



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



DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TÍTULO 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.*

Agradecimientos

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 llevó 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	3
2. Objetivos	7
2.1. Objetivo general	7
2.2. Objetivos específicos	7
O1. Sistematizar la evidencia científica sobre el uso de características de audio en entornos educativos.	8
O2. Caracterización de la actividad docente mediante el uso de descriptores paralingüísticos.	8
O3. Integración de características paralingüísticas y semánticas para la mejora de la clasificación y la interpretabilidad del discurso docente. . .	8
O4. Transferencia de resultados: desarrollo de un sistema integral de análisis y visualización para la reflexión docente.	8
3. Metodología	11
M1. Revisión sistemática de la literatura	11
M2. Clasificación de metodologías docentes mediante características paralingüísticas y machine learning	13
M3. Integración de características textuales y paralingüísticas mediante deep learning	13
M4. Arquitectura del sistema: diseño desacoplado de los servicios de análisis y la interfaz de usuario	14
4. Resultados	17
R1. Taxonomización de las características de audio, limitación de datos y falta de retroalimentación docente	17
R2. Caracterización de metodologías docentes, generalización y análisis de las principales características paralingüísticas	19

R3. Clasificación de intervenciones docentes mediante fusión multimodal y aprendizaje profundo	20
R4. Implementación y validación del desarrollo tecnológico para la retroalimentación docente.	21
5. Publicaciones	25
5.1. Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps	27
5.2. Exploring AI Techniques for Generalizable Teaching Practice Identification	29
5.3. Explaining Teacher Interventions in SRS-based Classrooms: A Classification Approach with BERT and Paralinguistic Cues	31
6. Conclusiones	33
C1. El audio incorpora múltiples niveles de información pedagógica	33
C2. La información paralingüística posee un amplio potencial pedagógico	34
C3. El conocimiento experto supera a los sistemas de caja negra	34
C4. Explicabilidad por diseño	35
C5. Cerrando el ciclo de retroalimentación docente	35
F1. Evaluación longitudinal del impacto de la plataforma en la práctica docente .	36
F2. Extensión del pipeline hacia características acústicas de bajo nivel	36
F3. Integración de modelos de lenguaje y arquitecturas agénticas para retroalimentación adaptativa	37
F4. Generalización multilingüe y multicultural	37
<i>Bibliografía</i>	39

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 previos 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.

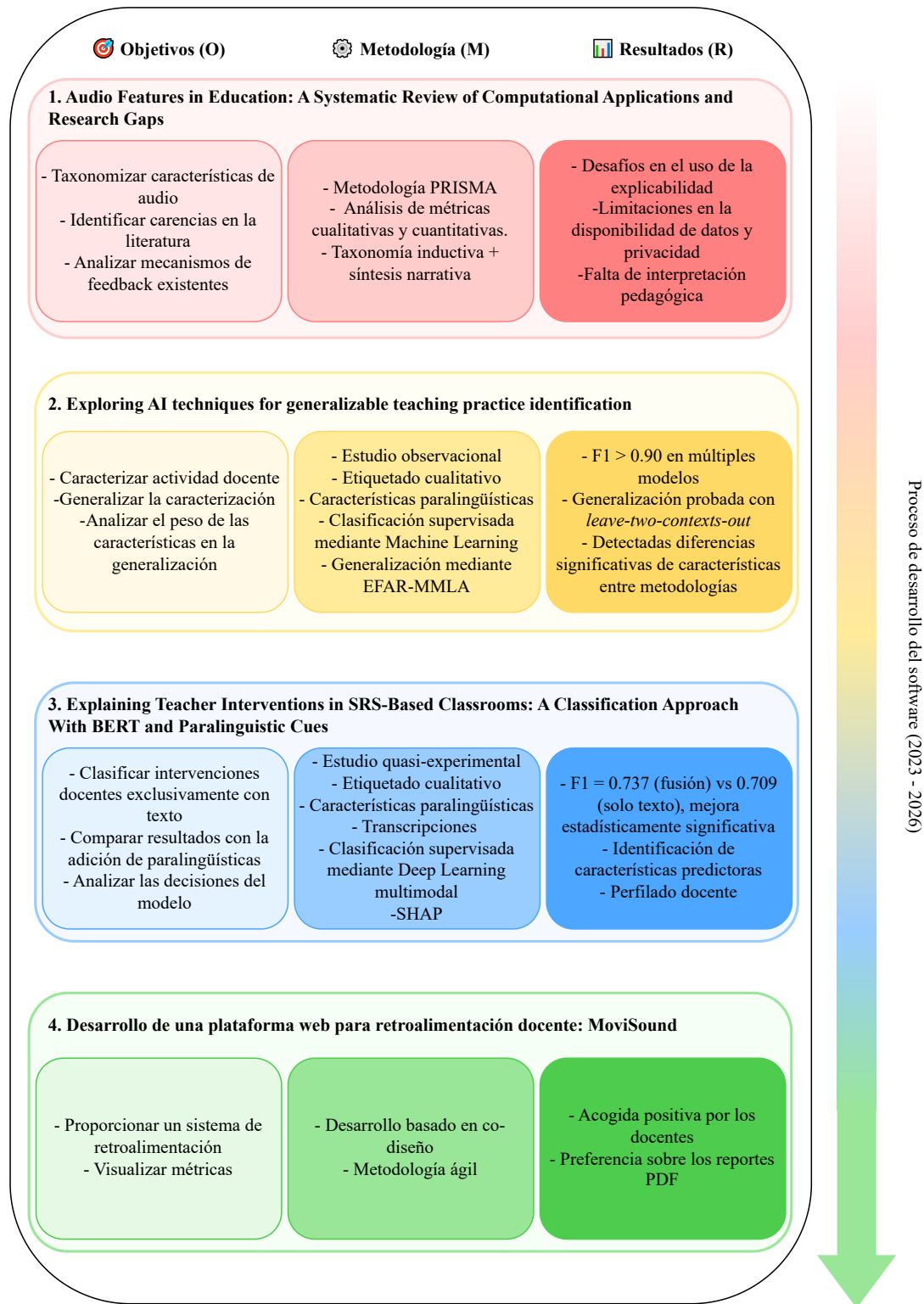


Figura 3.1: Arquitectura metodológica de la tesis: evolución incremental del enfoque de investigación a través de cuatro fases secuenciales. Cada fila representa un estudio con sus objetivos, diseño metodológico y hallazgos principales. La progresión muestra cómo los resultados de cada fase informaron 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 interdocente 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 opacidad 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 la 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 paralingüísticos, 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 realizar 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 sus 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 paralingüísticas

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.

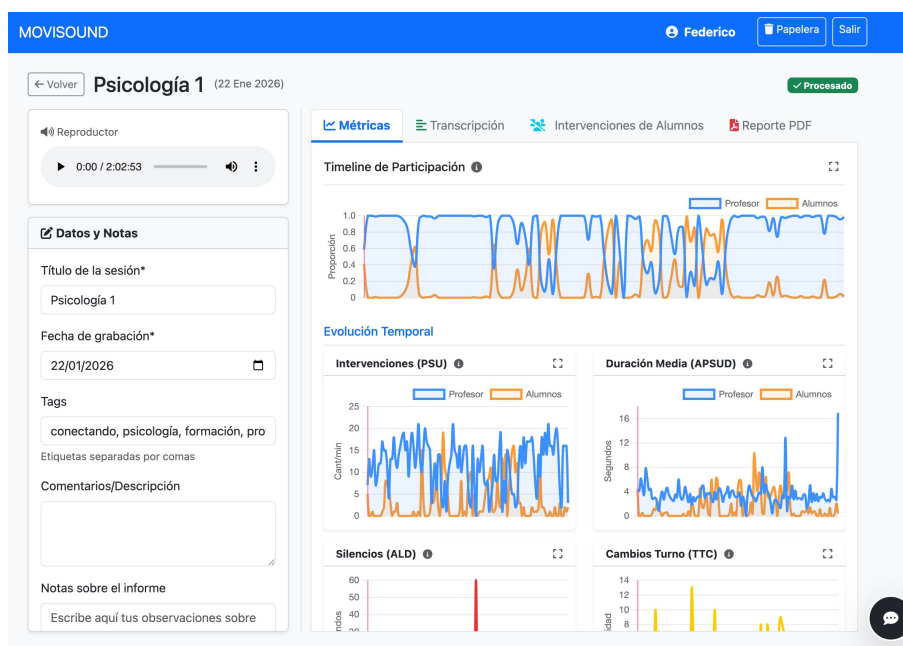


Figura 4.1: Interfaz de la web desarrollada para la retroalimentación docente. Se pueden visualizar los metadatos de la sesión a la izquierda. El panel central muestra la progresión temporal de algunas métricas extraídas como al PSR, representado por el ‘Timeline de Participación’. Se aprecian también 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 <i>Departamento de Ingeniería y Tecnología de Computadores Universidad de Murcia, España</i>			
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.			

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.			

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.			

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 sienta 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 es 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 características 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 trascender 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 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 Tecnologías del Aprendizaje* (2024).

