UNIVERSIDAD DE MURCIA

Escuela de Doctorado

TESIS DOCTORAL

Retroalimentación enriquecida de la dinámica del aula utilizando IA

AUTOR/A Federico Pardo García

DIRECTOR/ES Óscar Cánovas Reverte
Félix Jesús García Clemente

2026

UNIVERSIDAD DE MURCIA Escuela de Doctorado

TESIS DOCTORAL

Retroalimentación enriquecida de la dinámica del aula utilizando IA

AUTOR/A Federico Pardo García

DIRECTOR/ES Óscar Cánovas Reverte
Félix Jesús García Clemente

2026

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TITULO DE DOCTOR/A

Aprobado por la Comisión General de Doctorado el 19 de octubre de 2022.

Yo, D. Federico Pardo García, habiendo cursado el Programa de Doctorado de Informática de la Escuela Internacional de Doctorado de la Universidad de Murcia (EIDUM), como autor/a de la tesis presentada para la obtención del título de Doctor/a titulada:

Retroalimentación enriquecida de la dinámica del aula utilizando IA

y dirigida por:

D.: Óscar Cánovas Reverte D.: Félix Jesús García Clemente

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como compendio de publicaciones, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

Murcia, a 2 de febrero de 2026

D. Federico Pardo García

Información básica sobre protección de sus datos personales aportados:	
Responsable	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad	Gestionar su declaración de autoría y originalidad
Destinatarios	No se prevén comunicaciones de datos
Derechos	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la quinta hoja, después de la portada de la tesis presentada para la obtención del título de Doctor/a.

A mis padres, hermanos y familia
por acompañarme todo el camino.

Agradecimentos

A mis directores, Óscar Cánovas Reverte y Félix J. García Clemente, por su supervisión durante estos años. A Óscar especialmente, por no conformarse con menos de lo que yo era capaz de hacer, incluso cuando eso implicaba confrontaciones necesarias que, con perspectiva, agradezco profundamente.

A mis compañeros de laboratorio, Juanje, Emilio y Mengchen, por hacer del día a día algo más llevadero. Por las herramientas que me enseñasteis, las conversaciones que me salvaron en momentos de bloqueo, y por convertir el laboratorio en un espacio donde la amistad fue tan importante como la investigación.

A todas aquellas personas que, de una forma u otra, han formado parte de este camino: profesores, compañeros de departamento, amigos. Vuestro apoyo, aunque no siempre visible en estas páginas, ha sido fundamental.

Resumen

Tradicionalmente, la evaluación y mejora de la práctica docente se ha basado en la observación humana directa, un enfoque que, pese a ser el estándar de oro, conlleva altos costes logísticos y una inherente subjetividad. El campo de las Analíticas de Aprendizaje Multimodal (MMLA) ha experimentado un avance significativo en la capacidad de monitorizar y modelar entornos educativos. Sin embargo, existe una tendencia predominante en la literatura a perfeccionar las técnicas algorítmicas sin proporcionar mecanismos efectivos de retroalimentación hacia el docente, dejando un vacío crítico en la aplicación práctica de estos avances. Esta tesis doctoral aborda dicha desconexión, proponiendo una arquitectura computacional diseñada para el análisis de clases síncronas, con mecanismos de retroalimentación docente.

La investigación se estructura en cuatro fases metodológicas incrementales. En primer lugar, se realizó una revisión sistemática de la literatura (2014-2024) que permitió taxonomizar el uso de las características de audio y analizar tendencias en el campo de investigación. Esto llevo a la detección de una ausencia de sistemas de retroalimentación docente. En segundo lugar, se modeló la práctica docente utilizando exclusivamente características paralingüísticas derivadas de la diarización de hablantes (como la gestión de turnos y los silencios), capaces de clasificar metodologías (clase magistral, trabajo en grupo y uso de sistemas de respuesta de estudiantes (SRS)) con alta precisión. En tercer lugar, se implementó una fusión multimodal que integra características paralingüísticas con el análisis semántico de transcripciones mediante modelos de lenguaje, aplicando técnicas de Inteligencia Artificial Explicable (XAI) para desambiguar intervenciones docentes complejas en entornos SRS. Finalmente, los hallazgos se integraron en una plataforma web que permite a los docentes visualizar las métricas extraídas.

Los resultados demuestran que el audio contiene niveles estratificados de información pedagógica, donde elementos habitualmente descartados, como el ruido ambiental y el solapamiento de voces, actúan como indicadores válidos de colaboración en el aula. Asimismo, se confirma que la fusión de conocimiento experto (ingeniería de características) con modelos de aprendizaje profundo supera el rendimiento de enfoques puramente textuales en tareas de clasificación de intervenciones. Finalmente, el desarrollo de la plataforma proporciona indicios prometedores de que la entrega de métricas objetivas y transparentes tiene el potencial de activar procesos de reflexión docente.

Abstract

Traditionally, the evaluation and improvement of teaching practice have been based on direct human observation, an approach that, despite being the gold standard, entails high logistical costs and an inherent subjectivity. The field of Multimodal Learning Analytics (MMLA) has experienced a significant advancement in the capacity to monitor and model educational environments. However, there is a predominant trend in the literature toward perfecting algorithmic techniques without providing effective feedback mechanisms for the teacher, leaving a critical gap in the practical application of these advances. This doctoral thesis addresses this disconnection, proposing a computational architecture designed for the analysis of synchronous classes, with teacher feedback mechanisms.

The research is structured into four incremental methodological phases. First, a systematic literature review (2014-2024) was conducted, which allowed for the taxonomization of the use of audio features and the analysis of trends in the research field. This led to the detection of an absence of teacher feedback systems. Second, teaching practice was modeled using exclusively paralinguistic features derived from speaker diarization (such as turn-taking management and silences), capable of classifying methodologies (lecturing, group work, and the use of Student Response Systems (SRS)) with high precision. Third, a multimodal fusion was implemented that integrates paralinguistic features with the semantic analysis of transcriptions through language models, applying Explainable Artificial Intelligence (XAI) techniques to disambiguate complex teaching interventions in SRS environments. Finally, the findings were integrated into a web platform that allows teachers to visualize the extracted metrics.

The results demonstrate that audio contains stratified levels of pedagogical information, where elements usually discarded, such as ambient noise and overlapping voices, act as valid indicators of classroom collaboration. Likewise, it is confirmed that the fusion of expert knowledge (feature engineering) with deep learning models outperforms purely textual approaches in intervention classification tasks. Finally, the development of the platform provides promising indications that the delivery of objective and transparent metrics has the potential to activate teacher reflection processes.

Índice general

1
Introducción

Históricamente, la mejora de la práctica docente se ha sustentado en un modelo artesanal basado en la observación humana directa. El uso de mentores y pares expertos para codificar interacciones mediante protocolos estandarizados como CLASS [1] o COPUS [2] constituye el estándar de oro actual, pero este enfoque presenta limitaciones estructurales de escalabilidad. Los costes logísticos y económicos resultan prohibitivos para la mayoría de las instituciones [3], lo que restringe la evaluación a eventos episódicos y no sistemáticos. Más allá del coste, el método enfrenta un desafío de fiabilidad científica: la subjetividad intrínseca del observador. La variabilidad inter-observador y los sesgos cognitivos dificultan la generación de métricas longitudinales objetivas, impidiendo una evaluación del progreso docente basada en evidencias comparables a lo largo del tiempo.

Para mitigar estas carencias, las instituciones han adoptado las Analíticas de Aprendizaje (*Learning Analytics*). Este campo ha demostrado eficacia en entornos digitales al predecir el riesgo académico y optimizar itinerarios basados en registros de actividad (*logs*) [4-6]. Sin embargo, este paradigma sufre una desconexión crítica con el espacio físico del aula. Al centrarse exclusivamente en la huella digital, las métricas ignoran componentes centrales de la pedagogía como el debate espontáneo, la gestión de la oratoria y la interacción cara a cara. El resultado es un "punto ciego" informativo donde la actividad docente más significativa permanece invisible para los sistemas tradicionales de análisis de datos.

Como respuesta, el campo de Multimodal Learning Analytics (MMLA) propone integrar sensores heterogéneos, desde la biometría hasta el seguimiento ocular, para capturar la complejidad de la interacción real [7, 8]. Dentro de este ecosistema, el audio se posiciona como fuente de información prometedora y viable para entornos educativos. A diferencia del vídeo, que conlleva restricciones severas de privacidad y altos requisitos de almacenamiento, o de los sensores vestibles, que pueden alterar el comportamiento natural por su intrusión, el audio permite una captura ubicua y transparente de la actividad en el aula [9]. Esta señal permite monitorizar el clima y la participación sin comprometer la fluidez de la dinámica docente.

A pesar de que el análisis de audio puede caracterizar el clima del aula o medir la participación estudiantil [10-12], su aplicación actual sufre una fragmentación metodológica. Por un lado, la ingeniería acústica se centra en características de bajo nivel, como la energía y el tono, que a menudo carecen de una interpretación pedagógica directa. Por otro,

el Procesamiento de Lenguaje Natural (NLP) suele reducir el aula a un texto transcrito, eliminando la riqueza paralingüística de la prosodia y la dinámica de los turnos de palabra. Esta dicotomía ha impedido la creación de marcos unificados que vinculen la estructura de la interacción con su contenido semántico [13]. Sin esta integración, la capacidad para ofrecer una visión holística de la intención pedagógica es limitada, perdiéndose el matiz de cómo la forma del discurso influye en el aprendizaje.

Incluso con la madurez tecnológica de los modelos de extracción, persiste una brecha crítica en la traducción de estas capacidades en retroalimentación útil (*closing the loop*). Gran parte de la investigación en MMLA prioriza la optimización de modelos predictivos y la precisión técnica, relegando a un segundo plano el modo en que esta información retorna al profesorado [14]. La entrega de métricas crudas o resultados derivados de algoritmos de "caja negra" sin explicabilidad genera resistencia y desconfianza en los docentes. El desarrollo profesional requiere, por tanto, herramientas que no solo procesen datos, sino que fomenten la reflexión pedagógica mediante representaciones transparentes y contextualizadas [15, 16].

Esta tesis doctoral propone un marco metodológico diseñado para conectar las capacidades del procesamiento de audio con las necesidades reales del desarrollo docente. El objetivo central es definir y validar una arquitectura integral que abarque desde la captura de la señal hasta la generación de retroalimentación. La investigación explora el valor informativo de las características paralingüísticas como fuente independiente, investiga su enriquecimiento mediante técnicas semánticas y desarrolla mecanismos de interfaz para su visualización. El marco propuesto no busca sustituir el juicio docente, sino complementarlo mediante una arquitectura técnica que transforme el sonido del aula en una herramienta de análisis estructurada y accionable.

La validez práctica de esta propuesta se sustenta en su implementación dentro de tres Proyectos de Innovación Docente consecutivos en la Universidad de Murcia (2023-2026). El trabajo utiliza un corpus de 287 grabaciones reales (*in-the-wild*), obtenidas de nueve profesores voluntarios de diversas facultades. Estos datos capturan una amplia variedad de situaciones, incluyendo clases magistrales, trabajos en grupo y el uso de sistemas de respuesta de estudiantes, tanto en entornos presenciales como online. Esta diversidad de escenarios garantiza que el marco metodológico propuesto responda a la complejidad intrínseca de la interacción docente real, evitando simplificaciones teóricas que no se sostengan en la práctica cotidiana del aula.

Para garantizar el control metodológico, se ha realizado un desarrollo de software *end-to-end* encargado de la extracción de características y la generación de retroalimentación. Este desarrollo incluye modelos de clasificación para identificar metodologías docentes e intervenciones del profesorado, integrando técnicas de explicabilidad que permiten interpretar las decisiones de los modelos. El proceso culmina en una interfaz web diseñada para que los docentes analicen su propia práctica basándose en los resultados de la investigación. El rigor científico del trabajo queda respaldado por tres artículos publicados en revistas indexadas en JCR y revisados por pares, donde el doctorando figura como autor principal, conformando la estructura lógica y empírica de esta tesis:

1. Pardo, F., Cánovas, Ó., & García, F. J. (2025). **Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps.** Applied Sciences, 15(12), 6911. [17]
2. Pardo, F., Cánovas, Ó., & García, F. J. (2024). **Exploring AI techniques for ge-**

neralizable teaching practice identification. IEEE Access, vol. 12, pp. 134702-134713. [18]

3. Pardo, F., Cánovas, Ó., García, F. J., & Orenes, A. (2025). **Explaining Teacher Interventions in SRS-Based Classrooms: A Classification Approach With BERT and Paralinguistic Cues.** IEEE Access, vol. 13, pp. 208078-208093. [19]

La estructura de esta tesis refleja esta progresión metodológica. El Capítulo 2 presenta los objetivos generales y específicos que guían la investigación. El Capítulo 3 detalla el marco metodológico que conecta cada objetivo con las fases de desarrollo y los artículos publicados. El Capítulo 4 integra los hallazgos de los tres artículos junto con los resultados de la plataforma de transferencia, mostrando cómo la arquitectura propuesta responde al objetivo general planteado. El Capítulo 5 incorpora los artículos íntegros. Finalmente, el Capítulo 6 resume las conclusiones, limitaciones y líneas futuras de investigación.

2
Objetivos

Una vez identificada la desconexión entre la capacidad técnica de las analíticas actuales y la necesidad de una retroalimentación docente interpretable, este capítulo formaliza el propósito de la investigación. Se establecen los hitos necesarios para escalar desde la captura de señales acústicas hasta la validación de un sistema de transferencia pedagógica. A continuación, se detalla el objetivo general que guía la tesis y los cuatro objetivos específicos que estructuran la progresión técnica del trabajo.

2.1. Objetivo general

El propósito central de esta investigación es definir, desarrollar y validar una arquitectura computacional para el análisis automático de la práctica docente mediante el procesamiento de señales de audio. Esta arquitectura integra dimensiones prosódicas y semánticas para escalar desde la extracción de características de bajo nivel hacia el modelado de la práctica educativa. La propuesta no se limita a la clasificación técnica, sino que busca transformar datos acústicos en indicadores de retroalimentación interpretables. El fin último es proporcionar al profesorado una herramienta que facilite la reflexión estructurada sobre su propia intervención en el aula, sustentada en un marco metodológico riguroso y validado experimentalmente.

2.2. Objetivos específicos

Para alcanzar el propósito general, se definen cuatro objetivos específicos que progresan desde la fundamentación teórica hasta la transferencia práctica. La consecución de estos hitos está supeditada al marco experimental de la tesis, compuesto por un corpus de grabaciones obtenido en condiciones reales a través de tres proyectos de innovación docente de la Universidad de Murcia. Este escenario define los desafíos técnicos de los objetivos e impone restricciones severas. La arquitectura debe gestionar problemas como la reverberación acústica y el ruido ambiental propios de aulas grandes, además de cumplir con protocolos estrictos de privacidad.

O1. Sistematizar la evidencia científica sobre el uso de características de audio en entornos educativos.

La construcción de una arquitectura robusta requiere la identificación previa de las tendencias metodológicas y las brechas del estado del arte. Ante la ausencia de una revisión sistemática que abordase específicamente la explotación del audio en educación, este objetivo se centra en taxonomizar las características de audio empleadas por la literatura para derivar información pedagógica. El análisis no se limita a la extracción de datos, sino que identifica las técnicas utilizadas, las limitaciones reportadas y el modo en que las recomendaciones son compartidas con los docentes. Este estudio constituye el fundamento teórico indispensable sobre el que se asientan las decisiones técnicas de los objetivos posteriores, asegurando que la propuesta aborde las carencias reales del campo científico.

O2. Caracterización de la actividad docente mediante el uso de descriptores paralingüísticos.

Este hito aborda la caracterización de la actividad docente utilizando descriptores paralingüísticos derivados de la segmentación y diarización de hablantes. Se analiza la eficacia de atributos como los patrones de interacción verbal, la detección de actividad de voz y los eventos no verbales para identificar tres modalidades pedagógicas: clase magistral, trabajo en grupo y el uso de sistemas de respuesta de estudiantes (SRS). Dado que el corpus proviene de entornos reales, un requisito crítico es garantizar la robustez del sistema frente al ruido y las diferencias individuales. Por tanto, el objetivo es desarrollar un modelo computacional cuya estabilidad en las métricas de desempeño se mantenga constante frente a la variabilidad intrínseca del locutor y las condiciones acústicas del aula.

O3. Integración de características paralingüísticas y semánticas para la mejora de la clasificación y la interpretabilidad del discurso docente.

El objetivo O3 plantea la integración de la estructura interactiva (paralingüística) con el contenido instruccional (léxico) para resolver la clasificación de intervenciones docentes. Mientras que el análisis previo permitía distinguir modalidades generales, la fusión multimodal es necesaria para capturar la intencionalidad pedagógica de las intervenciones en sesiones mediadas por SRS. Se desarrolla un modelo capaz de clasificar intervenciones en cinco categorías pedagógicas, superando las limitaciones de los enfoques unimodales. Para evitar la opacidad de los algoritmos de caja negra, se aplican técnicas de inteligencia artificial explicable (XAI) que cuantifican la relevancia de cada descriptor. Esto proporciona una base técnica para la explicabilidad, permitiendo entender el sentido de la contribución de cada variable al reconocimiento de la categoría.

O4. Transferencia de resultados: desarrollo de un sistema integral de análisis y visualización para la reflexión docente.

El último objetivo transforma los avances técnicos en una herramienta de impacto práctico. Se desarrolla una plataforma web que integra la visualización de métricas de diarización y transcripciones de aula. Esta interfaz permite a los profesores interactuar con sus propios registros de manera privada y autónoma para analizar la evolución de su

práctica. La consecución de este objetivo constituye la validación final de la arquitectura en condiciones *in-the-wild*. De este modo, se demuestra la viabilidad del sistema para operar en entornos reales y se sientan las bases técnicas para futuros despliegues a gran escala.

3
Metodología

Este capítulo describe el marco metodológico desarrollado para abordar los cuatro objetivos específicos de la investigación. La Figura 3.1 sintetiza la arquitectura metodológica de la tesis estructurada en cuatro fases incrementales. Estas articulan objetivos específicos, diseños metodológicos diferenciados y resultados empíricos vinculados secuencialmente. La progresión inicia con una revisión sistemática (M1), la cual fundamenta el enfoque teórico y diagnostica brechas en el estado del arte. A continuación se desarrollan los modelos de clasificación, que transitan desde la validación de las características paralingüísticas (M2), hasta su enriquecimiento mediante la fusión léxica multimodal y uso de inteligencia artificial explicable (M3). Finalizamos con la transferencia de feedback al docente mediante el desarrollo de la plataforma web (M4) . Esta plataforma se basa en el software subyacente desarrollado durante la investigación.

El corpus recolectado contiene 287 grabaciones de clases reales en la Universidad de Murcia. Estas abarcan nueve profesores de Informática, Psicología, Veterinaria y Derecho, garantizando la diversidad disciplinar. Dado la complejidad de codificar manualmente los audios, el estudio limita su análisis a un conjunto de datos controlado y verificado. Esta decisión metodológica es necesaria para garantizar la precisión del *ground truth*. Por tanto, la disponibilidad de recursos humanos para la validación manual define el marco de actuación y la representatividad de los experimentos, priorizando la calidad del dato sobre el volumen masivo no supervisado.

M1. Revisión sistemática de la literatura

Se realizó una revisión sistemática para mapear el estado del procesamiento de audio en contextos educativos (O1). A diferencia de trabajos previas sobre *Multimodal Learning Analytics*, este trabajo prioriza el audio como fuente de datos primaria, incluyendo estudios unimodales y combinaciones con vídeo, logs o datos fisiológicos. El diseño se fundamenta en el protocolo PRISMA [20], adoptando un enfoque cualitativo y análisis inductivo para la construcción de taxonomías emergentes. En el caso de hallazgos heterogéneos se realiza una síntesis narrativa. La búsqueda en Scopus y Web of Science (2014-2024), resulta en una muestra final de 82 estudios tras aplicar criterios estrictos de inclusión.

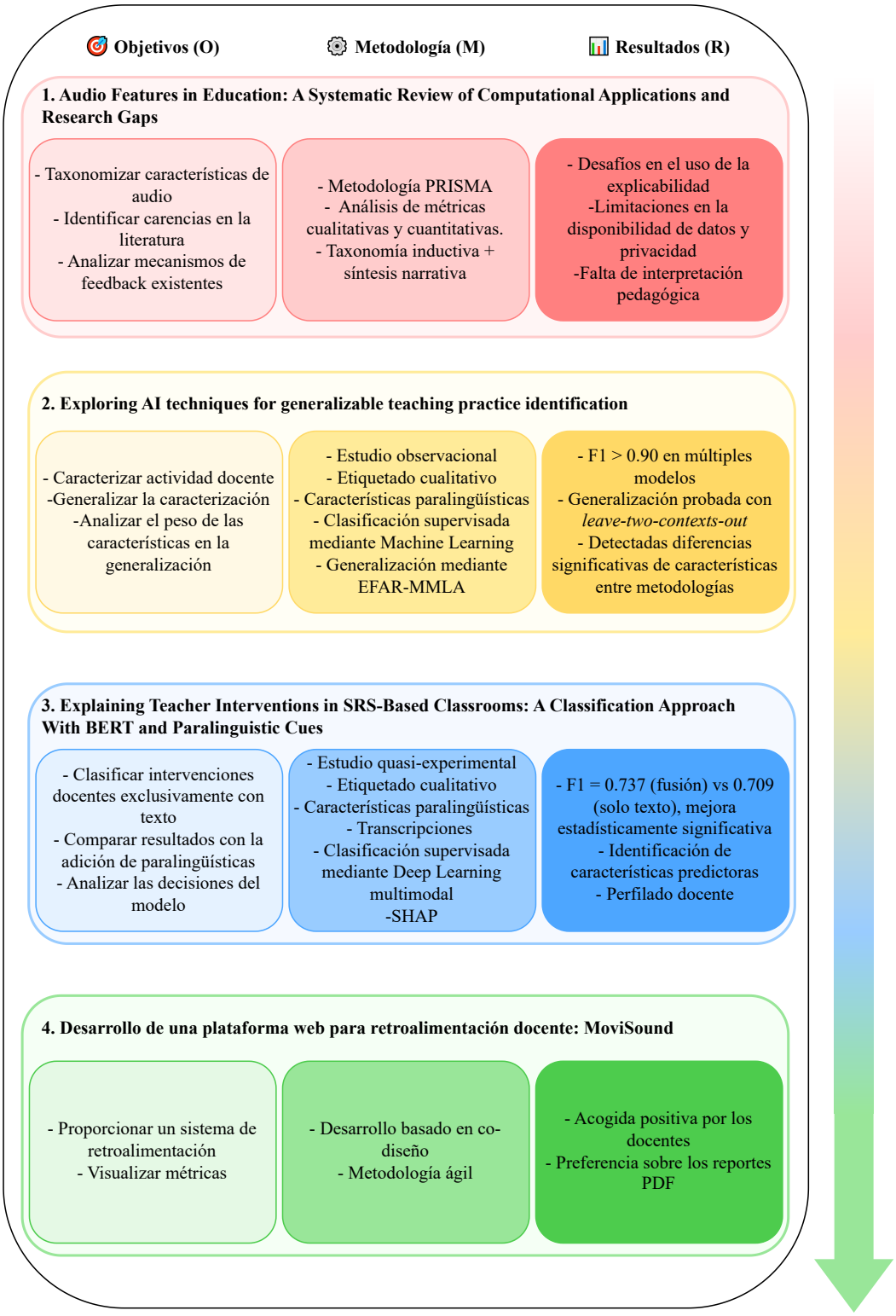


Figura 3.1: *Arquitectura metodológica de la tesis: evolución incremental del enfoque de investigación a través de cuatro fases secuenciales. Cada fila representa un estudio con sus objetivos, diseño metodológico y hallazgos principales. La progresión muestra cómo los resultados de cada fase informaron el diseño de las siguientes. El estudio evoluciona desde el análisis de la literatura hasta el desarrollo de la plataforma de retroalimentación. La flecha vertical derecha representa el desarrollo paralelo del software.*

El análisis se estructuró en cinco ejes: usos tecnológicos, características de audio, integración multimodal, técnicas de procesamiento y mecanismos de retroalimentación. Esta división permite explorar el ciclo de vida completo del audio en entornos educativos, facilitando el análisis de patrones comunes y carencias de la literatura. La Tabla 3.1 sintetiza los métodos e instrumentos empleados en esta fase.

Cuadro 3.1: *Métodos e instrumentos empleados en M1 (Revisión Sistemática).*

Componente	Método/Instrumento	Propósito
Cuantitativo	Protocolo PRISMA	Selección sistemática y replicable de literatura sobre audio educativo
Cualitativo	Análisis temático inductivo	Construcción de taxonomía de características de audio (bajo/medio/alto nivel)
Cualitativo	Síntesis narrativa	Diagnóstico de áreas subexploradas: paralingüística, feedback, interpretabilidad

M2. Clasificación de metodologías docentes mediante características paralingüísticas y machine learning

La segunda fase metodológica aborda el objetivo O2 mediante el uso de características paralingüísticas para clasificar metodologías educativas. El uso de rasgos extraídos de la diarización ofrece una ventaja crítica: Son agnósticos al contenido. Esto facilita la transferencia a nuevos entornos, pues las pausas, silencios y cambios de turno son intrínsecos a la comunicación humana. El estudio adopta un diseño observacional con enfoque mixto sobre un subconjunto de 26 horas de audio distribuidas en tres categorías: clase magistral, trabajo grupal y SRS. El dataset presenta un desbalance natural que refleja la distribución real de las clases observadas, aportando validez práctica. El etiquetado se realizó mediante codificación cualitativa experta.

Se extraen 13 características paralingüísticas cuantitativas utilizando diferentes tamaños de ventana para capturar las dinámicas de la interacción verbal. Estas variables entrenan múltiples flujos de clasificación supervisada que identifican la metodología utilizada. Para evaluar la robustez del sistema, se aplica el protocolo Evaluation Framework for Assessing Robustness in Multimodal Learning Analytics (EFAR-MMLA) mediante la estrategia *Leave-Two-Contexts-Out* [21]. Este método deja fuera del entrenamiento a la mitad de los docentes, garantizando que las métricas de desempeño sean estables a la variación de docentes y disciplinas. Adicionalmente, se emplea el análisis estadístico Kurskal-Wallis para cuantificar las diferencias en el peso de las características entre metodologías. La Tabla 3.2 sintetiza los datos analizados, sus métodos de obtención y su aportación al diseño metodológico.

M3. Integración de características textuales y paralingüísticas mediante deep learning

Esta fase aborda el objetivo O3, integrando información textual y paralingüísticas para la clasificación de intervenciones docentes específicas durante el uso de SRS. Esta

Cuadro 3.2: *Métodos e instrumentos empleados en M2 (Modelado Paralingüístico).*

Componente	Método/Instrumento	Propósito
Cuantitativo	Diarización automática de hablantes	Segmentación objetiva de turnos de habla docente/estudiantes
Cualitativo	Codificación experta	Ground truth para entrenamiento supervisado basado en juicio pedagógico
Cuantitativo	Extracción de características	Variables paralingüísticas representando la estructura del intercambio
Cuantitativo	EFAR-MMLA (Leave-Two-Contexts-Out)	Evaluación de robustez ante variabilidad inter-docente y disciplinar
Cuantitativo	Análisis estadístico Kruskal-Wallis	Confirmación de diferencias significativas en el peso de las características entre metodologías

integración responde a la insuficiencia intrínseca de los análisis unimodales para capturar la complejidad del discurso educativo. Mientras que el análisis léxico identifica el contenido explícito del mensaje, resulta insuficiente para discernir la intención pedagógica latente en la forma del habla. Por el contrario, la paralingüística captura la dinámica de interacción y el tono del discurso, pero es incapaz de procesar el significado instruccional. La fusión multimodal permite superar estas limitaciones individuales, enriqueciendo la capacidad de predicción al vincular el contenido del discurso con la manera en que se comunica.

El estudio emplea un diseño quasi-experimental con enfoque metodológico mixto sobre un corpus de tres horas y diecinueve minutos de audio, procedentes de diez sesiones basadas en SRS. La codificación de las intervenciones docentes se realizó mediante una adaptación del marco COPUS, simplificada a cinco categorías para el análisis de la interacción. Para asegurar la fiabilidad científica del *ground truth*, el etiquetado fue ejecutado por dos anotadores independientes, alcanzando un coeficiente kappa de Cohen de 0.97.

Se extraen las características paralingüísticas cualitativas previamente validadas para la clasificación de metodologías docentes, junto a las transcripciones. Estos datos alimentaron múltiples arquitecturas de *deep learning*, basadas en dos métodos de fusión. La fusión temprana o *early-fusion* consiste en la unión de los vectores de características en la etapa de entrada, permitiendo al modelo procesar la señal de forma conjunta desde el inicio del aprendizaje. Por el contrario, la fusión tardía o *late-fusion* procesa los datos de forma independiente antes de ser combinadas en una etapa final. Con el fin de superar la opcaidad e los modelos de caja negra, se aplican técnicas de explicabilidad (XAI) mediante SHAP. Este análisis permite cuantificar la relevancia de los elementos léxicos y paralingüísticos en la clasificación, facilitando la comprensión del modelo. La Tabla 3.3 sintetiza los datos analizados, sus métodos de obtención y su aportación al diseño metodológico.

M4. Arquitectura del sistema: diseño desacoplado de los servicios de análisis y la interfaz de usuario

La cuarta fase metodológica materializa la transferencia de hallazgos mediante el desarrollo de una arquitectura técnica orientada al usuario final. Este proceso se basa en un co-diseño cualitativo, fundamentado en ciclos iterativos de diseño, implementación

Cuadro 3.3: *Métodos e instrumentos empleados en M3 (Fusión Multimodal).*

Componente	**Método/Instrumento**	**Propósito**
Cuantitativo	Diarización automática de hablantes	Segmentación objetiva de turnos de habla docente/estudiantes
Cuantitativa	Transcripción automática	Obtención del texto hablado de la grabación
Cualitativo	Codificación experta independiente	Ground truth para clasificación de intervenciones pedagógicas
Cuantitativo	Extracción automática de características	Variables predictoras agnósticas al contenido y al idioma
Cuantitativo	Fusión tardía de modalidades	Integración complementaria de texto + paralingüística para clasificación mejorada
Cualitativo	Técnicas de explicabilidad (XAI)	Identificación de features relevantes por categoría para comprensión pedagógica

y evaluación mediante feedback docente. A diferencia de las etapas anteriores, esta fase prioriza la utilidad percibida y la adecuación a las necesidades reales del docente.

El desarrollo del *backend* se planteó como el núcleo central de procesamiento. Este sistema implementa una arquitectura desacoplada diseñada para gestionar la ejecución de los modelos de aprendizaje profundo y el procesamiento de señales acústicas. En lugar de seguir un proceso de desarrollo lineal, la lógica de procesamiento se integró en el sistema de manera paralela a la investigación. Esta sincronía permite que la experimentación técnica de las fases anteriores se ejecutara directamente sobre la arquitectura definitiva, asegurando que los modelos fuesen operativos dentro del flujo de datos de la plataforma desde su concepción. La robustez del *backend* garantiza la autonomía de los servicios de análisis frente a la interfaz de usuario.

Sobre la infraestructura del *backend* se construye la interfaz de visualización, cuyo diseño es el resultado de tres años de iteración con los docentes participantes. Durante este periodo se entregaron reportes periódicos en formato PDF que contenían los análisis de sus sesiones mediante las métricas validadas. El feedback obtenido a través de conversaciones informales y observaciones directas permitió identificar qué representaciones visuales resultaban útiles y cuáles generaban confusión. Este proceso de co-diseño aseguró que la plataforma final no fuera un desarrollo tecnológico aislado, sino una solución construida en colaboración con los usuarios.

La implementación de la plataforma siguió metodologías de desarrollo ágil, estructurando el avance en fases vinculadas a nuevas funcionalidades. En esta etapa participaron los mismos nueve docentes que conformaron el corpus inicial, garantizando así la continuidad metodológica y aprovechando su familiaridad previa con las información presentada. Se reconoce que esta decisión conlleva un sesgo positivo derivado de la disposición previa de los participantes a colaborar. No obstante, esta aproximación es la más adecuada para una fase de desarrollo inicial donde la prioridad reside en validar la viabilidad técnica y la estabilidad del sistema. A modo de resumen, la Tabla 3.4 sintetiza los datos analizados y métodos empleados en esta fase.

Cuadro 3.4: *Métodos e instrumentos empleados en M4 (Plataforma de Retroalimentación Docente).*

Componente	**Método/Instrumento**	**Propósito**
Cualitativo	Cuestionarios abiertos + conversaciones informales con docentes	Identificación iterativa de necesidades de visualización y comprensibilidad de métricas
Mixto	Métricas extraídas y validadas de M2/M3	Base empírica para alimentar las visualizaciones
Cualitativo	Metodología de Co-Diseño	Involucrar a los usuarios en el desarrollo de la plataforma
Cualitativo	Desarrollo ágil	Metodología ágil basada en sprints para el desarrollo del software

4
Resultados

Este capítulo estructura los hallazgos en cuatro bloques que corresponden directamente a los objetivos específicos planteados en el Capítulo 2, que a su vez se corresponden con cada una de las publicaciones y el desarrollo del software. Para facilitar la narrativa y organización de los resultados, se partirá de los hallazgos de la revisión sistemática, a pesar de que esta contribución no se corresponde con la primera publicación de la tesis.

R1. Taxonomización de las características de audio, limitación de datos y falta de retroalimentación docente

El análisis de los artículos seleccionados y su síntesis, presentada en la Sección 5.1 [17], facilitaron la clasificación de características en tres niveles de abstracción y la definición de sus usos principales. Asimismo, el estudio reveló problemas críticos como las restricciones de privacidad, disponibilidad de datos y la falta de retroalimentación pedagógica y explicabilidad.

Taxonomización de las características de audio

La base analítica de esta sección, así como una de las contribuciones centrales de la presente tesis, reside en la propuesta de una taxonomía para la categorización de las características de audio. Esta clasificación organiza las variables en tres niveles de abstracción: rasgos de bajo nivel, vinculados a la señal acústica como el tono o la reverberación; rasgos paralingüísticas, que definen la estructura de la interacción mediante los silencios y los turnos de palabra; y rasgos resultantes del procesamiento del lenguaje natural, centrados en el contenido léxico como muletillas o la velocidad del habla.

La jerarquía de esta clasificación responde a una dependencia funcional intrínseca: cada nivel de abstracción se construye necesariamente sobre el anterior. El procesamiento de la señal acústica constituye el requisito previo indispensable para realiza una diarización precisa de los hablantes. Del mismo modo, el análisis del lenguaje natural alcanza su máxima utilidad cuando se integra con los rasgos paralingüísticos, permitiendo discernir no solo el contenido del mensaje, sino la identidad del emisor y la dinámica temporal en

la que ocurre la interacción. Estableciendo este orden lógico, la taxonomía actúa como un marco pionero que dota de estructura el análisis de las contribuciones técnicas identificadas en la revisión sistemática

Restricciones de privacidad y disponibilidad de los datos

La investigación en este ámbito padece una marcada fragmentación estructural. La ausencia de un marco común de datos obliga a cada investigador a depender de grabaciones propias y privadas, lo que impide de raíz la validación cruzada de los resultados. Esta falta de estandarización en los protocolos de captura no solo genera una percepción de fragmentación, sino que constituye una barrera ontológica para la ciencia: sin datasets compartidos, la reproducibilidad es técnicamente imposible.

Esta dispersión de la evidencia se ve agravada por la naturaleza sensible de la voz, protegida como dato biométrico por marcos legales como el *General Data Protection Regulation* (GDPR). El envío de señales de audio crudas entre instituciones representa un conflicto legal cuya resolución excede a menudo las capacidades logísticas de los grupos de investigación. Ante este muro burocrático, la literatura científica muestra un silencio preocupante, eludiendo una discusión profunda sobre cómo conciliar la ética de la privacidad con la necesidad de transparencia en los datos.

Una solución viable para dar orden en esta disciplina es la publicación de conjuntos de datos pre-procesados y anonimizados. Si bien este enfoque implica la pérdida de control sobre la señal original, representa un compromiso necesario para el progreso de la comunidad. Al trabajar sobre una base de datos común, aunque sea derivada, los investigadores pueden comparar su modelos bajo las mismas condiciones, transformando una colección de estudios aislados en un cuerpo de conocimiento coherente y verificable.

Falta de retroalimentación pedagógica y explicabilidad

La investigación actual se ha estancado en un ciclo técnico que ignora la realidad del aula. Los estudios se limitan a optimizar modelos dentro de un entorno cerrado, olvidando que la tecnología educativa solo tiene sentido si impacta positivamente en la enseñanza. Esta obsesión por mejorar métricas de rendimiento ha invertido las prioridades del campo. Hoy se valora más la precisión de un algoritmo que su utilidad real para el docente. El resultado es un volumen masivo de publicaciones técnicas que, a pesar de su complejidad, carecen de trascendencia en la práctica diaria del profesorado.

Esta desconexión se agrava con la adopción del *deep learning*. Aunque estos modelos son extremadamente potentes, funcionan como cajas negras cuyo razonamiento es imposible de interpretar. En un entorno educativo, una herramienta que ofrece resultados sin justificarlos carece de autoridad y genera un rechazo natural entre los profesionales. Ante esta falta de transparencia, el uso de técnicas de inteligencia artificial explicable (XAI) es una necesidad ética y no solo una opción técnica. Solo si el docente comprende cómo decide la máquina, podrá confiar en ella para transformar los datos en una oportunidad de reflexión.

R2. Caracterización de metodologías docentes, generalización y análisis de las principales características paralingüísticas

La segunda publicación, detallada en la Sección 5.2 [18], se planteó basándose en una pregunta: ¿Es posible modelar la práctica docente ignorando deliberadamente lo que se dice? Los resultados obtenidos permiten responder afirmativamente, validando la existencia de patrones de interacción distintivos entre metodologías. Además, se validó la generalización dentro de las posibilidades de nuestro dataset, así como la confirmación de que existen diferencias significativas entre las principales características identificadas a la hora de clasificar cada metodología.

Clasificación de metodologías docentes mediante características paralingüísticas

Basándose en un *pipeline* desarrollado previamente [22], la investigación evaluó la capacidad de algoritmos como perceptrón multicapa (MLP), regresión logística (LR) o máquinas de soporte vectorial (SVM) para categorizar métodos de enseñanza. Los modelos alcanzaron un rendimiento general superior a un *F1-Score* de 0.90, una cifra notable considerando que se obtuvo sin realizar una búsqueda exhaustiva de hiperparámetros. Lejos de ser una carencia, esta falta de optimización técnica resalta la robustez de las variables paralingüísticas empleadas; es la calidad informativa del audio lo que sostiene la precisión del sistema.

Estos hallazgos confirman que es viable caracterizar la práctica docente utilizando exclusivamente rasgos paralingüísticos diseñados manualmente. Una vez validada la capacidad de estas variables, el estudio avanzó hacia el siguiente reto: verificar si esta capacidad predictiva es capaz de generalizarse ante la variabilidad de nuevos docentes y entornos fuera del conjunto de datos original.

Generalización de los resultados

Para verificar la robustez del sistema ante la variabilidad del locutor y el entorno, se aplicó el protocolo EFAR-MMLA mediante una estrategia de validación cruzada *leave-two-contexts-out*. Su propósito es discriminar si el modelo ha capturado patrones pedagógicos universales o si se ha limitado a memorizar las particularidades de sujetos específicos. En esta fase, se realizó una optimización exhaustiva de hiperparámetros sobre las seis combinaciones posibles de docentes, garantizando que el rendimiento alcanzado representara el límite superior de la arquitectura ante nuevos datos.

El *F1-Score* resultante de 0.92 no solo supera la línea base previa, sino que confirma la estabilidad de las características paralingüísticas diseñadas. El hecho de que el sistema mantenga su eficacia sobre docentes y aulas que no formaron parte del entrenamiento ratifica que las variables seleccionadas poseen una validez representativa real. Esta consistencia en entornos no vistos demuestra que es posible automatizar la caracterización docente sin depender de la identidad del hablante, sentando las bases para una retroalimentación escalable y objetiva.

Análisis de las características principales para la generalización

Las métricas ratio de silencio (SR), ratio de murmullo (MR) y número de intervenciones del docente (PSU_Teacher) emergieron como los predictores determinantes en la caracterización de las metodologías docentes. Estos descriptores no actúan como simples variables numéricas, sino que capturan la esencia rítmica y temporal de la interacción en el aula, traduciendo la dinámica pedagógica a una huella estructural medible. El test estadístico de Kruskal-Wallis confirmó diferencias estadísticas significativas, ratificando la potencia discriminatoria de cada rasgo frente a las distintas categorías.

La identificación de patrones de interacción altamente diferenciados valida la capacidad de estas métricas para representar la complejidad de la práctica docente de manera objetiva. Al demostrar que cada metodología posee una representación específica, el sistema deja de ser una "caja negra" para convertirse en un modelo transparente. Esta relevancia de los rasgos diseñados manualmente permite comprender los criterios subyacentes a la clasificación automática, asegurando que el sistema no solo sea preciso, sino también explicable y útil para la reflexión pedagógica del profesorado.

R3. Clasificación de intervenciones docentes mediante fusión multimodal y aprendizaje profundo

Tras validar que las características paralingüísticas proporcionan una representación robusta de la dinámica grupal, esta fase de la investigación se diseñó como un paso natural de enriquecimiento informativo [19]. El objetivo central no fue subsanar deficiencias de los modelos previos, sino evaluar el valor incremental de integrar el contenido léxico con la estructura de la interacción. Este enfoque aditivo permite capturar acciones pedagógicas específicas en entornos apoyados por SRS, donde la intención del docente a menudo trasciende el significado literal de sus palabras. Al transitar hacia una escala de intervención individual, el sistema adquiere la granularidad necesaria para ofrecer una retroalimentación precisa, complementada por un análisis de interpretabilidad que dota de transparencia a la arquitectura propuesta.

Clasificación de intervenciones basada en el contenido textual

Para establecer una referencia sobre el impacto de la información paralingüística, se diseñaron dos aproximaciones fundamentadas exclusivamente en transcripciones mediante modelos BERT. La primera configuración clasificó la intervención de forma aislada, mientras que la segunda incorporó el contexto de la intervención docente inmediata anterior. Los resultados reflejaron un *F1-Score* de 0.52 y 0.709 respectivamente. Esta diferencia subraya la importancia crítica del contexto histórico en la identificación de la intención pedagógica del docente. Estos hallazgos funcionaron como una línea base y confirmaron la viabilidad de clasificar las intervenciones en cinco categorías automáticamente.

Optimización de la clasificación mediante características paralingüísticos

La combinación de datos textuales y paralingüísticos planteó un desafío de disparidad dimensional. Mientras que el conjunto de variables paralingüísticas constaba de 13

descriptores principales, el modelo DeBERTa generaba representaciones de 768 dimensiones. Esta heterogenidad provocó que las fusiones tempranas tendieran a ignorar la información estructural de los turnos en favor del peso del texto. Para resolver este desequilibrio, se implementó una fusión tardía, permitiendo que ambas modalidades contribuyeran de forma equilibrada hasta alcanzar un *F1-Score* de 0.737 con el modelo *random forest* (RF). A pesar de que la mejora absoluta respecto al uso exclusivo de texto es de 0.028, el test de Bonferroni ratificó que dicha diferencia es estadísticamente significativa ($p < 0.05$). La robustez del modelo se confirmó mediante un conjunto de validación, manteniendo un rendimiento *F1-Score* de 0.707.

Análisis de explicabilidad

El análisis mediante valores SHAP permitió cuantificar la influencia relativa de cada modalidad en la decisión del sistema. Los resultados otorgan un peso del 64 % al contenido textual y un 36 % a los rasgos paralingüísticos, validando que la estructura de interacción es un pilar sustancial del razonamiento algorítmico y no un simple añadido marginal. Un hallazgo destacado es la correlación positiva entre la variable de solapamiento del habla OVR y la etiqueta de gestión (*Management*). Este patrón indica que, ante situaciones de habla simultánea, los docentes suelen intervenir con directrices organizativas para redirigir el flujo de la clase. Este ejemplo ilustra como las técnicas de explicabilidad actúan como un puente para interpretar los eventos pedagógicos a través de correlaciones aprendidas por el modelo.

R4. Implementación y validación del desarrollo tecnológico para la retroalimentación docente.

La cuarta fase se materializa como una respuesta a la falta de mecanismos de retroalimentación identificados en la revisión sistemática. Este desarrollo basado en un *backend* desarrollado en paralelo a la investigación y un *frontend* web para los docentes, integra las características paralingüísticas y transcripciones en un *pipeline* automatizado que procesa grabaciones de aula y genera retroalimentación en dos modalidades complementarias.

Backend desarrollado simultáneamente a la investigación

El desarrollo del *backend* que utiliza la plataforma web constituye un resultado de la investigación en sí mismo: una arquitectura que evolucionó orgánicamente en paralelo con las fases experimentales, incorporando progresivamente cada componente validado empíricamente. La estructura del sistema se fundamenta en una API REST que expone los modelos de clasificación de metodologías y de intervenciones docentes como servicios reutilizables. Esta modularidad no es meramente técnica, sino una decisión arquitectónica orientada a la escalabilidad y reproducibilidad científica.

Desarrollo de la plataforma web: MoviSound

La plataforma estructura la retroalimentación mediante visualizaciones gráficas de las métricas paralingüísticas validadas en R2 y R3, presentando tanto valores absolutos por sesión como su evolución temporal a lo largo de la misma. Esto incluye distribuciones de turnos de habla diferenciadas por rol (docente vs. estudiantes), ratios de silencio, métricas de solapamiento (OVR, MR, etc.) y ratio de participación del hablante (PSR) segregado por participantes. Adicionalmente, la interfaz presenta las transcripciones completas de las sesiones, junto a un apartado específico para intervenciones estudiantiles, respetando su anonimización. Una muestra de los gráficos en la interfaz web se presenta en la Figura 4.1.

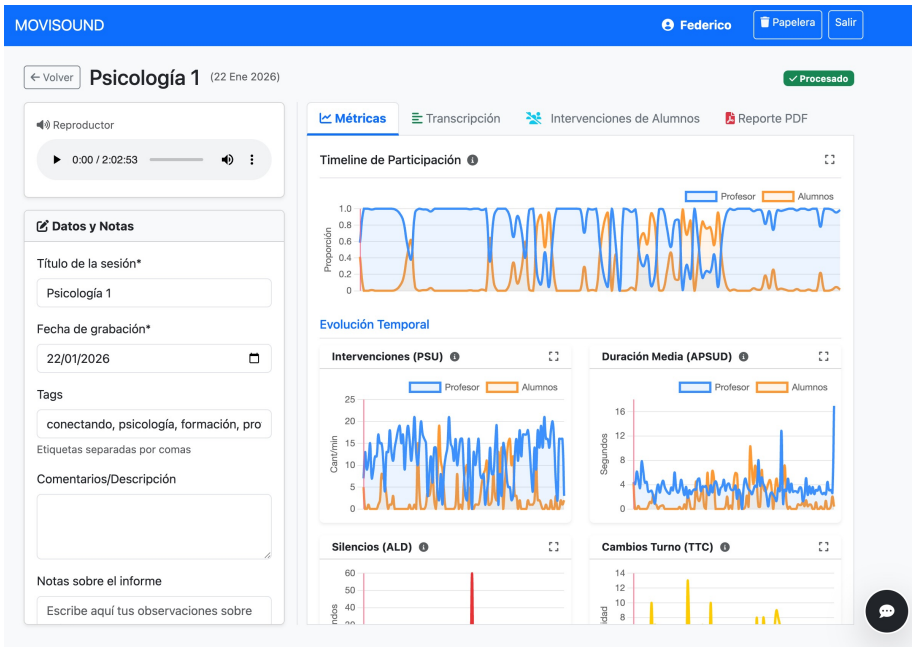


Figura 4.1: *Interfaz de la web desarrollada para la retroalimentación docente. Se pueden visualizar los metadatos de la sesión a la izquierda. El panel central muestra la progresión temporal de algunas métricas extraídas como al PSR, representado por el 'Timeline de Participación'. Se aprecian también las pestañas de transcripción y acceso al reporte en formato PDF.*

El análisis cualitativo del feedback docente reveló que el PSR constituyó la métrica más consultada, seguida de las visualizaciones temporales de turnos de habla diferenciadas por rol. Los docentes manifestaron especial interés por las transcripciones de intervenciones estudiantiles, empleándolas para identificar qué preguntas específicas formularon los alumnos durante la sesión, información que suele perderse en la dinámica temporal del aula. Esta relación entre el ratio de participación de los estudiantes y el contenido léxico de sus intervenciones específicas, refuerza la hipótesis de utilidad pedagógica de las características paralingüísticas: métricas aparentemente abstractas como el PSR resultaron interpretables e hicieron a los docentes consultar qué intervención hizo variar dichas métricas.

Percepción de la utilidad y potencial para la reflexión docente

Uno de los principales hallazgos de esta fase exploratoria es la capacidad de la herramienta para facilitar procesos de reflexión pedagógica. El análisis del feedback cualitativo

indica que la visualización de métricas objetivas permitió a los docentes identificar discrepancias significativas entre su autopercepción de la dinámica del aula y los patrones cuantificados por el sistema. Un caso ilustrativo fue la subestimación del tiempo de habla propio por parte de un participante, cuya percepción difería notablemente del registro generado. Esta confrontación entre la narrativa interna del docente y la evidencia basada en datos puede actuar como un catalizador necesario, aunque no suficiente por sí solo, para desencadenar la autocrítica constructiva.

La evaluación del impacto de la herramienta se realizó mediante un cuestionario de retroalimentación completado por ocho de los nueve docentes participantes. El análisis se estructuró en tres ejes fundamentales: la utilidad percibida de los informes, la concordancia de los datos con la autopercepción del profesor y el impacto real en su práctica profesional. Los resultados cuantitativos confirman la relevancia del sistema, alcanzando una valoración de utilidad superior al 80 % y evidenciando una alineación estrecha entre las métricas automatizadas y la percepción subjetiva de los docentes. El indicador de mayor trascendencia reside en la modificación de la conducta: siete de los ocho participantes afirmaron haber alterado su estrategia didáctica tras recibir los reportes, mientras que seis de ellos fueron capaces de verificar y cuantificar dichos cambios en informes posteriores.

La preferencia docente por la plataforma web sobre los reportes PDF previamente empleados resultó unánime. Aunque la web replica el contenido de los PDFs, la interactividad (navegación entre sesiones, consulta dinámica de transcripciones, visualizaciones ampliables) y la inmediatez del análisis post-clase fueron valoradas como ventajas sustanciales. Este hallazgo refuerza la relevancia de la automatización *end-to-end* implementada en el backend: no basta con generar análisis precisos (R2-R3), es necesario entregarlos de forma accesible y oportuna para maximizar su utilidad práctica.

5

Publicaciones

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

Título

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

Autores

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

Departamento de Ingeniería y Tecnología de Computadores Universidad de Murcia, España

Detalles de la publicación

Revista	Applied Sciences	**Editorial**	MDPI
Volumen	15	**Número**	12
Páginas	6911	**Año**	2025
JIF	2.5 (est. 2024)	**Ranking**	Q2
Estado	Publicado	**DOI**	10.3390/app15126911

Resumen

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.

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

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

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.

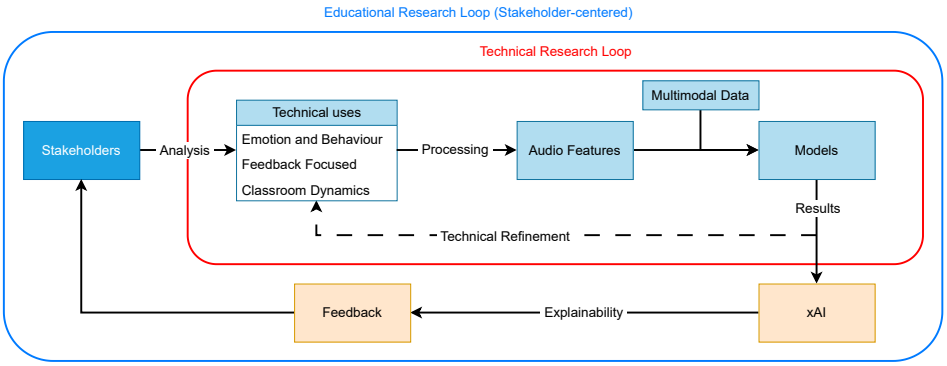


Figure 1. Conceptual overview of current and idealized research workflows involving audio features in educational contexts. The red loop represents the dominant technical pipeline, which focuses on extracting and modeling audio features but largely excludes explainability, feedback, and the participation 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:

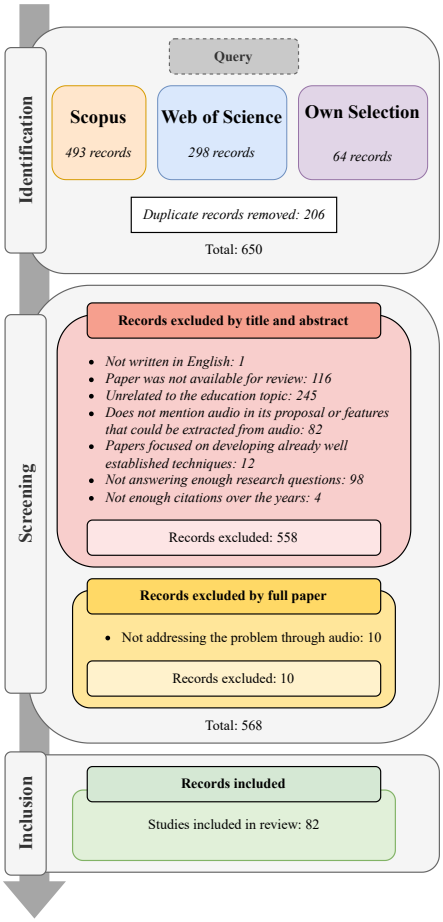


Figure 2. Flow diagram of the PRISMA methodology followed.

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].

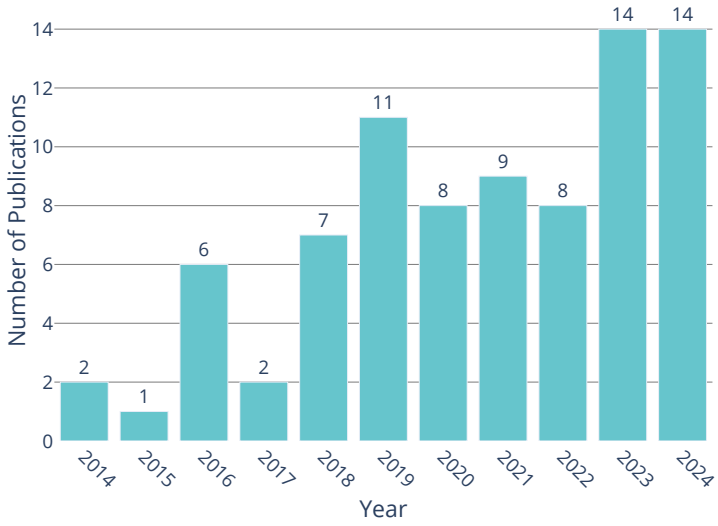


Figure 4. Number of selected studies per year of publication.

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.

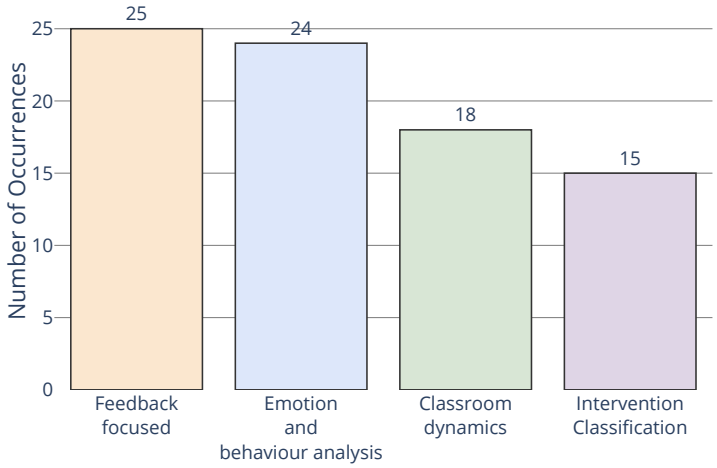


Figure 5. Main uses detected in the analyzed literature (RQ1).

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.

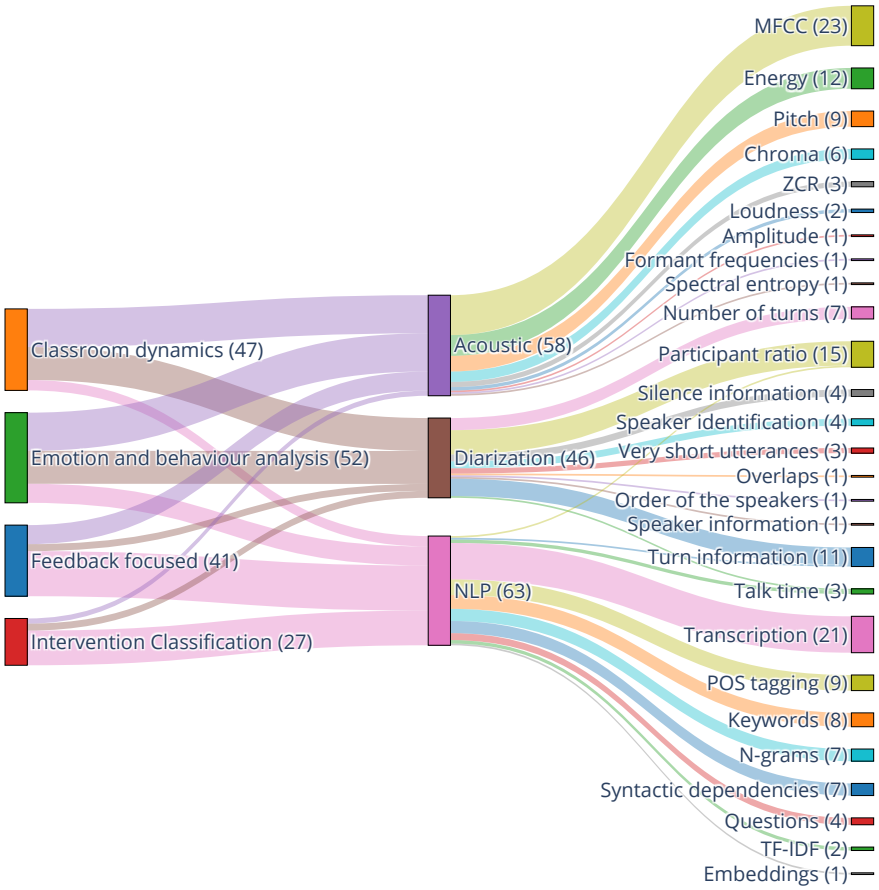


Figure 6. Relationships between types of usage categories (RQ1) and types of audio features identified (RQ2).

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, physiolog-

ical 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.

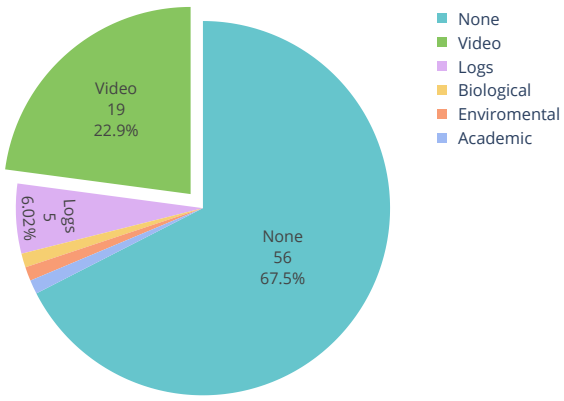


Figure 7. Distribution of analyzed papers that uses a multimodal approach (RQ3).

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.

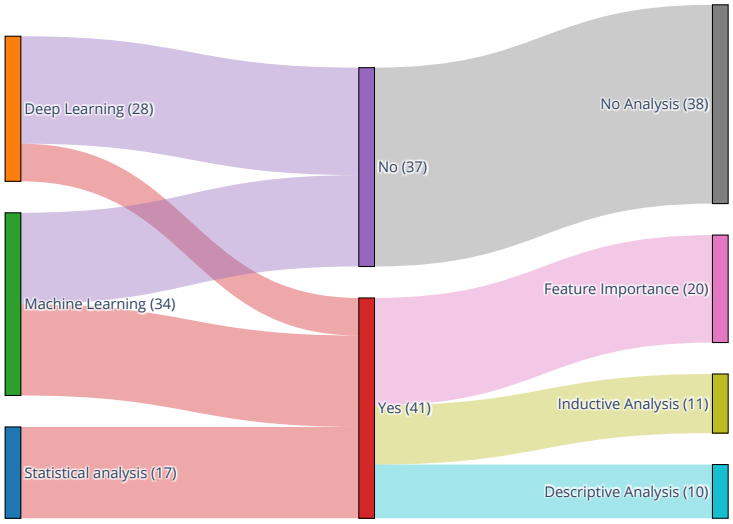


Figure 8. Sankey diagram illustrating the relationship between the analytical methods employed (deep learning, machine learning, and statistical analysis) and the presence and type of explainability techniques. This visualization highlights how interpretability varies across methodological families, 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.

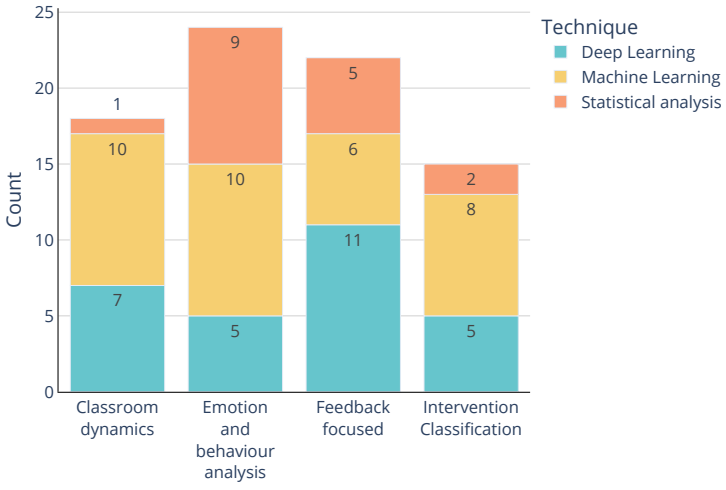


Figure 9. Relationship between RQ1 categories and RQ4 techniques used (RQ4).

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.

5.2. Exploring AI Techniques for Generalizable Teaching Practice Identification

Título

Exploring AI Techniques for Generalizable Teaching Practice Identification

Autores

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

Departamento de Ingeniería y Tecnología de Computadores Universidad de Murcia, España

Detalles de la publicación

Revista	IEEE Access	**Editorial**	IEEE
Volumen	12	**Número**	N/A
Páginas	134713–134726	**Año**	2024
JIF	3.4 (2023)	**Rank**	Q2
Estado	Publicado	**DOI**	10.1109/ACCESS.2024.3456915

Resumen

Using automated models to analyze classroom discourse is a valuable tool for educators to improve their teaching methods. In this paper, we focus on exploring alternatives to ensure the generalizability of models for identifying teaching practices across diverse teaching contexts. Our proposal utilizes artificial intelligence to analyze audio recordings of classroom activities. By leveraging deep learning for speaker diarization and traditional machine learning algorithms for classifying teaching practices, we extract features from the audio diarization using a processing pipeline to provide detailed insights into teaching dynamics. These features enable the classification of three distinct teaching practices: lectures, group discussions, and the use of audience response systems. Our findings demonstrate that these features effectively capture the nuances of teacher-student interactions, allowing for a refined analysis of teaching styles. To enhance the robustness and generalizability of our model, we explore various pipelines for audio processing, evaluating the model's performance across diverse contexts involving different teachers and students. By comparing these practices and their associated features, we illustrate how AI-driven tools can support teachers in reflecting on and improving their teaching strategies.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 July 2024, accepted 23 August 2024, date of publication 10 September 2024, date of current version 27 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3456915

APPLIED RESEARCH

Exploring AI Techniques for Generalizable Teaching Practice Identification

FEDERICO PARDO GARCÍA, ÓSCAR CÁNOVAS, AND FÉLIX J. GARCÍA CLEMENTE
Departamento de Ingeniería y Tecnología de Computadores, Universidad de Murcia, 30100 Murcia, Spain

Corresponding author: Federico Pardo García (federico.pardog@um.es)

This work was supported in part by MCIN/AEI/10.13039/501100011033 under Grant TED2021-129300B-I00; in part by European Union NextGenerationEU/PRTR; and in part by MCIN/AEI/10.13039/501100011033/FEDER, European Union, under Grant PID2021-122466OB-I00.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Murcia.

ABSTRACT Using automated models to analyze classroom discourse is a valuable tool for educators to improve their teaching methods. In this paper, we focus on exploring alternatives to ensure the generalizability of models for identifying teaching practices across diverse teaching contexts. Our proposal utilizes artificial intelligence to analyze audio recordings of classroom activities. By leveraging deep learning for speaker diarization and traditional machine learning algorithms for classifying teaching practices, we extract features from the audio diarization using a processing pipeline to provide detailed insights into teaching dynamics. These features enable the classification of three distinct teaching practices: lectures, group discussions, and the use of audience response systems. Our findings demonstrate that these features effectively capture the nuances of teacher-student interactions, allowing for a refined analysis of teaching styles. To enhance the robustness and generalizability of our model, we explore various pipelines for audio processing, evaluating the model's performance across diverse contexts involving different teachers and students. By comparing these practices and their associated features, we illustrate how AI-driven tools can support teachers in reflecting on and improving their teaching strategies.

INDEX TERMS Audio analysis, deep learning, machine learning, multi-modal learning analytics, speaker diarization, teaching practices.

I. INTRODUCTION

Effective feedback is crucial for teachers to become effective educators, enabling continuous learning, reflection on practice, and adaptation of teaching methods accordingly [1]. Tailored teaching analytics, which provide specific insights into the performance of teachers for each practice, can offer valuable guidance.

In this context, Multimodal Learning Analytics (MMLA) harnesses data from various modalities within the physical classroom environment to provide a comprehensive view of teaching and learning dynamics [2]. By integrating data from several sources (like audio, video, gesture recognition, and other sensory inputs) MMLA enables a nuanced understanding of teaching and learning processes.

For instance, research shows that automated feedback to teachers about the ratio of teacher to student talk can result in a notable increase in student participation [3]. This demonstrates that even basic insights into teachers' classroom discourse patterns can prompt beneficial changes in instructional practices.

To effectively analyze classroom dynamics and teaching practices, it is essential to first identify the specific teaching methods employed throughout different segments of a session, such as lecturing, group work, or the use of audience response systems. This initial identification facilitates a focused analysis tailored to each specific teaching practice. By understanding when and how each method is utilized, educators can gain specific insights relevant to enhancing

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

instructional strategies. For instance, segments of lecturing can be effectively described and analyzed using features such as speaking time, utterance duration, use of silence, and student participation. In contrast, work groups can be examined through the lens of turn-taking counts, overlapping utterances, and participation equality among group members.

Consequently, the ability to analyze teaching practices through automated models using classroom discourse represents a significant advancement in educational research and practice. This paper focuses on addressing the challenge of ensuring the generalizability of these models across various teaching contexts. This implies that the models are not only accurate within a specific environment but are also robust enough to be effective across diverse educational settings, encompassing various classroom dynamics, teacher styles, and student behaviors, even beyond the conditions they were originally trained on.

Our approach is based on several artificial intelligence (AI) methods to analyze audio recordings of classroom activities, employing deep learning for speaker diarization [4] and traditional machine learning algorithms for classifying teaching practices. By applying these methods, we extract comprehensive features from the audio data. Through our research, we show that the extracted features not only capture the subtleties of teacher-student interactions but also facilitate a refined analysis of three teaching styles: lectures, group discussions, and the utilization of audience response systems. These practices are relevant as they encompass different modes of interaction between teachers and students.

Following the EFAR-MMLA framework [5], an evaluation framework to assess and report generalizability of machine learning models in MMLA, our purpose is to evaluate the generalizability of our predictive models in a systematized way. We explore various pipelines for audio processing. This exploration includes evaluating the model's performance across diverse educational contexts, using unseen data. As we show, an appropriate selection of the modeling steps (noise reduction, hyperparameters, post-processing) is crucial to improve the generalizability of the models.

Furthermore, analyzing the features incorporated into the prediction models is essential for understanding which aspects drive the model's performance. This analysis helps extract information that can enhance the model's explainability, enabling a more generalized identification of various teaching practices.

These previously mentioned goals in our research led us to pose the following research questions:

- RQ1: What is the performance of a teaching practice identification model using a generic pipeline trained on data from diverse teaching contexts?
- RQ2: Which processing pipeline best enhances the generalizability of the model's performance across diverse teaching contexts for identifying teaching practices?
- RQ3: What features contribute to the generalizability of the model and effectively describe each teaching practice across diverse educational contexts?

To address these questions, we collected a dataset comprising classroom recordings from four distinct higher education teaching contexts, involving four different teachers and two different subjects. We utilized this dataset to develop classification models using a basic pipeline for RQ1. For RQ2, we identified high-performing pipelines in terms of cross-context generalizability. Finally, we investigate which features contribute to generalizability and examine their relationship to the different teaching practices for RQ3.

The remainder of this paper is organized as follows: Section II reviews the related work that frames our study. In Section III, we outline the methodology employed, followed by a detailed explanation of the processing pipelines and their application in Section IV. Section V presents the results derived from our experiments, which are then examined in Section VI, where we explore their broader implications. Finally, Section VII concludes the paper by summarizing our findings and suggesting avenues for future research.

II. RELATED WORK

In recent years, numerous studies have analyzed classroom climate [6] and discourse across various contexts. Our primary focus is the automated analysis of teacher discourse, particularly through recorded audio. Although recording classroom sessions for teacher assessment is not a new practice, the shift towards automatic analysis of these recordings has only recently gained momentum [7]. Most recent studies employ machine learning or deep learning techniques to examine different teaching practices and styles [8], [9]. This analysis often involves extracting nonverbal features or applying natural language processing techniques, which are not always necessary [10]. Initial research in this field focused on identifying single or multiple voices in classroom audio segments [11]. However, our proposal primarily relies on features derived from the diarization process, which involves labeling speakers and their respective speaking instances. This task can be approached using various methods, from traditional techniques to advanced neural networks [4], with some methods also incorporating video to enhance active speaker identification [12].

In relation to classroom discourse, multiple research teams have dedicated their efforts to the development and validation of automated models aimed at discerning fundamental discourse structures, such as lectures and group work. For instance, Donnelly et al. [13] trained supervised machine learning models to classify instructional segments, achieving F1 scores ranging from 0.64 to 0.78. Furthermore, [14] demonstrated the feasibility of employing automatic speech recognition and classification models to automatically segment classroom speech and identify instances where teachers' utterances contained questions. Other works based on automatic speech recognition have focused on segmenting teacher and student classroom speech [15] and leveraging low-level acoustic features [16], [17]. Most of the proposals pay particular attention to the role of the teacher [8] in order

to classify active learning tasks [18], [19], [20]. However, some works are beginning to shift their focus towards student speech as well [21].

To develop their models, researchers have often followed a standard machine learning process that includes data collection, pre-processing, feature extraction, model development, and evaluation [13], [16], [17]. This process involves multiple steps (e.g., pre-processing, outlier handling, fine-tuning) collectively known as the modeling pipeline. Each step usually offers several options. For instance, it is possible to reduce noise or to filter outliers. Since the choice of using a specific technique or not can affect the model's performance, it is crucial to understand how different choices impact the model's generalizability.

It is worth noting that these works predominantly focus on achieving high classification accuracy, often overlooking the provision of discourse features that can serve as descriptive and informative data [3], [19]. Informative data are crucial for capturing the nuances of teaching practices and providing meaningful insights. As we will introduce in Section IV-D, we defined new features and we also adapted some of the discourse features from previous works such as [22] and [23] which were originally designed for group meeting analysis.

Additionally, analyzing different teaching styles requires generalizability across different contexts, meaning the model's ability to perform on unseen data. To ensure this generalizability, we follow a specific evaluation framework to assess and report the generalizability of machine learning models in MMLA (EFAR-MMLA) [5], which has been used, for example, to evaluate collaboration models [24].We chose the EFAR-MMLA framework due to its specific focus on addressing the challenge of generalizability in MMLA contexts. Unlike traditional evaluation methods, which often fall short in assessing how well models perform across diverse learning environments, EFAR-MMLA is designed to rigorously evaluate and report a model's ability to generalize to unseen data across different contexts. By utilizing EFAR-MMLA, we aim to ensure that our models are not only accurate but also broadly applicable across various teaching styles and environments, thereby enhancing the reliability and applicability of our findings.

In addition to the non-verbal approach, traditional methods of automated teacher discourse analysis have relied on automatic speech recognition (ASR) transcripts [25]. While we are now working on a MMLA system based on non-verbal features and NLP (Natural Language Processing) techniques, this latter approach is out of the scope of this paper.

Finally, there is an increasing concern in the learning analytics field over the interpretability of the machine learning models used [26]. Explainable AI (XAI) should be a crucial part in learning analytics, as it enhances the transparency, trust, and effectiveness of AI systems used in educational settings. By providing clear and interpretable insights into AI-driven decisions, XAI helps educators understand and validate the models' recommendations, leading to more informed and actionable feedback [27], [28]. This is particularly important in education, where the implications of AI decisions directly affect teaching strategies and student outcomes.

III. METHOD

A. DATASET

We conducted our study using audio recordings representing different teaching contexts, obtained from two courses, namely Computer Networks and Computing Foundations, which are part of a bachelor's degree program in Computer Science taught in Spanish. To ensure a comprehensive dataset, we engaged one female and three male teachers to record their respective classes, resulting in a total of 30 audio files, each corresponding to an individual class taught by one of the teachers. The teachers in our sample averaged 20.5 years of experience. The cumulative duration of these audio files is approximately 27 hours, with each file ranging from 30 minutes to 2 hours in duration. All the data were collected with the approval of the teachers, students, and the Institutional Review Board.

Furthermore, each audio session included supplementary contextual data. This contextual data encompassed information such as the specific course to which the audio file pertained, the date of the recording, the duration of the audio file, the number of students present during the recording, the recording device utilized, the teaching methods employed during the class session, and the identity of the instructor delivering the lecture.

In relation to the annotation of the teaching practices implemented in the classrooms, there are four labels for the segments of the recordings:

- **Lecture**. This represents a traditional class where the teacher explains the course material, and the students listen to the teacher, take notes, and ask questions occasionally.
- **Wooclap**. This label is used when an Audience Response System [29] is employed in the classroom to pose questions to the students and engage with them through the use of mobile devices. Wooclap[1] is the platform utilized in all instances.
- **Group Work**. This represents a session where the students work in groups to solve problems.
- **Other**. This label is used when the audio segment does not align with any preceding description, such as instances when the teacher is getting ready to start or during a break in the class. The segments with this label are omitted for this work.

B. HUMAN CODING

The labeling method adopted consists of an audio timestamp indicating the start of the label, another timestamp denoting the end of the labeled segment, and the corresponding teaching method. This labeling process was manually conducted by a single independent coder from the teaching

[1] https://wooclap.com

TABLE 1. Approximate duration of audio per teacher and label in HH:MM.

Method \ Teacher	T1	T2	T3	T4	Total
Group Work	1:57	2:15	1:24	2:45	8:21
Lecture	3:39	2:57	3:15	3:12	13:03
Wooclap	3:06	0:33	0:45	1:09	5:33
Total	8:42	5:45	5:24	7:06	26:57

staff, using the ELAN software.[2] We believe that a single coder is appropriate in this case since the different teaching practices are easily identifiable by a teacher just listening to the audios. The resulting data serves as the ground truth for our experimentation. It is worth mentioning that it is not possible to have a completely balanced dataset where all the teachers follow the same teaching practices in the same proportion, as they were not influenced to change their teaching styles. However, we have attempted to use a dataset that is reasonably representative and contains sufficient data for every teaching practice across the different teaching contexts. The final dataset is composed of 13 hours of lectures, 5 and a half hours of Wooclap, and 8 hours of group work. A complete distribution of the dataset per teacher and teaching method is available in Table 1.

C. EXPERIMENTAL SETUP

During an in-person training session, teachers received instructions on how to record a class session. As part of this process, they conducted an initial recording in their respective classrooms to test and ensure the optimal placement of the digital recorder (TASCAM DR-07X). Each teacher was required to record a minimum of ten sessions that encompassed various teaching practices. For our analysis, we specifically selected recordings that exhibited high audio quality, discarding those that did not meet this criterion.

These classes were all held in traditional lecture classrooms with a seating capacity of 60 to 90 students. The classrooms were equipped with tables and chairs facing the teaching area, which included a desk with fully embedded technology. To ensure a seamless recording setup, the handheld digital recorder was discreetly placed on the teacher's desk, positioned at least 1.5 meters away from both the teacher and the students in the front row.

IV. GENERATION OF PROCESSING PIPELINES

Our approach consists of pipelines composed of different steps. These steps can be either included or skipped to generate various pipelines, though some steps are mandatory for every pipeline, such as speaker diarization, feature calculation, and model training, as shown in Figure 1. The boxes in yellow represent optional processing steps that can be combined in different ways to derive diverse pipelines.

A. NOISE REDUCTION

In our study, this step may involve applying a noise reduction algorithm to minimize background noise. This enhancement could benefit the accuracy of subsequent stages, notably affecting the diarization and feature extraction processes. For example, we have tested the noisereduce library [30] which employs a spectral subtraction-based noise reduction algorithm, which estimates the noise profile within the audio and subtracts this from the original signal. We have noticed that this preprocessing step sometimes alters the diarization process, particularly in identifying speakers. Recent proposals suggest filtering out segments that could lead to incorrect automated transcriptions using neural networks, which could be useful in future works to discard problematic audio segments [31].

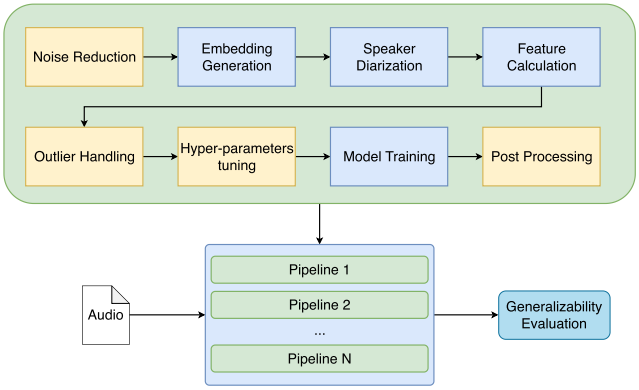


FIGURE 1. Pipeline of processing steps and data elements.

B. SPEAKER EMBEDDING GENERATION

To capture the distinctive features of the teacher's voice, we utilize speaker embeddings based on x-vectors [32]. These embeddings enable the subsequent identification of the teacher's speech within the diarization process for any given audio recording. The Pyannote-audio library [33] is employed to generate these embeddings, using a consistent model throughout the diarization process to ensure the embeddings are comparable. Currently, we do not create specific embeddings for students, as our feature extraction process groups them as a single entity.

C. SPEAKER DIARIZATION

Speaker diarization is the process of dividing audio recordings into segments and assigning speaker labels to determine "Who Speaks When". A diarization system comprises a Voice Activity Detection (VAD) model that identifies time intervals in the audio where speech is present while ignoring background noise. In our approach, we are able to differentiate the teacher's speech; however, we do not distinguish between individual students, labeling all other speakers as 'Students'. Achieving precise speaker diarization can be accomplished using tools like Pyannote-audio,[3] which is the option that we used in our experiments.

[2]https://archive.mpi.nl/tla/elan

[3]https://github.com/pyannote/pyannote-audio

TABLE 2. Description of features.

Feature Type	Feature Name	Description
Per Role	Participant Speaking Ratio (PSR)	The ratio of participation of each role during the recording segment.
Per Role	Participant Speaking Utterances (PSU)	The number of utterances in the current recording segment.
Per Role	Participant Speaking Utterances Ratio (PSUR)	The ratio of utterances of each role during the recording segment.
Per Role	Average Participant Speaking Utterance Duration (APSUD)	The average duration of the utterances of each role.
Global	Average Lapse Duration (ALD)	The average duration of the period of silence.
Global	Silence Ratio (SR)	The ratio of the period of silence during the recording segment.
Global	Average Pause Duration (APD)	The average duration of silence intervals between utterances by the same participant.
Global	Participation Equality (PEQ)	An indicator that assesses the balance of participation among different roles. It is calculated following the methodology outlined in [23]. Values close to 1 indicate an equal distribution of participation.
Global	Turn Taking Count (TTC)	The number of turn changes that occurred in the dialogue between students and the teacher.
Global	Very Short Utterances Ratio (VSUR)	The ratio of very short speech utterances (less than 2 seconds) over the total.
Global	Overlapping Rate (OVR)	The rate of overlapping time of different participants.
Global	Overlapping Utterances Rate (OVUR)	The ratio of utterances that are overlapped.
Global	Mumble Ratio (MR)	This evaluates the extent of voice activity that cannot be attributed to a particular speaker during the diarization process.

D. FEATURE CALCULATION

Considering previous works [22], [23], we defined several non-verbal features [34] extracted from the diarization. Our aim is to analyze whether these features can be used to train generalizable models for identifying teaching practices. Specifically, we defined four discourse features per role (teacher and students) and nine global discourse features, which are described in Table 2.

The primary objective behind extracting these features is to capture relevant information for training models that can reliably discern the employed teaching methodologies in a generalizable manner across contexts. Additionally, these features provide valuable insights into the dynamics of teacher discourse. Non-verbal cues offer a rich source of information about the interaction and dynamics within the classroom. For instance, Turn Taking Count (TTC) can indicate the distribution of speaking turns between the teacher and students, thereby helping to measure the balance of participation more accurately. Overlapping Rate (OVR) can reveal instances of interruption or overlapping speech, reflecting the level of engagement in group work.

E. OUTLIER HANDLING

Handling outliers is essential in data analysis to prevent distortion of statistical results, erroneous conclusions, and degraded performance of machine learning models.

Two prevalent methods for outlier handling are the clipping method and the Local Outlier Factor (LOF) method. The clipping method caps values exceeding a set threshold, offering simplicity and computational efficiency suitable for large datasets. However, it can be arbitrary, as the threshold choice significantly impacts the results and may not effectively address outliers in datasets with varying density distributions.

In contrast, the LOF method assesses the local density deviation of a data point relative to its neighbors, calculating a score that indicates the isolation of a point compared to its surrounding neighborhood. This makes LOF particularly effective for detecting outliers in datasets with complex structures and varying densities, without relying on a predefined threshold. Given the significant variations in density and complex patterns in our data, the LOF method's ability to provide a robust measure of outlierness makes it more suitable for our analysis. Therefore, we decided to employ the Local Outlier Factor method for outlier detection in our study.

F. HYPERPARAMETER TUNING

An optional step before model training is hyperparameter tuning. In this case, a comprehensive search for the best hyperparameters is conducted via grid search, focusing on optimizing the F1 weighted score. The F1 score is chosen as the optimization metric over alternatives like balanced accuracy due to its effectiveness in handling imbalanced datasets, optimizing both precision and recall. The various hyperparameters optimized for each classification method are detailed in Section V.

G. MODEL TRAINING

As we already mentioned, we perform a multi-class classification task involving three classes: Lecture, Group Work, and Wooclap. Our approach employs supervised machine learning algorithms to predict the appropriate label for a given recording segment. As Section V details, we explored a range of models, including Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Random Forest, Naive Bayes, Logistic Regression, Gradient Boosting and Multi-Layer Perceptron (MLP) to identify the most effective algorithm for our task. We employ the scikit-learn library [35] for training and evaluation.

Since a teacher can use multiple teaching methods within a single class, our classification stage aims to identify the start and end points of each method. To achieve this, the classification process involves analyzing small segments

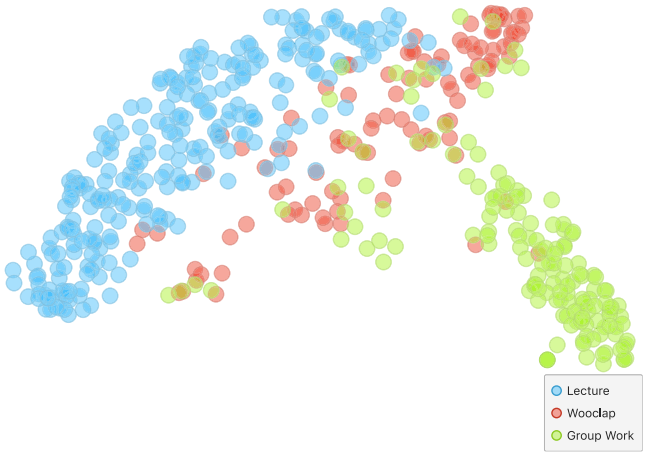


FIGURE 2. t-SNE visualization of three distinct clusters representing teaching practices.

of the recordings. We have examined overlapping window sizes ranging from 60 seconds to 300 seconds. Preliminary results indicate that the best trade-off between granularity and accuracy is achieved using 180-second windows, which will be used for all subsequent tests.

H. POST PROCESSING

The final stage in the pipeline is an optional step designed to account for temporal trends in classrooms. For instance, as a preliminary study of the data revealed, it is unlikely to have a short interval of group work between several segments labeled as lectures. Consequently, we designed a method using a sliding window filter (based on majority voting) that inputs the sequence of predicted labels for each individual recording. We aim to evaluate whether this filter can increase performance by eliminating noisy, often isolated predictions that diverge from neighboring predictions. We also need to assess if the overall generalizability is improved, considering that it may occasionally modify some correct predictions.

V. RESULTS

A. PRELIMINARY ANALYSIS OF FEATURES

Our models are based on the set of 13 distinct non-verbal features derived from the diarization process. Therefore, our first objective is to assess their suitability for automating the identification of teaching practices. To initiate this investigation, we utilize unsupervised techniques to explore the data. In particular, we employ dimensionality reduction techniques such as t-SNE (t-distributed Stochastic Neighbor Embedding) to reduce the data's dimensionality and facilitate visualization.

By applying t-SNE to the data (with three PCA components and exaggeration = 1), we obtain the visualization shown in Figure 2. As we observe, it confirms the potential separability of clusters within the data. The distribution of samples shows distinct groupings, suggesting promising prospects for automatically classifying different teaching practices, although the "Wooclap" label presents some challenges. Our next objective will be the generation of a model for a multiclass classification task that distinguishes among the three teaching methods.

B. MODEL USING A GENERIC PIPELINE TRAINED ON DATA FROM DIVERSE TEACHING CONTEXTS (RQ1)

For the development of this model we use a basic pipeline that was already tested in a previous work with a different dataset [34], composed of the following elements: embedding generation, speaker diarization, feature calculation, outlier handling, and model training. Utilizing supervised machine learning algorithms, we predict the teaching method for given recording segments. Since there are several alternatives, our exploration covers several models, including Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Logistic Regression, Gradient Boosting, and Multi-Layer Perceptrons (MLP).

The models were assessed using 10-fold stratified cross-validation using data from all the teaching contexts we introduced in Section III-A. Despite the fact that F1 score was used to optimize the cross validation, we have chosen to compare the models using F1, Precision, Recall, and Balanced Accuracy.[4] This decision is grounded in the recognition that the F1 score, being the harmonic mean of Precision and Recall, might not fully encapsulate the comparative performance across all models. Moreover, incorporating Balanced Accuracy allows for a more comprehensive overview of each model's classification efficacy. Therefore, we will showcase the models' results across all mentioned metrics.

Based on the results presented in Table 3, Multi-Layer Perceptron, Logistic Regression and Support Vector Classification models perform similarly, achieving F1 scores of 0.92, closely followed by the Gradient Boosting and k-Nearest Neighbors models. Overall, the high F1 scores across all models underscore their effectiveness in accurately identifying relevant instances while minimizing false positives and false negatives, even using this basic pipeline.

As reference points for the upper and lower bounds of the models' performance in the next subsection, we will use the results from this test and the knowledge from the collected dataset. The upper bound is set to an F1 score of 0.923 (corresponding to the MLP model). The lower bound is based on the majority class, which corresponds to a model that always outputs the predominant class, in this case, "Lecture," and provides an F1 score of 0.316.

Further analysis of the different models, detailed in Figure 3, elucidates the performance of the models for the various classes or teaching practices. The 'Lecture' class is is the easiest to classify, whereas 'Wooclap' presents a challenge since it has the lowest F1 score. This pattern is consistent with conclusions drawn from the t-SNE analysis,

[4]Balanced Accuracy will be referred to as Accuracy in the tables for simplicity.

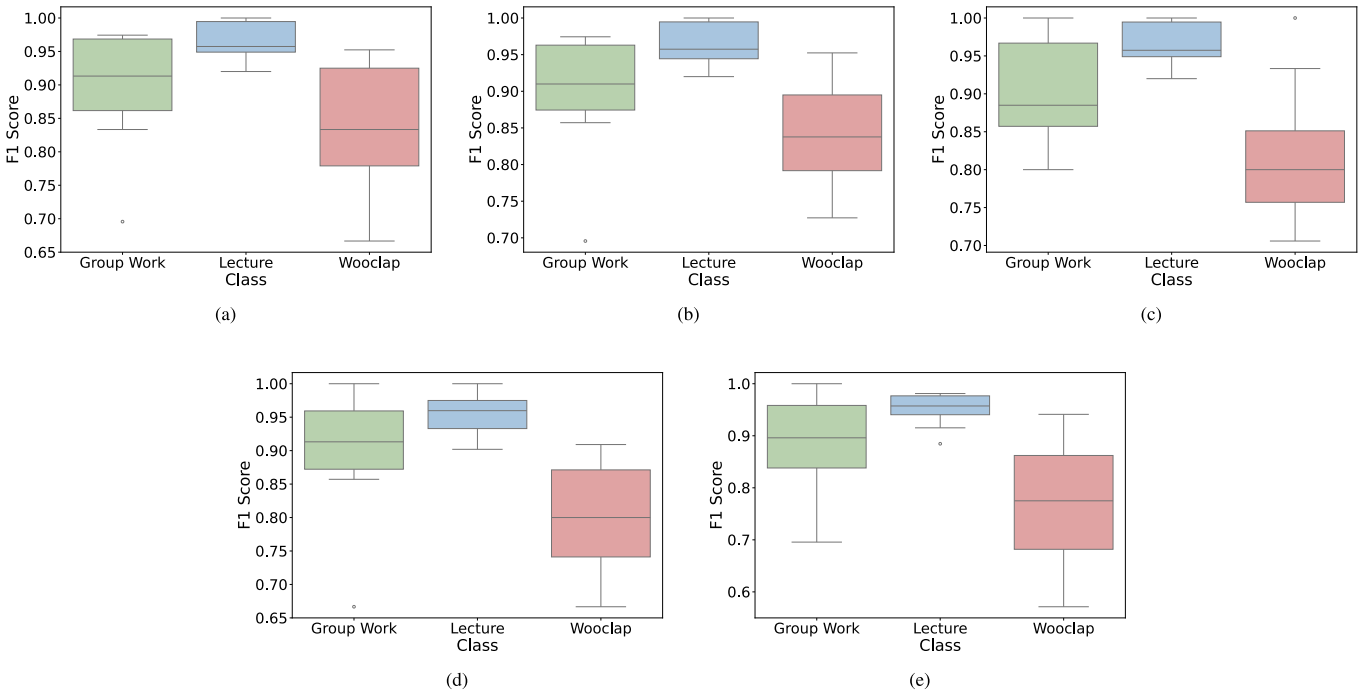


FIGURE 3. Detailed analysis of the different F1 scores for each method and teaching practice - (a) MLP, (b) Logistic regression, (c) SVM, (d) Gradient boosting, (e) k-Nearest neighbors.

TABLE 3. Test results for the models, sorted by F1 score.

Model	F1 Weighted Score	Balanced Accuracy	Precision	Recall
MLP Classifier	0.923	90.752%	0.932	0.923
Logistic Regression	0.923	90.667%	0.932	0.923
SVC	0.920	90.024%	0.925	0.921
Gradient Boosting Classifier	0.909	88.598%	0.918	0.909
k-Nearest Neighbors Classifier	0.897	86.502%	0.910	0.901

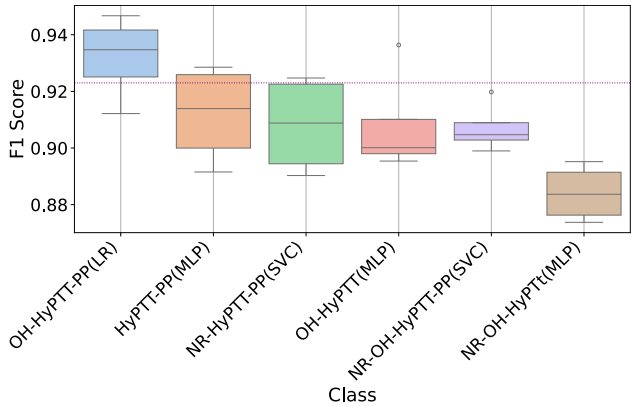


FIGURE 4. Developed pipelines F1 score.

which shows 'Wooclap' overlapping significantly with the other two classes. The proximity of 'Wooclap' to both other classes is expected, given that 'Wooclap' scenarios can fluctuate due to the tool's versatility, ranging from competitive environments—where students are more likely to actively participate—to non-competitive settings, which are quieter and involve fewer interactions. This variability highlights the challenges in accurately classifying teaching methods that display a broad spectrum of student participation dynamics and teacher behavior.

C. GENERALIZABILITY AND ROBUSTNESS OF THE MODEL'S PERFORMANCE ACROSS DIVERSE TEACHING CONTEXTS (RQ2)

In relation to generalizability and robustness, we defined six different pipelines, as Table 4 details. Robustness refers to the model's ability to consistently perform well across a wide range of teaching contexts, ensuring that variations in teaching styles, classroom environments, and audio conditions do not significantly impact the model's effectiveness. Our goal is to explore which pipelines provide better results at the context level. Since we have four different teaching contexts (four different teachers), we will test these pipelines using a leave-two-contexts-out strategy. This means we will use all combinations where two contexts are used for training and the other two for testing purposes. This results in six different combinations, which will also provide a measure of performance variation as an indication of oscillation in predictive ability.

TABLE 4. List of pipelines (NR: Noise reduction; OH: Outlier handling; HyPT: Hyperparameter tuning; PP: Post-processing).

Number	Pipeline
1	OH-HyPT
2	OH-HyPT-PP
3	HyPT-PP
4	NR-OH-HyPT
5	NR-OH-HyPT-PP
6	NR-HyPT-PP

TABLE 5. Optimized hyperparameters for the different classification methods (bold values provide the best results).

Pipeline	Hyperparameters
HyPTT-PP (MLP)	Activation: [**logistic**, tanh, relu] Regularization: [0.5, **1**] Hidden layers: [**(5, 3)**, (3, 3)] Learning rate: [**constant**, invscaling, adaptive] Solver: [**lbfgs**, sgd, adam]
OH-HyPTT-PP (LR)	Penalty: [L1, **L2**] Regularization: [0.001, 0.01, 0.1, **1**, 10, 100] Solver: [**liblinear**, saga] Max # of iterations: [**100**, 200, 300] Class Weight: [None, **balanced**]
NR-OH-HyPTT-PP (SVC)	Regularization: [1, 2, **5**, 10, 100] Kernel coeff.: [**scale**, auto, 0.1, 1, 2, 5, 10, 100] Kernel Type: [linear, **rbf**, poly, sigmoid] Probability: [**True**, False] Class weight: [**None**, balanced]
OH-HyPTT (MLP)	Activation: [**logistic**, tanh, relu] Regularization: [0.5, **1**] Hidden layers: [**(5, 3)**, (3, 3)] Learning rate: [**constant**, invscaling, adaptive] Solver: [**lbfgs**, sgd, adam]
NR-OH-HyPTT (MLP)	Activation: [**logistic**, tanh, relu] Regularization: [**0.5**, 1] Hidden layers: [**(5, 3)**, (3, 3)] Learning rate: [**constant**, invscaling, adaptive] Solver: [**lbfgs**, sgd, adam]
NR-HyPTT-PP (SVC)	Regularization: [**1**, 2, 5, 10, 100] Kernel coeff.: [**scale**, auto, 0.1, 1, 2, 5, 10, 100] Kernel Type: [linear, **rbf**, poly, sigmoid] Probability: [**True**, False] Class weight: [**None**, balanced]

For each pipeline, we did a thorough search for the best hyperparameters using grid search and stratified 10-fold cross-validation, focusing on optimizing the F1 weighted score. We chose F1 because it handles imbalanced datasets well, optimizing both precision and recall. The explored hyperparameters for each pipeline are shown in Table 5, with the selected values highlighted.

The results for each pipeline are presented in Figure 4. The horizontal dotted line represents the baseline F1 score from the pipeline developed in RQ1. Given the focus on generalization, where training is conducted on data from two teachers and predictions are made for the other two, a reduction in performance is anticipated. This expectation holds true for all pipelines, except for the OH-HyPTT-PP pipeline, which slightly outperforms the baseline. This improvement is likely due to hyperparameter tuning and post-processing steps, with the latter being particularly effective, as it corrects some of the wrong predictions. Nevertheless,

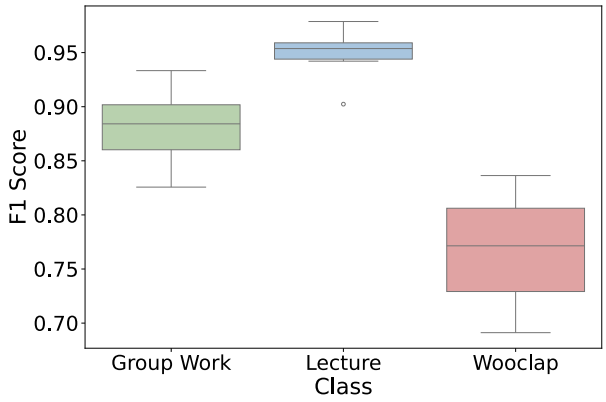


FIGURE 5. F1 score for the logistic regression classifier for the different classes using the pipeline OH-HyPTT-PP.

the F1 scores are similar across all pipelines and the baseline, except for the NR-OH-HyPTT pipeline, which falls below an F1 score of 0.9. In general, the difference in F1 score compared to the baseline is acceptable considering the focus on generalization.

Examining the performance of the best model for each method, as shown in Figure 5, we observe patterns similar to those noted for RQ1. The primary difference from the baseline methods is the reduced performance variation across all three methodologies, particularly with 'Wooclap'. Although the performance is slightly lower than before, it is more consistent across different contexts, which can lead to better generalization.

Finally, we aimed to analyze which features contribute most significantly to the model's predictions. To achieve this, we focused on the best-performing pipeline, specifically OH-HyPTT-PP with the Logistic Regression model, whose feature coefficients are presented in Figure 6. It should be noted that the coefficients in Logistic Regression can be negative; however, for visualization purposes, we present all coefficients as absolute values. We observe that 'Silence Ratio (SR)' is one of the most consistent and significant classifiers. This can lead to suboptimal performance in some cases, as the 'Silence Ratio' varies among teachers based on their teaching styles. The feature with the most variability is 'Mumble Ratio (MR),' which is highly valuable for classifying some teachers' methodologies but not for others. Nevertheless, it clearly provides meaningful information for classifying teaching methods. The remaining features vary depending on the combination of teachers, with notable contributions from 'ALD', 'PSU_Teacher', and 'PSUR_Teacher'.

When comparing these feature importances with those extracted from the SVM model for the same pipeline, presented in Figure 7, we observe some consistency. As before, we have presented all values as positive numbers. Similar to Logistic Regression, 'SR' remains the primary feature for classification, while 'MR' retains its variability. However, SVM appears to extract more meaningful information from 'PSU_Teacher'.

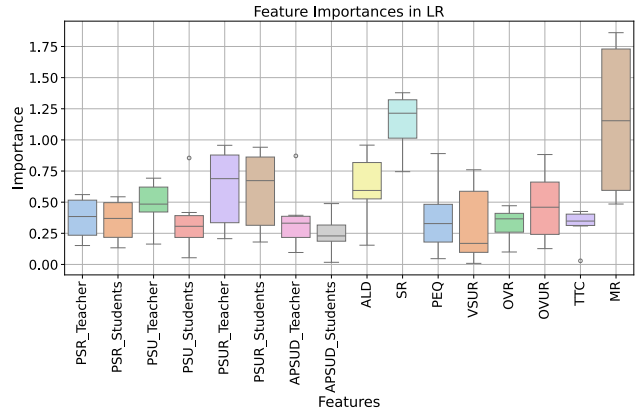


FIGURE 6. Logistic Regression feature importance in pipeline OH-HyPTT-PP.

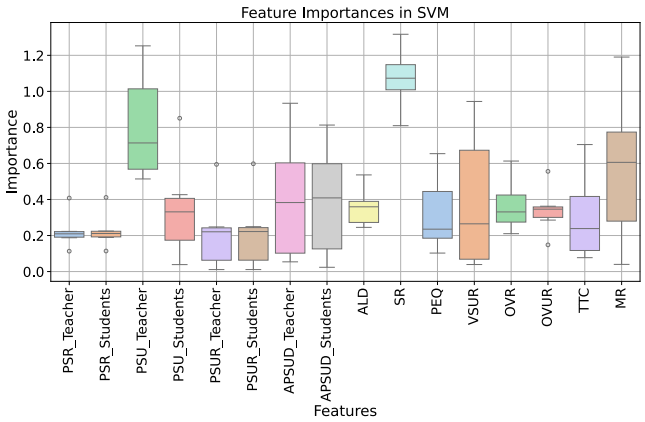


FIGURE 7. SVM feature importance in pipeline OH-HyPTT-PP.

When examining pipelines that utilize denoised audio, we find no significant performance gains. This suggests that the diarization models are effective across various acoustic settings, including noisy environments with reverberations and overlapping speech. Additionally, post-processing generally enhances performance, as indicated by higher F1 scores with post-processing compared to those without. The filter improves performance by removing noisy, isolated predictions that differ from neighboring predictions. Although it may occasionally exclude correct predictions, the overall effectiveness is significantly improved. Correct predictions are typically clustered with similar outcomes, protecting them from being filtered out.

D. FEATURES CONTRIBUTING TO THE GENERALIZABILITY OF THE MODEL (RQ3)

After analyzing the generalizability of the models for the identification of teaching practices using the features, it is essential to examine some of the features incorporated into the prediction models to understand which ones explain the model's performance. This analysis helps extract information that can enhance the model's explainability, enabling a more generalized identification of various teaching practices. Figure 7 showed the features for the best-performing pipeline. Therefore, we will focus our analysis on some of the most important features. Firstly, we will examine the statistical distribution of the features using windows of 180 seconds for all the audios in the dataset. Figure 8 displays the boxplots generated from the dataset, representing the data distribution of the selected features for each teaching practice. These boxplots provide a visual representation of how different features vary across teaching practices, highlighting their distinct characteristics.

TABLE 6. Kruskal-Wallis test results for three features among four teachers for the lecture sessions.

Variable	χ^2	gl	p	ϵ^2
SR	384	2	<.001	0.713
PSU Teacher	321	2	<.001	0.597
MR	374	2	<.001	0.694

To further validate the observed differences in the statistical distributions of the features for the different teaching practices, we also conducted a statistical test. The distributions are confirmed to be non-normal. Therefore, the non-parametric Kruskal-Wallis test was employed. Table 6 provides the details of this analysis, with the variables denoted along with their respective Chi-square (χ^2) values, degrees of freedom (gl), p-values (p), and effect sizes (ϵ^2). Each variable shows significant differences across the teaching practices, as indicated by the p-values being less than 0.001.

The Silence Ratio (SR) and Mumble Ratio (MR) exhibit large effect sizes (SR: 0.713, MR: 0.694). These results indicate substantial variation across teaching practices, highlighting the significant differences in the amount of silence and clarity of speech during different teaching methods.

Similarly, PSU Teacher (Speaking Utterances) shows a significant effect size of 0.597. This feature reflects considerable differences in the frequency and ratio of teacher speaking events across the various teaching practices, demonstrating varied levels of teacher engagement and interaction.

VI. DISCUSSION

This section presents the findings for each research question and highlights the limitations of the study.

A. RQ1:WHAT IS THE PERFORMANCE OF A TEACHING PRACTICE IDENTIFICATION MODEL USING A GENERIC PIPELINE TRAINED ON DATA FROM DIVERSE TEACHING CONTEXTS?

Considering the performance results obtained in our work, the selection of features for this study appears to be well-suited for addressing our research questions. Certain features provide valuable input for the classification models since the F1 score of the basic pipeline ranged from 0.897 to 0.923, depending on the machine learning method used. However, it is important to note that none of the teaching contexts was completely left out of the training process, as we were primarily interested in calculating an upper bound for performance. We discovered that the uneven distribution of classes in our dataset makes the weighted F1 score more

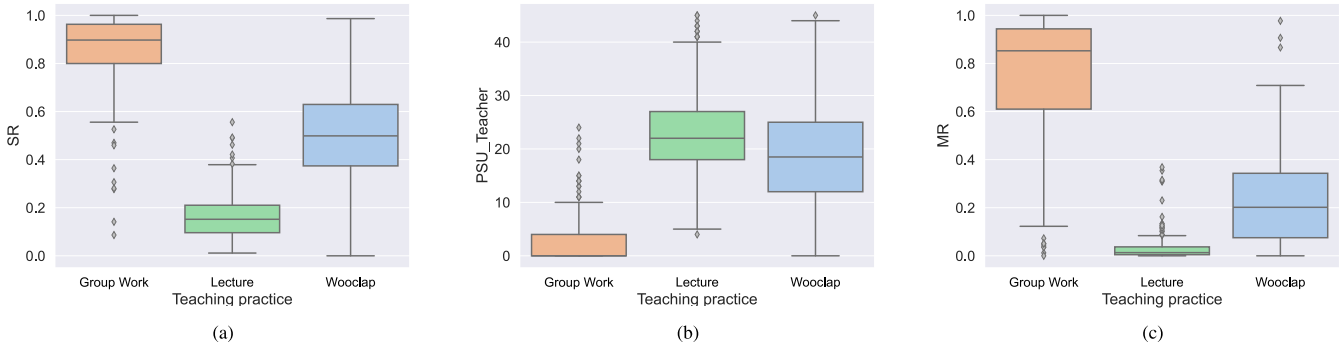


FIGURE 8. Distributions of the features for the different teaching practices - (a) SR (Silence ratio), (b) PSU teacher (Speaking utterances), (c) MR (Mumble ratio).

suitable than regular accuracy. Additionally, we found that certain teaching practices, particularly 'Wooclap', are more challenging and tend to accumulate most of the classification errors.

B. RQ2: WHICH PROCESSING PIPELINE BEST ENHANCES THE GENERALIZABILITY AND ROBUSTNESS OF THE MODEL'S PERFORMANCE ACROSS DIVERSE TEACHING CONTEXTS FOR IDENTIFYING TEACHING PRACTICES?

Based on our experiments, the optimal pipeline is OH-HyPTT-PP when using the Logistic Regression model. This pipeline slightly outperforms the baseline, achieving a mean F1 score of 0.928. Analyzing the F1 score for each category, we observe that the same issues from the baseline persist, with 'Wooclap' being the most challenging methodology to classify. However, the variability across the three classes is reduced, indicating more consistent performance across contexts. This consistency suggests that the models exhibit robust performance across diverse teaching contexts.

Regarding feature importance in the analyzed models, there is consensus on the use of 'Silence Ratio (SR)' as one of the main predictors and the significant variability of 'Mumble Ratio (MR)'. Additionally, features such as 'PSU_Teacher' and 'PSUR_Teacher' are identified as primary features in both models analyzed. The consistent identification of these features as important across different models highlights the robustness of the feature extraction process.

We also found that post-processing enhances performance compared to pipelines without it and does not adversely affect generalization. Conversely, noise reduction is not a necessary step, as the diarization models are already effective in capturing the relevant information for subsequent audio feature extraction.

C. RQ3:WHAT FEATURES CONTRIBUTE TO THE GENERALIZABILITY OF THE MODEL AND EFFECTIVELY DESCRIBE EACH TEACHING PRACTICE ACROSS DIVERSE EDUCATIONAL CONTEXTS?

As we mentioned earlier, a detailed analysis of the features, for example examining Figure 8, helps to extract information that can enhance the model's explainability.

In examining the Silence Ratio (SR), Group Work exhibits a very high ratio, with most values clustering close to 1. This suggests that group work sessions involve prolonged periods of silence, possibly indicating students working quietly or independently. Conversely, lectures show a much lower silence ratio, with values clustering around 0.2, implying continuous speaking with minimal silence. Wooclap sessions present a more variable silence ratio, ranging from 0.3 to 0.9, indicating a varied mix of speaking and silence intervals during these interactive sessions.

The number of speaking utterances reveals that during Group Work, the teacher's speaking utterances are minimal, typically below 10. This aligns with the nature of group activities where the teacher speaks less, allowing students more interaction time. Lectures, on the other hand, show a significantly higher number of speaking utterances by the teacher, ranging from 10 to 35, reflecting the teacher's dominant role in delivering content. In Wooclap sessions, the number of speaking utterances is moderately high, distributed around 20 to 30, suggesting a balance between teacher instructions and student interactions facilitated by the platform.

Finally, the Mumble Ratio (MR) feature indicates high values for Group Work, close to 1, suggesting a lot of indistinct or low-volume speech, likely from students discussing in groups. Lectures exhibit a very low mumble ratio, close to 0, indicating clear and distinct speech by the teacher. Wooclap sessions show a moderate mumble ratio, around 0.2-0.6, reflecting a mix of clear instructions and student responses, some of which might be mumbled.

The significant differences in individual features across teaching practices underscore the complexity of classroom dynamics, but these features alone cannot fully explain the nuances between different teaching methods. This limitation was addressed in previous sections using machine learning techniques to detect complex patterns and interactions within the data that are not immediately apparent through isolated feature analysis. However, this isolated analysis is really useful for explaining the patterns that ML techniques find in the data, and it is a first step to provide actionable information.

D. LIMITATIONS

Our work has four main limitations. The first limitation pertains to the placement of the digital recorder in the classrooms. Although care was taken to position the recorder optimally, this setup does not guarantee that all students are adequately recorded, especially when they are dispersed throughout the classroom or if the teacher moves around while speaking. Consequently, some audio segments may not fully capture interactions and contributions from all participants, potentially affecting the accuracy of the recorded data.

The second limitation is related to the linguistic scope of our dataset. The courses included in the study are exclusively taught in Spanish, which restricts the generalizability of our findings to other languages. Different languages can exhibit varying speech patterns, intonations, and classroom dynamics, which might influence the performance of the diarization and classification models. Thus, further research is necessary to validate the applicability of our models in multilingual settings.

The third limitation involves the specific configuration of some teaching contexts, particularly the use of Wooclap as the audience response system. Wooclap has unique characteristics that influence the audio patterns captured during its use. Other audience response tools may generate different types of interactions and audio signatures.

The fourth limitation is the diversity of teaching contexts. Although our study involves a set of different contexts, it would be beneficial to include a larger number of teachers to increase the reliability of our findings. A larger dataset with greater variability in teaching styles, classroom environments, and subjects would allow for a more thorough evaluation of the generalizability of our models.

VII. CONCLUSION AND FUTURE WORK

This study has demonstrated the efficacy of employing artificial intelligence methods, specifically deep learning for speaker diarization and machine learning for classification, to analyze and classify teaching practices based on audio recordings from classroom activities. Following the EFAR-MMLA framework, our findings confirm that the implemented AI techniques can effectively identify and differentiate between lectures, group discussions, and uses of audience response systems, capturing the subtle nuances of teacher-student interactions. These capabilities provide valuable insights into classroom dynamics, offering a tool for educators to reflect on and improve their teaching strategies. Crucially, the exploration into various audio processing pipelines highlighted the robustness of our models across diverse educational settings and configurations, affirming the generalizability of our approach which is critical for real-world deployment.

The use of AI-driven tools, as demonstrated in this study, suggests a promising direction for enhancing educational practices through technology. By offering a more nuanced understanding of classroom interactions, educators can receive tailored feedback that might be too subtle to detect through traditional observation methods alone. However, despite promising results, our study acknowledges certain limitations, including the dependency on high-quality audio recordings and the model's sensitivity to varying acoustic environments, which could affect scalability and practical application. Additionally, the cultural and linguistic diversity of the classroom settings was not fully explored, which may influence the generalizability of the findings.

Looking forward, we aim to extend our research to include more diverse educational settings and further refine our models to handle a wider range of acoustic conditions. We also plan to integrate multimodal data sources, such as video and textual feedback, to enhance the richness of the analysis and improve the model's accuracy and applicability. Another promising direction is the exploration of real-time feedback systems, where AI tools provide immediate insights during classroom sessions, thus allowing for dynamic adjustments in teaching strategies. By advancing our understanding and application of AI in educational settings, especially within an MMLA framework, we can better support teachers in their critical role of shaping learner experiences and outcomes.

ACKNOWLEDGMENT

The authors would like to thank the rater and teachers who participated in this study and helped to collect the data.

REFERENCES

[1] J. Archer, S. Cantrell, S. L. Holtzman, J. N. Joe, C. M. Tocci, and J. Wood," *Better FeedBack for Better Teaching: A Practical Guide to Improving Classroom Observations*. Hoboken, NJ, USA: Wiley, 2016.

[2] P. Blikstein and M. Worsley, "Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks," *J. Learn. Anal.*, vol. 3, no. 2, pp. 220–238, Sep. 2016.

[3] Z. Wang, X. Pan, K. F. Miller, and K. S. Cortina, "Automatic classification of activities in classroom discourse," *Comput. Educ.*, vol. 78, pp. 115–123, Sep. 2014.

[4] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101317.

[5] P. Chejara, L. P. Prieto, A. Ruiz-Calleja, M. J. Rodríguez-Triana, S. K. Shankar, and R. Kasepalu, "EFAR-MMLA: An evaluation framework to assess and report generalizability of machine learning models in MMLA," *Sensors*, vol. 21, no. 8, p. 2863, Apr. 2021.

[6] A. James, "Automated classification of classroom climate by audio analysis," in *Proc. 9th Int. Workshop Spoken Dialogue Syst. Technol.* Cham, Switzerland: Springer, 2019, pp. 41–49.

[7] M. E. Dale, A. J. Godley, S. A. Capello, P. J. Donnelly, S. K. D'Mello, and S. P. Kelly, "Toward the automated analysis of teacher talk in secondary ELA classrooms," *Teaching Teacher Educ.*, vol. 110, Feb. 2022, Art. no. 103584.

[8] D. Schlotterbeck, P. Uribe, R. Araya, A. Jimenez, and D. Caballero, "What classroom audio tells about teaching: A cost-effective approach for detection of teaching practices using spectral audio features," in *Proc. LAK21: 11th Int. Learn. Anal. Knowl. Conf.*, Apr. 2021, pp. 132–140.

[9] M. K. H. Kanchon, M. Sadman, K. F. Nabila, R. Tarannum, and R. Khan, "Enhancing personalized learning: AI-driven identification of learning styles and content modification strategies," *Int. J. Cognit. Comput. Eng.*, vol. 5, pp. 269–278, Jan. 2024.

[10] D. Wang and G. Chen, "Are perfect transcripts necessary when we analyze classroom dialogue using AIoT?" *Internet Things*, vol. 25, Apr. 2024, Art. no. 101105.
[11] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep learning for classroom activity detection from audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3727–3731.
[12] R. Ahmad, S. Zubair, H. Alquhayz, and A. Ditta, "Multimodal speaker diarization using a pre-trained audio-visual synchronization model," *Sensors*, vol. 19, no. 23, p. 5163, Nov. 2019.
[13] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, S. Kelly, M. Nystran, and S. K. D'Mello, "Automatic teacher modeling from live classroom audio," in *Proc. Conf. User Modeling Adaptation Personalization*, Jul. 2016, pp. 45–53.
[14] P. J. Donnelly, N. Blanchard, A. M. Olney, S. Kelly, M. Nystrand, and S. K. D'Mello, "Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 218–227.
[15] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly, "Multimodal capture of teacher–student interactions for automated dialogic analysis in live classrooms," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 557–566.
[16] D. Schlotterbeck, P. Uribe, A. Jiménez, R. Araya, J. Van der Molen Moris, and D. Caballero, "TARTA: Teacher activity recognizer from transcriptions and audio," in *Proc. Int. Conf. Artif. Intell. Educ.*, Utrecht, The Netherlands. Cham, Switzerland: Springer, Jun. 2021, pp. 369–380.
[17] E. Slyman, C. Daw, M. Skrabut, A. Usenko, and B. Hutchinson, "Fine-grained classroom activity detection from audio with neural networks," 2021, *arXiv:2107.14369*.
[18] H. Li, Y. Kang, W. Ding, S. Yang, S. Yang, G. Y. Huang, and Z. Liu, "Multimodal learning for classroom activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 9234–9238.
[19] M. T. Owens, "Classroom sound can be used to classify teaching practices in college science courses," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 12, pp. 3085–3090, 2017.
[20] H. Su, B. Dzodzo, X. Wu, X. Liu, and H. Meng, "Unsupervised methods for audio classification from lecture discussion recordings," in *Proc. Interspeech*, Sep. 2019, pp. 3347–3351.
[21] R. Southwell, W. Ward, V. A. Trinh, C. Clevenger, C. Clevenger, E. Watts, J. Reitman, S. D'Mello, and J. Whitehill, "Automatic speech recognition tuned for child speech in the classroom," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 12291–12295.
[22] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. F. Welles, and R. J. Radke, "A multimodal-sensor-enabled room for unobtrusive group meeting analysis," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 347–355.
[23] C. Lai, J. Carletta, and S. Renals, "Modelling participant affect in meetings with turn-taking features," in *Proc. Workshop Affect. Social Speech Signals*, 2013, pp. 1–5.
[24] P. Chejara, L. P. Prieto, M. J. Rodriguez-Triana, R. Kasepalu, A. Ruiz-Calleja, and S. K. Shankar, "How to build more generalizable models for collaboration quality? Lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics," in *Proc. LAK23: 13th Int. Learn. Anal. Knowl. Conf.*, Mar. 2023, pp. 111–121.
[25] T. Nazaretsky, J. N. Mikeska, and B. Beigman Klebanov, "Empowering teacher learning with AI: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion," in *Proc. LAK23: 13th Int. Learn. Anal. Knowl. Conf.*, Mar. 2023, pp. 122–132.
[26] S. Sohail, A. Alvi, and A. Khanum, "Interpretable and adaptable early warning learning analytics model," *Comput., Mater. Continua*, vol. 71, no. 2, pp. 3211–3225, 2022.
[27] B. Adrien and F. Benoît, "Interpretability of machine learning models and representations: An introduction," in *Proc. 24th Eur. Symp. Artif. Neural Netw., Comput. Intell. Machine Learn.*, Apr. 2016, pp. 77–82.
[28] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.
[29] J. I. Castillo-Manzano, M. Castro-Nuño, L. López-Valpuesta, M. T. Sanz-Díaz, and R. Yñiguez, "Measuring the effect of ARS on academic performance: A global meta-analysis," *Comput. Educ.*, vol. 96, pp. 109–121, May 2016.
[30] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLOS Comput. Biol.*, vol. 16, no. 10, Oct. 2020, Art. no. e1008228.
[31] J. Hewstone and R. Araya, "Neural network-based approach to detect and filter misleading audio segments in classroom automatic transcription," *Appl. Sci.*, vol. 13, no. 24, p. 13243, Dec. 2023.
[32] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.
[33] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.Audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7124–7128.
[34] O. Canovas, F. J. Garcia-Clemente, and F. Pardo, "AI-driven teacher analytics: Informative insights on classroom activities," in *Proc. IEEE Int. Conf. Teaching, Assessment Learn. Eng. (TALE)*, Nov. 2023, pp. 1–8.
[35] L. Buitinck, "API design for machine learning software: Experiences from the scikit-learn project," in *Proc. ECML PKDD Workshop Lang. Data Mining Mach. Learn.*, 2013, pp. 108–122.

FEDERICO PARDO GARCÍA received the degree in computer engineering from the University of Murcia, in 2020, and the master's degree in big data, in 2021. He is currently pursuing the Ph.D. degree with the University of Murcia. He has participated in projects involving sound classification using artificial intelligence and image recognition.

ÓSCAR CÁNOVAS received the degree in computer engineering from the University of Murcia, in 1998, and the Ph.D. degree in computer science, in 2003. He is currently an Associate Professor with the University of Murcia. He has participated in several research projects and published articles in various fields, such as information security, user authentication and authorization, software-defined networks, indoor positioning, educational research, and the application of educational technology.

FÉLIX J. GARCÍA CLEMENTE received the Ph.D. degree in computer science. He is currently a Full Professor in the area of computer architecture and technology with the Faculty of Computer Science, University of Murcia. His research interests include cybersecurity, cloud computing, and educational technology. As a result of his research, he has authored more than 120 publications, including journals and conference papers. He is an active member of various national and international research and development projects.

5.3. Explaining Teacher Interventions in SRS-based Classrooms: A Classification Approach with BERT and Paralinguistic Cues

Título

Explaining Teacher Interventions in SRS-based Classrooms: A Classification Approach with BERT and Paralinguistic Cues

Autores

Federico Pardo García, Óscar Cánovas Reverte, Félix J. García Clemente, Antonio Orenes Lucas

Departamento de Ingeniería y Tecnología de Computadores Universidad de Murcia, España

Detalles de la publicación

Revista	IEEE Access	**Editorial**	IEEE
Volumen	13	**Número**	N/A
Páginas	208078 - 208093	**Año**	2025
JIF	3.6 (2024)	**Ranking**	Q2
Estado	Publicado	**DOI**	10.1109/ACCESS.2025.3641484

Resumen

This article introduces an automated system designed to classify teacher interventions in classrooms where Student Response Systems (SRS) are actively used. SRS tools, increasingly common in higher education, are valued for promoting more active learning and student engagement through real-time questioning and feedback cycles. Our methodology processes raw audio recordings, employing automatic speech recognition (ASR) and speaker diarization to generate transcripts and to derive several quantitative features. These transcripts serve as input for a series of advanced BERT-based models, which progressively leverage textual content, conversational context, and additional features to classify interventions into predefined COPUS-based pedagogical categories. A key contribution of our work is the empirical demonstration that hybrid models, fusing textual content with paralinguistic cues, can improve the performance of highly-optimized text-only models. This fusion provides a complementary source of information, and we couple our analysis with Explainable AI (XAI) techniques to clarify the distinct influence of these new features. Crucially, Explainable Artificial Intelligence (XAI) techniques, particularly SHAP, are applied to elucidate how these models fuse textual cues with paralinguistic elements (e.g., speaker ratios, silence patterns), revealing their differential contributions to the classification process. Ultimately, this study demonstrates a viable approach for creating accurate, interpretable AI systems that offer teachers feedback on their instructional patterns.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 November 2025, accepted 27 November 2025, date of publication 8 December 2025,
date of current version 12 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3641484

RESEARCH ARTICLE

Explaining Teacher Interventions in SRS-Based Classrooms: A Classification Approach With BERT and Paralinguistic Cues

FEDERICO PARDO GARCÍA, ÓSCAR CÁNOVAS REVERTE, FÉLIX J. GARCÍA CLEMENTE, AND ANTONIO ORENES LUCAS

Departamento de Ingeniería y Tecnología de Computadores (DITEC), Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain

Corresponding author: Federico Pardo García (federico.pardog@um.es)

This work was supported in part by the strategic project "Development of Professionals and Researchers in Cybersecurity, Cyberdefense and Data Science (CDL-TALENTUM)" from the Spanish National Institute of Cybersecurity and the Recovery, Transformation and Resilience Plan, Next Generation EU; and in part by the Spanish Ministry of Science and Innovation under Project PID2021-122466OB-I00.

ABSTRACT This article introduces an automated system designed to classify teacher interventions in classrooms where Student Response Systems (SRS) are actively used. SRS tools, increasingly common in higher education, are valued for promoting more active learning and student engagement through real-time questioning and feedback cycles. Our methodology processes raw audio recordings, employing automatic speech recognition (ASR) and speaker diarization to generate transcripts and to derive several quantitative features. These transcripts serve as input for a series of advanced BERT-based models, which progressively leverage textual content, conversational context, and additional features to classify interventions into predefined COPUS-based pedagogical categories. A key contribution of our work is the empirical demonstration that hybrid models, fusing textual content with paralinguistic cues, can improve the performance of highly-optimized text-only models. This fusion provides a complementary source of information, and we couple our analysis with Explainable AI (xAI) techniques to clarify the distinct influence of these new features. Crucially, Explainable Artificial Intelligence (xAI) techniques, particularly SHAP, are applied to elucidate how these models fuse textual cues with paralinguistic elements (e.g., speaker ratios, silence patterns), revealing their differential contributions to the classification process. Ultimately, this study demonstrates a viable approach for creating accurate, interpretable AI systems that offer teachers feedback on their instructional patterns.

INDEX TERMS Artificial intelligence, audio features, explainable AI, NLP, student response systems.

I. INTRODUCTION

Student Response Systems (SRS), often referred to as Audience Response Systems, or simply as clickers, are powerful tools designed to make teaching and learning more interactive and effective [1]. Recent meta-analytic evidence confirms that these tools positively impact learning outcomes across a range of disciplines, particularly when aligned with well-structured pedagogical strategies and formative assessment practices [2]. By promoting active learning and real-time feedback, SRS tools contribute not only to knowledge retention but also to higher-order thinking and collaborative reasoning. Its effectiveness heavily relies on the teacher's skill and how they choose to implement it within their pedagogy [3]. Therefore, learning how to use SRS effectively requires teachers to develop a range of new skills and adapt their practices.

Overcoming internal factors, particularly teachers' beliefs and attitudes, appears to be more critical for fostering meaningful pedagogical change than simply acquiring

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

technical or procedural skills. This aligns with the principles of some pedagogies, like Technology-Enhanced Formative Assessment (TEFA) [4], which emphasize the importance of developing pedagogical strategies around formative questioning, discussion facilitation, and reflection. Sustained professional development (PD), along with structured support and opportunities for reflection and iterative practice, is essential. In this context, automated tools like the one presented in this study can provide valuable support. This work centers on a particular case: analyzing the nature and function of teacher interventions during classroom activities involving the use of Student Response Systems (SRS) by university instructors.

Extracting granular insights from complex classroom discourse presents significant technical challenges, as traditional text-based analysis often falls short in capturing the full richness of human interaction. To address this, our methodology leverages a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) [5] architecture to classify teacher interventions based on transcribed verbal content. A central objective of this work is to systematically investigate whether this fusion of verbal and paralinguistic features provides a quantifiable performance benefit over optimized text-only classifiers. We explore multiple fusion architectures, including early-fusion neural networks and late-fusion machine learning approaches, to test this hypothesis. A second key contribution lies in the application of Explainable Artificial Intelligence (xAI) to the most effective models identified, allowing us to clarify how both textual and paralinguistic features influence classification decisions.

The overall methodology followed in this study is depicted in Figure 1. The pipeline commences with audio recordings, which are processed through two parallel streams. The primary stream focuses on textual analysis, beginning with transcription via Whisper and subsequent manual correction. Concurrently, the second stream applies a diarization process to the audio to determine ''who spoke when''. The output of this diarization informs two subsequent steps: it is used to segment the corrected transcription by speaker, and it serves as the basis for extracting paralinguistic features. The speaker-segmented text is then filtered to isolate specific teacher interventions, which are labeled according to the COPUS framework [6]. Finally, these COPUS labels and the extracted paralinguistic features are integrated through data fusion. This unified dataset is then used to train multiple BERT-based classification models, ultimately generating the final results which incorporate explainable AI (xAI) components.

To evaluate the effectiveness and implications of our proposed approach, this study is structured around the following research questions:

- RQ1: To what extent can a BERT-based model accurately classify teacher interventions during SRS-supported classroom activities using only text from transcriptions?
- RQ2: How does the addition of paralinguistic features affect the performance of a BERT-based model for classifying teacher interventions from transcriptions?
- RQ3: To what extent can explainability techniques like xAI clarify the decision process of the developed models based on textual tokens and paralinguistic features?

The rest of the paper is organized as follows: Section II situates our work within the existing literature, reviewing prior research on Student Response Systems, speech-based AI in educational contexts, and Explainable Artificial Intelligence. Section III outlines the methodological foundations of our study, detailing the educational setting, the data collection process, and the construction of a labeled dataset that integrates both textual and paralinguistic features. Section IV introduces the classification models and explains the rationale behind their architectural design and evaluation strategy. In Section V, we present the results of our experiments, examining how different input configurations affect model performance and interpretability, and addressing the three research questions that guide this work. Section VI reflects on the broader implications of our findings for research and educational practice, while also acknowledging the study's limitations and identifying opportunities for future work. Finally, Section VII concludes with a summary of our main contributions.

II. RELATED WORK

A. STUDENT RESPONSE SYSTEMS

Student Response Systems (SRS) have been widely recognized for their pedagogical value, particularly in promoting student engagement, fostering active participation, and supporting formative assessment. Recent systematic literature reviews have confirmed the growing relevance of Student Response Systems in higher education. One meta-study [7] identified 77 papers across disciplines highlighting benefits such as enhanced engagement, motivation, and feedback, while also noting recurring technical and pedagogical challenges, especially when SRSs are not integrated within robust instructional designs.

Teachers engaging in the integration of SRSs into their instructional practice frequently encounter a range of challenges, including technical limitations, time constraints, difficulties in designing pedagogically effective questions, and tensions in aligning SRS use with existing curricular structures [8]. Addressing these obstacles requires more than general guidance. It calls for timely, specific, and context-aware feedback that supports teachers in refining their implementation strategies and advancing toward pedagogical proficiency.

Effective SRS-based pedagogies, particularly Technology-Enhanced Formative Assessment (TEFA) [4], are centered around a question cycle that involves posing questions, students responding (often via SRS), and crucially, whole-class discussion based on the distribution of responses. This discussion phase is vital for exploring diverse thinking, confronting ideas, clarifying misconceptions, and developing deeper understanding [9]. Teachers need to develop the

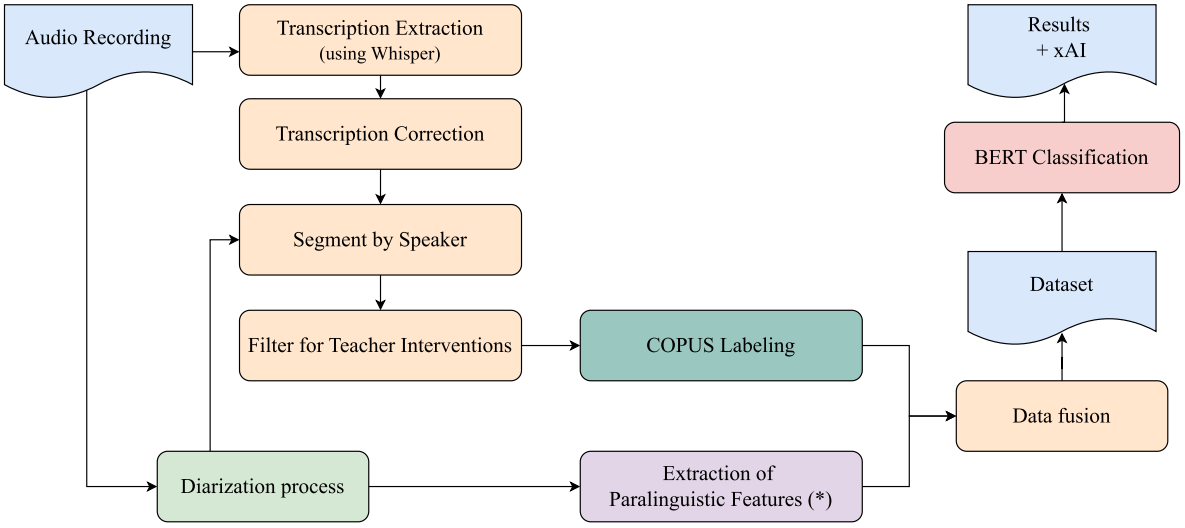


FIGURE 1. High-level overview of the experimental pipeline, from data collection and processing to model evaluation and xAI.

skill of orchestrating and facilitating productive classroom discourse. However, teachers often struggle with allocating sufficient class time for this essential discussion, sometimes feeling pressure to cover content quickly [10]. Feedback on the actual time spent engaging students in post-response discussion provides teachers with objective data to reflect on whether they are dedicating enough focus to this critical pedagogical phase and help them manage pace effectively.

Furthermore, managing student behavior and ensuring that all students are engaged in meaningful participation during interactive activities can be challenging, particularly in larger classes. Time can be lost during transitions between activities or due to student distractions [11]. While SRS can help increase participation and change the classroom atmosphere, teachers still need to be skilled at orchestrating the discourse and managing interactions. Feedback that highlights the proportion of time spent on classroom management issues (as opposed to instructional discussion or activity) can reveal specific areas where a teacher might need to refine their strategies for student engagement or behavioral norms during SRS use.

Additionally, some platforms like Kahoot! or Wooclap, utilizes gamification elements (points, leaderboards, visual triggers) to boost student engagement, motivation, and participation [12]. Gamified systems are designed to actively create a more energetic and engaging atmosphere. Teachers interested in leveraging the motivational power of gamification need to understand how to integrate these elements effectively. Objective feedback on the use of gamification elements during SRS activities [13] can help teachers reflect on their adoption patterns and assess whether these strategies are effectively supporting student motivation and engagement.

B. AI-DRIVEN SPEECH ANALYSIS

The analysis of classroom discourse fundamentally relies on converting spoken language into accurate, speaker-attributed text. Automatic Speech Recognition (ASR) in authentic educational settings presents significant challenges due to high ambient background noise, overlapping speech, and the acoustic variability of speakers, particularly children [14], [15]. OpenAI's Whisper model marks a significant advancement in ASR, offering robust performance even in noisy environments by learning to condition its transcription on background noise, trained on an extensive and diverse dataset [16], [17], [18].

Beyond mere transcription, knowing 'who spoke when' is critical for meaningful discourse analysis. Speaker diarization, which attributes audio segments to specific speakers, faces similar challenges in classroom environments, where standard models exhibit substantially higher error rates compared to clean audio [19]. A pivotal innovation in recent years has been the tight integration of ASR and diarization, mitigating the cascading nature of errors prevalent in sequential pipelines [14], [20].

Once a reliable, speaker-attributed transcript is obtained, the subsequent challenge is to automatically derive pedagogical meaning from the text. Modern Natural Language Processing (NLP) models, particularly those based on the Transformer architecture like Bidirectional Encoder Representations from Transformers (BERT) [5], have proven indispensable for this task, setting new benchmarks in language understanding [21].

Prior work has demonstrated the efficacy of BERT-based models in classifying teacher and student utterances within educational contexts, often employing a contextual classification approach where the input to the model includes conversational turns or sequences to better capture relational dynamics. For instance, [22] utilized BERT-based models to identify focusing questions and uptake of student ideas, showing that automated feedback on talk time increased student talk ratios in tutoring sessions. Similarly, [23] applied fine-tuned BERT models to analyze teacher attention to student ideas in simulated classroom discussions, demonstrating

their ability to accurately score teacher performance and provide justification for those scores. Furthermore, recent work by Wang and Chen [24] has directly investigated the relationship between AI model accuracy and educational outcomes in the context of classroom dialogue analysis for teacher professional development, providing crucial insights into the practical utility of such systems.

However, while these approaches effectively classify various interventions based on textual content, differentiating certain intervention types based solely on linguistic cues can remain complex. Our methodology addresses this limitation by incorporating additional contextual information derived from audio features [25], [26]. This integration of paralinguistic features with transcriptions provides a richer context, which can significantly enhance the precision of classifying teacher interventions, particularly where textual content alone might be ambiguous, as we show in this work.

C. EXPLAINABLE ARTIFICIAL INTELLIGENCE (xAI)

The application of xAI in educational settings is a growing area of research, with various studies exploring its utility in understanding complex learning processes and AI system behaviors. For instance, the xAI-ED framework [27] highlights key aspects for the design of educational AI tools, directly addressing concerns such as fairness, accountability, transparency, and ethics (FATE) within AI interventions. Beyond general frameworks, xAI techniques have been applied to interpret specific AI models in educational contexts. Authors in [28], for example, leveraged xAI to understand the quality of online tutoring sessions by analyzing feature importance in audio classification models. Furthermore, xAI has been employed to enhance intelligent tutoring systems, such as the ExAIT system by Ogata et al. [29], which combines AI-generated explanations with learner self-explanation to foster mutual understanding. In the context of large language models, authors in [24] investigated the use of BERT and Llama for analyzing classroom dialogue and their impact on teacher learning, implicitly touching upon the interpretability of such models' outputs.

Several techniques have been developed to help interpret how machine learning models make decisions [30]. For deep learning models used in Natural Language Processing (NLP), two widely used methods are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP [31], [32] assigns an importance value to each input feature, helping to explain both individual predictions and overall model behavior. LIME [33], on the other hand, builds simple local models around each prediction to highlight the most influential words.

Unlike most prior work, which applies explainability techniques to purely textual inputs, our challenge lies in using these methods, particularly SHAP, to interpret models that combine textual and paralinguistic features. This requires analyzing the joint contribution of heterogeneous inputs within a single fused architecture.

III. METHOD

A. CONTEXT

This study was developed within the framework of a specific educational innovation project. We initially collected 113 class recordings encompassing a variety of pedagogical methodologies, including traditional lectures, collaborative group work, and activities supported by Student Response Systems (SRS). For the purposes of this research, our focus was exclusively on classes that incorporated the use of an Student Response System, particularly Wooclap.[1] From this subset, we specifically selected recordings in which Wooclap was used as a distinct and self-contained activity focused on reviewing concepts from a previously covered content block, resulting in 10 final audios. This criterion was essential, as many other recordings featured more sporadic SRS use intermingled with other teaching formats. This deliberate selection allowed us to construct a homogeneous and representative dataset centered on a prevalent use case of the tool.

We conducted our study using audio recordings representing different teaching contexts, obtained from three courses: Computer Networks, Computing Foundations, and Microbiology I. The first two courses are part of a bachelor's degree program in Computer Science, while the third is from a Veterinary degree. Both degree programs are taught in Spanish. To ensure a comprehensive dataset, we engaged two female and two male teachers to record their respective classes. The teachers in our sample averaged 20.5 years of experience. The cumulative duration of these audio files is 3 hours and 19 minutes, with each file ranging from 10 minutes to 30 minutes in duration. All the data were collected with the approval of the teachers, students, and the Institutional Review Board.[2]

Furthermore, each audio session was augmented with supplementary contextual metadata, encompassing details such as the specific course, audio file duration, student cohort size, recording instrumentation, and instructor identification.

For this analysis, we adapted categories from the established COPUS system [6], a standardized protocol for labeling teaching activities and interventions. The adoption of a recognized, standard classification mechanism was crucial to enhance the interpretability and comparability of our findings. These adapted labels, assigned to individual teacher interventions, are detailed in Table 1.

B. CODING AND DATASET

Upon receiving the recordings, transcriptions were extracted using Whisper [16], a model recognized for its robust transcription capabilities and accurate textual conversion. Subsequently, the textual data was segmented, leveraging outputs from PyAnnote [34], a specialized tool for speaker diarization that identifies and differentiates individual speakers within an audio stream, and spaCy [35], which was

[1] https://wooclap.com
[2] Code 2024/341.

TABLE 1. Description of teacher intervention labels.

Label	Description	Example (Translated)
MG	Moving through class guiding ongoing student work during active learning task.	"*Alright, silence!*"
FUp	Follow-up/feedback on clicker question or activity to entire class.	"*Look, 192 cannot be assigned by a DHCP because it's not a valid address.*"
CQ	Asking a clicker question.	"*What is the address assigned to PPP0 on router 1?*"
AnQ	Listening to and answering student questions with entire class listening.	"*You told it to release, it releases, but you have to face the consequences.*"
Ga	Referencing student performance metrics (e.g., scores) or the challenge level of a question. (A non-COPUS code).	"*Come on, let's continue! This one is easy!*"

utilized for linguistic analysis and further segmentation. This integrated approach, combining Whisper's corrected transcription with PyAnnote's precise speaker attribution, was instrumental in enriching the dataset by allowing for not only the content transcription but also the accurate assignment of speech to specific individuals through their respective timestamps. Following this comprehensive segmentation of interventions, these were then labeled according to the coding scheme detailed in Section III-A.

Two independent coders annotated the dataset with reference to the audio recordings. To measure annotation agreement among coders, we used the Cohen's Kappa, achieving an initial value of 0.6995. The initial discrepancies were mainly due to differing interpretations of the subtle distinctions between labels. Several contentious interventions were used as reference cases to foster discussion, allowing annotators to clarify their reasoning and progressively converge toward a shared understanding of the labeling criteria. This discussion led to the generation of more straight labeling guidelines, which significantly improved concordance to a Cohen's Kappa of 0.9706.

Interventions deemed contextually insufficient, such as monosyllabic utterances (e.g., 'yes', 'no'), were entirely excluded from the dataset. These brief instances were considered to lack the necessary contextual information for meaningful classification within the predefined categories. The remaining 3% of interventions lacking consensus encompassed a variety of ambiguous cases, mostly instances that could be dually classified. Furthermore, certain lengthy interventions presented challenges due to their multifaceted nature, allowing for plausible categorization into different schemes depending on the emphasis placed on various parts of the utterance.

Table 2 provides a detailed overview of the dataset's distribution, presenting the raw counts of each intervention label per teacher. For clarity and anonymization, the four participating teachers are designated as T1, T2, T3, and T4. This table provides a crucial representation of our dataset, from which the frequency of different teacher interventions within each instructor's unique teaching context can be understood. A clear imbalance is observable across categories, as certain intervention types are inherently more common in classroom discourse. For instance, 'FUp' (Follow-up) interventions are far more frequent than 'AnQ' (Answering Questions), a pattern that reflects authentic teaching dynamics. This class imbalance is therefore not a limitation of our dataset but rather a representative feature of the pedagogical context under study. Consequently, the dataset's sufficiency is not based on its size, but on its ability to serve as a challenging and realistic testbed for our central hypothesis.

TABLE 2. Count of interventions by label teacher (number of cases).

Label \ Teacher	T1	T2	T3	T4	Total
AN	63	8	2	1	74
CQ	37	55	61	161	314
FU	142	221	179	44	586
GA	81	33	68	52	234
MG	94	137	44	191	466
Total	417	454	354	449	1674

C. EXTRACTED PARALINGUISTIC FEATURES

Following the transcription and diarization stages, the data from each class session is consolidated into a structured format. For each identified teacher intervention, a record is created that unifies the text transcript, its conversational context, and a set of 18 key paralinguistic features. After processing and filtering, this consolidation resulted in a final dataset comprising 1674 distinct teacher interventions, each forming a single record, exemplified in Table 3. The final data structure for each intervention is as follows:

- **Textual Data:** The transcribed text of the current intervention (`text`) and the text of the teacher's immediately preceding intervention (`previous_text`).
- **Paralinguistic Features:** A vector of numerical features calculated over the duration of the intervention, detailed in Figure 2. Details about features meaning are detailed in Table 4.
- **Metadata and Labels:** Timestamps, speaker identification, and the pedagogical label assigned by human annotators (`label`) as defined in Section III-A.

The selection of these eighteen features was guided by their potential to serve as quantitative proxies for key pedagogical dynamics in a classroom environment. The features related to participant's ratios were chosen to directly measure conversational dominance and turn-taking, which are fundamental to distinguishing teacher-led exposition from student-centered interaction. Metrics related to silence were included to capture the rhythm and pacing of the discourse;

TABLE 3. An anonymized example of a single consolidated data record. For visualization purposes, some features are omitted.

previous_text	text	PSR_prof	PSR_Oth	SR	OVR	MR	WPS	SBI	label
(in Spanish) Vale, pues vamos a ver esta.	(in Spanish) ¿Cuál es la dirección que se le asigna a PPPO en el router uno?	0.85	0.05	0.10	0.02	0.0	3.1	1.5	CQ
(Translated) Okay, let's see this one.	(Translated) What is the address assigned to PPPO on router one?	0.85	0.05	0.10	0.02	0.0	3.1	1.5	CQ

TABLE 4. Description of the calculated paralinguistic features.

Type	Acronym	Feature	Description and Calculation
Per Role	PSR	Participant Speaking Ratio	Proportion of speech time for a specific role (e.g., teacher) over the total duration of the segment.
Per Role	PSU	Participant Speaking Utterances	Absolute number of interventions (utterances) made by a role during the segment.
Per Role	PSUR	Participant Speaking Utterances Ratio	Proportion of the number of interventions from one role over the total number of interventions in the segment.
Per Role	APSUD	Average Participant Speaking Utterance Duration	Average duration of a role's interventions. Measures if a role tends to make long or short interventions.
Global	ALD	Average Lapse Duration	Average duration of silences that occur between interventions of different roles, i.e., a transitional silence.
Global	SR	Silence Ratio	Proportion of total silence time (no participant speaking) over the total duration of the segment.
Global	PEQ	Participation Equality	Indicator that measures the balance in participation among roles. A value close to 1 indicates a perfectly equitable distribution of speech.
Global	TTC	Turn Taking Count	Total number of turn changes between roles (e.g., from teacher to student or vice versa) during the segment.
Global	VSUR	Very Short Utterances Ratio	Proportion of very short interventions (defined as those less than 2 seconds) over the total number of interventions. May indicate quick responses or interjections.
Global	OVR	Overlapping Rate	Proportion of time during which two or more participants speak simultaneously (overlapping) over the total duration of the segment.
Global	OVUR	Overlapping Utterances Rate	Proportion of the number of interventions that contain some overlap over the total number of interventions.
Global	MR	Mumble Ratio	Proportion of voice activity that the diarization system could not confidently attribute to a specific speaker. May indicate mumbles, background noise, or unintelligible speech.
Global	WPS	Words Per Second	Words Per Second for the current speaker.
Global	SBI	Silence Between Interventions	The duration of silence preceding the current intervention.

for example, a long silence preceding an intervention (SBI) might differentiate a thoughtful response from a rapid-fire question. The Overlap Ratio (OVR) and Mumble Ratio (MR) provide insights into the level of classroom interactivity, helping to distinguish a single speaker's clear utterance from the conversational buzz of group work or a lively, multi-speaker discussion. Finally, Words Per Second (WPS) was included as an indicator of speaking pace, which can correlate with the pedagogical function of an utterance, such as giving quick instructions versus providing a slower, more deliberate explanation.

To clarify how these paralinguistic features were computed, Figure 2 provides a schematic of the temporal windows used for their extraction.

Two specific features are derived from their unique segments: Words Per Second (WPS) is computed exclusively from the 'current intervention' period [t3, t4], and features

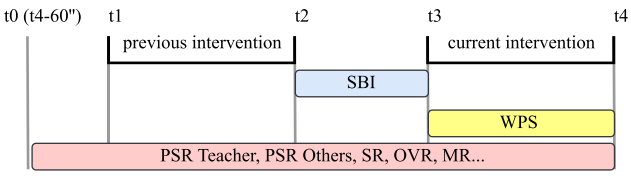


FIGURE 2. Schematic illustrating the temporal segmentation for paralinguistic feature extraction. The diagram delineates the "previous intervention" [t1, t2], the "silent break interval" (SBI) [t2, t3], and the "current intervention" [t3, t4].

related to the 'silent break interval' (SBI) are derived from its window [t2, t3]. All the other features are calculated for the time windows [t0, t4].

This integrated dataset, structured to unify textual transcripts, conversational context, and a rich array of

paralinguistic features, thus serves as the foundational input for a series of BERT-based models.

IV. MODELS AND PROCESSING

This study is founded on the hypothesis that the pedagogical role of a teacher's speech cannot be fully understood from an isolated utterance. To test this, we move beyond simple text-based analysis and explore the predictive power of two additional information sources: conversational context and paralinguistic features. We developed a series of DeBERTa-based classification models designed to systematically probe this assumption. The following sections detail these models, which progress from a text-only baseline to more sophisticated architectures that leverage preceding interventions and paralinguistic cues.

A. EXPERIMENTAL SETUP

All experiments were conducted on a workstation equipped with an Intel Core i7-10700F CPU, 32GB of RAM, and two NVIDIA GPUs (an RTX 3080 with 10GB VRAM and an RTX 2070 with 8GB VRAM). A fixed random seed (42) was used throughout all stages of data splitting, model training, and evaluation to ensure the full reproducibility of our results.

B. DATASET SPLIT

For hyperparameter optimization, the models were trained using 5-fold stratified cross-validation. The optimal hyperparameter combination for each model family was determined by selecting the configuration that yielded the highest average macro-averaged F1 score across these cross-validation folds. To provide a more robust and generalizable estimate of our final models' performance, we subsequently performed a 10-fold stratified cross-validation on the entire dataset, using the best-found hyperparameters for each model.

It is important to note that no artificial balancing (e.g., oversampling or undersampling) or data augmentation techniques were applied to the training set. The observed class imbalance is an authentic feature of the pedagogical context, where certain interventions are naturally more frequent than others. We made a deliberate decision to train the models on this natural distribution to ensure their performance reflects the challenges of a real-world classroom environment.

C. CLASSIFICATION MODELS

We developed and evaluated three families of models, denoted as Textual, Early-Fusion and Late-Fusion, all built upon a pre-trained DeBERTa architecture. The choice of 'mdeberta-v3-base' was motivated by several key factors: its state-of-the-art performance on various NLP benchmarks, often outperforming previous BERT iterations like RoBERTa due to architectural enhancements such as disentangled attention; its crucial multilingual capability, essential for processing Spanish classroom discourse; and its robust adaptation to real-world, automatically transcribed data. This ''base'' version also strikes an optimal balance between high performance and computational efficiency. These models are designed to incrementally leverage more context to classify a teacher's utterance according to our predefined labels. A visualization of all the developed models is shown in Figure 3.

1) TEXTUAL MODELS: CURRENT AND CONTEXTUAL INTERVENTIONS

Textual models serve as our baseline. The first one is designed to classify an intervention based solely on its own textual content. The input to this model is the tokenized text from the 'text' field. The objective is to establish a performance benchmark using only the semantic information contained within a single intervention, without any broader context. After this model, we extended the textual context by incorporating immediate conversational history. The input to this model is the concatenation of the previous teacher intervention ('previous_text') and the current intervention ('text'), separated by a special '[SEP]' token. This approach tests the hypothesis that the model can make more accurate predictions by understanding the sequential relationship between consecutive teacher interventions.

2) EARLY-FUSION: DEEP LEARNING WITH DIARIZATION FEATURES

Early-fusion represents our most advanced approach, creating a multimodal architecture that fuses textual information with quantitative paralinguistic features. The goal is to determine if these acoustic-derived cues about the classroom environment can further enhance classification accuracy. All these variants use the same contextual text input as in the previous model but also incorporate the 18-dimensional vector of paralinguistic features.

The textual input is processed by a DeBERTa encoder, which, in its standard configuration, outputs embeddings with a dimension of 768 features. Given that there is no single standard method for fusing textual and numerical data, and to balance the influence of the high-dimensional text embeddings (768 features) with the relatively low-dimensional paralinguistic features (18 features), we experimented with three distinct architectural variants. These variants aimed to identify the most effective fusion strategy by adjusting the dimensionality of either the text embeddings or the paralinguistic features. Dimensionality adjustments (both reductions and expansions) were performed using a small Multi-Layer Perceptron (MLP) [36], as shown in Figure 3. The paralingustic features and the BERT embedding go through to different projection layers before being concatenated, resulting in double the sizes shown in the projection boxes. Meanwhile, the 'Naive Concatenation' just concatenates both vectors with their current dimensions, resulting in a (1, 786) vector for classification. An MLP is a suitable choice for this task due to its ability to learn non-linear transformations between input and output layers, allowing it to effectively project features into different dimensional spaces while capturing complex

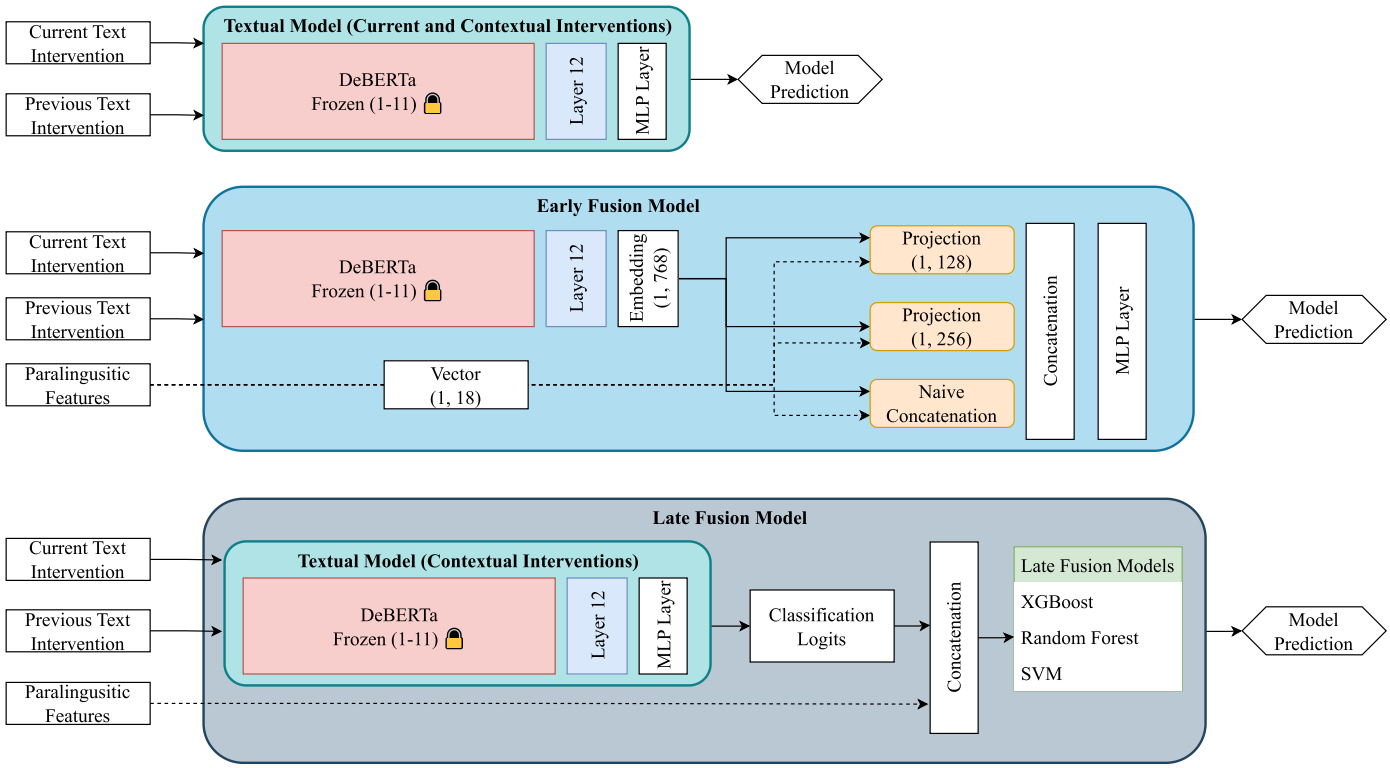


FIGURE 3. Diagram summarizing the tested model's architectures.

relationships, which is crucial for integrating diverse data modalities [37].

The hyperparameters used to train all the DeBERTa-based models, which includes these early-fusion and textual models are detailed in Table 5. A critical methodological decision for all DeBERTa-based models was the freezing of the first 11 encoder layers. This aggressive transfer learning strategy was found to be essential for mitigating the rapid overfitting observed during initial experiments. This approach preserves the model's powerful, generalized linguistic knowledge (contained in the frozen layers) while allowing only the final, task-specific layer to adapt to our specialized dataset. To further control for overfitting, this layer-freezing strategy was complemented by explicit regularization techniques, namely Dropout and L2 Weight Decay.

All models were trained using the AdamW optimizer and a cross-entropy loss function, which are standard choices for multi-class classification tasks with transformer-based models. The results of these experiments are presented in the following section.

3) LATE-FUSION HYBRID MODELS

As a final investigative branch, we explore an alternative late-fusion architecture. This approach tests whether paralinguistic cues are more effective at refining a classification after the text model has fully processed the semantic content, a contrast to the early-fusion models.

TABLE 5. Hyperparameter and training configuration for DeBERTa models (textual and early-fusion).

Parameter	Value / Search Space
Model & Tokenizer	
Base Model Name	microsoft/mdeberta-v3-base
Max Token Length	256
Frozen Encoder Layers	11
Grid Search Space	
Learning Rate	{1e-5, 5e-6}
Dropout Rate (in MLP)	{0.2, 0.4}
Weight Decay	{0.01, 0.1}
Training Configuration	
Optimizer	AdamW
Loss Function	CrossEntropyLoss (with class weights)
Max Epochs	100
Early Stopping	Patience of 10 (on validation F1-score)
Random Seed	42

The methodology for this approach consists of two stages. First, the best-performing Contextual model is used to process the textual input (text + previous text) and generate the 5-dimensional logit vector for each intervention. Second, this logit vector is concatenated with the 18-dimensional paralinguistic feature vector, resulting in a 23-dimensional feature set, as shown in Figure 3. This combined vector is then used to train and evaluate several classical machine learning classifiers known for their robustness on structured tabular data. The models selected for this task are:

- **Support Vector Machine (SVM):** A kernel-based model effective for classification.

TABLE 6. Hyperparameter grid search space for late-fusion models. Best hyperparameters are in bold.

Model	Hyperparameter	Search Space
Random Forest	n_estimators	{**100**, 200, 300}
	max_depth	{**10**, 20, None}
	min_samples_leaf	{**1**, 2, 4}
XGBoost	n_estimators	{**100**, 200, 300}
	learning_rate	{0.05, 0.1, **0.2**}
	max_depth	{3, **5**, 7}
SVM	C	{0.1, 1, **10**}
	gamma	{'scale', 'auto', **0.1**, 1}
	kernel	{**'rbf'**, 'poly'}

- **Random Forest (RF):** An ensemble method based on decision trees that is robust to overfitting.
- **XGBoost:** A highly efficient and powerful implementation of gradient-boosted decision trees.

This two-stage process leverages the DeBERTa model as a powerful semantic feature extractor, tasking the subsequent, simpler ML models with finding the optimal decision boundary based on the combined semantic and paralinguistic signals.

The distinct hyperparameter search space for the late-fusion machine learning models is detailed subsequently in Table 6.

V. RESULTS

Our investigation into the classification of teacher interventions was guided by three primary research questions. This section presents the empirical findings, beginning with the evaluation of models based exclusively on textual features, and subsequently incorporating contextual information derived from the discourse features.

A. RQ1: TO WHAT EXTENT CAN A BERT-BASED MODEL ACCURATELY CLASSIFY TEACHER INTERVENTIONS DURING SRS-SUPPORTED CLASSROOM ACTIVITIES USING ONLY TEXT FROM TRANSCRIPTIONS?

To establish a foundational understanding of classification performance, we first evaluated the two text-only models detailed in Section IV-C1. The performance metrics for all models are presented in Table 7.

Our initial baseline, the Text Model (Current) model, classifies interventions using only the text of the current utterance. This model achieved a Macro-F1 score of 0.520 ± 0.222. While demonstrating some classification capability, the extremely high standard deviation indicates significant instability and high variance across the cross-validation folds.

We then evaluated the Text Model (Contextual), which incorporates the immediate conversational history by concatenating the previous intervention's text. This model yielded a notable improvement as shown in Table 7, achieving a Macro-F1 score of 0.709 ± 0.035. This result represents a substantial increase in mean performance over the baseline, affirming our hypothesis (detailed in Section IV-C1) that sequential context is critical for this task.

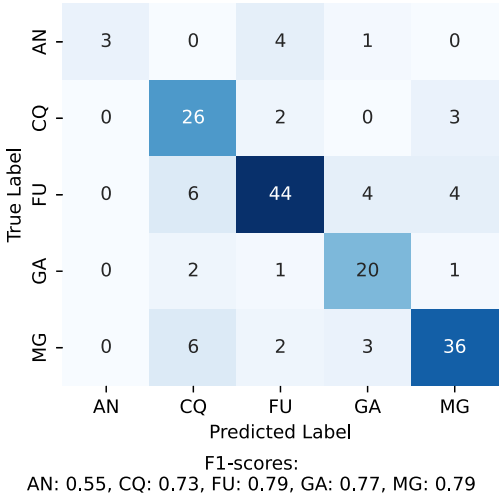


FIGURE 4. Confusion Matrix for Text Model (Contextual).

Furthermore, the standard deviation dropped from 0.222 to 0.035, signifying a dramatic increase in model stability and robustness. This context-aware model is more generalizable and serves as the strong, stable baseline for evaluating the advanced fusion models in RQ2.

To illustrate the model's typical error patterns, Figure 4, presents a representative confusion matrix from a single validation fold. While this view is not aggregated, it highlights the consistent challenges the mode faced during cross-validation. The matrix shots that while most categories are classified with high fidelity (all F1-scores ≥ 0.73), the model struggles significantly with the label "AN (Answering Question)" class. The primary source of erros is the model frequenly misclassifying 'An' interventions as 'FUp'. This specific ambiguity appears to be the model's key limitation.

We therefore proceed to evaluate the utility of incorporating richer, non-textual contextual cues to enhance overall performance.

B. RQ2: HOW DOES THE ADDITION OF PARALINGUISTIC FEATURES AFFECT THE PERFORMANCE OF A BERT-BASED MODEL FOR CLASSIFYING TEACHER INTERVENTIONS FROM TRANSCRIPTIONS?

Having established a strong text-only baseline with the "Text Model (Contextual)", our second research question investigates whether the integration of paralinguistic features can provide a statistically significant performance benefit. To answer this, we systematically evaluated the two distinct fusion architectures detailed in Section IV: Early-Fusion and Late-Fusion. All comparative results are presented in Table 7, with the Macro-F1 score distributions visualized in the boxplot in Figure 5.

First, we evaluated the Early-Fusion variants. These models, which fuse features at the input level of the DeBERTa architecture, failed to yield a performance improvement. All

TABLE 7. Comprehensive 10-fold cross-validation performance metrics (Mean ± Std. Dev.)

Model	Macro-F1	Precision (Macro)	Recall (Macro)	AUC (Macro)
Text Model (Current)	0.520 ± 0.222	0.549 ± 0.257	0.538 ± 0.180	0.825 ± 0.129
Text Model (Contextual)	0.709 ± 0.035	0.721 ± 0.039	0.708 ± 0.036	0.885 ± 0.022
Early-Fusion (Naive)	0.676 ± 0.039	0.686 ± 0.043	0.686 ± 0.033	0.868 ± 0.021
Early-Fusion (256)	0.700 ± 0.030	0.709 ± 0.041	0.702 ± 0.025	0.868 ± 0.026
Early-Fusion (128)	0.693 ± 0.034	0.701 ± 0.041	0.699 ± 0.030	0.877 ± 0.026
Late-Fusion (RF)	0.737 ± 0.036	0.747 ± 0.034	0.736 ± 0.037	0.902 ± 0.021
Late-Fusion (SVM)	0.714 ± 0.034	0.721 ± 0.031	0.714 ± 0.034	0.885 ± 0.021
Late-Fusion (XGB)	0.735 ± 0.046	0.744 ± 0.045	0.737 ± 0.044	0.898 ± 0.023

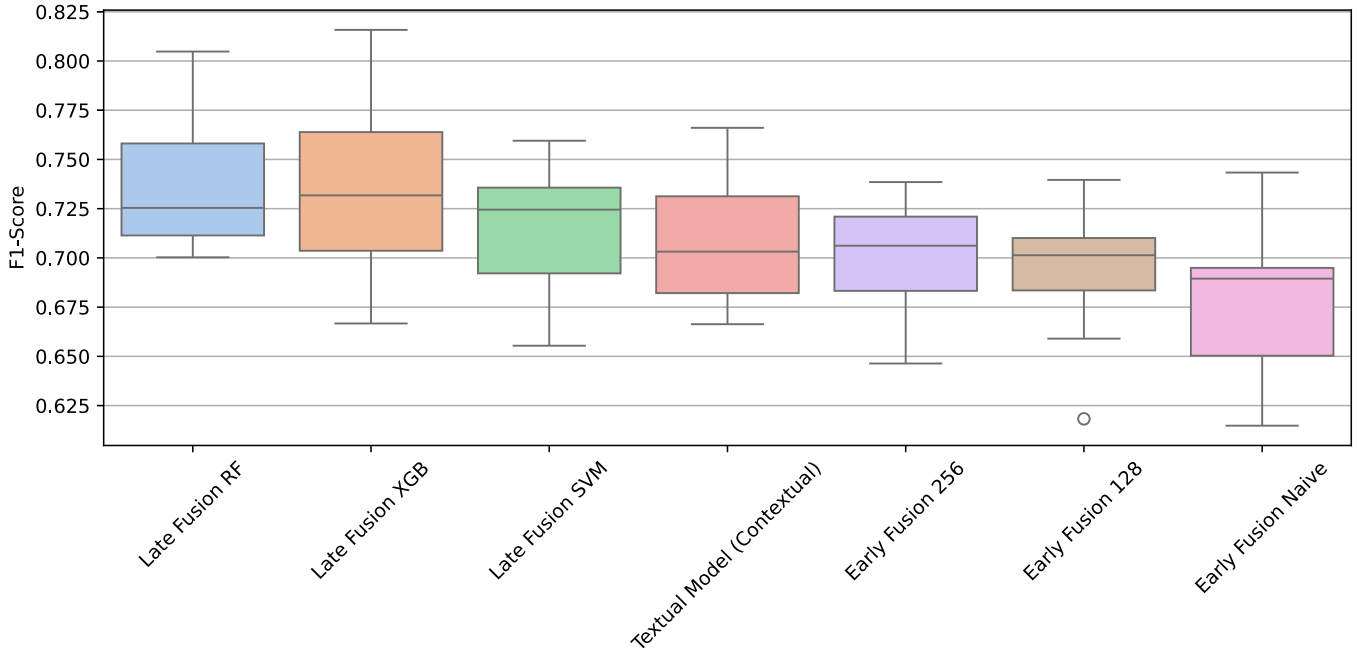


FIGURE 5. Macro-F1 score distribution from the 10-fold stratified cross-validation. This plot visualizes the median (line), interquartile range (box), and variance (whiskers) of each model's performance.

three variants performed at or slightly below the text-only baseline, as shown in Table 7.

We hypothesize two primary reasons for this failure. First, the aggressive 11-layer freezing (as noted in Section IV-C2), while essential for mitigating overfitting, likely prevented the model's single trainable layer from adequately learning the complex, non-linear interactions between the two distinct modalities.

Second, this challenge is compounded by the extreme dimensional imbalance between the text embedding (768 dimensions) and the paralinguistic feature vector (18 dimensions). For a deep learning model, the text signal is information-rich and dimensionally dominant, making it difficult for the architecture to assign sufficient weight to the comparatively small paralinguistic vector. The architectural variants designed to mitigate this imbalance (Models 128 and 256) also failed to improve performance, suggesting that the process of compressing the 768-dimensional text embedding resulted in a critical loss of semantic information that outweighed any potential benefit from the fused features.

TABLE 8. Post-Hoc pairwise comparisons (Bonferroni) for macro-f1 scores against the "Text Model (Contextual)" baseline.

Model Name	Mean Diff.	p-value (Bonferroni)
Late-Fusion (RF)	0.028	0.006
Late-Fusion (XGB)	0.026	0.005
Late-Fusion (SVM)	0.006	1.000

Then, we evaluated the Late-Fusion variants. This two-stage approach, which uses the "Text Model (Contextual)" as a semantic feature extractor and combines its logits with the paralinguistic features, proved to be the most effective architecture.

The "Late-Fusion (RF)" model, using a Random Forest classifier, emerged as the best-performing model in the entire study, as shown in Table 7, achieving a Macro-F1 score of 0.737. The "Late-Fusion (XGB)" model was also highly competitive, achieving 0.735. Both models represent a clear and robust performance gain over the text-only baseline. This superiority of the top late-fusion models is also evident in Figure 5, which shows their higher median F1-scores and stable distributions.

To statistically validate these findings, we conducted a one-way repeated measures ANOVA on the Macro-F1 scores of the four most competitive models: "Text Model (Contextual)", "Late-Fusion (RF)", "Late-Fusion (SVM)", and "Late-Fusion (XGB)". The Greenhouse-Geisser corrected test confirmed a significant overall difference between the models ($F(1.53, 13.80) = 4.96$, $p = 0.031$). We then performed post-hoc pairwise comparisons with Bonferroni correction to identify the specific sources of this difference. The key comparisons against our baseline "Text Model (Contextual)", are detailed in Table 8. The post-hoc analysis provides statistical evidence that both the Late-Fusion (RF) ($p = 0.006$) and Late-Fusion (XGB) ($p = 0.005$) models significantly outperformed the "Text Model (Contextual)" baseline.

To evaluate generalizability, we compiled a hold-out test set comprising 426 interventions from teachers T01 and T03, who were excluded from both the training and validation phases. Crucially, this dataset relied on uncorrected transcripts to simulate a real-world environment. The Late Fusion RF model achieved a Macro-F1 score of 0.707 on this independent dataset, demonstrating its ability to generalize to unseen subjects even without manual transcription correction.

The answer to RQ2 is therefore clear and statistically validated: the addition of paralinguistic features provides a statistically significant enhancement to classification performance, but only when integrated via a robust, two-stage late-fusion architecture. The deeply integrated early-fusion approach was not effective for this task.

C. RQ3: TO WHAT EXTENT CAN EXPLAINABILITY TECHNIQUES LIKE XAI CLARIFY THE DECISION PROCESS OF THE DEVELOPED MODELS BASED ON TEXTUAL TOKENS AND PARALINGUISTIC FEATURES?

Having identified the "Late-Fusion (RF)" model as the most robust architecture in RQ2, we now address RQ3 by applying explainability techniques to understand how it achieved its superior performance. For this, we employed the computationally efficient `shap.TreeExplainer` to analyze the contributions of both its inputs: the DeBERTa-derived logits and the paralinguistic features.

Our first and most critical finding is the global importance of each input type. The DeBERTa-derived logits (the text-based component) account for 64.02% of the model's predictive power. The paralinguistic features account for the remaining 35.98%.

This result demonstrates that the model's decisions are based on a substantial combination of both textual and non-textual information. Given this clear importance split, a comprehensive analysis requires that we investigate both components separately. We will first analyze the lexical cues that inform the textual component, followed by an analysis of the behavioral patterns captured by the most influential paralinguistic features.

1) ANALYSIS OF TEXTUAL COMPONENT

For this first part of the analysis, we applied the SHAP framework to the "Text Model (Contextual)". Specifically, we utilized 'shap.Explainer', which applies a partition-based explainer suitable for transformer models, to analyze the contribution of individual text tokens. While the analysis was conducted for all five categories, we present the findings for the FUp (Follow-Up) and Ga (Gamification) classes. These two categories provide exemplary and highly distinct insights into how the model differentiates these categories based purely on lexical content.

a: FUP (FOLLOW-UP) TEXTUAL ANALYSIS

The 'FUp' category is defined as the teacher providing follow-up or feedback on a clicker question to the entire class. The SHAP analysis, shown in Figure 6, reveals that the model learned to identify this category by associating it with highly specific, technical vocabulary. The figure displays the 25 most influential tokens, a count we selected as it is sufficient to illustrate the dominant lexical patterns and analyze the model's alignment with a human-level, semantic understanding of the category.

The most influential tokens are not generic conversational words, but terms directly related, for example, to the Computer Networks course content. For example, we see tokens such as *gateway*, *ARP*, *broadcast*, *MAC*, *router*, and *red* (network). Other technical tokens like *calent* (likely from "calentamiento" or heating), *barra* (bar/slash), and *bits* are also highly ranked. This provides clear evidence that the model correctly learned that when the teacher is using specific, technical jargon, it is a very strong indicator of a 'FUp' intervention, where they are explaining the details of a problem or its solution.

b: GA (GAMIFICATION/GUIDING) TEXTUAL ANALYSIS

In contrast, the 'Ga' category is defined as referencing student performance or the challenge level of a question, often to manage the activity's pace and motivation. The textual cues for this category, shown in Figure 7, are entirely procedural and motivational, containing almost no technical jargon.

The most influential tokens are directly related to the management of the SRS activity itself. These include *Bien* (Good/Well), *última* (Last), *Respuesta* (Answer), and *vot* (from "votar", to vote). Other high-impact tokens like *gamos* (e.g., from "jugamos", let's play), *paro* (I stop), *pregunta* (question), *vamos* (let's go), and *Cuando* (When) are all

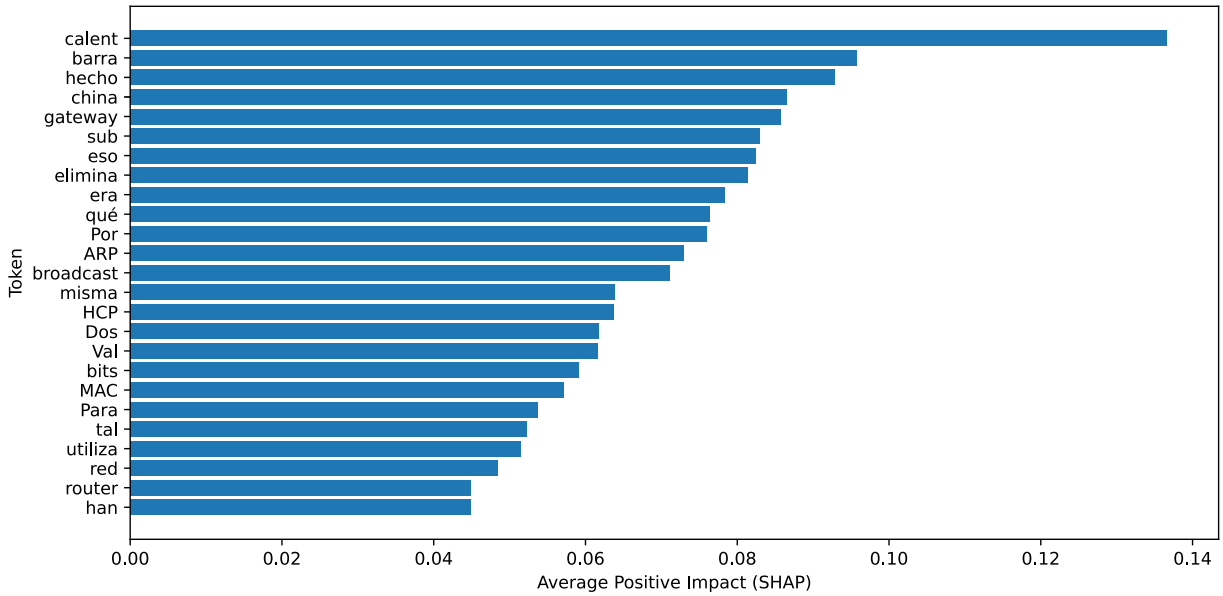


FIGURE 6. Top 25 most influential tokens for the FUp class in the "Text Model (Contextual)".

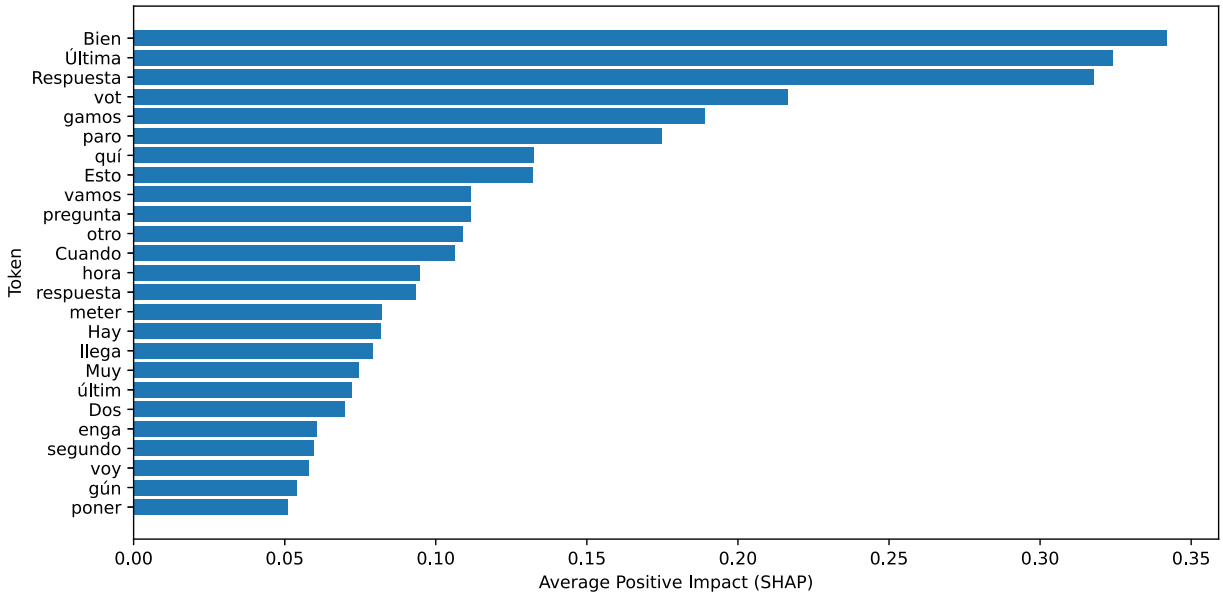


FIGURE 7. Top 25 most influential tokens for the Ga class in the "Text Model (Contextual)".

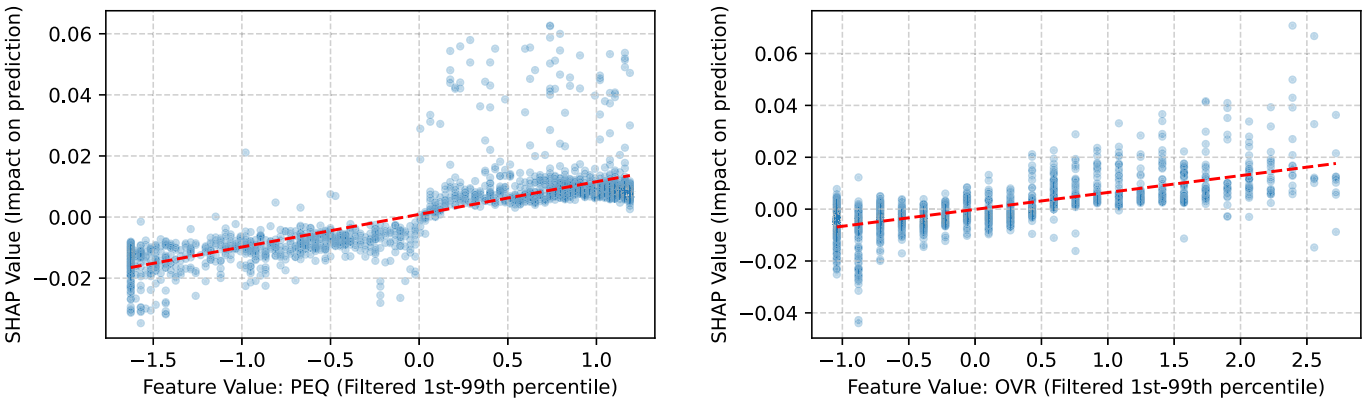
indicative of a teacher managing the flow of the activity, encouraging students, or commenting on the voting process.

2) LATE-FUSION (RF) MODEL ANALYSIS

Having established the textual baseline, we now analyze the best-performing model, the "Late-Fusion (RF)". For this tree-based model, we used the computationally efficient 'shap.TreeExplainer' to quantify how the model fuses the DeBERTa-derived logits with the paralinguistic features to make its final predictions.

To understand how these features contribute, we first analyzed the SHAP dependence plots for the most influential features. We present three exemplary findings in Figure 8, which reveal that the model learned to associate specific, acoustic patterns with the intervention categories.

The first plot, Figure 8a, shows a strong positive correlation (0.588) between the Participation Equality (PEQ) feature and the AN (Answering Question) class. A higher PEQ indicates a more balanced, less-monopolized dialogue. The model correctly learned that an intervention is more likely to be an 'AN' when this "dialogue equality" is high, a behavior that



(a) Class AN vs. PEQ (r^2: 0.588)

(b) Class MG vs. OVR (r^2: 0.480)

FIGURE 8. SHAP dependence plots for three influential paralinguistic features, showing their correlation with the SHAP value (impact on prediction) for specific classes.

perfectly aligns with the back-and-forth nature of a question-and-answer exchange.

Perhaps the most compelling finding is in Figure 8b, which shows a strong positive correlation (0.480) between the Overlapping Rate (OVR) and the MG (Moving/Guiding) class. The 'MG' label corresponds to the teacher moving through the class to guide students as they work. A high 'OVR' indicates significant overlapping speech. The model learned to associate this overlapping speech of a busy classroom—the sound of multiple speakers talking at the same time—as a powerful predictor that the teacher's intervention is one of guidance within that active learning environment, like asking for silence among students.

This finding suggests that the statistically significant performance boost documented in RQ2 is achieved because these features provide tangible and quantifiable performance that offer a source of evidence not captured by the text-only model.

VI. DISCUSSION

A. TECHNICAL IMPLICATIONS: ENHANCING AI MODELS THROUGH FUSED DATA

The enhanced performance observed with the integration of paralinguistic features clearly demonstrates a significant advancement in classification capabilities. This superiority of fused data is not unprecedented; indeed, it has been previously established that the use of multimodal information surpasses unimodal systems, including those based on BERT, in educational contexts [38]. This aligns with the inherent multidimensionality of human conversations, where complete information cannot be solely extracted from textual transcripts. Furthermore, multimodal data not only contributes to improved model performance but also enhances generalization and robustness, as a greater diversity of data is considered for classification, as comprehensively reviewed in the literature [39].

Contextualizing these findings within our work, prior research, such as [40], has compared the performance of classical machine learning techniques and deep learning models like BERT, showing that the latter tend to benefit more clearly from larger datasets. Although our dataset is relatively modest in size, the observed performance gains through the inclusion of paralinguistic features suggest that further improvements could potentially be achieved as more data becomes available.

B. IMPLICATIONS FOR EDUCATIONAL PRACTICE: ANALYSIS OF TEACHING PRACTICES DURING SRS

The utilization of Student Response Systems (SRS) is well established in educational literature and has been analyzed across a range of disciplines. However, their implementation does not always align with pedagogically sound practices. The Technology-Enhanced Formative Assessment (TEFA) framework [4] emphasizes a structured question cycle—comprising questioning, response, and whole-class discussion—supported by classroom response technology. This cycle reflects an ideal pedagogical flow that moves beyond knowledge assessment toward active student engagement and formative dialogue. Crucially, the effective use of SRS requires fostering participation and discussion, rather than simply posing questions and collecting answers [41]. When well implemented, SRS tools create opportunities for students to explain and justify their reasoning while the teacher adopts a facilitative role, encouraging collaborative reflection on possible responses [42].

Our contribution lies in the system's capacity to measure the time teachers dedicate to different phases of this cycle throughout their classes, providing objective data for self-reflection. Furthermore, thanks to the implemented explainability techniques, we can not only identify how teachers allocate their time, but also understand what leads the system to believe that time is being dedicated to those particular sections. This contextualizes the information provided by the model, thereby increasing confidence in the system's decisions and offering insights into why certain parts of the cycle receive more attention than others.

C. PRACTICAL APPLICATIONS IN REAL-WORLD EDUCATIONAL SETTINGS

The analytical framework developed in this study translates into several practical applications designed to enhance pedagogical practice in real-world educational settings. The primary contribution lies in its potential to create a tighter feedback loop for teacher professional development, moving from subjective self-assessment to objective, data-informed reflection.

At the individual teacher level, the system can function as a tool for reflective practice. Following a class session, a teacher could receive an automated report detailing the temporal distribution of their interventions. For instance, a report might reveal that posing 'Clicker Questions' (CQ) consumed 40% of the activity time, while substantive 'Follow-Up' (FUp) discussions constituted only 15%. Such concrete, quantitative evidence provides a specific entry point for reflection, prompting the teacher to consider whether their practice aligns with their pedagogical intentions. This process operationalizes the concept of the ''reflective practitioner'' by providing the evidentiary basis needed for critical self-assessment and targeted instructional adjustment [43].

Within professional development, the framework can facilitate deeper, evidence-based dialogue among peers or in training workshops. For example, a group of instructors could use the analytics to compare their implementation patterns of the Technology-Enhanced Formative Assessment (TEFA) model [4]. The data makes the abstract TEFA cycle—questioning, response, and whole-class discussion—visible and measurable. A teacher might notice they excel at the questioning phase but consistently allocate minimal time to discussion, a crucial component for fostering deeper understanding [44]. This data-driven insight can guide peer-to-peer mentoring and help focus professional learning on specific, high-leverage practices.

Finally, at an institutional level, aggregated and anonymized data could inform the strategic design of faculty development initiatives. This approach enables institutions to move beyond generic training and allocate resources to address demonstrated, context-specific needs, thereby fostering a culture of continuous improvement supported by learning analytics [45].

D. LIMITATIONS

While our study presents significant contributions, it is important to acknowledge certain limitations that warrant consideration for future research.

Firstly, the dataset employed is relatively constrained, comprising 10 audio files collected from a limited number of teachers and courses. This inherent constraint on sample size and diversity may impact the generalizability of our findings to broader educational contexts. Different teaching styles, subject matters, classroom acoustics, and student demographics could introduce variations not fully captured by our current dataset, thus limiting the direct applicability of our model's performance to other pedagogical environments. Nevertheless, despite this constrained dataset size, the significant performance improvements observed, particularly with the integration of paralinguistic features, strongly indicate the viability and potential of this multimodal approach for automated teacher intervention analysis.

Secondly, while our approach leverages the robustness of modern automatic speech recognition (ASR) and speaker diarization processes, these upstream stages are not infallible, especially within the acoustically challenging environment of a classroom. Factors such as overlapping speech, varying vocal qualities, and ambient background noise can lead to misrecognized words or incorrect speaker attributions. These inherent errors, although partially mitigated by our multimodal approach, could potentially propagate and influence the accuracy of downstream classification. For instance, a misrecognized word or a wrongly attributed utterance might lead to the misclassification of a teacher's pedagogical move. Furthermore, in its current form, the system is not entirely automatic, as initial ASR transcriptions were subjected to manual review and correction to ensure high accuracy for teacher interventions. While this manual curation established a high-quality ground truth essential for model development and validation, it represents a reliance on human intervention in the data preparation pipeline, impacting scalability in a fully automated deployment. It is noteworthy, however, that recent research suggests classification systems can maintain high performance even when confronted with less-than-perfect ASR outputs [46], providing further confidence that the observed benefits of integrating paralinguistic features would likely persist even with purely automatic, uncorrected transcripts.

VII. CONCLUSION AND FUTURE WORK

The effective integration of Student Response Systems (SRS) in educational settings requires feedback for instructors to refine their pedagogical approaches. This study addressed this imperative by developing and evaluating an automated system for classifying teacher interventions within SRS-supported classrooms, aiming to provide granular insights into complex classroom dynamics.

Our methodology, leveraging robust speech recognition and speaker diarization, served as input for advanced DeBERTa-based models. Through a series of progressively complex models, we demonstrated that while text-only approaches provide a foundational understanding of teacher interventions, they often fall short in capturing nuanced pedagogical acts. Crucially, the integration of paralinguistic features alongside textual content in our multimodal models significantly enhanced classification performance, validating our hypothesis that human conversation's multidimensionality extends beyond lexical content.

Furthermore, the application of Explainable Artificial Intelligence (xAI) techniques, particularly SHAP, proved instrumental in elucidating the decision-making processes of these multimodal models. xAI not only clarified how the

models fused textual and paralinguistic cues but also revealed the differential importance of these features across various intervention categories. This interpretability allows the generation of feedback, enabling teachers to understand 'why' an intervention was classified in a particular way, fostering trust in the system, and guiding targeted professional development.

As a statement of direction, we are currently exploring two complementary research lines. The first focuses on extending the classification framework to analyze pedagogical moves within student collaborative learning scenarios, where the dynamics of interaction are more complex and may benefit even more from the integration of paralinguistic features. The second line of work involves leveraging Large Language Models (LLMs) both to enhance the classification process and to generate interpretable explanations based on their internal reasoning capabilities, opening the door to new forms of model transparency and adaptive feedback.

REFERENCES

[1] M. J. Micheletto, "Using audience response systems to encourage student engagement and reflection on ethical orientation and behavior," *Contemp. Issues Educ. Res. (CIER)*, vol. 4, no. 10, pp. 9–18, Sep. 2011.

[2] O. Özdemir, "Kahoot! Game-based digital learning platform: A comprehensive meta-analysis," *J. Comput. Assist. Learn.*, vol. 41, no. 1, p. 13084, Feb. 2025.

[3] I. D. Beatty, W. J. Gerace, W. J. Leonard, and R. J. Dufresne, "Designing effective questions for classroom response system teaching," *Amer. J. Phys.*, vol. 74, no. 1, pp. 31–39, Jan. 2006.

[4] I. D. Beatty and W. J. Gerace, "Technology-enhanced formative assessment: A research-based pedagogy for teaching science with classroom response technology," *J. Sci. Educ. Technol.*, vol. 18, no. 2, pp. 146–162, Apr. 2009.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[6] M. K. Smith, F. H. M. Jones, S. L. Gilbert, and C. E. Wieman, "The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices," *CBE—Life Sci. Educ.*, vol. 12, no. 4, pp. 618–627, Dec. 2013.

[7] O. Kocak, "A systematic literature review of Web-based student response systems: Advantages and challenges," *Educ. Inf. Technol.*, vol. 27, no. 2, pp. 2771–2805, Mar. 2022.

[8] P. Diaz, S. Hrastinski, and P. Norström, "How teacher educators use response systems—An interview study," *Interact. Learn. Environ.*, vol. 32, no. 7, pp. 3652–3664, 2024.

[9] J. E. Caldwell, "Clickers in the large classroom: Current research and best-practice tips," *CBE—Life Sci. Educ.*, vol. 6, no. 1, pp. 9–20, Mar. 2007.

[10] H. Lee, A. Feldman, and I. D. Beatty, "Factors that affect science and mathematics teachers' initial implementation of technology-enhanced formative assessment using a classroom response system," *J. Sci. Educ. Technol.*, vol. 21, no. 5, pp. 523–539, Oct. 2012.

[11] A. I. Wang, "The wear out effect of a game-based student response system," *Comput. Educ.*, vol. 82, pp. 217–227, Mar. 2015.

[12] Á. Tóth, P. Lógó, and E. Lógó, "The effect of the kahoot quiz on the student's results in the exam," *Periodica Polytechnica Social Manage. Sci.*, vol. 27, no. 2, pp. 173–179, Jan. 2019.

[13] Ó. Canovas, P. González-Férez, F. J. García-Clemente, and F. Pardo-García, "Analyzing Wooclap's competition mode with AI through classroom recordings," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 19, pp. 220–229, 2024.

[14] R. Southwell, S. L. Pugh, E. M. Perkoff, C. Clevenger, J. O. Bush, R. Lieber, W. D. Ward, P. W. Foltz, and S. K. D'Mello, "Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms," *Int. Educ. Data Mining Soc.*, vol. 2022, pp. 1–14, Jan. 2022.

[15] A. S. Khan, T. Ogunremi, A. A. Attia, and D. Demszky, "Multi-stage speaker diarization for noisy classrooms," 2025, *arXiv:2505.10879*.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, *arXiv:2212.04356*.

[17] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," 2023, *arXiv:2307.13008*.

[18] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers," 2023, *arXiv:2307.03183*.

[19] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. Interspeech*, Sep. 2018, pp. 2808–2812.

[20] T. Park, I. Medennikov, K. Dhawan, W. Wang, H. Huang, N. R. Koluguri, K. C. Puvvada, J. Balam, and B. Ginsburg, "Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems," 2024, *arXiv:2409.06656*.

[21] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, Dec. 2020.

[22] D. Demszky, R. Wang, S. Geraghty, and C. Yu, "Does feedback on talk time increase student engagement? Evidence from a randomized controlled trial on a math tutoring platform," in *Proc. 14th Learn. Anal. Knowl. Conf.*, Mar. 2024, pp. 632–644.

[23] T. Nazaretsky, J. N. Mikeska, and B. Beigman Klebanov, "Empowering teacher learning with AI: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion," in *Proc. 13th Int. Learn. Anal. Knowl. Conf.*, Mar. 2023, pp. 122–132.

[24] D. Wang and G. Chen, "Evaluating the use of BERT and llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy," *Brit. J. Educ. Technol.*, vol. 56, no. 6, pp. 2671–2704, Nov. 2025.

[25] F. Pardo, Ó. Cánovas, and F. J. G. Clemente, "Audio features in education: A systematic review of computational applications and research gaps," *Appl. Sci.*, vol. 15, no. 12, p. 6911, Jun. 2025.

[26] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-S. Lee, "SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering," 2019, *arXiv:1910.11559*.

[27] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević, "Explainable artificial intelligence in education," *Comput. Educ., Artif. Intell.*, vol. 3, Jan. 2022, Art. no. 100074.

[28] L. Zhang, L. Deng, S. Zhang, and L. Chen, "How well can tutoring audio be autoclassified and machine explained with XAI: A comparison of three types of methods," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1290–1300, 2024.

[29] H. Ogata, B. Flanagan, K. Takami, Y. Dai, R. Nakamoto, and K. Takii, "EXAIT: Educational eXplainable artificial intelligent tools for personalized learning," *Res. Pract. Technol. Enhanced Learn.*, vol. 19, p. 19, Aug. 2023.

[30] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[31] L. S. Shapley, "A value for N-person games," in *Contributions to Theory Games*, 1952, pp. 307–317.

[32] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[34] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. Audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7124–7128.

[35] M. Honnibal, I. Montani, S. V. Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," Tech. Rep., 2020.

[36] J. Li and Y. Wang, "NPCA: A linear dimensionality reduction method using a multilayer perceptron," *Frontiers Genet.*, vol. 14, Jan. 2024, Art. no. 1290447.

[37] A. C. Wardhana, "Dimension-expanding MLP in transformer: Inappropriate sentences and paragraph digital content filtering," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 1202–1213, May 2025.

[38] Y. Chen, C. Huang, S. Gao, Y. Lyu, X. Chen, S. Liu, D. Bao, and C. Lv, "A multimodal deep learning approach for legal English learning in intelligent educational systems," *Sensors*, vol. 25, no. 11, p. 3397, May 2025.
[39] J. D. T. Guerrero-Sosa, F. P. Romero, V. H. Menéndez-Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. A. Olivas, "A comprehensive review of multimodal analysis in education," *Appl. Sci.*, vol. 15, no. 11, p. 5896, May 2025.
[40] E. Jensen, S. L. Pugh, and S. K. D'Mello, "A deep transfer learning approach to modeling teacher discourse in the classroom," in *Proc. 11th Int. Learn. Anal. Knowl. Conf.*, Apr. 2021, pp. 302–312.
[41] G. Pai, "Using formative assessment and feedback from student response systems (SRS) to revise statistics instruction and promote student growth for all," *J. Statist. Data Sci. Educ.*, vol. 33, no. 1, pp. 16–25, Jan. 2025.
[42] D. K. Anderson, M. Schoenleber, and S. Korshavn, "Higher-order clicker questions engage students and prepare them for higher-order thinking activities," *J. Microbiol. Biol. Educ.*, vol. 24, no. 1, Apr. 2023.
[43] J. Hattie and H. Timperley, "The power of feedback," *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007.
[44] T. Vickrey, K. Rosploch, R. Rahmanian, M. Pilarz, and M. Stains, "Research-based implementation of peer instruction: A literature review," *CBE-Life Sci. Educ.*, vol. 14, no. 1, p. es3, Mar. 2015.
[45] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Hum. Behav.*, vol. 89, pp. 98–110, Dec. 2018.
[46] D. Wang and G. Chen, "Are perfect transcripts necessary when we analyze classroom dialogue using AIoT?" *Internet Things*, vol. 25, Apr. 2024, Art. no. 101105.

FEDERICO PARDO GARCÍA received the bachelor's degree in computer engineering and the master's degree in big data from the University of Murcia, Spain, in 2020 and 2021, respectively, where he is currently pursuing the Ph.D. degree. His research interests include sound classification and image recognition using artificial intelligence.

ÓSCAR CÁNOVAS REVERTE received the bachelor's degree in computer engineering and the Ph.D. degree in computer science from the University of Murcia, Spain, in 1998 and 2003, respectively. He is currently an Associate Professor with the University of Murcia. He has participated in several research projects and published articles in diverse fields, such as information security, user authentication and authorization, software-defined networks, indoor positioning, and educational technology research and applications.

FÉLIX J. GARCÍA CLEMENTE is currently pursuing the Ph.D. degree in computer science with the University of Murcia, Spain. He is also a Full Professor in computer architecture and technology with the Faculty of Computer Science, University of Murcia. As a result of this activity, he is the author of more than 120 publications, including journals and conferences. He is an active member of various national and international research and development projects. His research interests include cybersecurity, dynamic service management, Industry 4.0, and educational technology.

ANTONIO ORENES LUCAS received the bachelor's degree in computer engineering (software engineering), the first master's degree in artificial intelligence applications in medicine, and the second master's degree in artificial intelligence from the University of Murcia, Spain, in 2022, 2023, and 2025, respectively. He developed his master's thesis on explainable AI for teacher evaluation. He is currently a Software Engineer with IMIB, Murcia, combining software development with AI research in healthcare applications.

6
Conclusiones

Esta sección presenta las principales conclusiones derivadas del desarrollo de esta tesis doctoral. A continuación, se identifican las direcciones futuras que emergen como extensiones naturales de la investigación realizada.

C1. El audio incorpora múltiples niveles de información pedagógica

El inicio de esta tesis fue la revisión sistemática, motivada por el objetivo O1. La principal aportación a este objetivo fue la organización de las posibles características extraídas del audio en tres niveles: bajo nivel, paralingüísticas y procesamiento del lenguaje natural. Esta taxonomización original, delineó el desarrollo de la tesis en etapas posteriores, marcando los objetivos O2 y O3.

Previo a la publicación de la revisión sistemática no existía una taxonomía disponible para las características de audio dividida por niveles y con detalles sobre la información que aportan. Centralizar en un solo análisis las diferentes formas de procesar el audio, así como ejemplos específicos sobre aplicaciones y combinaciones que se han realizado de dichas características, constituye una importante aportación al campo. Esto permite comprender el audio como un fenómeno estratificado donde cada nivel captura diferentes aspectos pedagógicos y posibilita un diseño de investigación más sistemático al seleccionar características según el nivel de análisis deseado. Esta selección puede orientar a los investigadores para formular sus preguntas de investigación en base a la posible información que se puede extraer de cada nivel de características.

Además, esta taxonomía invita a no ceñirse a un único tipo de rasgo para una tarea concreta. Las carencias de un tipo de característica pueden ser subsanadas por otras, ya que estas pueden comportarse de forma complementaria, como planteamos en la Sección M3. Gracias a esta decisión se obtuvieron mejoras en rendimiento e interpretabilidad que no habrían sido posibles con un solo tipo de característica.

C2. La información paralingüística posee un amplio potencial pedagógico

El objetivo O2 planteaba la caracterización de metodologías docentes ignorando deliberadamente el uso del contenido verbal disponible. Esta decisión, lejos de constituir una limitación, abrió un amplio abanico de posibilidades para el análisis de la práctica docente. No solo resulta posible clasificar metodologías docentes basándose únicamente en la estructura de los turnos, como ya se validó en R2, sino que la evidencia también sientan las bases para el perfilado de los docentes en función de la distribución temporal de sus metodologías [23].

Además, dado que este tipo de características son agnósticas al contenido léxico, resulta posible generalizarlas a nuevos entornos no contemplados inicialmente, facilitando la transferencia de conocimiento a nuevas áreas y campos. Esta implicación desafía la asunción de que el contenido verbal es imprescindible a la hora de realizar un análisis pedagógico. Lo que podría percibirse como una limitación puede subsanarse con lo mencionado en la conclusión C1, integrando información de múltiples niveles cuando la situación lo requiera.

C3. El conocimiento experto supera a los sistemas de caja negra

En el momento de plantear la metodología del objetivo O3, especialmente en la fusión de características, se enfrentó un problema de sobredimensionamiento. El contenido léxico requiere de un dimensionamiento alto (768 dimensiones) para poder contener toda la información verbal que se presenta en la clase. Esto contrastaba con el modesto vector de características paralingüísticas (13 dimensiones). Trabajando con los sistemas de *early-fusion*, las características paralingüísticas eran ignoradas por los modelos, independientemente de la técnica que se utilizase para combatir este efecto. Sin embargo, al pasar a sistemas de *late-fusion*, no solo las características diseñadas manualmente adquirieron valor, sino que superaron al sistema basado exclusivamente en texto.

Este efecto en el que las características diseñadas por humanos superan a las que pueden extraer internamente los sistemas de *deep learning* ya se empezó a observar con la integración del MR en la fase metodológica M2. El diseño de esta característica surgió del análisis del comportamiento de los datos extraídos en la fase de diarización. Sin este análisis previo no habría sido posible extraer esta información, que además habría sido descartada por las características de bajo nivel y NLP, al ser considerada ruido que no permite un procesamiento útil ni una transcripción adecuada.

Este hecho lleva a reflexionar sobre el uso que se hace de las técnicas de inteligencia artificial más contemporáneas, donde la principal metodología consiste en alimentar con un volumen suficientemente grande de datos en bruto hasta que el modelo sepa diferenciar por sí mismo lo útil del ruido para la tarea que se le encomienda. La ingeniería de características y la implicación humana más allá de unir las piezas de un puzzle tecnológico siguen siendo relevantes, especialmente para comprender las relaciones que establecen los modelos. Estas relaciones por sí mismas tienen valor y pueden constituir la base de nuevos hallazgos, pero no se dedican suficientes recursos a esta tarea, lo que conduce a la necesidad de abordar

la explicabilidad.

C4. Explicabilidad por diseño

En la misma revisión sistemática se analizaron las principales limitaciones que enfrenta el campo para la adopción generalizada de los desarrollos, siendo la principal de ellas la retroalimentación docente. Sin embargo, desarrollar únicamente herramientas para los profesores no resulta suficiente, puesto que también es necesario ceñirse a prácticas comprometidas con la privacidad y especialmente con el uso de técnicas de XAI por diseño. Ambas promesas son necesarias para convencer a los docentes de que estas herramientas no constituyen sistemas de vigilancia forzosa o desarrollos que no aportan ningún tipo de información útil, sino que pueden servir para mejorar sus prácticas.

La XAI juega un papel fundamental en el desarrollo de herramientas para los docentes. Comprender las decisiones de una máquina genera confianza en el usuario y permite entender por qué la máquina toma ciertas decisiones que en un primer momento pueden no haberse comprendido o carecer de sentido. Particularmente en la metodología M3 se hizo especial hincapié en el uso de la XAI para entender las decisiones del modelo, dado que se trataba de un sistema complejo basado en distintos tipos de caracaterísticas para realizar la clasificación. Las técnicas de explicabilidad permitieron comprender las relaciones que hacía el modelo con cada tipo de característica, describiendo los patrones aprendidos y qué información aportaban.

Sin embargo, las técnicas de explicabilidad actuales presentan algunas limitaciones. Por ejemplo, los valores SHAP analizados para la explicabilidad en la Sección M3 pueden no resultar apropiados para perfiles no técnicos, puesto que solo presentan información numérica que es necesario interpretar para comprender adecuadamente. Esta limitación puede conectarse con la conclusión C3 respecto al uso de conocimiento experto, pues la ingeniería de características que tienen sentido humano por sí mismas permite reducir la brecha con las técnicas de explicabilidad habituales en modelos de caja negra.

C5. Cerrando el ciclo de retroalimentación docente

La principal motivación de la tesis es el cierre del ciclo de retroalimentación docente. Este cierre podría haberse materializado únicamente con la entrega de los reportes PDF mencionados en la Sección R4. No obstante, la plataforma web materializa la aplicación práctica del sistema, permitiendo que el profesorado acceda a la retroalimentación de forma autónoma. Esta implementación transforma el software de análisis en una herramienta accesible que los docentes pueden utilizar de manera independiente y según sus necesidades.

La democratización en el acceso a esta herramienta, al prescindir de mediadores externos, abre la posibilidad a fomentar una autonomía que estimule la curiosidad del docente por su propio desempeño. Esta desintermediación sentaría las bases para que el aula pueda concebirse como un entorno de experimentación pedagógica, donde el profesorado tendría la capacidad de implementar nuevas metodologías y contrastar su impacto de manera empírica. Bajo esta premisa, la disponibilidad de métricas objetivas proporciona un marco de referencia para la innovación, facilitando que el proceso reflexivo se transforme

en una dinámica de exploración iterativa orientada al descubrimiento de estrategias de enseñanza más eficaces.

La plataforma web no solo cierra el ciclo analítico-reflexivo identificado como principal laguna en la revisión sistemática, sino que representa una prueba de concepto de que la investigación en MMLA puede y debe transcender el ámbito académico para ofrecer herramientas reales que empoderen a los docentes en su práctica cotidiana.

A partir de los resultados y conclusiones presentados, emergen cuatro direcciones de investigación que permitirían extender y consolidar las contribuciones de esta tesis.

F1. Evaluación longitudinal del impacto de la plataforma en la práctica docente

La limitación más significativa del resultado R4 es la ausencia de una evaluación rigurosa del impacto de la retroalimentación automática en la práctica docente real. Aunque se recogió evidencia cualitativa de reflexión y cambio intencional, no se establecieron medidas objetivas de mejora pedagógica ni se controló la evolución longitudinal de las métricas extraídas.

Una investigación futura debería diseñar un estudio cuasi-experimental que compare grupos de docentes con y sin acceso a la plataforma de retroalimentación, midiendo tanto indicadores de proceso (evolución de las métricas paralingüísticas a lo largo del curso) como indicadores de resultado (rendimiento estudiantil, satisfacción, percepción del clima de aula). Complementariamente, estudios de usabilidad con protocolos de *think-aloud* permitirían identificar qué elementos de la interfaz resultan más útiles, qué métricas son mejor comprendidas y qué barreras frenan la adopción sostenida. Esta línea conecta directamente con la necesidad, identificada en C4, de validar empíricamente la comprensibilidad de las explicaciones generadas por XAI.

F2. Extensión del pipeline hacia características acústicas de bajo nivel

La arquitectura de software desarrollada durante esta tesis extrae características acústicas de bajo nivel que no fueron explotadas activamente en los modelos de clasificación presentados. Esta decisión metodológica, justificada por la prioridad de validar primero el nivel paralingüístico, deja abierta la exploración del potencial de estas señales.

La literatura sugiere que características como la variabilidad del tono (F0) pueden correlacionarse con estados emocionales del docente o con estrategias retóricas como el énfasis y la modulación. Investigaciones futuras podrían integrar estas métricas de bajo nivel con las características paralingüísticas validadas, evaluando si la fusión de múltiples estratos acústicos mejoran los sistemas de clasificación. Un enfoque particularmente prometedor sería la detección de momentos de alta carga emocional que podrían enriquecer la retroalimentación con información sobre el clima afectivo del aula, una dimensión actualmente no capturada por los modelos implementados.

F3. Integración de modelos de lenguaje y arquitecturas agénticas para retroalimentación adaptativa

El componente de generación narrativa implementado en R4 empleó modelos de lenguaje ejecutados localmente con restricciones computacionales significativas. Las narrativas generadas, aunque funcionales, evidenciaron limitaciones derivadas del desconocimiento de los modelos sobre el dominio específico de las métricas paralingüísticas y las dinámicas de interacción educativa.

El avance exponencial de los grandes modelos de lenguaje y las arquitecturas agénticas ofrece oportunidades para superar estas limitaciones. Un agente conversacional especializado podría actuar como mediador pedagógico entre las métricas extraídas y el docente: respondiendo preguntas específicas ("¿Es problemático que hable el 80 % del tiempo?"), sugiriendo estrategias concretas basadas en la literatura pedagógica, o incluso proponiendo experimentos de aula ("La próxima sesión, intente aumentar el tiempo de espera tras cada pregunta y observe si cambia el ratio de participación del hablante"). Investigaciones futuras deberían explorar arquitecturas de Retrieval-Augmented Generation (RAG) que anclen las respuestas del LLM en conocimiento pedagógico validado, mitigando el riesgo de alucinaciones y garantizando la fundamentación de las recomendaciones.

F4. Generalización multilingüe y multicultural

El corpus empleado en esta tesis comprende exclusivamente grabaciones en español, procedentes de una única universidad y un conjunto limitado de disciplinas. Aunque las características paralingüísticas fueron diseñadas para ser agnósticas al idioma, esta propiedad teórica no ha sido validada empíricamente en contextos multilingües.

Investigaciones futuras deberían replicar los experimentos de clasificación con grabaciones en otros idiomas y culturas educativas. Los patrones de interacción verbal podrían manifestar variaciones significativas entre tradiciones pedagógicas: culturas con mayor deferencia jerárquica podrían exhibir distribuciones de turnos más asimétricas; contextos donde el debate abierto es norma podrían mostrar mayores tasas de solapamiento sin que esto implique necesariamente trabajo colaborativo. Comprender estas variaciones es requisito previo para desarrollar soluciones genuinamente generalizables. Colaboraciones internacionales que aporten grabaciones de aula de distintos contextos lingüísticos y culturales constituirían una contribución valiosa al campo.

En síntesis, esta tesis doctoral ha demostrado que el audio educativo constituye una fuente de información pedagógica estratificada con un gran potencial. Las características paralingüísticas derivadas de la diarización de hablantes codifican una huella estructural de la práctica docente que puede capturarse de forma generalizable y agnóstica al contenido. La fusión multimodal con representaciones semánticas potencia la clasificación de intervenciones, mientras que las técnicas de explicabilidad transforman las métricas abstractas en señales interpretables. La plataforma de retroalimentación desarrollada materializa el cierre del ciclo entre el análisis computacional y la retroalimentación docente, aunque su impacto longitudinal permanece como una pregunta abierta para futuras investigaciones. Las direcciones propuestas delinean un programa de investigación que podría consolidar

las analíticas de audio educativo como un campo maduro con aplicaciones institucionales reales.

Bibliografía

[1] Karen M La Paro et al. «The classroom assessment scoring system: Findings from the prekindergarten year». En: *The elementary school journal* 104.5 (2004), págs. 409-426.

[2] Michelle K Smith et al. «The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices». En: *CBE—Life Sciences Education* 12.4 (2013), págs. 618-627.

[3] S Kelly et al. «Automatically Measuring Question Authenticity in Real-World Classrooms». En: *Educational Researcher* 47 (7 2018), págs. 451-464.

[4] Doug Clow. «An overview of learning analytics». En: *Teaching in Higher Education* 18.6 (2013), págs. 683-695.

[5] Naif Radi Aljohani et al. «Predicting at-risk students using clickstream data in the virtual learning environment». En: *Sustainability* 11.24 (2019), pág. 7238.

[6] Olga Viberg et al. «The current landscape of learning analytics in higher education». En: *Computers in human behavior* 89 (2018), págs. 98-110.

[7] Paulo Blikstein. «Multimodal learning analytics». En: *Proceedings of the third international conference on learning analytics and knowledge*. 2013, págs. 102-106.

[8] Marcelo Worsley. «Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces». En: *Proceedings of the 14th ACM international conference on Multimodal interaction*. 2012, págs. 353-356.

[9] Xavier Ochoa et al. «Augmenting learning analytics with multimodal sensory data». En: *Journal of Learning Analytics* 3.2 (2016), págs. 213-219.

[10] A Ramakrishnan et al. «Toward automated classroom observation: Predicting positive and negative climate». En: *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*. 2019.

[11] P J Donnelly et al. «Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context». En: *ACM International Conference Proceeding Series*. 2017, págs. 218-227.

[12] Dorottya Demszky et al. «M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes». En: *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20â • fi22, 2023, Copenhagen, Denmark* 1 (2023), págs. 23-759. ISSN: 9798400700255.

[13] Deliang Wang et al. «Artificial intelligence in classroom discourse: A systematic review of the past decade». English. En: *International Journal of Educational Research* 123 (2024). ISSN: 08830355.

[14] Qiujie Li et al. «How instructors use learning analytics: the pivotal role of pedagogy». En: *Journal of Computing in Higher Education* (2025), págs. 1-29.

[15] R Kasepalu et al. «Do Teachers Find Dashboards Trustworthy, Actionable and Useful? A Vignette Study Using a Logs and Audio Dashboard». En: *Technology, Knowledge and Learning* 27 (3 2022), págs. 971-989.

[16] Benedikt Wisniewski et al. «The power of feedback revisited: A meta-analysis of educational feedback research». En: *Frontiers in psychology* 10 (2020), pág. 487662.

[17] Federico Pardo et al. «Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps». En: *Applied Sciences* 15.12 (2025), pág. 6911.

[18] Federico Pardo et al. «Exploring AI techniques for generalizable teaching practice identification». En: *IEEE Access* (2024).

[19] Federico Pardo García et al. «Explaining Teacher Interventions in SRS-Based Classrooms: A Classification Approach With BERT and Paralinguistic Cues». En: *IEEE Access* 13 (2025), págs. 208078-208093. DOI: 10.1109/ACCESS.2025.3641484.

[20] Matthew J Page et al. «PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews». En: *bmj* 372 (2021).

[21] P Chejara et al. «Efar-mmla: An evaluation framework to assess and report generalizability of machine learning models in mmla». En: *Sensors* 21 (8 2021).

[22] Oscar Cánovas et al. «AI-driven Teacher Analytics: Informative Insights on Classroom Activities». En: *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2023 - Conference Proceedings*. 2023. ISBN: 978-1-66545-331-8.

[23] Oscar Cánovas et al. «Analyzing Wooclap's competition mode with AI through classroom recordings». En: *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* (2024).