ZCR), speech speed/rate (words per minute at the acoustic level), and higher-order statistics (e.g., skewness and kurtosis) of the waveform.
Chroma	Represents the intensity of each of the 12 distinct pitch classes in music/speech, useful for tonal or harmonic analysis.
Diarization Features	These focus on identifying "who spoke when," measuring how speech is distributed among individuals, and capturing dynamics of turn-taking.
Turn-Taking and Number of Turns	Measures each change of speaker or turn in the conversation (e.g., turn counts, very short utterances, and participant order).
Speaking Time/Talk Ratio	Quantifies how long each individual (or group) speaks, useful for comparing teacher vs. student speech.
Speaker Identification/Uniqueness	Detects how many distinct voices appear and how often each participant speaks.
Silence Detection	Tracks periods of no speech (silent pauses, pause duration, and silence ratio), which can indicate reflection or inactivity.
Participation Equality/-Participant Ratio	Reflects whether speech is evenly distributed or dominated by a single speaker.
Speech Overlap/Interruptions	Monitors when multiple speakers talk simultaneously, showing interaction flow.
Direction of Arrival	Locates the position of a speaker in the physical environment, used in multi-microphone setups.
NLP Features	These derive from textual representations of speech (i.e., after transcription) and capture linguistic, semantic, or conversational structures.
Transcription-Based Lexical Features	Direct use of transcribed text (including raw word tokens, word counts, and words per minute).
Keyword/Key-Phrase Detection	Identifies specific terms or question stems (e.g., "why" and "how" domain-related keywords).
POS Tagging and Grammatical Analysis	Uses part-of-speech tags, syntactic dependencies, named entities, or discourse relations.
N-grams, TF-IDF, and Embeddings	Captures local word sequences (n-grams), term-frequency distributions (TF-IDF), or semantic overlap (embedding-based comparisons and pointwise Jensen–Shannon divergence).
Semantic/Pedagogical Indicators	Focuses on features like "teacher uptake of student ideas", sentiment or emotion in text, and question authenticity.

3.3. How Do Researchers Combine Audio Features with Other Data Sources? (RQ3)

Audio features, while rich in information, are often insufficient to capture the full complexity of classroom interactions on their own. Educational settings are inherently multimodal, involving not only spoken language but also gestures, visual cues, physiological signals, and contextual data. To address this, researchers frequently integrate audio with other data sources, a practice widely recognized under the umbrella of multimodal learning analytics (MMLA) [11,77].

This integration is not merely additive but complementary: while audio may capture intonation or turn-taking, video can provide facial expressions and body posture, and log data can contextualize student actions and engagement. Machine learning models, particularly those based on deep learning, have enabled the joint modeling of such heterogeneous signals [78].

In our review, we found that nearly one-third of the analyzed studies adopt a multimodal approach, most commonly combining audio with video features. Other studies also incorporate contextual, environmental, biological, or academic performance data. Figure 7 shows the proportion of multimodal studies, while Table 4 summarizes the frequency and function of each data type across the reviewed literature.

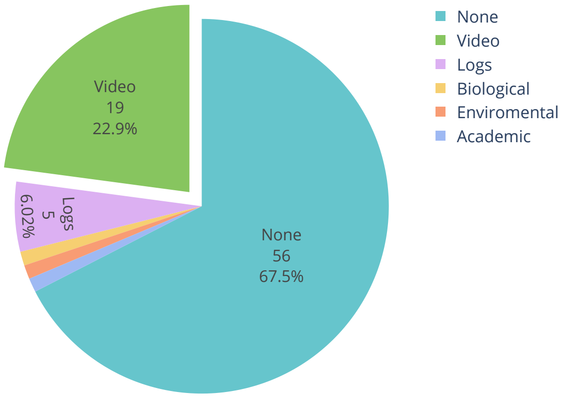

Figure 7. Distribution of analyzed papers that uses a multimodal approach (RQ3).

It is important to note that not every educational task requires multimodal data. In many cases, audio alone suffices to answer the research questions, for example, identifying teacher question types or measuring talk-time distributions. Adding video, physiological signals, or logs could enrich the analysis but also raise the cost and complexity of data collection (e.g., camera setup, synchronization, and classroom consent) and processing. Therefore, although multimodal pipelines are on the rise thanks to modern deep-learning models, a majority of studies remain unimodal because their core goals can be met with audio-only features, which are faster, cheaper, and easier to obtain in real-world school settings.

3.3.1. Combination of Audio and Video Features

Several studies integrate audio and video features to build richer representations of classroom dynamics. While audio captures elements such as speech content, turn-taking, and prosody, video provides complementary signals, including facial expressions, gaze, and posture, offering insights that are otherwise inaccessible through audio alone.

Some works use video primarily as a tool to support the labeling or segmentation of audio data. For example, D'Angelo and Rajarathinam [57] combine diarized audio with video labels to analyze teaching assistant interventions during collaborative problem-solving, allowing for detailed mapping of intervention timing and response. Similarly, refs. [56,60,79,80] use video to segment turn-taking episodes in engineering classrooms, contextualizing audio-based interaction markers with visual cues.

Other studies extract concrete video features that are combined with audio in multimodal machine learning models. For instance, Ramakrishnan et al. [48] use face count and facial emotion recognition to estimate classroom climate alongside acoustic features. Ma et al. [53] incorporate facial action units (FAUs), eye gaze, and body pose to detect confusion and conflict during pair collaboration. Likewise, Heng et al. [9] fuse audio features with pose estimation data to identify teaching methodologies, while Chan et al. [81] include face detection and actions like note-taking to assess student engagement.

Overall, the combination of audio and video data enables a multidimensional view of classroom interactions, supporting the identification of affective states, collaborative dynamics, and pedagogical patterns that would be challenging to infer from a single modality.

3.3.2. Combination of Audio and Contextual Data

Beyond video, several studies integrate audio features with contextual data to deepen their understanding of educational processes. These contextual signals, ranging from log traces of student activity to physiological or environmental measurements, offer complementary information that enriches audio-based analyses.

Log data are some of the most commonly used forms of contextual information. For example, refs. [59,82] combine audio features with student-level editing traces (e.g., number of characters added or deleted) to estimate collaboration quality. Similarly, studies like [83,84] integrate student interaction logs with audio inputs to evaluate the accuracy and usability of teacher dashboards and socially shared regulation.

Some studies also incorporate biological or environmental signals. Prieto et al. [47] combines audio with physiological indicators such as pupil diameter and blink rate to classify instructional formats. In another approach, Uzelac et al. [35] use environmental data (e.g., CO_2 levels, noise, temperature, and air pressure) alongside audio to assess lecture quality. Academic data are also used as a secondary context; for instance, Cánovas et al. [40] integrate students' academic performance with audio-derived group behavior indicators to contextualize responses in audience response systems.

The distribution of feature types used across multimodal studies is summarized in Table 4. Video features dominate the landscape, with 13 instances used in modeling tasks and 6 as support information. In contrast, academic, biological, and environmental data appear more sporadically (each only once), reflecting their specialized use cases. Log data, meanwhile, play a more balanced role, contributing to both modeling and contextual enrichment.

Table 4. Types of non-audio features used in multimodal studies and their role in modeling or support tasks (RQ3).

Feature Type	Model	Support
Academic	0	1
Biological	1	0
Environmental	1	0
Logs	3	2
Video	13	6

3.4. *What Techniques Are Employed to Process Audio Features in Educational Studies, and to What Extent Are These Solutions Interpretable? (RQ4)*

In analyzing how audio features are computationally processed in educational research, we identified three broad methodological categories: statistical analysis, machine learning, and deep learning. These techniques are applied across the full spectrum of audio features discussed in RQ2, including acoustic properties, speaker diarization metrics, and linguistic indicators derived from transcriptions.

This section examines both the nature of these techniques and the degree to which they incorporate strategies for interpretability or explainable AI (xAI). While technical details of the models are not revisited here (see [85] for a comprehensive overview), our focus lies in categorizing the methodological choices made by researchers and evaluating how transparently these models support pedagogical interpretation.

Figure 8 presents an overview of the relationship between each methodological category and its typical level of interpretability. Interpretability was assessed based on whether studies included correlation-based insights, feature importance rankings, or advanced xAI techniques (e.g., SHAP and LIME). As expected, statistical approaches offer the greatest transparency, while deep learning models, though powerful, generally lack interpretive mechanisms.

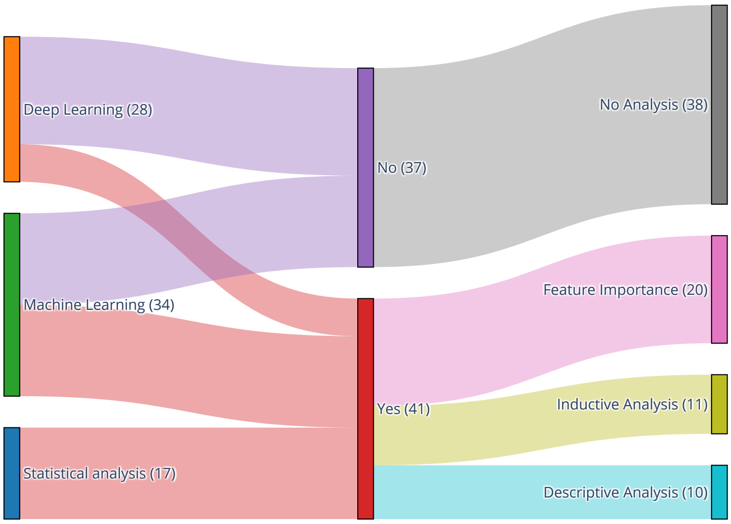

Figure 8. Sankey diagram illustrating the relationship between the analytical methods employed (deep learning, machine learning, and statistical analysis) and the presence and type of explainability techniques. This visualization highlights how interpretability varies across methodological families, showing which types of explainability (if any) are applied in each case (RQ4).

To explore how these methodological choices map to educational purposes, Figure 9 presents the distribution of techniques across the usage categories defined in RQ1. A clear trend emerges: deep learning and machine learning are more frequently applied in domains like classroom dynamics and emotion and behavior analysis, where patterns are abstract or subtle. In contrast, studies categorized under feedback-focused and intervention classification continue to predominantly employ statistical approaches, likely reflecting their closer alignment with the principles of transparency and the generation of actionable feedback.

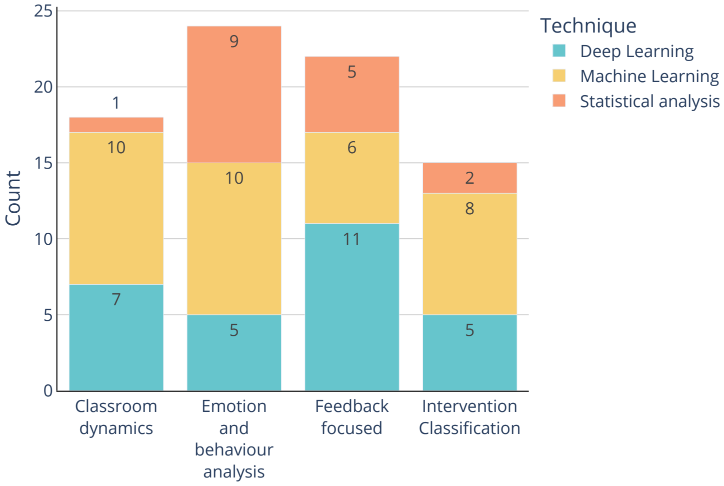

Figure 9. Relationship between RQ1 categories and RQ4 techniques used (RQ4).

For the purposes of this review, we distinguish machine learning algorithms from deep learning models using two objective criteria. First, traditional ML methods (e.g., support vector machines, random forests, and XGBoost) operate on carefully engineered feature vectors (e.g., MFCCs, prosodic statistics, and diarization parameters) and generally contain on the order of 10,000 trainable parameters or fewer. In contrast, deep-learning models ingest raw or minimally processed audio representations, such as spectrograms or learned embeddings, and learn hierarchical features end-to-end. These architectures typically surpass several million trainable parameters and achieve high performance without relying on manual feature selection or preprocessing.

3.4.1. Machine Learning Techniques

Traditional machine learning (ML) pipelines typically rely on engineered features extracted from raw audio, such as prosodic attributes, spectral coefficients, or lexical indicators. These features are then fed into algorithms such as random forests, support vector machines (SVMs), decision trees, or logistic regression to perform classification, regression, or clustering tasks. A core strength of these approaches lies in their relative transparency compared to deep learning approaches: many models allow for the direct inspection of feature importance, aiding both explainability and pedagogical interpretation.

Numerous studies in this category focus on classifying aspects of classroom discourse and instructional style. For instance, Tsalera et al. [86] introduces a PCA-based feature selection strategy to manage high-dimensional audio in noisy classrooms, subsequently comparing multiple classifiers. Random forests are especially popular: Donnelly et al. [87] uses them to infer teaching methodologies (e.g., lecture vs. group work) by identifying the most salient acoustic features, while Prieto et al.[47] applies a similar strategy to detect instructional formats, discussing the most predictive feature sets for each activity type.

Some researchers go further by combining audio with other data streams. Donnelly et al. [88], for example, fuses audio with wearable sensor data to segment teacher–student interactions using naive Bayes classifiers. In contexts with limited technological resources, simpler classifiers such as KNN or naive Bayes have been tested for low-cost classroom monitoring and analysis [89,90]. Similarly, Sandanayake and Bandara [91] proposes a multimodal summarization pipeline combining KNN, speech-to-text models, and textual clustering to automatically generate lecture summaries.

Feature engineering remains central to these ML-based studies, and authors frequently report which features (e.g., pitch variation and lexical tokens) contribute most to the model's predictions [32]. This emphasis on transparency distinguishes ML from deep learning approaches. However, more advanced explainability techniques, such as SHAP [92] and LIME [93], are rarely used in practice. One notable exception is [29], which leverages SHAP to visualize the contribution of emotional features to model decisions.

Overall, while ML pipelines strike a useful balance between performance and interpretability, the integration of explicit xAI practices remains inconsistent. There is considerable room for improvement in using these tools to strengthen transparency and trust, especially when outputs are intended to guide pedagogical decision-making.

3.4.2. Deep Learning Techniques

Deep learning (DL) approaches, such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and transformers, typically learn representations of audio directly from spectrograms or even raw waveforms. This reduces reliance on handcrafted features but often comes at the cost of increased model opacity, given the layered and non-linear nature of these architectures.

In educational contexts, CNNs and LSTMs are frequently applied to tasks such as classroom sound classification and speech transcription under noisy conditions. For example, Mou et al. [94] trains both CNN and LSTM architectures on spectrogram inputs to categorize classroom events, while Siddhartha et al. [36] designs a custom network to handle the interruptions and acoustic challenges typical of early childhood environments. Transformers, particularly BERT, are used in studies involving transcribed classroom discourse. For example, Wang and Chen [64] investigates the impact of removing specific word categories on classification accuracy, while Alic et al. [95] employs BERT to differentiate between funneling and focusing questions.

More complex architectures are also explored. Alkhamali et al . [46], for instance, combines CNNs, LSTMs, and transformers in an ensemble to predict emotional states from classroom audio, reflecting a growing interest in modeling subtle affective cues. Multimodal integration is another emerging trend, as discussed in Section 3.3: Heng et al. [9] fuses audio and video streams to model classroom climate, showcasing how DL architectures naturally accommodate multimodal inputs.

Despite these technical advances, interpretability remains a major limitation. Most DL studies prioritize performance metrics (e.g., accuracy and F1-score) over explainability. When attempts are made, they are typically superficial, such as inspecting attention weights in transformers or correlating predictions with input features. Rigorous explainability frameworks, such as SHAP or integrated gradients, are rarely used.

This lack of transparency is particularly problematic in educational settings, where stakeholders must understand and trust system outputs to make informed decisions. Without interpretable outputs, DL models risk becoming “black boxes” that may perform well technically but lack pedagogical legitimacy.

3.4.3. Statistical Analyses

A third methodological category relies on classical statistical methods, which often focus on describing, correlating, or regressing measured audio variables against educational outcomes. Unlike machine learning or deep learning models, these approaches typically avoid predictive modeling. Instead, they aim to uncover interpretable relationships between acoustic or discourse-related metrics and constructs such as student engagement or teacher–student dynamics.

Several studies exemplify this approach. For instance, D’Angelo and Rajarathinam [57] analyzes the talk-time of teaching assistants and relates it to collaborative group performance using basic correlation measures. The same authors [56] similarly links instructor talk patterns to student responses through correlation coefficients. Others apply regression-based methods: Demszky et al. [58] models how real-time measures of teacher talk-time predict engagement levels. Experimental comparisons also appear, such as in [63], where an “engagement index” derived from acoustic and visual cues is compared statistically across baseline and intervention conditions. Studies like [37] use descriptive and correlational statistics to examine the structure of instructional discourse, while Hardman [96] interprets teacher-to-student talk ratios as indicators of classroom authority and control. Along similar lines, France [97] analyzes how teachers value dialogue and how they implement it in classroom practice, drawing on statistical analyses to explore its perceived importance and actual use.

These methods offer a key advantage: immediate interpretability. Coefficients, effect sizes, and p-values provide clear, direct evidence of how specific features relate to educational variables. This transparency makes statistical analysis particularly useful when communicating findings to educators and stakeholders who may lack technical expertise.

However, this interpretability often comes at the expense of modeling complexity. Statistical approaches may overlook nuanced patterns or interactions that more sophisticated machine learning techniques can detect. Nonetheless, their use remains widespread, especially in studies prioritizing theoretical validation or descriptive insight over predictive accuracy.

3.4.4. Extent of Interpretability and xAI Practices

Overall, most articles focus on improving predictive accuracy or illustrating empirical relationships rather than detailing how each audio feature drives model decisions. Nevertheless, three broad approaches to interpretability appear with varying frequencies:

1. Inductive (Correlation-Based) Analysis: Several studies employ correlation and regression analyses to explore associations between audio features and educational outcomes. While not true explainability methods in the xAI sense, these inductive approaches are used as proxies to validate whether model decisions align with expected behavioral patterns. For example, Chejara et al. [82] correlates audio-based collaboration features with model outcomes to check generalizability, while Chejara et al. [98] similarly relies on correlation metrics to validate whether ML models remain robust across different classroom environments. These methods provide initial evidence of construct validity but fall short of revealing how individual predictions are made.
2. Feature Importance: Tree-based ML algorithms such as random forests or gradient-boosted trees enable direct analysis of feature contributions, often through built-in importance metrics. These have been used to highlight which acoustic or linguistic features drive model predictions. For instance, Donnelly et al. [99] identifies para verbal signals as key markers for detecting teacher questions, and James et al. [32] uses feature importance to analyze contributors to perceived classroom climate. In a more advanced case, Hou et al. [29] applies SHAP values to clarify how emotional features contribute to warm or encouraging feedback. While useful, such practices are applied inconsistently across studies and often without methodological transparency or justification for the selected xAI technique.
3. Descriptive Analysis: Some studies enhance interpretability by comparing model outputs with human-annotated ground truth or classroom observations. These comparisons, while not algorithmic explanations, provide qualitative insight into how predictions align with real-world phenomena. For instance, Cook et al. [18] illustrates discrepancies between its regression-tree predictions and human-coded discourse segments, while Kelly et al. [19] analyzes how the model compares with the performance of human annotators. These approaches can build trust among end users, particularly educators, by revealing whether the system offers pedagogical value in their reasoning. However, they remain anecdotal and rarely constitute a systematic framework for interpretability.

Across the reviewed literature, statistical models offer inherent transparency by design, while ML models provide moderate interpretability depending on the availability of feature attribution methods. In contrast, DL approaches tend to sacrifice transparency in favor of performance, with only rare instances incorporating formal xAI tools. This trade-off between predictive power and explainability is a key concern for the deployment of these systems in real educational settings, where understanding the basis of decisions is often as critical as the output itself. In our corpus, only 7 out of 27 deep learning studies (26%) explicitly reported the use of explainability techniques, compared to 17 out of 34 machine learning studies (50%). These figures highlight a significant interpretability gap in the most complex and widely adopted approaches, one that deserves more attention if educational stakeholders are to trust and adopt AI-powered systems in practice.

3.5. Which Studies Provide Feedback for Participants Derived from Obtained Results? (RQ5)

Providing feedback to teachers and students is a key mechanism for translating analytical insights into pedagogical action. In educational settings, feedback serves as a bridge between data and practice, allowing teachers to refine their instructional strategies, students to reflect on their learning behaviors, and both groups to engage in more effective classroom interactions. Without such mechanisms, the potential of audio-based analytics remains largely theoretical, disconnected from the real-world contexts they are intended to support.

In this question, we examine how studies in our corpus integrate feedback into their design. Specifically, we analyze who receives the feedback, what type of information is shared, and when it is delivered. This approach allows us to map the practical utility of audio-derived data across educational scenarios.

However, it is important to highlight a critical limitation in the current literature: very few studies discuss the perception, acceptance, or impact of feedback mechanisms from the perspective of the actual stakeholders–teachers and students. While some papers mention user-facing dashboards or post-session summaries, most do not include empirical evaluations of how these outputs were received or used in practice, nor whether they resulted in actual pedagogical changes. This lack of user-centered assessment restricts our ability to draw conclusions about the effectiveness, usability, or educational value of these systems. We return to this gap in the discussion as a key avenue for future research.

Of the 82 reviewed articles, only 11 (approximately 13%) explicitly report delivering feedback derived from audio features to participants. Despite being a minority, these studies provide valuable insights into the current state of feedback integration in audio-based educational research. To structure our analysis, we organize the findings according to three dimensions: who receives the feedback, what kind of information is shared, and when the feedback is delivered.

3.5.1. Who Receives the Feedback?

Most studies delivering feedback target teachers as the primary recipients. The goal is typically to help them refine pedagogical strategies using insights derived from their own classroom discourse, for example, their use of specific talk moves, questioning patterns, or instructional vocabulary. In one case, teachers received personalized statistics on their use of mathematics terminology and the distribution of teacher–student talk-time, with comparisons against their own past data and the behavior of other platform users [100]. Other studies similarly provide post-session feedback summarizing discourse patterns or question types, encouraging gradual, data-informed pedagogical refinement [41,71,101,102].

While teacher-focused feedback dominates, a smaller group of studies extends feedback to students, either directly or through mediated teacher actions. Some systems benefit both teachers and students simultaneously by visualizing classroom participation in real time, for example, dashboards that display talk-time proportions or overlapping speech, prompting more balanced turn-taking [58]. Other systems offer individualized feedback to students, such as metrics on pronunciation accuracy or fluency scores. In these cases, teachers may also receive alerts when students fall below performance thresholds, allowing for timely instructional interventions [44,45]. This dual-feedback approach can support not only student reflection but also responsive teaching.

3.5.2. What Kind of Feedback Is Delivered?

Most feedback systems in the reviewed literature rely on quantitative metrics to make classroom dynamics visible. These include measures such as talk-time, frequency of authentic questions, or discipline-specific vocabulary use, which serve as interpretable baselines for reflection and instructional refinement. For example, one study provides real-time speaking ratios that help teachers and students rebalance participation mid-lesson [58]. Another tracks how often teachers pose authentic questions and examine the relationship between these frequencies and student engagement levels [101].

Beyond raw counts, several platforms enhance quantitative feedback with interpretative guidance. They highlight specific moments where teachers used effective discourse strategies and offer actionable suggestions, such as rephrasing or elaboration moves, that may promote deeper student reasoning [41,100]. Dashboards frequently incorporate color-coded visualizations to identify zones of high or low engagement, helping educators focus attention where it is most needed [27,83].

Notably, while quantitative feedback dominates, few studies attempt to provide qualitative or normative feedback, for example, judgments about whether a teacher's interaction style aligns with pedagogical best practices. This absence may reflect an implicit reluctance to define what constitutes 'good teaching.' Educational contexts vary widely, and there is little consensus on ideal instructional behavior. In practice, providing such evaluative guidance would require not only technical robustness but also normative frameworks capable of accounting for differences in subject matter, age group, and cultural setting. As a result, most systems remain focused on reporting metrics rather than interpreting them in light of pedagogical theory or instructional goals.

3.5.3. When Is Feedback Delivered?

Feedback timing varies considerably across studies, revealing trade-offs between immediate interventions and reflective practice.

Real-Time Feedback

A small subset of studies implement real-time feedback, offering live insights during classroom sessions. In these scenarios, both teachers and students make immediate adjustments: one system updates a display of talk-time balances every 20 min, leading to a quick rebalancing of classroom discourse [58]. Another approach visualizes student interest levels in real time, allowing instructors to quickly pivot if engagement appears to wane [27].

Post-Session Feedback

More commonly, feedback is delivered after a session or across multiple sessions. Teachers might receive summaries of their questioning techniques, discourse moves, or engagement metrics after each class [71,100,102]. This setup favors reflective practice, allowing educators to review and adapt without the pressure of real-time classroom management. Longitudinal feedback, where data is collected and returned across weeks or months, has also been explored to track instructional change over time [41,101].

Single Exposure or Irregular Delivery

A few studies deliver feedback only once or at irregular intervals, often within pilot trials or prototype demonstrations. For instance, [83] uses vignette-based dashboards to gauge educators' trust in feedback systems but does not implement sustained use. These instances serve more as proof-of-concept explorations than fully integrated classroom tools.

4. Discussion of Findings and Implications

This section critically synthesizes the main findings of our review, highlighting key limitations and opportunities in the current use of audio features within educational research. While the reviewed studies demonstrate considerable technical sophistication, spanning acoustic, diarization, and linguistic features, multimodal integration, and advanced processing techniques, we identify three recurring issues that constrain the field's practical impact.

4.1. Challenges in Explainability

This section addresses a critical pattern that cuts across multiple research questions: despite the richness of audio-derived data (RQ2), its combination with other modalities (RQ3), the sophistication of processing techniques (RQ4), and even its intended use for feedback (RQ1 and RQ5), few studies succeed in translating analytics into pedagogical action.

As seen in RQ2, researchers extract a wide array of features, from low-level acoustic metrics and speaker diarization to sophisticated NLP indicators, offering a detailed representation of classroom discourse. These features are then processed using machine learning (RQ4), including explainable models such as decision trees and feature attribution methods or with deep learning pipelines that promise high performance but limited interpretability. For audio analytics in particular, misidentified prosodic cues or diarization errors can quickly erode teacher confidence; without transparent rationales, educators are unlikely to trust or act on the model's recommendations. While a few studies do attempt to provide actionable feedback, often through visual dashboards or targeted recommendations, these remain exceptions. The vast majority stop at reporting descriptive indicators, shifting the responsibility for interpretation and pedagogical action entirely to the user.

This limitation is particularly evident in RQ1, where feedback provision emerges as one of the most common use cases. However, as shown in RQ5, the feedback offered is typically generic, quantitative, and delivered post hoc. While teachers may receive dashboards displaying metrics such as talk ratios or the frequency of authentic questions, these indicators are seldom accompanied by an explanation of their relevance or instructional implications. Moreover, when feedback is provided, it often lacks evidence of real-world deployment or user adoption, as most systems are confined to academic prototypes and remain untested in practical classroom environments [103].

Embedding a human-in-the-loop (HITL) methodology can address both the need for explainability and meaningful stakeholder involvement. By design, HITL systems require transparent, interpretable outputs so that educators can review, correct, and refine model predictions in real time. This process ensures that each step, from feature extraction to final recommendations, is accompanied by an explanation that teachers understand and trust. Simultaneously, involving teachers (and, when possible, students) in iterative model refinement grounds development in actual classroom practices, leading to tools that align with pedagogical goals and have higher adoption rates. For example, Qui et al. [104] demonstrate usage of GenAI for teaching and learning systems in which instructors monitor and adjust LLM-generated feedback during live sessions, ensuring that suggested interventions remain contextually appropriate.

A further challenge lies in the widespread reluctance to make normative claims about what constitutes "good" teaching. This hesitation is understandable, given the diversity of educational contexts and pedagogical philosophies. However, in the absence of interpretive scaffolds or evaluative frameworks, feedback risks becoming either meaningless or misinterpreted. Translating analytical outputs into actionable insights would require the integration of explainable models with established pedagogical principles, an approach that remains rare in the current literature.

Yet the problem may not lie in the individual contributions of each study but in their isolation. When viewed collectively, the field has already developed many of the building blocks needed for real-world applications. Tools like TeachFX (https://teachfx.com) demonstrate that it is possible to deliver personalized, real-time feedback on teacher discourse using audio analytics. These tools draw upon many of the same techniques and features found in our reviewed corpus, suggesting that research findings are indeed translatable if properly integrated.

4.2. Data Availability and Privacy Constraints

One of the most striking patterns across the reviewed literature is the widespread use of locally collected, non-public datasets, often recorded in specific classrooms or institutions to support experimental prototypes [105]. While this is understandable given the logistical and ethical complexities of educational data collection, it has resulted in a highly fragmented research landscape, particularly when it comes to audio and multimodal data (RQ2 and RQ3).

There are currently no widely adopted standards for how audio data should be recorded in educational settings. Some studies use a single microphone placed centrally in the classroom; others distribute multiple microphones across the room, and some attach individual recorders to each participant. These decisions, driven by convenience, resources, or technical constraints, profoundly affect the quality and type of features that can be extracted, such as signal-to-noise ratios, speaker separability for diarization, or prosodic fidelity (RQ2). As a consequence, models trained under one recording setup may struggle to generalize to another (RQ4), limiting the applicability of proposed solutions.

This issue is further compounded by the near-total absence of dataset publication. Very few studies share their recordings and even fewer offer access to accompanying metadata or extracted features. As a result, it is almost impossible to replicate findings or benchmark new methods on common grounds. This problem is especially critical in a field that increasingly relies on complex AI models whose performance can be sensitive to minor variations in data distribution.

Notably absent from most papers is a structured consideration of privacy. Across the 82 reviewed studies, explicit discussions of data protection, anonymization, or ethical frameworks were rare despite the sensitive nature of the data involved. This omission is particularly striking given the frequent use of multimodal inputs (e.g., audio, transcriptions, and video), often without an accompanying explanation of how personally identifiable information is safeguarded. Although some studies propose using anonymized diarization features to preserve privacy [26], most treat privacy as an implicit assumption, not a methodological constraint.

However, this invisibility of privacy does not make the problem disappear. Transcripts can expose personal opinions or sensitive content; diarization can link speech to individuals; video can reveal faces and physical environments. Longitudinal studies (e.g., RQ5) raise the stakes even further by enabling behavioral tracking over time. In this context, informed consent is necessary but insufficient. Compliance with data protection frameworks like the General Data Protection Regulation (GDPR) also requires data minimization, purpose limitation, and the right to erasure. Voice is a biometric identifier protected under GDPR Articles 4 and 9, and FERPA treats student speech as personally identifiable information; consequently, large repositories of raw classroom audio cannot be shared across institutions without extensive anonymization or legal agreements.

A few notable exceptions do exist. For instance, the TalkMoves dataset [106] provides access to annotated classroom transcriptions focused on mathematics instruction in K–12 settings. While it does not include raw audio, thus limiting its use for acoustic or diarization analysis, it represents an important step toward sharing structured, educational data in a privacy-conscious way. However, such efforts remain isolated and domain-specific.

Given these constraints, the field must find a middle ground between analytical ambition and ethical responsibility [107]. Directly sharing raw classroom audio or video may be unfeasible, but a compelling alternative lies in the standardized extraction and anonymized release of feature sets. These may include acoustic measures, diarization statistics, and NLP-derived indicators curated to strip away identifiable information while retaining pedagogical relevance (RQ2 and RQ4). If accompanied by rich metadata about the educational context (i.e., subject, age group, cultural background), such datasets could enable meaningful generalization testing across settings (RQ4, RQ5).

Ultimately, advancing the field will require us to move beyond isolated case studies and toward a collective infrastructure that is reproducible, privacy-conscious, and pedagogically focused. The creation of shared, anonymized feature-level benchmarks is not just a technical necessity, it is a foundation for trustworthy, ethical, and scalable educational research.

4.3. Lack of Pedagogical Interpretation in Analytical Results

A pervasive limitation across the reviewed literature is the tendency to report quantitative results, such as talk-time distributions, question types, or engagement indices, without translating them into pedagogically actionable recommendations. While these studies offer detailed analytics, they rarely address the practical question that educators face: what does this mean for my teaching? Instead, the interpretive burden is implicitly placed on the teacher, who must infer whether the measured patterns are desirable, problematic, or contextually appropriate. The feedback, even when provided, is often generic and descriptive, lacking the guidance necessary to inform instructional decision-making [108].

This disconnect reveals an epistemological mismatch between the precision of computational models and the situated complexity of educational practice. For example, detecting that a teacher asked few “authentic questions” may highlight a pattern, but it does not clarify whether this pattern was pedagogically appropriate for that lesson’s objectives, student level, or classroom culture. In other cases, studies flag high or low teacher talk-time yet offer no interpretive baseline against which to judge these values. As a result, many outputs remain technically sophisticated but pedagogically inert, becoming precise measurements with unclear implications [109].

In many areas of educational analytics, advanced models have already outperformed classical approaches. For instance, Aljohani, et al. [110] demonstrated that a model based on neural networks trained on clickstream data from a virtual learning environment achieved higher accuracy in predicting at-risk students than traditional logistic regression. This suggests that adopting more complex architectures to learn hierarchical audio representations directly, rather than relying on manually engineered features, could similarly yield more robust and generalizable insights in classroom audio analytics.

Furthermore, integrating analytics within established formative assessment models [111] can transform raw metrics into actionable guidance. For example, if audio analytics reveal that students speak less than 10% of class time, a teacher can use that evidence to adjust instruction in real time, perhaps by posing more open-ended questions or incorporating think–pair–share activities to boost engagement. By aligning audio-derived insights with the feedback-for-learning cycle (gather evidence, interpret, adjust instruction, and gather new evidence), educators receive concrete steps, rather than abstract data, to enhance teaching and learning outcomes.

To bridge this gap, a more integrated research design that includes the active participation of pedagogical experts from the outset is required. These experts can help determine which features are educationally meaningful and how they should be interpreted within varied classroom contexts [112]. Moreover, their input is essential for translating numeric results into normative statements: not just what is happening, but whether it should be happening, and if not, what might be done differently, as reflected in approaches such as design-based research and human-centered learning analytics [113].

Generative AI offers a promising complement in this regard. When fine-tuned on domain-specific corpora, including classroom recordings, transcriptions, and expert commentary, LLMs could support teachers by delivering conversational, context-aware interpretations of feedback, improving what other works already do [114,115]. Instead of receiving abstract metrics or visualizations, teachers could ask questions like "Is it a problem that I spoke 80% of the time?" or "What does it mean that I used mostly funneling questions?" and receive grounded, nuanced answers. By embedding expert pedagogical reasoning into the model's responses, LLMs could act as a bridge between analytic outputs and instructional sense-making.

Recent papers offer concrete classroom demonstrations of LLM-driven feedback. Ref. [116] reports on "Feedback on Feedback", where few-shot prompting with a large language model produces personalized writing feedback that students rate as more specific and useful than teacher-only comments. Likewise, ref. [104] presents a human-in-the-loop GenAI dashboard that lets instructors monitor and tweak real-time LLM guidance during lessons, ensuring suggestions remain pedagogically sound. If LLM systems can already support students in these ways, similar architectures could be adapted for teaching analytics on audio-derived metrics and actionable ideas embedded in formative assessment cycles or other instructional design models.

5. Conclusions

This systematic review provides a focused examination of how audio features are used within educational research, encompassing low-level acoustic features, speaker diarization metrics, and linguistic indicators derived via NLP. These features are analyzed across different levels of abstraction and use, providing a structured account of their role in modeling classroom discourse and supporting educational analysis.

Beyond this mapping, our synthesis reveals several systemic limitations that constrain the practical impact of these technologies. Despite significant technical progress, many studies fall short of translating their outputs into insights that are pedagogically actionable. While some recent studies offer promising approaches, e.g., interpretable models or privacy-aware feature extraction, these remain exceptions. Most contributions still operate in fragmented data silos with limited pedagogical scaffolding or replicability. These challenges reflect a persistent misalignment between the analytical capabilities of current systems and the practical needs of educators and learners.

To advance the field, we identify three strategic priorities. First, enhancing the explainability of analytic systems is essential to ensure that stakeholders can understand, trust, and make informed use of model outputs. Second, the development and publication of standardized, anonymized feature-level datasets is critical for improving reproducibility, enabling cross-context evaluation, and ensuring ethical use of sensitive data. Third, future research should prioritize the active involvement of educational experts throughout the design process, fostering systems that are not only technically robust but also aligned with real-world pedagogical needs.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app15126911/s1, PRISMA 2020 Main Checklist.

Funding: This work was 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.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Summary of all the analyzed papers, with the year of publication, educational level analyzed and the context of the research.

Author	Title	Year	Level	Context
Dang, Belle and Nguyen, Andy and Järvelä, Sanna	The Unspoken Aspect of Socially Shared Regulation in Collaborative Learning: AI-Driven Learning Analytics Unveiling 'Silent Pauses'	2024	K12	In-person
Jacobs, Jennifer and Scornavacco, Karla and Clevenger, Charis and Suresh, Abhijit and Sumner, Tamara	Automated feedback on discourse moves: teachers' perceived utility of a professional learning tool	2024	K12	In-person
Alkhamali, Eman Abdulrahman and Allinjawi, Arwa and Ashari, Rehab Bahaaddin	Combining Transformer, Convolutional Neural Network, and Long Short-Term Memory Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms	2024	Higher education	Online
D'Angelo, Cynthia M. and Rajarathinam, Robin Jephthah	Speech analysis of teaching assistant interventions in small group collaborative problem solving with undergraduate engineering students	2024	Higher education	In-person
Wang, Deliang and Chen, Gaowei	Are perfect transcripts necessary when we analyze classroom dialogue using AIoT?	2024	K12	In-person
Chejara, Pankaj and Kasepalu, Reet and Prieto, Luis P. and Rodríguez-Triana, Mar\ía Jesús and Ruiz Calleja, Adolfo and Schneider, Bertrand	How well do collaboration quality estimation models generalize across authentic school contexts?	2024	Higher education	In-person
Liu, Xiaoting and Gu, Wen and Ota, Koichi and Hasegawa, Shinobu	Design of Voice Style Detection of Lecture Archives	2023	Higher education	In-person
Chejara, Pankaj and Prieto, Luis P. and Rodriguez-Triana, Maria Jesus and Kasepalu, Reet and Ruiz-Calleja, Adolfo and Shankar, ShashiKant Kant and Jesús Rodríguez-Triana, María and Calleja, Adolfo-Ruiz and Kasepalu, Reet and Shankar, ShashiKant Kant and Rodriguez-Triana, Maria Jesus and Kasepalu, Reet and Ruiz-Calleja, Adolfo and Shankar, ShashiKant Kant	How to Build More Generalizable Models for Collaboration Quality? Lessons Learned from Exploring Multi-Context Audio-Log Datasets using Multimodal Learning Analytics	2023	Higher education	In-person
Cosbey, Robin and Wusterbarth, Allison and Hutchinson, Brian	Deep Learning for Classroom Activity Detection from Audio	2019	Higher education	In-person
Ma, Yingbo and Celepkolu, Mehmet and Boyer, Kristy Elizabeth and Lynch, Collin F. and Wiebe, Eric and Israel, Maya	How Noisy is Too Noisy? The Impact of Data Noise on Multimodal Recognition of Confusion and Conflict During Collaborative Learning	2023	K12	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Rajarathinam, Robin Jephthah and D'Angelo, Cynthia M.	Description of Instructor Intervention Using Individual Audio Data in Co-Located Collaboration	2023	Higher education	In-person
Solopova, Veronika and Rostom, Eiad and Cremer, Fritz and Gruszczynski, Adrian and Witte, Sascha and Zhang, Chengming and López, Fernando Ramos and Plößl, Lea and Hofmann, Florian and Romeike, Ralf and Gläser-Zikuda, Michaela and Benzmüller, Christoph and Landgraf, Tim	PapagAI: Automated Feedback for Reflective Essays	2023	Higher education	In-person
Canovas, Oscar and Garcia, Felix J.	Analysis of Classroom Interaction Using Speaker Diarization and Discourse Features from Audio Recordings	2023	Higher education	In-person
Demszky, Dorottya and Wang, Rose and Geraghty, Sean and Yu, Carol	Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform	2024	K12	Online
Jensen, Emily and Dale, Meghan and Donnelly, Patrick J. and Stone, Cathlyn and Kelly, Sean and Godley, Amanda and D'Mello, Sidney K.	Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning	2020	Multiple	In-person
Demszky, Dorottya and Liu, Jing	M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes	2023	K12	Online
Demszky, Dorottya and Liu, Jing and Hill, Heather C. and Jurafsky, Dan and Piech, Chris	Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course	2024	Higher education	Online
Rajarathinam, Robin Jephthah and Dangelo, Cynthia M.	Turn-taking analysis of small group collaboration in an engineering discussion classroom	2023	Higher education	In-person
Nazaretsky, Tanya and Mikeska, Jamie N. and Beigman Klebanov, Beata and Mikeska, Jamie N. and Beigman Klebanov, Beata	Empowering Teacher Learning with AI: Automated Evaluation of Teacher Attention to Student Ideas during Argumentation-focused Discussion	2023	K12	Simulation
Cv, Siddhartha and Rao, Preeti and Velmurugan, Rajbabu and Siddhartha, C. V. and Rao, Preeti and Velmurugan, Rajbabu	Classroom Activity Detection in Noisy Preschool Environments with Audio Analysis	2023	Toddlers	In-person
Albaladejo-González, Mariano and Gaspar-Marco, Rubén and Mármol, Félix Gómez and Reich, Justin and Ruipérez-Valiente, José A	Improving Teacher Training Through Emotion Recognition and Data Fusion	2024	K12	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Canovas, Oscar and Garcia-Clemente, Felix J. and Pardo, Federico	AI-driven Teacher Analytics: Informative Insights on Classroom Activities	2023	Higher education	In-person
Li, Zongxi and Xie, Haoran and Wang, Minhong and Wu, Bian and Hu, Yiling	Automatic Coding of Collective Creativity Dialogues in Collaborative Problem Solving Based on Deep Learning Models	2022	K12	In-person
Kasepalu, Reet and Chejara, Pankaj and Prieto, Luis P. and Ley, Tobias	Do Teachers Find Dashboards Trustworthy, Actionable and Useful? A Vignette Study Using a Logs and Audio Dashboard	2022		
Schlotterbeck, Danner and Jiménez, Abelino and Araya, Roberto and Caballero, Daniela and Uribe, Pablo and Van der Molen Moris, Johan	Teacher, Can You Say It Again? Improving Automatic Speech Recognition Performance over Classroom Environments with Limited Data	2022	K12	In-person
Southwell, Rosy and Pugh, Samuel and Perkoff, E. Margaret and Clevenger, Charis and Bush, Jeffrey B. and Lieber, Rachel and Ward, Wayne and Foltz, Peter and DMello, Sidney	Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms	2022	K12	In-person
Zhang, Shaoyun and Li, Chao	Research on Feature Fusion Speech Emotion Recognition Technology for Smart Teaching	2022	Higher education	Online
Hunkins, Nicholas and Kelly, Sean and DMello, Sidney	"beautiful work, youre rock stars!": Teacher Analytics to Uncover Discourse that Supports or Undermines Student Motivation, Identity, and Belonging in Classrooms	2022	K12	In-person
Alic, Sterling and Demszky, Dorottya and Mancenido, Zid and Liu, Jing and Hill, Heather and Jurafsky, Dan	Computationally Identifying Funneling and Focusing Questions in Classroom Discourse	2022	K12	In-person
Dale, Meghan E. and Godley, Amanda J. and Capello, Sarah A. and Donnelly, Patrick J. and DMello, Sidney K. and Kelly, Sean P.	Toward the automated analysis of teacher talk in secondary ELA classrooms	2022	K12	In-person
Yuzhong, Hou	Students emotional analysis on ideological and political teaching classes based on artificial intelligence and data mining	2021		
Emara, Mona and Hutchins, Nicole M. and Grover, Shuchi and Snyder, Caitlin and Biswas, Gautam	Examining student regulation of collaborative, computational, problem-solving processes in openended learning environments	2021	K12	In-person
France, Ann	Teachers Using Dialogue to Support Science Learning in the Primary Classroom	2021	K12	In-person
Cánovas Reverte, Óscar and González Férez, Pilar and García Clemente, Félix J. and Pardo García, Federico	Analyzing Wooclap's Competition Mode with AI Through Classroom Recordings	2024	Higher education	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Albaladejo-González, Mariano and Gaspar-Marco, Rubén and Mármol, Félix Gómez and Reich, Justin and Ruipérez-Valiente, José A	Improving Teacher Training Through Emotion Recognition and Data Fusion	2024		Simulation
Hou, Ruikun and Fütterer, Tim and Bühler, Babette and Bozkir, Efe and Gerjets, Peter and Trautwein, Ulrich and Kasneci, Enkelejda	Automated Assessment of Encouragement and Warmth in Classrooms Leveraging Multimodal Emotional Features and ChatGPT	2024	K12	In-person
Sun, Anchen and Londono, Juan J. and Elbaum, Batya and Estrada, Luis and Lazo, Roberto Jose and Vitale, Laura and Villasanti, Hugo Gonzalez and Fusaroli, Riccardo and Perry, Lynn K. and Messinger, Daniel S.	Who Said what? An Automated Approach to Analyzing Speech in Preschool Classrooms	2024	Toddlers	In-person
Canovas, Oscar and Garcia, Felix J.	Analysis of Classroom Interaction Using Speaker Diarization and Discourse Features from Audio Recordings	2023		In-person
García, Federico Pardo and Cánovas, Óscar and García Clemente, Félix J.	Exploring AI Techniques for Generalizable Teaching Practice Identification	2024	Higher education	In-person
Schlotterbeck, Danner and Uribe, Pablo and Jiménez, Abelino and Araya, Roberto and van der Molen Moris, Johan and Caballero, Daniela	TARTA: Teacher Activity Recognizer from Transcriptions and Audio	2021	K12	In-person
Jensen, Emily and Pugh, Samuel L. and Dmello, Sidney K.	A deep transfer learning approach to modeling teacher discourse in the classroom	2021	K12	In-person
Li, Zongxi and Xie, Haoran and Wang, Minhong and Wu, Bian and Hu, Yiling	Automatic Coding of Collective Creativity Dialogues in Collaborative Problem Solving Based on Deep Learning Models	2022	K12	In-person
Schlotterbeck, Danner and Uribe, Pablo and Araya, Roberto and Jimenez, Abelino and Caballero, Daniela	What classroom audio tells about teaching: A cost-effective approach for detection of teaching practices using spectral audio features	2021	K12	In-person
Tsalera, Eleni and Papadakis, Andreas and Samarakou, Maria	Novel principal component analysis-based feature selection mechanism for classroom sound classification	2021	Higher education	In-person
Demszky, Dorottya and Liu, Jing and Mancenido, Zid and Cohen, Julie and Hill, Heather and Jurafsky, Dan and Hashimoto, Tatsunori	Measuring conversational uptake: A case study on student-teacher interactions	2021	K12	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Chejara, Pankaj and Prieto, Luis P. and Ruiz-Calleja, Adolfo and Rodríguez-Triana, María Jesús and Shankar, Shashi Kant and Kasepalu, Reet	Quantifying collaboration quality in face-to-face classroom settings using mmla	2020	K12	In-person
Jensen, Emily and Dale, Meghan and Donnelly, Patrick J. and Stone, Cathlyn and Kelly, Sean and Godley, Amanda and DMello, Sidney K.	Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning	2020	K12	In-person
Khan, Muhammed S. and Zualkernan, Imran	Using Convolutional Neural Networks for Smart Classroom Observation	2020	K12	In-person
Sharma, Archana and Mansotra, Vibhakar	Multimodal decision-level group sentiment prediction of students in classrooms	2019	K12	Multiple
Varatharaj, Ashvini and Botelho, Anthony F. and Lu, Xiwen and Heffernan, Neil T.	Supporting teacher assessment in Chinese language learning using textual and tonal features	2020	Higher education	In-person
Jie, Liang and Zhao, Xiaoyan and Zhang, Zhaohui	Speech Emotion Recognition of Teachers in Classroom Teaching	2020		
Yang, Bohong and Yao, Zeping and Lu, Hong and Zhou, Yaqian and Xu, Jinkai	In-classroom learning analytics based on student behavior, topic and teaching characteristic mining	2020	Higher education	In-person
Sharma, Archana and Mansotra, Vibhakar	Multimodal decision-level group sentiment prediction of students in classrooms	2019	K12	In-person
Suresh, Abhijit and Sumner, Tamara and Jacobs, Jennifer and Foland, Bill and Ward, Wayne	Automating analysis and feedback to improve mathematics teachers classroom discourse	2019	K12	In-person
Jensen, Emily and Dale, Meghan and Donnelly, Patrick J. and Stone, Cathlyn and Kelly, Sean and Godley, Amanda and DMello, Sidney K.	Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning	2020	Multiple	In-person
Su, Hang and Dzodzo, Borislav and Wu, Xixin and Liu, Xunying and Meng, Helen	Unsupervised methods for audio classification from lecture discussion recordings	2019	Higher education	In-person
James, Anusha and Chua, Yi Han Victoria and Maszczyk, Tomasz and Núñez, Ana Moreno and Bull, Rebecca and Lee, Kerry and Dauwels, Justin	Automated classification of classroom climate by audio analysis	2019	Toddlers	In-person
Ahuja, Karan and Kim, Dohyun and Xhakaj, Franceska and Varga, Virag and Xie, Anne and Zhang, Stanley and Townsend, Jay Eric and Harrison, Chris and Ogan, Amy and Agarwal, Yuvraj	EduSense: Practical Classroom Sensing at Scale	2019	Higher education	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Barbadekar, Ashwinee and Gaikwad, Vijay and Patil, Sanjay and Chaudhari, Tushar and Deshpande, Shardul and Burad, Saloni and Godbole, Rohini	Engagement Index for Classroom Lecture using Computer Vision	2019		In-person
Viswanathan, Sree Aurovindh and VanLehn, Kurt	Collaboration detection that preserves privacy of students speech	2019	Higher education	In-person
James, Anusha and Chua, Yi Han Victoria and Maszczyk, Tomasz and Núñez, Ana Moreno and Bull, Rebecca and Lee, Kerry and Dauwels, Justin	Automated classification of classroom climate by audio analysis	2019	Higher education	In-person
Sharma, Archana and Mansotra, Vibhakar	Multimodal decision-level group sentiment prediction of students in classrooms	2019	Higher education	In-person
Cosbey, Robin and Wusterbarth, Allison and Hutchinson, Brian	Deep Learning for Classroom Activity Detection from Audio	2019	Multiple	In-person
Gerard, Libby and Kidron, Ady and Linn, Marcia C.	Guiding collaborative revision of science explanations	2019	K12	In-person
Kelly, Sean and Olney, Andrew M. and Donnelly, Patrick and Nystrand, Martin and DMello, Sidney K.	Automatically Measuring Question Authenticity in Real-World Classrooms	2018	K12	In-person
Shapsough, Salsabeel and Zualkernan, Imran	Using Machine Learning to Automate Classroom Observation for Low-Resource Environments	2018	K12	In-person
Howard, Sarah K. and Yang, Jie and Ma, Jun and Ritz, Chrisian and Zhao, Jiahonz and Wynne, Kylie	Using Data Mining and Machine Learning Approaches to Observe Technology-Enhanced Learning	2018	K12	In-person
James, Anusha and Kashyap, Mohan and Chua, Yi Han Victoria and Maszczyk, Tomasz and Nunez, Ana Moreno and Bull, Rebecca and Dauwels, Justin	Inferring the Climate in Classrooms from Audio and Video Recordings: A Machine Learning Approach	2018	Toddlers	In-person
Lugini, Luca and Litman, Diane	Argument Component Classification for Classroom Discussions	2018	K12	In-person
Cook, Connor and Olney, Andrew M. and Kelly, Sean and DMello, Sidney K.	An open vocabulary approach for estimating teacher use of authentic questions in classroom discourse	2018	K12	In-person
Uzelac, Ana and Gligoric, Nenad and Krco, Srdan and Gligorić, Nenad and Krčo, Srđan	System for recognizing lecture quality based on analysis of physical parameters	2018	Higher education	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Owens, Melinda T. and Seidel, Shannon B. and Wong, Mike and Bejines, Travis E. and Lietz, Susanne and Perez, Joseph R. and Sit, Shangheng and Subedar, Zahur-Saleh Saleh and Acker, Gigi N. and Akana, Susan F. and Balukjian, Brad and Benton, Hilary P. et al.	Classroom sound can be used to classify teaching practices in college science courses	2017	Higher education	In-person
Donnelly, Patrick J. and Kelly, Sean and Blanchard, Nathaniel and Nystrand, Martin and Olney, Andrew M. and DMello, Sidney K.	Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context	2017	K12	In-person
Donnelly, Patrick and Blanchard, Nathan and Samei, Borhan and Olney, Andrew M. and Sun, Xiaoyi and Ward, Brooke and Kelly, Sean and Nystrand, Martin and DMello, Sidney K.	Automatic teacher modeling from live classroom audio	2016	K12	In-person
Blanchard, Nathaniel and Donnelly, Patrick J. and Olney, Andrew M. and Samei, Borhan and Ward, Brooke and Sun, Xiaoyi and Kelly, Sean and Nystrand, Martin and DMello, Sidney K.	Identifying Teacher Questions Using Automatic Speech Recognition in Classrooms	2016	K12	In-person
Donnelly, Patrick J. and Blanchard, Nathaniel and Samei, Borhan and Olney, Andrew M. and Sun, Xiaoyi and Ward, Brooke and Kelly, Sean and Nystrand, Martin and DMello, Sidney K.	Multi-Sensor modeling of teacher instructional segments in live classrooms	2016	K12	In-person
Hardman, Jan	Tutor–student interaction in seminar teaching: Implications for professional development	2016	Higher education	In-person
Prieto, Luis P. and Sharma, Kshitij and Dillenbourg, Pierre and Jesús, María	Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors	2016	K12	In-person
Blanchard, Nathaniel and Donnelly, Patrick J. and Olney, Andrew M. and Samei, Borhan and Ward, Brooke and Sun, Xiaoyi and Kelly, Sean and Nystrand, Martin and DMello, Sidney K.	Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms	2016	K12	In-person
Gligoric, Nenad and Uzelac, Ana and Krco, Srdjan and Kovacevic, Ivana and Nikodijevic, Ana	Smart classroom system for detecting level of interest a lecture creates in a classroom	2015	Higher education	In-person

Table A1. *Cont.*

Author	Title	Year	Level	Context
Li, Zongxi and Xie, Haoran and Wang, Minhong and Wu, Bian and Hu, Yiling	Automatic Coding of Collective Creativity Dialogues in Collaborative Problem Solving Based on Deep Learning Models	2022	K12	In-person
Samei, Borhan and Li, Haiying and Keshtkar, Fazel and Rus, Vasile and Graesser, Arthur C.	Context-based speech act classification in intelligent tutoring systems	2014	K12	Online

References

1. Elkins, D.; Hickerson, T. The use of the tape recorder in teacher education. *J. Teach. Educ.* **1964**, *15*, 432–438. [CrossRef]
2. Ochoa, X.; Worsley, M. Augmenting learning analytics with multimodal sensory data. *J. Learn. Anal.* **2016**, *3*, 213–219. [CrossRef]
3. Praharaj, S.; Scheffel, M.; Specht, M.; Drachsler, H. Measuring collaboration quality through audio data and learning analytics. In *Unobtrusive Observations of Learning in Digital Environments: Examining Behavior, Cognition, Emotion, Metacognition and Social Processes Using Learning Analytics*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 91–110.
4. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. *SciPy* **2015**, *2015*, 18–24.
5. Bredin, H.; Yin, R.; Coria, J.M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; Gill, M.P. Pyannote. audio: Neural building blocks for speaker diarization. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7124–7128.
6. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
7. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-Strength Natural Language Processing in Python. 2020. Available online: https://spacy.io/ (accessed on 17 June 2025).
8. Lee, L.K.; Cheung, S.K.; Kwok, L.F. Learning analytics: Current trends and innovative practices. *J. Comput. Educ.* **2020**, *7*, 1–6. [CrossRef]
9. Heng, C.H.; Toyoura, M.; Leow, C.S.; Nishizaki, H. Analysis of Classroom Processes Based on Deep Learning With Video and Audio Features. *IEEE Access* **2024**, *12*, 110705–110712. [CrossRef]
10. Schlotterbeck, D.; Uribe, P.; Araya, R.; Jimenez, A.; Caballero, D. What classroom audio tells about teaching: A cost-effective approach for detection of teaching practices using spectral audio features. In Proceedings of the LAK21: LAK21: 11th International Learning Analytics and Knowledge Conference, Stanford, CA, USA, 12–16 April 2021; pp. 132–140.
11. Worsley, M. Multimodal learning analytics: Enabling the future of learning through multimodal data analysis and interfaces. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; pp. 353–356.
12. Wang, J.; Dudy, S.; He, X.; Wang, Z.; Southwell, R.; Whitehill, J. Speaker Diarization in the Classroom: How Much Does Each Student Speak in Group Discussions? In Proceedings of the 17th International Conference on Educational Data Mining, Long Beach, CA, USA, 9–12 July 2024; pp. 360–367.
13. Wisniewski, B.; Zierer, K.; Hattie, J. The power of feedback revisited: A meta-analysis of educational feedback research. *Front. Psychol.* **2020**, *10*, 487662. [CrossRef]
14. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
15. Chadegani, A.A.; Salehi, H.; Yunus, M.M.; Farhadi, H.; Fooladi, M.; Farhadi, M.; Ebrahim, N.A. A comparison between two main academic literature collections: Web of Science and Scopus databases. *arXiv* **2013**, arXiv:1305.0377. [CrossRef]
16. Fang, S.; Gao, B.; Wu, Y.; Teoh, T.T. Unibrivl: Robust universal representation and generation of audio driven diffusion models. *arXiv* **2023**, arXiv:2307.15898.
17. Blanchard, N.; Donnelly, P.J.; Olney, A.M.; Samei, B.; Ward, B.; Sun, X.; Kelly, S.; Nystrand, M.; D'Mello, S.K. Identifying Teacher Questions Using Automatic Speech Recognition in Classrooms. In Proceedings of the SIGDIAL 2016—17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference, Los Angeles, CA, USA, 13–15 September 2016; pp. 191–201.

18. Cook, C.; Olney, A.M.; Kelly, S.; D'Mello, S.K. An open vocabulary approach for estimating teacher use of authentic questions in classroom discourse. In Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018); International Educational Data Mining Society, Buffalo, NY, USA, 16–20 July 2018; pp. 493–498.
19. Kelly, S.; Olney, A.M.; Donnelly, P.; Nystrand, M.; D'Mello, S.K. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educ. Res.* **2018**, *47*, 451–464. [CrossRef]
20. Schaffalitzky, C. What Makes Authentic Questions Authentic? *Dialogic Pedagog.* **2022**, *10*, A30–A42. [CrossRef]
21. Liu, X.; Gu, W.; Ota, K.; Hasegawa, S. Design of Voice Style Detection of Lecture Archives. In Proceedings of the IEEE Region 10 Annual International Conference, TENCON 2023, Perth, Australia, 31 October–3 November 2023; pp. 1139–1144.
22. Lugini, L.; Litman, D. Argument Component Classification for Classroom Discussions. In Proceedings of the EMNLP 2018—5th Workshop on Argument Mining, Brussels, Belgium, 31 October–4 November 2018; pp. 57–67.
23. Khan, M.S.; Zualkernan, I. Using Convolutional Neural Networks for Smart Classroom Observation. In Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication Fukuoka, Japan, 21–24 February 2020; pp. 608–612.
24. Dang, B.; Nguyen, A.; Järvelä, S. The Unspoken Aspect of Socially Shared Regulation in Collaborative Learning: AI-Driven Learning Analytics Unveiling 'Silent Pauses'. In Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24), Kyoto, Japan, 18–22 March 2024; ACM Press: New York, NY, USA, 2024; pp. 231–240.
25. Li, Z.; Xie, H.; Wang, M.; Wu, B.; Hu, Y. Automatic Coding of Collective Creativity Dialogues in Collaborative Problem Solving Based on Deep Learning Models. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), International Conference on Artificial Intelligence in Education (AIED 2022), Durham, UK, 27–31 July 2022; Volume 13357 LNCS, pp. 123–134.
26. Viswanathan, S.A.; VanLehn, K. Collaboration detection that preserves privacy of students' speech. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, International Conference on Artificial Intelligence in Education (AIED 2019), Chicago, IL, USA, 25–29 June 2019; Volume 11625 LNAI, pp. 507–517.
27. Gligoric, N.; Uzelac, A.; Krco, S.; Kovacevic, I.; Nikodijevic, A. Smart classroom system for detecting level of interest a lecture creates in a classroom. *J. Ambient Intell. Smart Environ.* **2015**, *7*, 271–284. [CrossRef]
28. Yang, B.; Yao, Z.; Lu, H.; Zhou, Y.; Xu, J. In-classroom learning analytics based on student behavior, topic and teaching characteristic mining. *Pattern Recognit. Lett.* **2020**, *129*, 224–231. [CrossRef]
29. Hou, R.; Fütterer, T.; Bühler, B.; Bozkir, E.; Gerjets, P.; Trautwein, U.; Kasneci, E. Automated Assessment of Encouragement and Warmth in Classrooms Leveraging Multimodal Emotional Features and ChatGPT. In *Proceedings of the Lecture Notes in Computer Science*; Springer Science and Business Media Deutschland GmbH: Berlin/Heidelberg, Germany, 2024; Volume 14829 LNAI, pp. 60–74.
30. Hou, Y. Students' emotional analysis on ideological and political teaching classes based on artificial intelligence and data mining. *J. Intell. Fuzzy Syst.* **2021**, *40*, 3801–3809.
31. Jie, L.; Zhao, X.; Zhang, Z. Speech Emotion Recognition of Teachers in Classroom Teaching. In Proceedings of the 32nd Chinese Control and Decision Conference (CCDC 2020), Hefei, China, 22–24 August 2020; pp. 5045–5050.
32. James, A.; Kashyap, M.; Chua, Y.H.V.; Maszczyk, T.; Nunez, A.M.; Bull, R.; Dauwels, J. Inferring the Climate in Classrooms from Audio and Video Recordings: A Machine Learning Approach. In Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2018), Wollongong, NSW, Australia, 4–7 December 2018; pp. 983–988.
33. James, A.; Chua, Y.H.V.; Maszczyk, T.; Núñez, A.M.; Bull, R.; Lee, K.; Dauwels, J. Automated classification of classroom climate by audio analysis. In Proceedings of the 9th International Workshop on Spoken Dialogue System Technology, Singapore, 18–20 April 2018; Volume 579, pp. 41–49.
34. Ramakrishnan, A.; Ottmar, E.; LoCasale-Crouch, J.; Whitehill, J. Toward automated classroom observation: Predicting positive and negative climate. In Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019.
35. Uzelac, A.; Gligorić, N.; Krčo, S. System for recognizing lecture quality based on analysis of physical parameters. *Telemat. Inform.* **2018**, *35*, 579–594. [CrossRef]
36. Siddhartha, C.V.; Rao, P.; Velmurugan, R. Classroom Activity Detection in Noisy Preschool Environments with Audio Analysis. In Proceedings of the 2023 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES), Sivakasi, India, 6–7 April 2023; pp. 1–6.
37. Canovas, O.; Garcia, F.J. Analysis of Classroom Interaction Using Speaker Diarization and Discourse Features from Audio Recordings. In Proceedings of the Learning in the Age of Digital and Green Transition (ICL 2022), Vienna, Austria, 27–30 September 2022; Volume 634 LNNS, pp. 67–74.
38. Demszky, D.; Liu, J.; Mancenido, Z.; Cohen, J.; Hill, H.; Jurafsky, D.; Hashimoto, T. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. *arXiv* **2021**, arXiv:2106.03873.

39. Demszky, D.; Liu, J.; Hill, H.C.; Jurafsky, D.; Piech, C. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educ. Eval. Policy Anal.* **2024**, *46*, 483–505. [CrossRef]
40. Cánovas, O.; González, P.; Clemente, F.J.G.; Pardo, F. Analyzing Wooclap's competition mode with AI through classroom recordings. *IEEE Rev. Iberoam. Tecnol. Aprendiz.* **2024**, *19*, 220–229.
41. Liu, J.; Hill, H.C.; Sanghi, S.; Chung, A.; Demszky, D. *Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms*; Annenberg Institute for School Reform at Brown University: Providence, RI, USA, 2023.
42. Hunkins, N.; Kelly, S.; D'Mello, S. "Beautiful work, you're rock stars!": Teacher Analytics to Uncover Discourse that Supports or Undermines Student Motivation, Identity, and Belonging in Classrooms. In Proceedings of the ACM International Conference Proceeding Series, International Conference on Learning Analytics and Knowledge (LAK '22), Evry, France, 21–25 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 230–238.
43. Dale, M.E.; Godley, A.J.; Capello, S.A.; Donnelly, P.J.; D'Mello, S.K.; Kelly, S.P. Toward the automated analysis of teacher talk in secondary ELA classrooms. *Teach. Teach. Educ.* **2022**, *110*, 103584. [CrossRef]
44. Gerard, L.; Kidron, A.; Linn, M.C. Guiding collaborative revision of science explanations. *Int. J. Comput.-Support. Collab. Learn.* **2019**, *14*, 291–324. [CrossRef]
45. Varatharaj, A.; Botelho, A.F.; Lu, X.; Heffernan, N.T. Supporting teacher assessment in Chinese language learning using textual and tonal features. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), International Conference on Artificial Intelligence in Education (AIED 2020), Ifrane, Morocco, 6–10 July 2020; Volume 12163 LNAI, pp. 562–573.
46. Alkhamali, E.A.; Allinjawi, A.; Ashari, R.B. Combining Transformer, Convolutional Neural Network, and Long Short-Term Memory Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms. *Appl. Sci.* **2024**, *14*, 5050. [CrossRef]
47. Prieto, L.P.; Sharma, K.; Dillenbourg, P.; Jesús, M. Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors. In Proceedings of the ACM International Conference Proceeding Series, Niagara Falls, ON, Canada, 6–9 November 2016; Volume 25-29-Apri, pp. 148–157.
48. Ramakrishnan, A.; Zylich, B.; Ottmar, E.; Locasale-Crouch, J.; Whitehill, J. Toward Automated Classroom Observation: Multimodal Machine Learning to Estimate CLASS Positive Climate and Negative Climate. *IEEE Trans. Affect. Comput.* **2023**, *14*, 664–679. [CrossRef]
49. Sharma, A.; Mansotra, V. Multimodal decision-level group sentiment prediction of students in classrooms. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 4902–4909. [CrossRef]
50. Zhang, S.; Li, C. Research on Feature Fusion Speech Emotion Recognition Technology for Smart Teaching. *Mob. Inf. Syst.* **2022**, *2022*, 82–92. [CrossRef]
51. Albaladejo-González, M.; Gaspar-Marco, R.; Mármol, F.G.; Reich, J.; Ruipérez-Valiente, J.A. Improving Teacher Training Through Emotion Recognition and Data Fusion. *Expert Syst.* **2024**, *17*, 200171. [CrossRef]
52. Li, H.; Kang, Y.; Ding, W.; Yang, S.; Yang, S.; Huang, G.Y.; Liu, Z. Multimodal Learning for Classroom Activity Detection. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Virtual, 4–8 May 2020; pp. 9234–9238.
53. Ma, Y.; Celepkolu, M.; Boyer, K.E.; Lynch, C.F.; Wiebe, E.; Israel, M. How Noisy is Too Noisy? The Impact of Data Noise on Multimodal Recognition of Confusion and Conflict During Collaborative Learning. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023.
54. Su, H.; Dzodzo, B.; Wu, X.; Liu, X.; Meng, H. Unsupervised methods for audio classification from lecture discussion recordings. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 3347–3351.
55. Cosbey, R.; Wusterbarth, A.; Hutchinson, B. Deep Learning for Classroom Activity Detection from Audio. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 3727–3731.
56. Rajarathinam, R.J.; D'Angelo, C.M. Description of Instructor Intervention Using Individual Audio Data in Co-Located Collaboration. In Proceedings of the Computer-Supported Collaborative Learning Conference, CSCL, Montreal, QC, Canada, 10–14 July 2023; Volume 2023-June, pp. 317–320.
57. D'Angelo, C.M.; Rajarathinam, R.J. Speech analysis of teaching assistant interventions in small group collaborative problem solving with undergraduate engineering students. *Br. J. Educ. Technol.* **2024**, *55*, 1583–1601 [CrossRef]
58. Demszky, D.; Wang, R.; Geraghty, S.; Yu, C. *Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform*; EdWorkingPaper No. 23-891; Annenberg Institute for School Reform at Brown University: Providence, RI, USA, 2023
59. Chejara, P.; Prieto, L.P.; Ruiz-Calleja, A.; Rodríguez-Triana, M.J.; Shankar, S.K.; Kasepalu, R. Quantifying collaboration quality in face-to-face classroom settings using mmla. In Proceedings of the Lecture Notes in Computer Science (Including Subseries

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), International Conference on Artificial Intelligence in Education (AIED 2020), Ifrane, Morocco, 6–10 July 2020; Volume 12324 LNCS, pp. 159–166.

60. Rajarathinam, R.J.; D'Angelo, C.M. Turn-taking analysis of small group collaboration in an engineering discussion classroom. In Proceedings of the ACM International Conference Proceeding Series, Association for Computing Machinery, International Conference on Learning Analytics and Knowledge (LAK '23), Arlington, TX, USA, 13–17 March 2023; pp. 650–656.
61. Pardo, F.; Cánovas, O.; Clemente, F.J.G. Exploring AI techniques for generalizable teaching practice identification. *IEEE Access* **2024**, *12*, 134702–134713.
62. Wang, Z.; Pan, X.; Miller, K.F.; Cortina, K.S. Automatic classification of activities in classroom discourse. *Comput. Educ.* **2014**, *78*, 115–123. [CrossRef]
63. Barbadekar, A.; Gaikwad, V.; Patil, S.; Chaudhari, T.; Deshpande, S.; Burad, S.; Godbole, R. Engagement Index for Classroom Lecture using Computer Vision. In Proceedings of the 2019 Global Conference for Advancement in Technology (GCAT 2019), Bengaluru, India, 4–6 October 2019.
64. Wang, D.; Chen, G. Are perfect transcripts necessary when we analyze classroom dialogue using AIoT? *Internet Things* **2024**, *25*, 101105 [CrossRef]
65. Sun, A.; Londono, J.J.; Elbaum, B.; Estrada, L.; Lazo, R.J.; Vitale, L.; Villasanti, H.G.; Fusaroli, R.; Perry, L.K.; Messinger, D.S. Who Said what? An Automated Approach to Analyzing Speech in Preschool Classrooms. In Proceedings of the 2024 IEEE International Conference on Development and Learning (ICDL 2024), Pittsburgh, PA, USA, 15–18 July 2024.
66. Southwell, R.; Pugh, S.; Perkoff, E.M.; Clevenger, C.; Bush, J.B.; Lieber, R.; Ward, W.; Foltz, P.; D'Mello, S. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. In Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022), Durham, UK, 27–31 July 2022.
67. Blanchard, N.; Donnelly, P.J.; Olney, A.M.; Samei, B.; Ward, B.; Sun, X.; Kelly, S.; Nystrand, M.; D'Mello, S.K. Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. In Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016), Raleigh, NC, USA, 4–7 July 2016; pp. 288–291.
68. Schlotterbeck, D.; Jiménez, A.; Araya, R.; Caballero, D.; Uribe, P.; der Molen Moris, J.V. "Teacher, Can You Say It Again?" Improving Automatic Speech Recognition Performance over Classroom Environments with Limited Data. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), International Conference on Artificial Intelligence in Education (AIED 2022), Durham, UK, 27–31 July 2022; Volume 13355 LNCS, pp. 269–280.
69. Samei, B.; Li, H.; Keshtkar, F.; Rus, V.; Graesser, A.C. Context-based speech act classification in intelligent tutoring systems. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), International Conference on Artificial Intelligence in Education (AIED 2014), Memphis, TN, USA, 1–5 July 2014; Volume 8474 LNCS, pp. 236–241.
70. Suresh, A.; Sumner, T.; Jacobs, J.; Foland, B.; Ward, W. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 2019), Honolulu, HI, USA, 27 January–1 February 2019; pp. 9721–9728.
71. Demszky, D.; Liu, J. M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes. In Proceedings of the Tenth ACM Conference on Learning @ Scale, Copenhagen, Denmark, 20–22 July 2023; pp. 23–759.
72. Song, Y.; Lei, S.; Hao, T.; Lan, Z.; Ding, Y. Automatic Classification of Semantic Content of Classroom Dialogue. *J. Educ. Comput. Res.* **2021**, *59*, 496–521. [CrossRef]
73. Jensen, E.; Pugh, S.L.; D'mello, S.K. A deep transfer learning approach to modeling teacher discourse in the classroom. In Proceedings of the LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA, 12–16 April 2021; pp. 302–312.
74. Solopova, V.; Rostom, E.; Cremer, F.; Gruszczynski, A.; Witte, S.; Zhang, C.; Lopez, F.R.; Ploessl, L.; Hofmann, F.; Romeike, R.; et al. PapagAI: Automated Feedback for Reflective Essays. In Proceedings of the Advances in Artificial Intelligence (KI 2023), Berlin, Germany, 26–29 September 2023; Volume 14236, pp. 198–206.
75. Nazaretsky, T.; Mikeska, J.N.; Klebanov, B.B. Empowering Teacher Learning with AI: Automated Evaluation of Teacher Attention to Student Ideas during Argumentation-focused Discussion. In Proceedings of the LAK23: 13th International Learning Analytics and Knowledge Conference, Arlington, TX, USA, 13–17 March 2023; Volume 1, pp. 122–132.
76. Schlotterbeck, D.; Uribe, P.; Jiménez, A.; Araya, R.; van der Molen Moris, J.; Caballero, D. TARTA: Teacher Activity Recognizer from Transcriptions and Audio. In Proceedings of the Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, 14–18 June 2021; Volume 12748 LNAI, pp. 369–380.
77. Blikstein, P. Multimodal learning analytics. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium, 8–13 April 2013; pp. 102–106.

78. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the International Conference on Machine Learning (ICML), Bellevue, WA, USA, 28 June–2 July 2011; Volume 11, pp. 689–696.
79. Ahuja, K.; Kim, D.; Xhakaj, F.; Varga, V.; Xie, A.; Zhang, S.; Townsend, J.E.; Harrison, C.; Ogan, A.; Agarwal, Y. EduSense: Practical Classroom Sensing at Scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 1–26. [CrossRef]
80. Howard, S.K.; Yang, J.; Ma, J.; Ritz, C.; Zhao, J.; Wynne, K. Using Data Mining and Machine Learning Approaches to Observe Technology-Enhanced Learning. In Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2018), Wollongong, NSW, Australia, 4–7 December 2018; pp. 788–793.
81. Chan, M.C.E.; Ochoa, X.; Clarke, D. Multimodal learning analytics in a laboratory classroom. In *Machine Learning Paradigms: Advances in Learning Analytics*; Springer: Cham, Switzerland, 2020; Volume 158, pp. 131–156.
82. Chejara, P.; Kasepalu, R.; Prieto, L.P.; Rodríguez-Triana, M.J.; Calleja, A.R.; Schneider, B. How well do collaboration quality estimation models generalize across authentic school contexts? *Br. J. Educ. Technol.* **2023**, *55*, 1602–1624. [CrossRef]
83. Kasepalu, R.; Chejara, P.; Prieto, L.P.; Ley, T. Do Teachers Find Dashboards Trustworthy, Actionable and Useful? A Vignette Study Using a Logs and Audio Dashboard. *Technol. Knowl. Learn.* **2022**, *27*, 971–989. [CrossRef]
84. Emara, M.; Hutchins, N.M.; Grover, S.; Snyder, C.; Biswas, G. Examining student regulation of collaborative, computational, problem-solving processes in openended learning environments. *J. Learn. Anal.* **2021**, *8*, 49–74. [CrossRef]
85. Wang, D.; Tao, Y.; Chen, G. Artificial intelligence in classroom discourse: A systematic review of the past decade. *Int. J. Educ. Res.* **2024**, *123*, 102275 [CrossRef]
86. Tsalera, E.; Papadakis, A.; Samarakou, M. Novel principal component analysis-based feature selection mechanism for classroom sound classification. *Comput. Intell.* **2021**, *37*, 1827–1843. [CrossRef]
87. Donnelly, P.; Blanchard, N.; Samei, B.; Olney, A.M.; Sun, X.; Ward, B.; Kelly, S.; Nystrand, M.; D'Mello, S.K. Automatic teacher modeling from live classroom audio. In Proceedings of the UMAP 2016—2016 Conference on User Modeling Adaptation and Personalization, Halifax, NS, Canada, 13–16 July 2016; pp. 45–53.
88. Donnelly, P.J.; Blanchard, N.; Samei, B.; Olney, A.M.; Sun, X.; Ward, B.; Kelly, S.; Nystrand, M.; D'Mello, S.K. Multi-Sensor modeling of teacher instructional segments in live classrooms. In Proceedings of the ICMI 2016—18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 177–184.
89. Shapsough, S.; Zualkernan, I. Using Machine Learning to Automate Classroom Observation for Low-Resource Environments. In Proceedings of the GHTC 2018—IEEE Global Humanitarian Technology Conference, San Jose, CA, USA, 18–21 October 2018.
90. Cánovas, O.; Clemente, F.J.G.; Pardo, F. AI-driven Teacher Analytics: Informative Insights on Classroom Activities. In Proceedings of the 2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE 2023), Auckland, New Zealand, 28 November–1 December 2023
91. Sandanayake, T.C.; Bandara, A.M. Automated classroom lecture note generation using natural language processing and image processing techniques. *Int. J. Adv. Trends Comput. Sci. Eng.* **2019**, *8*, 1920–1926.
92. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
93. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
94. Mou, A.; Milanova, M.; Baillie, M. Deep Learning Approaches for Classroom Audio Classification Using Mel Spectrograms. In Proceedings of the New Approaches for Multidimensional Signal Processing (NAMSP 2022), Sofia, Bulgaria, 23–25 June 2022.
95. Alic, S.; Demszky, D.; Mancenido, Z.; Liu, J.; Hill, H.; Jurafsky, D. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse. *arXiv* **2022**, arXiv:2208.04715.
96. Hardman, J. Tutor–student interaction in seminar teaching: Implications for professional development. *Act. Learn. High. Educ.* **2016**, *17*, 63–76. [CrossRef]
97. France, A. Teachers Using Dialogue to Support Science Learning in the Primary Classroom. *Res. Sci. Educ.* **2021**, *51*, 845–859. [CrossRef]
98. Chejara, P.; Prieto, L.P.; Rodriguez-Triana, M.J.; Kasepalu, R.; Ruiz-Calleja, A.; Shankar, S.K. 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 the ACM International Conference Proceeding Series, International Conference on Learning Analytics and Knowledge (LAK '23), Arlington, TX, USA, 13–17 March 2023; pp. 111–121.
99. Donnelly, P.J.; Kelly, S.; Blanchard, N.; Nystrand, M.; Olney, A.M.; D'Mello, S.K. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In Proceedings of the ACM International Conference Proceeding Series, International Conference on Learning Analytics and Knowledge (LAK '17), Vancouver, BC, Canada, 13–17 March 2017; pp. 218–227.
100. Jacobs, J.; Scornavacco, K.; Clevenger, C.; Suresh, A.; Sumner, T. Automated feedback on discourse moves: Teachers' perceived utility of a professional learning tool. *Educ. Technol. Res. Dev.* **2024**, *72*, 1307–1329. [CrossRef]

101. Owens, M.T.; Seidel, S.B.; Wong, M.; Bejines, T.E.; Lietz, S.; Perez, J.R.; Sit, S.; Subedar, Z.-S.; Acker, G.N.; Akana, S.F.; et al. Classroom sound can be used to classify teaching practices in college science courses. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3085–3090. [CrossRef] [PubMed]
102. Jensen, E.; Dale, M.; Donnelly, P.J.; Stone, C.; Kelly, S.; Godley, A.; D'Mello, S.K. Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning. In Proceedings of the Conference on Human Factors in Computing Systems (CHI 2020), Honolulu, HI, USA, 25–30 April 2020.
103. Topali, P.; Ortega-Arranz, A.; Rodríguez-Triana, M.J.; Er, E.; Khalil, M.; Akçapınar, G. Designing human-centered learning analytics and artificial intelligence in education solutions: A systematic literature review. *Behav. Inf. Technol.* **2025**, *44*, 1071–1098. [CrossRef]
104. Qiu, W.; Thway, M.; Lai, J.W.; Lim, F.S. GenAI for teaching and learning: A Human-in-the-loop Approach. In Proceedings of the Companion Proceedings 15th International Conference on Learning Analytics & Knowledge (LAK25), Dublin, Ireland, 3–7 March 2025; pp. 33–36.
105. Worsley, M. Framing the future of multimodal learning analytics. In *The Multimodal Learning Analytics Handbook*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 359–369.
106. Suresh, A.; Jacobs, J.; Harty, C.; Perkoff, M.; Martin, J.H.; Sumner, T. The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. *arXiv* **2022**, arXiv:2204.09652.
107. Ahn, J.; Campos, F.; Nguyen, H.; Hays, M.; Morrison, J. Co-designing for privacy, transparency, and trust in K-12 learning analytics. In Proceedings of the LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA, 12–16 April 2021; pp. 55–65.
108. Wiedbusch, M.; Sonnenfeld, N.; Henderson, J. Pedagogical Companions to Support Teachers' Interpretation of Students' Engagement from Multimodal Learning Analytics Dashboards. In Proceedings of the International Conference on Computers in Education, Kuching, Malaysia, 28 November–2 December 2022; pp. 432–437.
109. Li, Q.; Jung, Y.; Wise, A.F. How instructors use learning analytics: The pivotal role of pedagogy. *J. Comput. High. Educ.* **2025**, 1–29. [CrossRef]
110. Aljohani, N.R.; Fayoumi, A.; Hassan, S.U. Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability* **2019**, *11*, 7238. [CrossRef]
111. Black, P.; Wiliam, D. Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* **1998**, *5*, 7–74. [CrossRef]
112. Tsai, Y.S.; Singh, S.; Rakovic, M.; Lim, L.A.; Roychoudhury, A.; Gasevic, D. Charting design needs and strategic approaches for academic analytics systems through co-design. In Proceedings of the LAK22: 12th International Learning Analytics and Knowledge Conference, Online, 21–25 March 2022; pp. 381–391.
113. Ouhaichi, H.; Bahtijar, V.; Spikol, D. Exploring design considerations for multimodal learning analytics systems: An interview study. In *Proceedings of the Frontiers in Education*; Frontiers Media SA: Lausanne, Switzerland, 2024; Volume 9, p. 1356537.
114. Yan, L.; Zhao, L.; Echeverria, V.; Jin, Y.; Alfredo, R.; Li, X.; Gaševi'c, D.; Martinez-Maldonado, R. VizChat: Enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots. In Proceedings of the International Conference on Artificial Intelligence in Education, Racife, Brazil, 8–12 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 180–193.
115. Mazzullo, E.; Bulut, O.; Wongvorachan, T.; Tan, B. Learning analytics in the era of large language models. *Analytics* **2023**, *2*, 877–898. [CrossRef]
116. Rüdian, S.; Podelo, J.; Kužílek, J.; Pinkwart, N. Feedback on Feedback: Student's Perceptions for Feedback from Teachers and Few-Shot LLMs. In Proceedings of the 15th International Learning Analytics and Knowledge Conference, Dublin, Ireland, 3–7 March 2025; pp. 82–92.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.