PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE CIENCIAS EXACTAS NATURALES Y AMBIENTALES

CARRERA DE CIENCIA DE DATOS

Redes neuronales profundas en bioacústica: precisión, limitaciones y aplicabilidad en entornos naturales

Monografía previa a la obtención del título de Ingeniero en Ciencia de Datos

Sofía Milena Aguilar Córdova

Quito, 2026

Certifico que la Monografía de Ciencia de Datos de **Sofía Milena Aguilar Córdova** ha sido concluida de conformidad con las normas establecidas; por lo tanto, puede ser presentada para la calificación correspondiente.



Firmado digitalmente por ANDRES ESTEBAN MERINO TOAPANTA
Fecha: 2026.06.17 15:45:10 -05'00'

Andrés Esteban Merino Toapanta

Tutor de la monografía

Quito, 26 de Abril de 2026

TABLA DE CONTENIDOS

1. Introducción **3**

2. Objetivos **9**

2.1. Objetivo general 9

2.2. Objetivos específicos 9

3. Fundamento teórico **11**

3.1. Monitoreo acústico pasivo y problemática del reconocimiento automático 11

3.2. Representación y procesamiento de señales acústicas 13

3.3. Modelos de aprendizaje automático y profundo para reconocimiento acústico . . 21

3.3.1. Enfoques basados en aprendizaje automático 22

3.3.2. Aprendizaje profundo y modelos avanzados 23

3.3.3. Modelos preentrenados y aprendizaje por transferencia 24

3.3.4. Pipeline de reconocimiento acústico 25

3.3.5. Evaluación del desempeño 26

3.4. Factores que afectan el rendimiento en entornos naturales 28

3.5. Enfoques avanzados y análisis de errores en entornos reales 31

4. Metodología **38**

4.1. Diseño metodológico y criterios de selección 38

4.2. Estrategia de búsqueda 41

4.3. Identificación y cribado de registros 43

4.4. Elegibilidad por texto completo 44

4.5. Extracción, normalización y síntesis de la información 46

5. Resultados y discusión **50**

5.1. RQ1. ¿Qué arquitecturas y métodos de preprocesamiento predominan o reportan mejor desempeño en el reconocimiento acústico de especies? 52

5.2. RQ2. ¿Cómo varía la precisión de los modelos ante ruido ambiental, calidad del dataset y heterogeneidad acústica? 58

5.3. RQ3. ¿Qué alcances y limitaciones presentan los métodos de reconocimiento acústico para estimar presencia, ocupación o abundancia? 61

5.4. RQ4. ¿Qué métricas, protocolos de validación y tendencias metodológicas predominan en la literatura reciente? 65

5.4.1. Contraste crítico entre propuestas metodológicas representativas 71

5.5. Síntesis integral de las RQ 74

6. Conclusiones **77**

7. Referencias **81**

A. Matriz de identificación metodológica del corpus **91**

LISTA DE TABLAS

4.1. Criterios de inclusión y exclusión aplicados en la revisión 40
4.2. Palabras clave y sinónimos utilizados en la estrategia de búsqueda 42
4.3. Resultados del cribado de registros 44
4.4. Escala de evaluación utilizada en el proceso de Quality Assessment 45
4.5. Resultados de la evaluación de elegibilidad por texto completo 46

5.1. Variables metodológicas y métricas detectadas en el corpus final 51
5.2. Distribución del método principal asignado a cada artículo 53
5.3. Tareas bioacústicas detectadas en el corpus 62
5.4. Métricas y recursos de evaluación más frecuentes en el corpus 66

A.1. Matriz resumida de identificación de arquitecturas y enfoques metodológicos en el corpus final de artículos. 92

LISTA DE FIGURAS

3.1. *Flujo de transformación de una señal acústica cruda hacia el reconocimiento automático de especies mediante CNN* 16
3.2. Flujo de procesamiento acústico aplicado al reconocimiento automático de especies 20
3.3. Flujo general de reconocimiento acústico automático de especies 26
3.4. Esquema conceptual de los factores que condicionan el rendimiento de sistemas de reconocimiento bioacústico en entornos naturales 31
3.5. Esquema de causas potenciales de errores de clasificación en reconocimiento bioacústico 33

4.1. Diagrama PRISMA 2020 del proceso de identificación, cribado, elegibilidad e inclusión de estudios 48

5.1. Frecuencia de familias metodológicas detectadas en el corpus 54
5.2. Representaciones acústicas más frecuentes detectadas en el corpus 55
5.3. Síntesis cualitativa de comparaciones de desempeño reportadas en estudios representativos del corpus 57
5.4. Distribución de tareas bioacústicas detectadas en el corpus 63
5.5. Frecuencia de métricas y recursos de evaluación detectados en el corpus 67
5.6. Factores que condicionan el rendimiento en escenarios reales y respuestas metodológicas reportadas en la literatura 70
5.7. Síntesis gráfica de los aportes y límites interpretativos de propuestas metodológicas representativas del corpus 74

RESUMEN

La presente monografía analiza el estado del arte del reconocimiento acústico automatizado de especies en entornos naturales mediante redes neuronales profundas, a partir de una revisión sistemática basada en lineamientos PRISMA de estudios publicados entre 2017 y 2025. Se conformó un corpus final de 67 estudios, examinado según arquitecturas, representaciones acústicas, métodos de preprocesamiento, métricas de evaluación, estrategias de validación y alcance ecológico. Los resultados evidencian el predominio de arquitecturas convolucionales aplicadas sobre representaciones tiempo–frecuencia, especialmente espectrogramas y mel-espectrogramas; sin embargo, no se identifica una arquitectura universalmente superior, ya que el desempeño depende de la calidad del conjunto de datos, el ruido ambiental, el balance de clases, las condiciones de grabación y los protocolos de evaluación. Se concluye que las redes neuronales profundas son métodos útiles para la bioacústica computacional y el monitoreo ecológico, especialmente en tareas de detección de presencia o actividad temporal, aunque su uso para inferir abundancia relativa, número de individuos u otros indicadores ecológicos requiere validaciones más robustas, métricas comparables y protocolos que consideren la incertidumbre ecológica.

Palabras clave: reconocimiento acústico automatizado, bioacústica, monitoreo acústico pasivo, redes neuronales profundas, CNN, aprendizaje profundo, monitoreo ecológico.

ABSTRACT

This monograph analyzes the state of the art of automated acoustic species recognition in natural environments using deep neural networks, based on a PRISMA-guided systematic review of studies published between 2017 and 2025. A final corpus of 67 studies was examined according to architectures, acoustic representations, preprocessing methods, evaluation metrics, validation strategies, and ecological scope. The results show a predominance of convolutional architectures applied to time–frequency representations, especially spectrograms and mel-spectrograms; however, no universally superior architecture was identified, since performance depends on dataset quality, environmental noise, class balance, recording conditions, and evaluation protocols. It is concluded that deep neural networks are a promising tool for computational bioacoustics and ecological monitoring, particularly for detecting species presence or temporal activity, although their use to infer relative abundance, number of individuals, or other ecological indicators requires caution, more robust validation, comparable metrics, and protocols that consider ecological uncertainty.

Keywords: automated acoustic recognition, bioacoustics, passive acoustic monitoring, deep neural networks, CNN, deep learning, ecological monitoring.

1. INTRODUCCIÓN

En los últimos años, el reconocimiento automático de especies a partir de registros acústicos se ha consolidado como una línea de investigación central en la bioacústica computacional y en la Ciencia de Datos aplicada al monitoreo ambiental (Kershenbaum et al., 2025; Stowell et al., 2019). Este desarrollo está vinculado con la expansión del monitoreo acústico pasivo (*Passive Acoustic Monitoring*, PAM), una estrategia que permite registrar sonidos de forma continua mediante sensores instalados en campo y conservar evidencia acústica de especies vocales sin requerir observación humana permanente (Kershenbaum et al., 2025; Teixeira et al., 2024). Su aporte ecológico radica en ampliar la cobertura temporal y espacial del muestreo. Esa misma capacidad de registro, no obstante, introduce un reto metodológico considerable: la generación de grandes volúmenes de audio ambiental no estructurado, cuya revisión manual resulta poco viable en campañas extensas, múltiples sitios de monitoreo o periodos prolongados de grabación (Stowell et al., 2019; Teixeira et al., 2024).

Desde una perspectiva de Ciencia de Datos, el registro sonoro constituye el punto de partida de un proceso analítico más amplio, orientado a convertir grabaciones ambientales en información ecológica interpretable, comparable y útil para la investigación y la conservación (Kershenbaum et al., 2025; Saoud et al., 2024). Esta transformación requiere procesar señales acústicas variables, segmentarlas, representarlas numéricamente, asociarlas con etiquetas biológicas y evaluarlas mediante criterios compatibles con el objetivo del estudio (Knight et al., 2017b; Stowell et al., 2019). En ambientes naturales, estas grabaciones forman parte de *soundscapes*, entendidos como paisajes acústicos donde interactúan sonidos biológicos, geofísicos y antrópicos (Kershenbaum et al., 2025; Teixeira et al., 2024). Por ello, los modelos deben reconocer vocalizaciones de interés dentro de señales afectadas por ruido ambiental, solapamiento entre especies, eventos no objetivo y variaciones asociadas con sensores, hábitat, distancia al micrófono o condiciones meteorológicas (Knight et al., 2017b; Lostanlen et al., 2024).

El aprendizaje profundo, conocido en la literatura como *deep learning*, ha ganado importancia en bioacústica porque permite modelar patrones complejos a partir de grandes conjuntos de datos (Saoud et al., 2024; Stowell et al., 2019). En el reconocimiento acústico de especies,

estas técnicas hacen referencia a representaciones visuales del audio, como espectrogramas, mel-espectrogramas o transformadas wavelet, que convierten la señal sonora en una representación tiempo-frecuencia (Jeantet y Dufourq, 2023; Kiskin et al., 2020). Esta conversión permite observar cómo cambia la energía del sonido a lo largo del tiempo y de las frecuencias, facilitando que los modelos identifiquen patrones asociados con cantos, llamados o vocalizaciones específicas (Kiskin et al., 2020). En particular, las redes neuronales convolucionales, conocidas como *Convolutional Neural Networks* (CNN), han sido ampliamente empleadas porque pueden extraer patrones locales y jerárquicos en imágenes o representaciones bidimensionales (Kiskin et al., 2020; Stowell et al., 2019).

El uso de redes neuronales profundas no asegura, por sí solo, un desempeño elevado ni una aplicación ecológica confiable. En bioacústica, la capacidad predictiva de los modelos depende de la interacción entre la arquitectura empleada, la calidad del audio, la representación acústica, el balance entre clases, el ruido ambiental, el protocolo de validación y el propósito ecológico del análisis (Kershenbaum et al., 2025; Knight et al., 2017b). La arquitectura neuronal debe examinarse dentro de un *pipeline* completo que articula preprocesamiento, segmentación, representación acústica, entrenamiento, validación e interpretación de resultados (Kershenbaum et al., 2025). Esta lectura evita atribuir el rendimiento únicamente al modelo: un sistema puede alcanzar resultados favorables en condiciones controladas y perder estabilidad al aplicarse en otro hábitat, con sensores distintos o bajo escenarios acústicos diferentes (Kershenbaum et al., 2025; Lostanlen et al., 2024).

La literatura reciente también muestra que algunos modelos incorporan información contextual para mejorar la clasificación acústica (Jeantet y Dufourq, 2023). Esta información puede incluir variables temporales o espaciales que ayudan al modelo a interpretar la probabilidad de presencia de una especie en un sitio o momento determinado (Jeantet y Dufourq, 2023; Teixeira et al., 2024). Esta estrategia es relevante porque las vocalizaciones animales pueden depender de horarios de actividad, distribución espacial y condiciones ambientales (Teixeira et al., 2024). Sin embargo, el uso de información contextual exige precaución metodológica, ya que el modelo podría aprender patrones asociados al sitio de grabación o al periodo de muestreo sin necesariamente generalizar bien a nuevos escenarios (Knight et al., 2017b; Lostanlen et al., 2024).

Uno de los principales problemas para comparar los estudios existentes es la heterogeneidad metodológica (Kershenbaum et al., 2025; Knight et al., 2017b). Los trabajos pueden diferir en el número de especies analizadas, la duración de los audios, la forma de construir las etiquetas, la proporción de clases, el tipo de sensor utilizado, las condiciones ambientales, las métricas reportadas y la forma en que se separan los datos de entrenamiento y prueba (Kershenbaum et al., 2025). Esta diversidad hace que dos valores de desempeño aparentemente similares no siempre signifiquen lo mismo (Knight et al., 2017b). Por ejemplo, una alta *accuracy*, entendida como la proporción de predicciones correctas sobre el total de casos evaluados, puede ser poco informativa si el conjunto de datos está desbalanceado o si existen muchas más grabaciones de una clase que de otra (Knight et al., 2017b). En tareas de monitoreo ecológico, además, los falsos positivos pueden sugerir la presencia de una especie cuando esta no está realmente presente, mientras que los falsos negativos pueden ocultar eventos acústicos relevantes (Knight et al., 2017b). Por esta razón, métricas como precisión, sensibilidad, especificidad, F1-score, AUC u otras medidas deben seleccionarse de acuerdo con el objetivo del estudio, el tipo de tarea y las consecuencias de los errores en el contexto ecológico analizado (Kershenbaum et al., 2025; Stowell et al., 2019).

El reconocimiento acústico automatizado trasciende la clasificación de especies y puede contribuir a preguntas ecológicas de mayor alcance (Saoud et al., 2024; Teixeira et al., 2024). En determinados contextos, las detecciones acústicas permiten aproximar patrones de presencia, actividad temporal, comportamiento, abundancia relativa o cambios en la composición de comunidades biológicas (Kershenbaum et al., 2025; Saoud et al., 2024). Esta utilidad requiere una distinción metodológica clara: registrar una vocalización no equivale a estimar con precisión el número de individuos ni la abundancia real de una especie (Knight et al., 2017b; Runge, 2023). Para sostener inferencias ecológicas más amplias, es necesario considerar la detectabilidad, el diseño de muestreo, la dependencia entre observaciones, la incertidumbre y la relación entre variables ambientales y respuestas biológicas (Runge, 2023; Teixeira et al., 2024).

En este marco, la presente monografía examina el uso de redes neuronales profundas en el reconocimiento acústico de especies en entornos naturales, con énfasis en los factores que condicionan su precisión y aplicabilidad (Jeantet y Dufourq, 2023; Stowell et al., 2019). El estudio adopta una perspectiva integral para analizar la relación entre arquitecturas, representaciones

acústicas, preprocesamiento, condiciones de los datos, ruido ambiental, métricas, validación y alcance ecológico de las aplicaciones propuestas, evitando una lectura reducida a listados de modelos o valores aislados de desempeño (Knight et al., 2017b; Lostanlen et al., 2024). Esta orientación es pertinente para la Ciencia de Datos porque permite interpretar el rendimiento de los modelos como resultado de un sistema completo de procesamiento y evaluación y no como una propiedad independiente de la arquitectura neuronal (Kershenbaum et al., 2025; Kiskin et al., 2020).

A partir de este problema general, la investigación se orienta mediante la siguiente pregunta central:

> **¿Cómo se han aplicado las redes neuronales profundas en el reconocimiento de especies y qué factores determinan su precisión y aplicabilidad en entornos naturales?**

Esta pregunta articula dos dimensiones del problema. La primera es metodológica y se relaciona con las arquitecturas de aprendizaje profundo, las representaciones acústicas, las estrategias de preprocesamiento y los procedimientos de evaluación empleados en la literatura reciente (Kershenbaum et al., 2025; Stowell et al., 2019). La segunda es aplicada y se refiere a la capacidad de estos modelos para sostener tareas de monitoreo en ambientes naturales caracterizados por ruido, heterogeneidad acústica y variación en la calidad de los datos (Lostanlen et al., 2024; Teixeira et al., 2024). Con el fin de delimitar el análisis, se plantean las siguientes preguntas de investigación:

- **RQ1.** ¿Qué arquitecturas de aprendizaje profundo y técnicas de preprocesamiento acústico se han empleado en la identificación de especies por audio y bajo qué configuraciones reportan mayores valores de desempeño?
- **RQ2.** ¿Cómo varía el desempeño de los modelos ante condiciones de ruido ambiental, diferencias en la calidad del conjunto de datos y heterogeneidad acústica de los hábitats?
- **RQ3.** ¿Qué métodos basados en reconocimiento acústico permiten aproximar variables

ecológicas como número de individuos o abundancia relativa de especies y qué limitaciones presentan sus estimaciones?

- **RQ4.** ¿Qué métricas, enfoques de validación y líneas de desarrollo predominan en los estudios recientes sobre reconocimiento acústico de fauna con redes neuronales profundas?

Estas preguntas cumplen una función organizadora dentro de este trabajo. La RQ1 permite analizar la relación entre arquitecturas neuronales, representaciones acústicas y técnicas de preprocesamiento. La RQ2 examina la sensibilidad de los modelos frente a ruido ambiental, variabilidad entre hábitats y diferencias en la calidad de los datos. La RQ3 evalúa hasta qué punto el reconocimiento acústico puede apoyar inferencias ecológicas más amplias, distinguiendo entre detección de vocalizaciones y estimaciones sobre número de individuos o abundancia relativa. Finalmente, la RQ4 revisa cómo se mide y valida el desempeño en la literatura, aspecto central para determinar si los resultados reportados son comparables entre estudios (Kershenbaum et al., 2025; Knight et al., 2017b).

Metodológicamente, el trabajo se apoya en una revisión sistemática de literatura centrada en estudios publicados entre 2017 y 2025 sobre reconocimiento acústico de especies en entornos naturales mediante redes neuronales profundas. A partir del corpus final seleccionado, se examinan comparativamente las arquitecturas reportadas, las representaciones acústicas utilizadas, las técnicas de preprocesamiento, las condiciones de los conjuntos de datos, el tratamiento del ruido, las métricas de evaluación, los enfoques de validación y las aplicaciones ecológicas derivadas del reconocimiento automático. Este enfoque permite construir una síntesis crítica del campo, identificar regularidades y vacíos persistentes y evitar asumir que los valores de desempeño son directamente comparables cuando provienen de condiciones experimentales distintas (Kershenbaum et al., 2025; Knight et al., 2017b; Lostanlen et al., 2024).

El documento se organiza de la siguiente manera. En el capítulo de fundamento teórico se presentan los conceptos principales relacionados con monitoreo acústico pasivo, procesamiento de señales de audio, aprendizaje profundo y evaluación de modelos. Posteriormente, en la metodología, se describe el proceso seguido para la identificación, selección y análisis de los estudios incluidos en la revisión. A continuación, los resultados y la discusión se desarrollan en

función de las preguntas de investigación planteadas, de modo que cada sección responda de manera explícita a uno de los ejes analíticos del estudio. Finalmente, las conclusiones detallan los hallazgos más relevantes, sus implicaciones metodológicas y aplicadas y las posibles líneas de trabajo futuro.

2. OBJETIVOS

Los objetivos delimitan el alcance de la revisión sistemática y organizan el análisis en tres dimensiones: técnica, evaluativa y aplicada. La dimensión técnica aborda las arquitecturas de aprendizaje profundo, las representaciones acústicas y las estrategias de preprocesamiento empleadas en los estudios revisados. La dimensión evaluativa considera las métricas, los esquemas de validación y los factores que afectan la precisión y generalización de los modelos. Finalmente, la dimensión aplicada examina el alcance ecológico del reconocimiento acústico y sus limitaciones al utilizarse para inferir presencia, actividad, número de individuos o abundancia relativa en entornos naturales.

2.1 OBJETIVO GENERAL

Analizar el estado actual, la precisión y la aplicabilidad de las redes neuronales profundas en el reconocimiento acústico de especies en entornos naturales, integrando la evidencia publicada entre 2017 y 2025 sobre el desempeño de modelos CNN y no-CNN, las metodologías empleadas, los factores que afectan su rendimiento y las principales tendencias de desarrollo.

2.2 OBJETIVOS ESPECÍFICOS

Los objetivos específicos que orientan esta revisión sistemática buscan analizar las metodologías, el desempeño y las aplicaciones del reconocimiento acústico de especies mediante redes neuronales profundas, considerando tanto los aspectos técnicos como los factores que influyen en su precisión y utilidad en entornos naturales.

- Examinar las arquitecturas de redes neuronales profundas y los métodos de preprocesamiento acústico que reportan mayores niveles de desempeño en la identificación de especies por audio.
- Evaluar cómo varía la precisión de los modelos ante diferentes niveles de ruido ambiental, calidad del conjunto de datos y heterogeneidad acústica de los hábitats.
- Comparar los métodos de reconocimiento acústico empleados para estimar número de

individuos o abundancia relativa, identificando sus alcances y limitaciones.

- Identificar las métricas, enfoques de validación y tendencias metodológicas predominantes en las investigaciones recientes sobre reconocimiento acústico con redes neuronales profundas.

3. FUNDAMENTO TEÓRICO

Este capítulo presenta los fundamentos que permiten entender el reconocimiento acústico de especies como un flujo completo de datos. El punto de partida es el monitoreo acústico pasivo, como una fuente de registros ambientales obtenidos en campo. A partir de esas grabaciones, se revisa cómo la señal sonora se transforma en representaciones tiempo–frecuencia, cómo estas representaciones pueden ser procesadas por modelos de aprendizaje automático y profundo y qué condiciones afectan la precisión cuando el sistema se aplica en entornos naturales. Esta base teórica permite interpretar los resultados de la revisión desde una lógica de pipeline, donde la captura del audio, el preprocesamiento, la arquitectura del modelo, las métricas de evaluación y el contexto ecológico intervienen conjuntamente en la confiabilidad del reconocimiento automático.

3.1 MONITOREO ACÚSTICO PASIVO Y PROBLEMÁTICA DEL RECONOCIMIENTO AUTOMÁTICO

El monitoreo acústico pasivo (*Passive Acoustic Monitoring*, PAM) permite registrar señales sonoras de forma continua mediante sensores autónomos instalados en entornos naturales (Kershenbaum et al., 2025; Teixeira et al., 2024). Su valor metodológico radica en que amplía la capacidad de observación ecológica en contextos donde el trabajo de campo directo está limitado por costos, accesibilidad, restricciones temporales o por el comportamiento esquivo de las especies monitoreadas (Teixeira et al., 2024). Al conservar registros acústicos revisables, el PAM aporta evidencia sobre presencia, actividad vocal y patrones temporales de especies sonoras, siempre que los datos generados puedan procesarse mediante procedimientos analíticos consistentes (Kershenbaum et al., 2025; Teixeira et al., 2024).

El principal desafío del PAM surge del volumen y la complejidad de las grabaciones producidas. Una campaña extensa puede generar archivos acústicos continuos que exceden la capacidad de revisión manual, lo que exige métodos automáticos para detectar, clasificar o filtrar eventos sonoros relevantes (LeBien et al., 2020; Stowell et al., 2019). En este marco, el reconocimiento acústico automatizado opera como una cadena de análisis que transforma señales crudas en representaciones numéricas, predicciones y resultados interpretables para el monitoreo ecológico (Kershenbaum et al., 2025; Knight et al., 2017b). Su desempeño depende de

la interacción entre la calidad del registro, el procesamiento previo, la representación acústica y el modelo de clasificación, no únicamente de la arquitectura algorítmica utilizada.

Las grabaciones obtenidas en campo no contienen únicamente la vocalización de una especie. Forman parte de paisajes sonoros complejos, conocidos como *soundscapes*, donde coinciden sonidos biológicos, geofísicos y antrópicos (Kershenbaum et al., 2025; Teixeira et al., 2024). Esta mezcla acústica dificulta separar la señal objetivo del fondo ambiental, especialmente cuando varias fuentes sonoras se superponen en el tiempo o comparten rangos de frecuencia similares (Knight et al., 2017b; Lostanlen et al., 2019). Una misma especie también puede producir registros acústicos variables según el contexto de grabación, la distancia al sensor, la intensidad vocal, el hábitat o la presencia de otras fuentes sonoras, lo que reduce la separabilidad entre clases y aumenta la dificultad de aprender patrones discriminativos estables (Lostanlen et al., 2019; Stowell et al., 2019).

La generalización constituye uno de los problemas centrales de estos sistemas, entendida como la capacidad de mantener un desempeño confiable cuando el modelo se aplica a datos distintos de aquellos usados durante el entrenamiento (Goodfellow et al., 2016; Knight et al., 2017b). En aplicaciones reales, las condiciones acústicas pueden cambiar entre el conjunto de entrenamiento y el conjunto de evaluación, situación que puede representarse como

$$P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y), \tag{3.1}$$

donde $P_{\text{train}}(x, y)$ representa la distribución conjunta de entradas acústicas xy etiquetas y observada durante el entrenamiento, mientras que $P_{\text{test}}(x, y)$ corresponde a la distribución encontrada durante la evaluación o aplicación del modelo (Goodfellow et al., 2016). Esta desigualdad formaliza el cambio de distribución que puede aparecer cuando el sistema enfrenta condiciones acústicas distintas a las del entrenamiento, sin asumir que se trata de una medición directa de un estudio específico (Knight et al., 2017b; Lostanlen et al., 2019). En consecuencia, un modelo con alto desempeño en condiciones controladas no necesariamente conserva la misma precisión en escenarios naturales no observados durante su ajuste (Lostanlen et al., 2019; Stowell et al., 2019).

La confiabilidad del reconocimiento acústico automático depende de la interacción entre

el dato, su procesamiento y el contexto de evaluación. Una predicción no adquiere validez ecológica únicamente por provenir de una arquitectura profunda; requiere grabaciones con calidad suficiente, segmentaciones coherentes, representaciones acústicas informativas, etiquetas consistentes y protocolos de validación capaces de reflejar condiciones naturales distintas del entrenamiento (Kershenbaum et al., 2025; LeBien et al., 2020). Cuando estos componentes son débiles, incluso modelos avanzados pueden producir resultados inestables ante ruido ambiental, desbalance de clases o cambios en la distribución acústica del hábitat (Knight et al., 2017b; Lostanlen et al., 2019).

3.2 REPRESENTACIÓN Y PROCESAMIENTO DE SEÑALES ACÚSTICAS

El reconocimiento automático de especies requiere transformar la señal acústica cruda en una representación estructurada que pueda ser procesada por un modelo computacional (Oppenheim y Schafer, 2009; Stowell et al., 2019). Una grabación en el dominio temporal describe la variación de amplitud de la señal, pero no muestra directamente cómo se distribuye la energía entre frecuencias ni cómo esa distribución cambia durante una vocalización (Oppenheim y Schafer, 2009). Por esta razón, el procesamiento tiempo–frecuencia resulta central en bioacústica, especialmente cuando las señales contienen modulaciones, pulsos o cambios rápidos de frecuencia (Kiskin et al., 2020; Stowell et al., 2019).

Dentro de estas representaciones, el espectrograma permite observar cómo se distribuye la energía del sonido en el tiempo y en la frecuencia; por ello, resulta útil para identificar patrones espectrotemporales generales de cantos, llamados o pulsos acústicos (Oppenheim y Schafer, 2009; Stowell et al., 2019). El mel-espectrograma parte de esta misma lógica, pero reorganiza el eje de frecuencia mediante la escala Mel, generando una representación más compacta de la energía acústica en bandas perceptualmente informativas (Davis y Mermelstein, 1980; LeBien et al., 2020). Las transformadas *wavelet*, en cambio, analizan la señal con resolución variable, lo que permite capturar con mayor detalle eventos breves, transitorios o cambios rápidos de frecuencia (Kiskin et al., 2020). Estas diferencias son relevantes porque cada representación determina qué rasgos de la señal quedan disponibles para el modelo durante el entrenamiento y la predicción.

Desde la perspectiva del aprendizaje estadístico, el reconocimiento automático de especies

puede formularse como un problema supervisado en el que el modelo aprende una relación entre entradas acústicas y etiquetas conocidas a partir de ejemplos observados durante el entrenamiento (Goodfellow et al., 2016). Sea un conjunto de datos

$$(X_i, Y_i)_{i=1}^{N}, \tag{3.2}$$

donde N representa el número total de ejemplos disponibles para entrenamiento o evaluación, X_i corresponde a la entrada acústica del ejemplo i y Y_i denota la etiqueta asociada a dicha entrada (Goodfellow et al., 2016). En reconocimiento bioacústico, X_i suele representarse como una matriz $X_i \in \mathbb{R}^{T \times F}$, donde T indica los intervalos temporales de análisis y F las bandas o componentes frecuenciales extraídas de la señal (Oppenheim y Schafer, 2009; Stowell et al., 2019). La etiqueta Y_i puede corresponder a una especie, un evento acústico, una clase de presencia o ausencia, o una categoría multietiqueta cuando varias fuentes sonoras aparecen en un mismo fragmento (Knight et al., 2017b; LeBien et al., 2020).

Cada observación del conjunto de datos constituye un par supervisado (X_i, Y_i), compuesto por una representación acústica y una respuesta conocida. La entrada X_i contiene el patrón sonoro que será procesado por el modelo, mientras que Y_i define la categoría biológica asociada al fragmento. En la práctica, un canto o llamado puede transformarse en un espectrograma cuyas filas representan bandas de frecuencia, cuyas columnas corresponden a intervalos temporales y cuyos valores reflejan la intensidad de la señal acústica. El aprendizaje supervisado consiste en ajustar los parámetros del modelo a partir de múltiples pares (X_i, Y_i), de modo que pueda reconocer asociaciones entre configuraciones acústicas y etiquetas en registros no observados durante el entrenamiento.

Esta transformación desplaza el reconocimiento acústico desde la percepción auditiva hacia el análisis computacional de la señal. El modelo no procesa el sonido como lo haría un observador humano; opera sobre una representación matemática que codifica propiedades físicas del audio (Goodfellow et al., 2016; Oppenheim y Schafer, 2009). Al expresar la grabación como una matriz tiempo–frecuencia, el análisis se centra en la distribución de energía entre intervalos temporales y bandas frecuenciales, dejando en segundo plano la variación directa de presión sonora en el tiempo (Oppenheim y Schafer, 2009; Stowell et al., 2019). Bajo esta representación, muchas vocalizaciones exhiben patrones espectrotemporales distintivos que pueden separarse

con mayor nitidez en el espectrograma que en la onda cruda (Kiskin et al., 2020; Stowell et al., 2019).

Desde una perspectiva formal, una señal acústica $x(t)$ puede analizarse en el dominio de la frecuencia mediante la Transformada de Fourier, que permite expresar una señal temporal como una combinación de componentes sinusoidales de distintas frecuencias (Oppenheim y Schafer, 2009). La energía espectral asociada a la señal puede representarse como

$$X(f) = |\mathcal{F}x(t)|^2, \tag{3.3}$$

donde $\mathcal{F}\cdot$ representa la Transformada de Fourier, f corresponde a la frecuencia analizada y $X(f)$ expresa la energía relativa de la señal en cada componente frecuencial (Oppenheim y Schafer, 2009). Esta representación permite identificar qué frecuencias dominan en una señal, aunque pierde información sobre el momento exacto en que esas frecuencias aparecen (Oppenheim y Schafer, 2009).

En bioacústica, esta pérdida temporal es problemática porque las vocalizaciones suelen tener estructura dinámica. Un canto, llamado o pulso acústico puede incluir variaciones rápidas de frecuencia, pausas, modulaciones o patrones repetitivos que no se describen adecuadamente mediante un espectro global de toda la grabación (Kiskin et al., 2020). Por esta razón, se utiliza con frecuencia la Transformada de Fourier de Tiempo Corto (*Short-Time Fourier Transform*, STFT), que divide la señal en ventanas temporales sucesivas y calcula el contenido frecuencial dentro de cada una de ellas (Oppenheim y Schafer, 2009). Conceptualmente, la STFT permite observar tanto las frecuencias presentes en la señal como el momento en que aparecen (Oppenheim y Schafer, 2009).

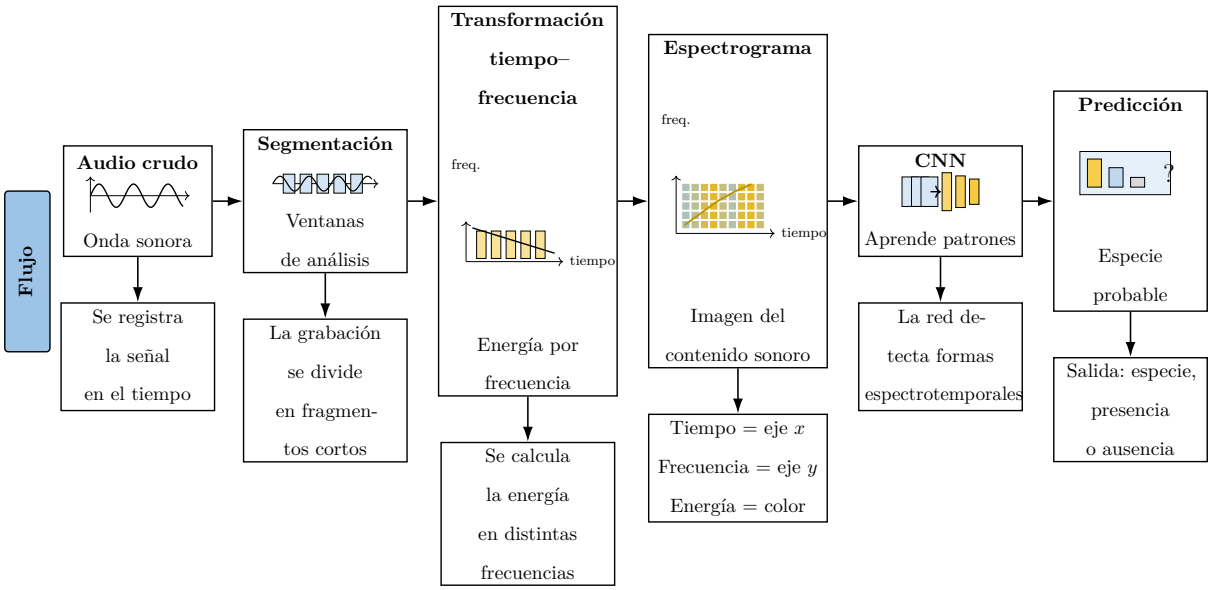
El resultado de este procedimiento es el espectrograma, una representación bidimensional donde el eje horizontal corresponde al tiempo, el eje vertical corresponde a la frecuencia y la intensidad o color representa la energía de la señal en cada punto tiempo–frecuencia (Oppenheim y Schafer, 2009; Stowell et al., 2019). Esta representación es especialmente útil para redes neuronales convolucionales porque convierte el audio en una estructura visual donde las vocalizaciones pueden aparecer como trazos, bandas, pulsos, curvas o formas espectrotemporales características (Kiskin et al., 2020; Stowell et al., 2019). En este sentido, el modelo trabaja con una matriz

compatible con operaciones de aprendizaje sobre imágenes, en lugar de procesar la onda sonora como experiencia auditiva directa (Goodfellow et al., 2016; Kahl et al., 2021).

La Figura 3.1 resume esta transformación conceptual desde la señal cruda hasta la predicción automática. La primera etapa representa el audio como una onda temporal; la segunda divide la grabación en fragmentos analizables; la tercera calcula la distribución de energía por frecuencia; la cuarta organiza esa información como espectrograma; la quinta introduce el modelo convolucional; y la sexta representa la salida del sistema, que puede expresarse como especie probable, presencia o ausencia, o probabilidad asociada a una clase (Goodfellow et al., 2016; Oppenheim y Schafer, 2009; Stowell et al., 2019). Esta secuencia muestra que la predicción final surge de una cadena de decisiones de procesamiento que condicionan la información disponible para el modelo, más que del sonido crudo tomado de forma directa (Kershenbaum et al., 2025; Knight et al., 2017b).

Figura 3.1

Flujo de transformación de una señal acústica cruda hacia el reconocimiento automático de especies mediante CNN



Nota. La figura representa el proceso mediante el cual una señal acústica cruda se segmenta, se transforma en una representación tiempo–frecuencia y se analiza mediante una red neuronal convolucional para obtener una predicción asociada a especie, presencia o ausencia. Elaboración propia.

En muchos sistemas bioacústicos, el espectrograma se transforma posteriormente a escala

Mel para generar mel-espectrogramas (Davis y Mermelstein, 1980; LeBien et al., 2020). Esta escala reorganiza las frecuencias de manera no lineal, comprime las frecuencias altas y otorga mayor resolución relativa a frecuencias bajas, lo que permite representar la energía acústica en bandas perceptualmente informativas (Davis y Mermelstein, 1980). La transformación clásica puede expresarse como

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f_{\text{Hz}}}{700}\right), \tag{3.4}$$

donde f_{Hz} representa la frecuencia original medida en Hertz y f_{mel} representa su valor transformado en la escala Mel (Davis y Mermelstein, 1980). El parámetro 700 Hz define la región en la que la representación pasa gradualmente de un comportamiento aproximadamente lineal a uno logarítmico, mientras que la constante 2595 normaliza la escala para que una frecuencia de 1000 Hz corresponda aproximadamente a 1000 mels, preservando la referencia perceptual utilizada en los experimentos psicoacústicos originales. En términos generales, la operación logarítmica comprime el crecimiento de las frecuencias físicas, de modo que el modelo recibe una representación más compacta de la distribución energética del sonido (Davis y Mermelstein, 1980; LeBien et al., 2020).

En reconocimiento bioacústico, los mel-espectrogramas transforman una grabación en una matriz bidimensional donde un eje representa el tiempo y el otro las bandas de frecuencia en escala Mel. Cada celda indica cuánta energía acústica se concentra en una banda específica durante un intervalo breve de la señal (Kahl et al., 2021; LeBien et al., 2020). Aunque cada mel-espectrograma puede entenderse como una matriz, durante el entrenamiento suele organizarse dentro de tensores que agrupan múltiples ejemplos o canales de entrada para ser procesados por la CNN. Esta representación permite que el modelo trate el sonido como una imagen y detecte patrones espectro-temporales asociados con las vocalizaciones de distintas especies.

LeBien et al. (2020) emplean mel-espectrogramas en un sistema de clasificación multietiqueta para sonidos de aves y anfibios, mientras que Kahl et al. (2021) muestran que la resolución temporal de la representación acústica modifica la capacidad del modelo para reconocer vocalizaciones aviares. Esta influencia ocurre porque la resolución temporal define qué tan finamente se divide la señal en el eje del tiempo: una resolución baja puede suavizar o perder notas breves, cambios rápidos de frecuencia y estructuras silábicas cortas; una resolución más

detallada conserva mejor esos rasgos, aunque también puede aumentar la sensibilidad al ruido y a variaciones pequeñas de la grabación. Por ello, la representación de entrada condiciona los rasgos acústicos que el modelo puede aprender durante el entrenamiento y utilizar durante la predicción (Kiskin et al., 2020; Knight et al., 2017b).

Además de los mel-espectrogramas, algunos estudios han explorado representaciones tiempo–frecuencia alternativas o complementarias para capturar distintos componentes de la señal acústica (Kahl et al., 2021; Kiskin et al., 2020). Las transformadas *wavelet* permiten analizar patrones con resolución variable en tiempo y frecuencia, lo que resulta útil cuando las señales presentan estructuras breves, transitorias o distribuidas en bandas específicas (Kiskin et al., 2020). Otros enfoques separan componentes armónicos y percusivos o combinan varias representaciones acústicas para proporcionar al modelo descripciones complementarias del mismo evento sonoro (Kahl et al., 2021). Esta diversidad confirma que el preprocesamiento acústico forma parte del diseño metodológico del sistema y no debe tratarse como una operación puramente automática (Kershenbaum et al., 2025; Knight et al., 2017b).

Antes del uso extendido de redes neuronales profundas, el reconocimiento acústico se apoyaba con frecuencia en características diseñadas manualmente, como los coeficientes cepstrales en frecuencia Mel (*Mel-Frequency Cepstral Coefficients*, MFCC) (Davis y Mermelstein, 1980). Los MFCC resumen la envolvente espectral de una señal después de aplicar filtros en escala Mel y una transformación cepstral, lo que permite obtener descriptores compactos del contenido acústico (Davis y Mermelstein, 1980). Aunque los enfoques actuales basados en CNN privilegian el aprendizaje automático de representaciones, estas características siguen siendo relevantes porque muestran la continuidad entre el procesamiento clásico de audio y las representaciones modernas usadas en aprendizaje profundo (Goodfellow et al., 2016; Stowell et al., 2019).

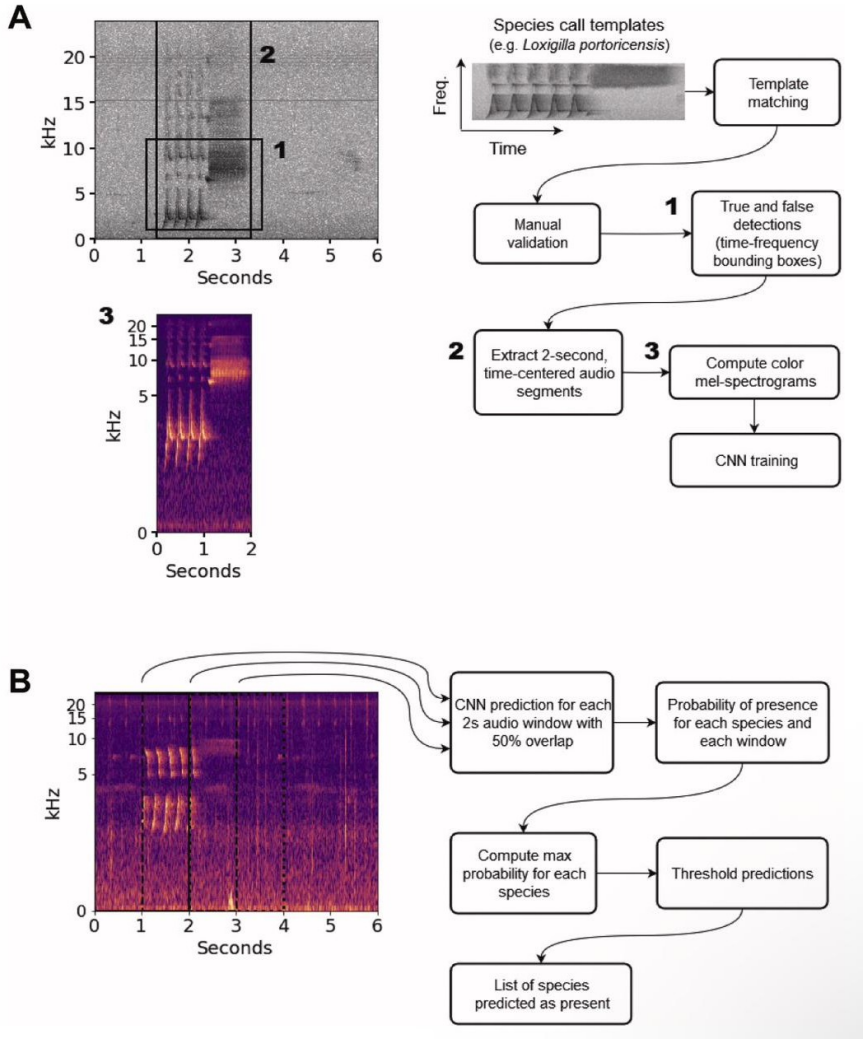
El rendimiento de los modelos depende también de la forma en que se construye el conjunto de datos (Kershenbaum et al., 2025; Knight et al., 2017b). LeBien et al. (2020) proponen un flujo donde primero se emplea coincidencia con plantillas acústicas, luego se validan manualmente detecciones verdaderas y falsas, después se extraen segmentos de audio centrados en el evento y finalmente se calculan mel-espectrogramas en color para entrenar una CNN. Este proceso evidencia que el entrenamiento parte de una cadena de preparación donde cada ejemplo

debe ser detectado, validado, recortado y transformado antes de ingresar al modelo, en lugar de utilizar grabaciones crudas sin estructura (Knight et al., 2017b; LeBien et al., 2020).

La Figura 3.2 presenta el pipeline reportado por LeBien et al. (2020) y permite distinguir dos fases metodológicas. En la sección A de la figura, se construye el conjunto de entrenamiento: las plantillas de llamadas de especies se usan para localizar posibles eventos acústicos, las detecciones se validan manualmente, se extraen segmentos de dos segundos centrados en el evento y se calculan mel-espectrogramas que sirven como entrada para la CNN (LeBien et al., 2020). En la sección B, el modelo entrenado se aplica sobre ventanas de dos segundos con (50 %) de solapamiento, calcula probabilidades de presencia por especie y por ventana, resume la probabilidad máxima y aplica umbrales para producir una lista de especies predichas como presentes (LeBien et al., 2020).

Figura 3.2

Flujo de procesamiento acústico aplicado al reconocimiento automático de especies



Nota. La figura muestra el proceso de transformación de una señal acústica en una representación útil para el reconocimiento automático de especies mediante redes neuronales convolucionales. Adaptado de LeBien et al. (2020).

La segmentación en ventanas de duración fija permite que la CNN reciba matrices comparables durante el entrenamiento y la predicción (Goodfellow et al., 2016; LeBien et al., 2020). El solapamiento entre ventanas reduce la posibilidad de perder eventos ubicados cerca de los bordes de los fragmentos, mientras que los umbrales convierten salidas probabilísticas en decisiones discretas de presencia o ausencia (LeBien et al., 2020). Esta etapa debe evaluarse cuidadosamente porque modifica la relación entre falsos positivos y falsos negativos (Knight et al., 2017b; LeBien et al., 2020).

El procesamiento acústico determina cómo la señal sonora se convierte en una entrada interpretable para los modelos de reconocimiento. Las transformaciones tiempo–frecuencia reorganizan la energía del audio en matrices donde pueden observarse patrones espectrotemporales asociados con las vocalizaciones, mientras que decisiones como la segmentación, la escala Mel, la definición de ventanas y la normalización modifican la resolución, estabilidad y comparabilidad de esas representaciones (Davis y Mermelstein, 1980; Kiskin et al., 2020; Lostanlen et al., 2019; Oppenheim y Schafer, 2009). En grabaciones de campo, la energía del fondo sonoro varía entre sensores, sitios, horarios y condiciones meteorológicas; por ello, una misma especie puede registrarse con intensidades relativas distintas según el contexto de captura (Lostanlen et al., 2019; Teixeira et al., 2024). Procedimientos como la normalización de amplitud, la reducción de ruido o la normalización energética pueden estabilizar las entradas usadas por el modelo, pero también atenuar rasgos acústicos relevantes cuando se aplican sin considerar la estructura de la señal objetivo (Knight et al., 2017b; Lostanlen et al., 2019). Por esta razón, la evaluación de modelos no debe centrarse únicamente en la arquitectura empleada, ya que la representación acústica y las etapas de preprocesamiento condicionan qué información queda disponible antes del entrenamiento y la validación (Kershenbaum et al., 2025; Knight et al., 2017b). La normalización debe entenderse, entonces, como una decisión metodológica que puede modificar la separabilidad entre clases, no como una operación neutral de limpieza previa al entrenamiento (Kershenbaum et al., 2025; Kiskin et al., 2020).

3.3 MODELOS DE APRENDIZAJE AUTOMÁTICO Y PROFUNDO PARA RECONOCIMIENTO ACÚSTICO

El reconocimiento acústico de especies ha evolucionado desde enfoques basados en características diseñadas manualmente hacia modelos de aprendizaje profundo capaces de aprender representaciones directamente a partir de los datos (Goodfellow et al., 2016; Stowell et al., 2019). Esta evolución responde a la necesidad de procesar grandes volúmenes de audio y detectar patrones acústicos en condiciones donde el análisis manual resulta poco escalable (Kershenbaum et al., 2025; Teixeira et al., 2024). Sin embargo, el aumento de complejidad del modelo no elimina la necesidad de evaluar la calidad de los datos, la representación de entrada, las métricas usadas y la capacidad de generalización en ambientes naturales (Knight et al., 2017b; Lostanlen et al., 2019).

3.3.1 ENFOQUES BASADOS EN APRENDIZAJE AUTOMÁTICO

Los métodos tradicionales de aprendizaje automático dependen en gran medida de la extracción manual de características acústicas, es decir, de descriptores definidos previamente para resumir propiedades temporales, espectrales o cepstrales de la señal (Davis y Mermelstein, 1980; Goodfellow et al., 2016). Estas características pueden incluir MFCC, energía espectral, distribución de frecuencia, duración de eventos, tasa de cruce por cero o medidas derivadas del espectro, dependiendo de la tarea y del tipo de señal analizada (Davis y Mermelstein, 1980; Knight et al., 2017b). La finalidad de estos descriptores es transformar la señal acústica en un conjunto de variables numéricas que puedan ser usadas por clasificadores supervisados (Goodfellow et al., 2016).

Sobre estas representaciones pueden aplicarse modelos como k-vecinos más cercanos, máquinas de soporte vectorial, árboles de decisión, bosques aleatorios o redes neuronales clásicas (Goodfellow et al., 2016). Estos métodos pueden funcionar adecuadamente cuando las características seleccionadas capturan diferencias discriminativas entre clases y cuando las condiciones de entrenamiento y evaluación son relativamente consistentes (Goodfellow et al., 2016; Knight et al., 2017b). No obstante, su desempeño depende de la calidad de la ingeniería de características, porque el modelo solo puede aprender a partir de la información que esos descriptores logran representar (Goodfellow et al., 2016; Kershenbaum et al., 2025).

La principal limitación de estos enfoques es que las características manuales pueden no generalizar bien en paisajes sonoros complejos (Knight et al., 2017b; Lostanlen et al., 2019). En grabaciones naturales, las vocalizaciones pueden superponerse con ruido ambiental, variar entre individuos o cambiar por condiciones de hábitat, distancia al micrófono y tipo de sensor (Lostanlen et al., 2019; Teixeira et al., 2024). Si los descriptores fueron diseñados para un conjunto acústico específico, pueden perder capacidad discriminativa cuando se aplican a nuevas especies, sitios o condiciones ambientales (Kershenbaum et al., 2025; Knight et al., 2017b). Por esta razón, los enfoques tradicionales son relevantes como antecedente metodológico, pero presentan restricciones cuando el objetivo es construir sistemas robustos para monitoreo ecológico en entornos heterogéneos (Lostanlen et al., 2019; Stowell et al., 2019).

3.3.2 APRENDIZAJE PROFUNDO Y MODELOS AVANZADOS

El aprendizaje profundo ha reducido la dependencia de la ingeniería manual de características al permitir que redes neuronales con múltiples capas aprendan representaciones jerárquicas a partir de los datos (Goodfellow et al., 2016). En reconocimiento bioacústico, esta capacidad es especialmente útil cuando el audio se transforma en espectrogramas, mel-espectrogramas u otras matrices tiempo–frecuencia, porque esas representaciones permiten que el modelo detecte patrones locales asociados con vocalizaciones (Kiskin et al., 2020; Stowell et al., 2019). Las redes neuronales convolucionales, conocidas como *Convolutional Neural Networks* (CNN), han sido ampliamente usadas en este contexto porque aplican filtros locales que identifican patrones espaciales y los combinan progresivamente en representaciones más abstractas (Goodfellow et al., 2016; Kahl et al., 2021).

El proceso general de predicción de una red neuronal puede representarse como,

$$\hat{y} = f_\theta(X)\,, \tag{3.5}$$

donde X corresponde a la representación de entrada, f_θ representa la función aprendida por la red neuronal y θ agrupa los parámetros entrenables del modelo, como pesos y sesgos (Goodfellow et al., 2016). La salida $\hat{y}$ puede interpretarse como una clase predicha, una puntuación de confianza o un vector de probabilidades asociado a las clases posibles, dependiendo de si la tarea es binaria, multiclase o multietiqueta (Goodfellow et al., 2016; LeBien et al., 2020). Esta formulación resume que el modelo ajusta parámetros para aproximar una relación entre representaciones acústicas y etiquetas observadas durante el entrenamiento, en lugar de memorizar reglas explícitas escritas por el investigador (Goodfellow et al., 2016).

Para entrenar modelos de clasificación se emplean funciones de pérdida que cuantifican la diferencia entre la etiqueta real y la predicción del modelo (Goodfellow et al., 2016). Una función habitual en clasificación multiclase es la entropía cruzada, que puede expresarse como,

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)\,, \tag{3.6}$$

donde C representa el número total de clases, y_i indica si la clase i corresponde a la etiqueta verdadera y $\hat{y}_i$ representa la probabilidad estimada por el modelo para esa clase (Goodfellow

et al., 2016). El signo negativo aparece porque el logaritmo de una probabilidad entre 0 y 1 es menor o igual que cero, de modo que la pérdida se expresa como un valor no negativo que aumenta cuando el modelo asigna baja probabilidad a la clase correcta (Goodfellow et al., 2016). En términos intuitivos, esta función penaliza con mayor intensidad las predicciones confiadas pero equivocadas, lo que la vuelve adecuada para ajustar modelos probabilísticos de clasificación (Goodfellow et al., 2016).

Los modelos basados en CNN son ampliamente utilizados en clasificación y detección bioacústica porque pueden aprender patrones espectrotemporales directamente desde representaciones tiempo–frecuencia, con menor dependencia de características acústicas definidas manualmente (Kahl et al., 2021; Stowell et al., 2019). En un espectrograma, las vocalizaciones aparecen como bandas, pulsos, modulaciones o trazos de energía; las capas convolucionales detectan estas formas locales y las combinan progresivamente en patrones más complejos asociados con especies o eventos acústicos (Goodfellow et al., 2016; Kiskin et al., 2020). Esta capacidad explica el uso frecuente de espectrogramas como entrada visual para el clasificador (LeBien et al., 2020; Stowell et al., 2019). Aun así, el rendimiento no depende únicamente de la arquitectura: factores como la calidad de los datos, la resolución del espectrograma, la segmentación, el balance entre clases y el procedimiento de validación condicionan la estabilidad del modelo (Kershenbaum et al., 2025; Knight et al., 2017b). Aunque la literatura reciente incorpora arquitecturas híbridas, modelos recurrentes, mecanismos de atención y transformadores para capturar dependencias temporales más extensas o relaciones distribuidas en la señal, las CNN conservan un papel central por su capacidad para explotar de manera eficiente la estructura bidimensional de los espectrogramas (Goodfellow et al., 2016; Kahl et al., 2021; Kershenbaum et al., 2025; Stowell et al., 2019).

3.3.3 MODELOS PREENTRENADOS Y APRENDIZAJE POR TRANSFERENCIA

En el estado del arte destacan modelos preentrenados de gran escala, como BirdNET, diseñados para reconocer miles de especies de aves a partir de vocalizaciones (Kahl et al., 2021; Pérez-Granados, 2023). Estos modelos permiten analizar grandes volúmenes de grabaciones y ofrecen una base práctica para aplicaciones de monitoreo ecológico en diferentes regiones, aunque

su desempeño puede variar según la cobertura taxonómica, la calidad del audio y las condiciones locales de grabación (Kershenbaum et al., 2025; Pérez-Granados, 2023). La utilidad de este tipo de modelos se explica por su entrenamiento previo sobre grandes colecciones de datos, lo que les permite aprender representaciones generales de señales acústicas antes de ser aplicados en un contexto particular (Goodfellow et al., 2016; Kahl et al., 2021).

El aprendizaje por transferencia consiste en reutilizar conocimiento aprendido en una tarea o dominio amplio para mejorar el desempeño en una tarea o dominio más específico (Goodfellow et al., 2016). En bioacústica, se hace uso de esta estrategia dado que muchas especies, regiones o hábitats tienen pocos datos etiquetados disponibles, lo que dificulta entrenar modelos profundos desde cero (Schiavo et al., 2025; Ziegenhorn et al., 2025). Mediante ajuste fino, conocido como *fine-tuning*, un modelo preentrenado puede actualizar parte de sus parámetros con datos locales para adaptarse mejor a especies o condiciones acústicas particulares (Goodfellow et al., 2016; Schiavo et al., 2025). Esta idea permite aprovechar representaciones generales ya aprendidas y reducir la cantidad de datos nuevos necesarios para obtener un sistema funcional (Ghani et al., 2023; Ziegenhorn et al., 2025).

No obstante, el uso de modelos preentrenados no elimina la necesidad de validación local (Knight et al., 2017b; Pérez-Granados, 2023). Un modelo entrenado con datos de ciertas regiones, especies o condiciones acústicas puede presentar sesgos cuando se aplica en un nuevo entorno, especialmente si cambian el paisaje sonoro, el sensor, la comunidad biológica o la intensidad del ruido (Ghani et al., 2025; Lostanlen et al., 2019). Por ello, el ajuste fino, la calibración de umbrales y la evaluación con datos locales son pasos necesarios para evitar que la aparente escalabilidad del modelo oculte errores de transferencia (Knight et al., 2017b; Schiavo et al., 2025). En términos metodológicos, un modelo preentrenado debe considerarse un punto de partida robusto, no una garantía automática de validez ecológica en cualquier escenario (Kershenbaum et al., 2025; Pérez-Granados, 2023).

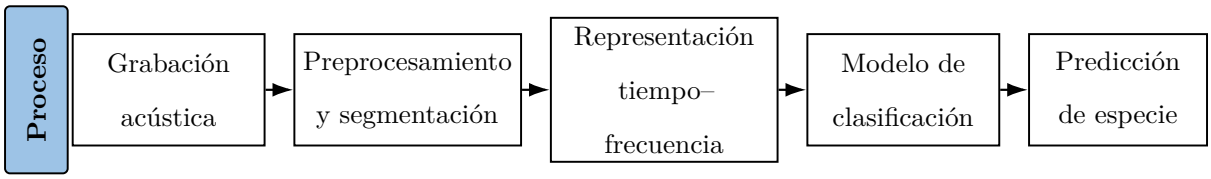
3.3.4 PIPELINE DE RECONOCIMIENTO ACÚSTICO

El reconocimiento acústico automatizado puede organizarse como un flujo de etapas interdependientes que conecta la grabación en campo con la predicción ecológicamente inter-

pretable (Kershenbaum et al., 2025; LeBien et al., 2020). Primero se registra la señal acústica mediante sensores; luego, la grabación se preprocesa y segmenta; posteriormente, se transforma en una representación tiempo–frecuencia; después, se entrena o aplica un modelo de clasificación; finalmente, se obtiene una predicción asociada a una especie, evento sonoro o probabilidad de presencia (LeBien et al., 2020; Stowell et al., 2019). Este flujo se sintetiza en la Figura 3.3.

Figura 3.3

Flujo general de reconocimiento acústico automático de especies



Nota. Elaboración propia con base en (Goodfellow et al., 2016; Kershenbaum et al., 2025; Knight et al., 2017b; LeBien et al., 2020).

La Figura 3.3 divide el proceso en cinco etapas porque cada una introduce decisiones metodológicas distintas (Kershenbaum et al., 2025; Knight et al., 2017b). La grabación determina la calidad inicial de la señal; el preprocesamiento define qué ruido se atenúa o qué fragmentos se conservan; la representación tiempo–frecuencia determina qué patrones serán visibles para el modelo; el clasificador aprende relaciones entre entradas y etiquetas; y la predicción traduce la salida del modelo en una categoría o probabilidad ecológicamente interpretable (Goodfellow et al., 2016; LeBien et al., 2020). Por esta razón, una falla en cualquiera de estas etapas puede afectar el desempeño final, incluso cuando el modelo usado sea técnicamente avanzado (Knight et al., 2017b; Lostanlen et al., 2019).

3.3.5 EVALUACIÓN DEL DESEMPEÑO

La evaluación del desempeño en reconocimiento acústico debe ajustarse a la tarea analizada, la distribución de clases y las consecuencias ecológicas asociadas a cada tipo de error (Kershenbaum et al., 2025; Knight et al., 2017b). En clasificación bioacústica, un falso positivo ocurre cuando el sistema predice la presencia de una especie o evento acústico que no está realmente presente, mientras que un falso negativo corresponde a una vocalización o evento real que el modelo no detecta (Knight et al., 2017b). Esta diferencia tiene implicaciones directas para

la interpretación ecológica: los falsos positivos pueden inflar estimaciones de presencia, mientras que los falsos negativos pueden ocultar actividad biológica relevante y subestimar la ocurrencia de especies en el paisaje sonoro (Knight et al., 2017b; Teixeira et al., 2024).

Una métrica ampliamente utilizada es el F1-score, que combina precisión y sensibilidad o *recall* mediante una media armónica (Goodfellow et al., 2016; Knight et al., 2017b). La precisión mide qué proporción de las detecciones positivas del modelo son correctas, mientras que el *recall* mide qué proporción de los eventos reales positivos fueron detectados (Knight et al., 2017b). El F1-score puede expresarse como,

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3.7}$$

donde el numerador Precision·Recall combina ambas dimensiones de desempeño y el denominador Precision + Recall normaliza la relación entre ellas (Goodfellow et al., 2016; Knight et al., 2017b). El factor 2 aparece porque el F1-score corresponde a la media armónica de dos cantidades, lo que penaliza valores muy bajos en cualquiera de las dos métricas e impide que un resultado alto en una compense completamente un resultado bajo en la otra (Goodfellow et al., 2016). Sobre todo cuando la data se encuentra en contextos con desbalance de clases o cuando se busca equilibrar falsas detecciones, asociadas con errores de comisión y detecciones perdidas, asociadas con errores de omisión (Knight et al., 2017b).

En modelos probabilísticos también se emplea el área bajo la curva ROC, conocida como AUC, para evaluar la capacidad del modelo de separar clases a través de distintos umbrales de decisión (Fawcett, 2006). La curva ROC relaciona la tasa de verdaderos positivos (*True Positive Rate*, TPR) con la tasa de falsos positivos (*False Positive Rate*, FPR), y el AUC puede expresarse como:

$$AUC = \int_0^1 TPR(FPR), d(FPR), \tag{3.8}$$

donde (TPR(FPR)) representa la tasa de verdaderos positivos observada para cada nivel de falsos positivos, y (d(FPR)) indica que el área se calcula acumulando el desempeño a lo largo del eje de falsos positivos (Fawcett, 2006). De forma intuitiva, el AUC indica qué tan bien el modelo asigna puntajes más altos a los registros donde la especie está presente que a los registros donde está ausente. Un valor cercano a (1) refleja una separación alta entre clases, mientras que un valor cercano a (0.5) sugiere un comportamiento similar al azar. Esta métrica resume el

desempeño del clasificador a través de múltiples umbrales, en lugar de limitar la evaluación a un punto fijo de decisión (Fawcett, 2006). En bioacústica, esta característica es relevante porque los umbrales de detección pueden ajustarse según se priorice reducir falsos positivos o minimizar falsos negativos (Kershenbaum et al., 2025; Knight et al., 2017b).

Las métricas de desempeño adquieren sentido únicamente cuando se interpretan en relación con la distribución de clases, el objetivo ecológico del estudio y el costo asociado a cada tipo de error (Fawcett, 2006; Knight et al., 2017b). En conjuntos desbalanceados, una exactitud elevada puede ofrecer una imagen artificialmente favorable del modelo, ya que el acierto frecuente sobre la clase dominante puede coexistir con fallos sistemáticos en especies raras, poco representadas o de mayor interés para la conservación (Kershenbaum et al., 2025; Knight et al., 2017b). Bajo estas condiciones, indicadores como F1-score, sensibilidad, especificidad, AUC, AUPRC y matrices de confusión deben seleccionarse según la pregunta de investigación y la estructura del dataset, no como criterios intercambiables de rendimiento (Fawcett, 2006; Knight et al., 2017b). En reconocimiento bioacústico, evaluar un modelo implica determinar si predice correctamente, qué errores concentra, qué especies resultan más vulnerables a la confusión y bajo qué condiciones acústicas se reduce su confiabilidad (Kershenbaum et al., 2025; Knight et al., 2017b).

3.4 FACTORES QUE AFECTAN EL RENDIMIENTO EN ENTORNOS NATURALES

El rendimiento de los sistemas de reconocimiento bioacústico en entornos naturales depende de la arquitectura del modelo, la señal registrada, la distribución del conjunto de datos, la calidad del etiquetado, el procedimiento de evaluación y la capacidad de generalización (Kershenbaum et al., 2025; Knight et al., 2017b). Esta relación resulta clave porque los modelos suelen entrenarse bajo condiciones limitadas, mientras las aplicaciones reales de monitoreo acústico pasivo incorporan sensores distintos, hábitats heterogéneos, ruido ambiental variable y especies con representación desigual en los datos (Lostanlen et al., 2019; Teixeira et al., 2024). Por ello, el desempeño observado en un conjunto de prueba debe interpretarse como resultado de una configuración metodológica específica, no como una propiedad absoluta del modelo (Kershenbaum et al., 2025; Knight et al., 2017b).

El ruido ambiental es una fuente principal de variación en campo. Las grabaciones pueden incluir viento, lluvia, insectos, actividad humana, sonidos geofísicos y vocalizaciones de especies no objetivo (Lostanlen et al., 2019; Teixeira et al., 2024). Estos componentes reducen la relación entre la señal de interés y el fondo acústico, dificultando la identificación de patrones espectrotemporales propios de una especie (Kiskin et al., 2020; Lostanlen et al., 2019). La evidencia revisada indica que la baja robustez frente a condiciones meteorológicas y acústicas cambiantes limita la confiabilidad de los detectores automáticos en escenarios reales (Stowell et al., 2019). Lostanlen et al. (2019) señalan que el ruido de fondo en redes de sensores bioacústicos varía temporal y espacialmente; por ello, un detector entrenado en ciertas condiciones puede perder estabilidad al aplicarse en otros sitios, periodos de muestreo o contextos acústicos.

La distribución de los datos también condiciona el desempeño. Los conjuntos bioacústicos suelen presentar desbalance de clases: algunas especies o eventos sonoros aparecen con alta frecuencia y otros cuentan con pocos ejemplos (Assarzadeh et al., 2025; Knight et al., 2017b). Esta estructura puede ocultar un bajo rendimiento en especies minoritarias cuando las métricas globales se sostienen en las clases más abundantes (Knight et al., 2017b). La evaluación requiere complementar la *accuracy*, entendida como proporción total de predicciones correctas, con precisión, *recall*, F1-score, AUC, AUPRC y matrices de confusión (Fawcett, 2006; Knight et al., 2017b). Estas métricas muestran distintos tipos de error y permiten valorar si el modelo funciona de manera equilibrada entre especies frecuentes y poco frecuentes (Assarzadeh et al., 2025; Knight et al., 2017b).

La confiabilidad de las etiquetas afecta directamente el aprendizaje. El ruido de etiquetas, o *label noise*, ocurre cuando las anotaciones no describen con precisión el contenido acústico registrado (Assarzadeh et al., 2025). Puede originarse en errores humanos, segmentaciones imprecisas, sonidos no objetivo, similitud acústica entre especies o baja calidad de grabación (Assarzadeh et al., 2025; Knight et al., 2017b). En reconocimiento bioacústico, las etiquetas inconsistentes pueden inducir patrones equivocados o reforzar confusiones sistemáticas entre clases (Assarzadeh et al., 2025; Kershenbaum et al., 2025). La curación del dataset, la validación manual y la revisión de ejemplos problemáticos deben formar parte del diseño metodológico orientado a mejorar la confiabilidad del sistema (Assarzadeh et al., 2025; LeBien et al., 2020).

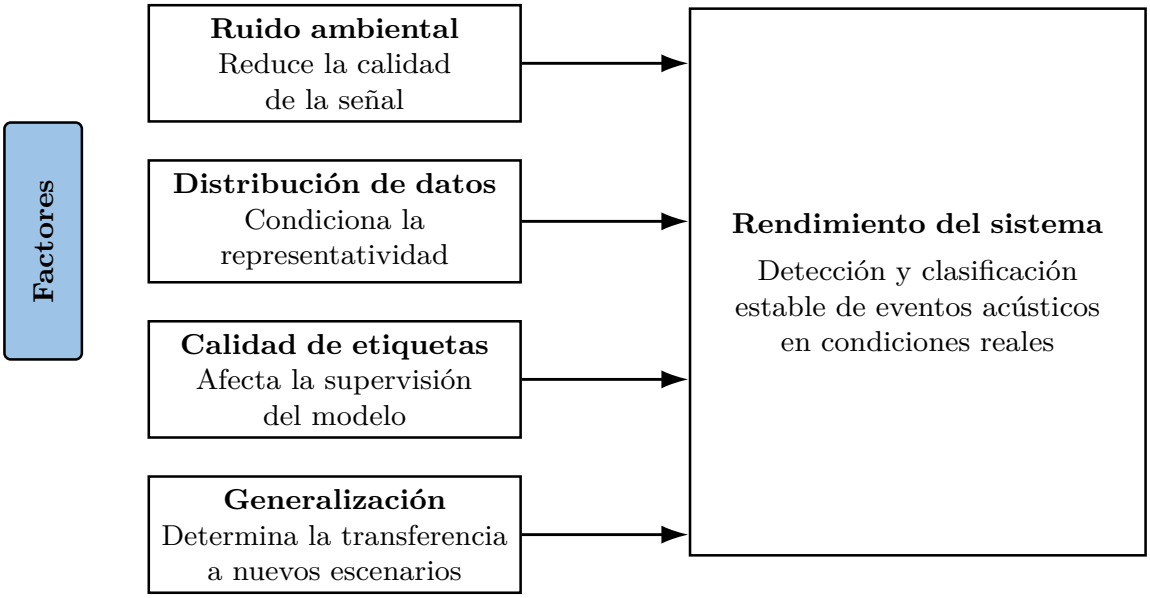
La generalización permite evaluar la estabilidad del modelo fuera del entorno inicial. Un sistema puede alcanzar buen desempeño cuando los datos de prueba se parecen a los de entrenamiento y perder precisión al aplicarse en otro hábitat, con sensores distintos, en otra temporada o frente a una comunidad acústica diferente (Knight et al., 2017b; Lostanlen et al., 2019). Este cambio de distribución puede surgir por diferencias en el fondo sonoro, la distancia al micrófono, la intensidad de las vocalizaciones, el solapamiento acústico o la presencia de especies no observadas durante el entrenamiento (Goodfellow et al., 2016; Kershenbaum et al., 2025; Lostanlen et al., 2019; Teixeira et al., 2024). Las evaluaciones *cross–site* y *cross–sensor*, basadas en probar el modelo en sitios o dispositivos distintos, estiman con mayor rigor la estabilidad del sistema fuera del contexto experimental inicial (Knight et al., 2017b; Lostanlen et al., 2019).

Estos factores actúan de forma interdependiente. El ruido ambiental degrada la señal de entrada y reduce la separabilidad entre eventos acústicos y fondo sonoro; la distribución de datos concentra el aprendizaje en especies frecuentes y debilita el reconocimiento de clases minoritarias; la calidad de etiquetas introduce errores de supervisión; la generalización define si el modelo conserva su desempeño cuando cambian el sitio, el sensor o las condiciones acústicas de aplicación (Assarzadeh et al., 2025; Kershenbaum et al., 2025; Knight et al., 2017b; Lostanlen et al., 2019).

Esta interacción se resume en la Figura 3.4. La figura se organiza en cuatro bloques porque cada factor afecta una dimensión distinta del sistema: el ruido ambiental incide en la calidad de la señal, la distribución de datos en la representatividad del aprendizaje, la calidad de etiquetas en la confiabilidad de la supervisión y la generalización en la transferencia a contextos no observados durante el entrenamiento (Assarzadeh et al., 2025; Knight et al., 2017b). Las flechas convergen hacia el rendimiento porque el desempeño final depende del efecto conjunto de estas condiciones sobre la detección, clasificación y estabilidad del modelo, junto con la arquitectura neuronal (Kershenbaum et al., 2025; Lostanlen et al., 2019).

Figura 3.4

Esquema conceptual de los factores que condicionan el rendimiento de sistemas de reconocimiento bioacústico en entornos naturales



Nota. Elaboración propia con base en Assarzadeh et al. (2025), Lostanlen et al. (2019), Knight et al. (2017b) y Kershenbaum et al. (2025).

De esta forma no se atribuye el bajo desempeño únicamente a la arquitectura neuronal. En aplicaciones reales, un error puede originarse en una grabación ruidosa, un conjunto de datos desbalanceado, etiquetas inconsistentes o una diferencia entre el dominio de entrenamiento y el dominio de aplicación (Assarzadeh et al., 2025; Lostanlen et al., 2019). El desarrollo de modelos robustos requiere evaluar conjuntamente la calidad de la señal, la estructura del dataset, la consistencia del etiquetado y la estabilidad del desempeño frente a condiciones no observadas durante el entrenamiento (Kershenbaum et al., 2025; Knight et al., 2017b).

3.5 ENFOQUES AVANZADOS Y ANÁLISIS DE ERRORES EN ENTORNOS REALES

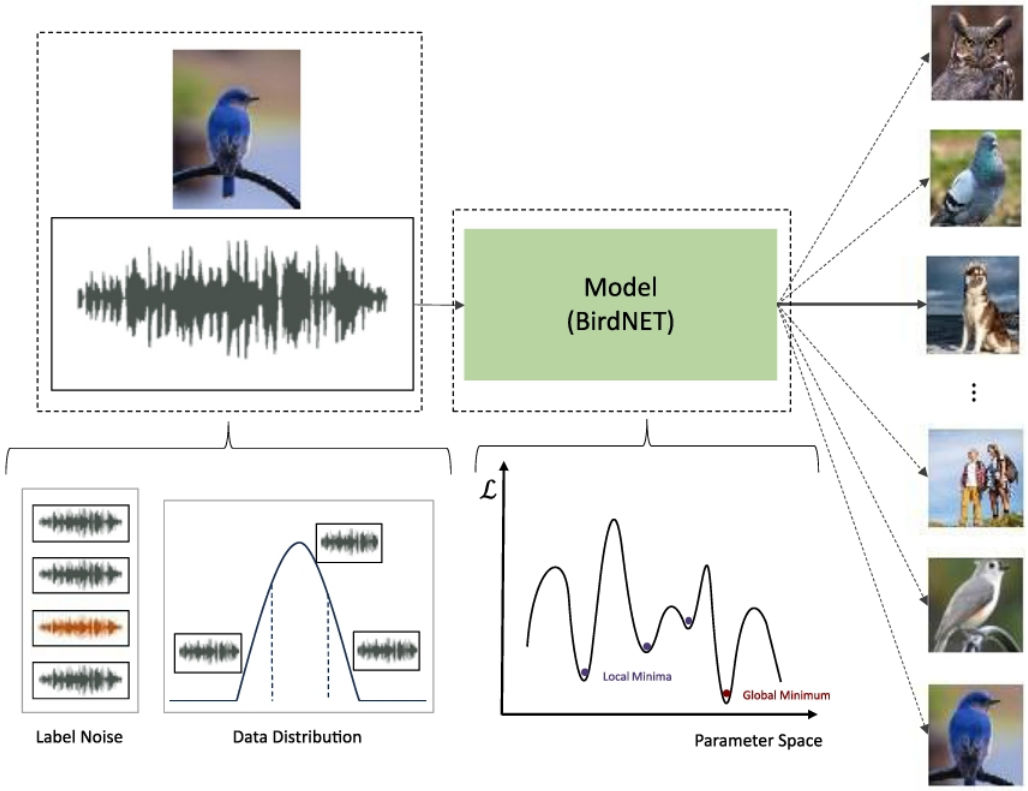
Los errores de clasificación en reconocimiento bioacústico no deben atribuirse de forma inmediata a una falla del algoritmo. Un sistema puede equivocarse porque la representación aprendida no separa adecuadamente las clases, porque una especie cuenta con pocas observaciones, porque la señal está degradada por ruido ambiental o porque la etiqueta asignada durante el

entrenamiento no describe con precisión el evento acústico registrado (Assarzadeh et al., 2025; Kershenbaum et al., 2025). Esta lectura desplaza la evaluación desde la comparación directa de métricas hacia una pregunta metodológica más informativa: qué componente del sistema produce o amplifica el error observado (Assarzadeh et al., 2025; Knight et al., 2017b). Bajo este enfoque, el análisis de errores de clasificación aporta una forma aplicada de explicabilidad, ya que vincula las fallas predictivas con posibles causas en los datos, las etiquetas, la representación acústica y el espacio de parámetros, aunque no permita reconstruir por completo el funcionamiento interno del modelo (Assarzadeh et al., 2025; Goodfellow et al., 2016).

La Figura 3.5 representa los errores de clasificación en BirdNET. El esquema vincula la señal acústica de entrada con problemas asociados al ruido de etiquetas, la distribución de los datos y el ajuste del modelo en el espacio de parámetros, Permite entender que una predicción incorrecta puede originarse en anotaciones imprecisas, ejemplos acústicos ambiguos, clases representadas de forma desigual o mínimos locales durante la optimización de la función de pérdida. La presencia de varias clases candidatas en la salida ilustra la dificultad del reconocimiento multiclase, especialmente cuando los patrones acústicos disponibles no permiten una separación clara entre categorías. En este sentido, la figura ayuda a interpretar el error como un efecto combinado entre datos, etiquetas y entrenamiento del modelo, antes que como una falla atribuible únicamente a la arquitectura (Assarzadeh et al., 2025).

Figura 3.5

Esquema de causas potenciales de errores de clasificación en reconocimiento bioacústico



Nota. Adaptado de Assarzadeh et al. (2025).

La Figura 3.5 representa las causas potenciales de errores de clasificación en BirdNET como resultado de la interacción entre el dato acústico, la distribución del conjunto de entrenamiento, la calidad de las etiquetas y el ajuste del modelo. El esquema permite interpretar la predicción incorrecta como un efecto que puede surgir de muestras mal etiquetadas, clases representadas de forma desigual, ejemplos acústicos difíciles de separar o problemas propios del entrenamiento, como mínimos locales o límites de decisión poco adecuados. Esta lectura coincide con las hipótesis planteadas por Assarzadeh et al. (2025), quienes analizan los errores de clasificació desde tres frentes principales: problemas relacionados con el modelo, problemas asociados a los datos de entrada y problemas derivados del ruido de etiquetas.

Uno de los aportes conceptuales más relevantes del análisis de errores es la distinción entre distintos tipos de ruido de etiquetas (Assarzadeh et al., 2025). El error de etiquetado puede

modelarse mediante tres esquemas de probabilidad,

$$p_e = p(e), \tag{3.9}$$

$$p_e = p(e \mid y), \tag{3.10}$$

$$p_e = p(e \mid x, y), \tag{3.11}$$

donde e representa el evento de error de etiquetado, y representa la clase verdadera y x representa las características de entrada o la señal acústica observada (Assarzadeh et al., 2025). La primera expresión, $p_e = p(e)$, representa ruido aleatorio, porque la probabilidad de error no depende ni de la clase ni de las características de la señal (Assarzadeh et al., 2025). La segunda expresión, $p_e = p(e \mid y)$, representa ruido dependiente de la clase, ya que algunas especies o categorías pueden ser más propensas a recibir etiquetas incorrectas que otras (Assarzadeh et al., 2025). La tercera expresión, $p_e = p(e \mid x, y)$, representa ruido dependiente de la clase y de la instancia, porque la probabilidad de error cambia según la especie y según las propiedades concretas del audio, como ruido, baja energía, solapamiento o ambigüedad acústica (Assarzadeh et al., 2025).

Estas tres formulaciones muestran que el ruido de etiquetado no afecta al sistema de una sola manera. Cuando el ruido es aleatorio, los errores aparecen sin un patrón definido; cuando depende de la clase, ciertas especies concentran mayor ambigüedad; y cuando depende simultáneamente de (x) y de (y), la dificultad surge de la interacción entre la identidad de la especie y las condiciones acústicas particulares de cada grabación (Assarzadeh et al., 2025). En bioacústica, este último caso resulta especialmente plausible: una especie puede ser reconocible en una señal limpia, pero volverse difícil de identificar cuando aparece con baja intensidad, superpuesta con otras vocalizaciones o mezclada con ruido ambiental (Assarzadeh et al., 2025; Lostanlen et al., 2019). Una evaluación basada únicamente en el rendimiento promedio puede ocultar estos patrones, ya que los errores tienden a concentrarse en especies específicas, ejemplos acústicamente ambiguos o clases con escasa representación en el conjunto de datos (Assarzadeh et al., 2025; Knight et al., 2017b).

La distribución del conjunto de datos condiciona de forma directa los errores de clasificación. Un sistema puede mostrar un rendimiento global elevado y, al mismo tiempo, fallar en especies poco representadas o en sonidos que se alejan del dominio acústico observado durante el

entrenamiento (Assarzadeh et al., 2025; Knight et al., 2017b). Este comportamiento responde a que el modelo aprende con mayor estabilidad los patrones de las clases abundantes, mientras que las categorías con pocos ejemplos tienden a generar representaciones menos robustas y límites de decisión más frágiles (Assarzadeh et al., 2025; Goodfellow et al., 2016). El error, por tanto, no debe atribuirse únicamente a una limitación del algoritmo; también refleja la estructura estadística del dataset y la manera en que dicha estructura orienta el proceso de aprendizaje (Assarzadeh et al., 2025; Kershenbaum et al., 2025).

El nivel de ruido de la señal no explica por sí mismo todos los errores de clasificación observados. La relación señal–ruido, o *signal-to-noise ratio* (SNR), permite estimar qué tan distinguible resulta una vocalización frente al fondo acústico y, por tanto, ofrece un criterio útil para evaluar la calidad relativa de las grabaciones (Kiskin et al., 2020; Lostanlen et al., 2019). Assarzadeh et al. (2025) muestran que las distribuciones de SNR entre clases bien resueltas, clases problemáticas y sonidos no objetivo no siempre bastan para explicar las predicciones incorrectas. Por lo general, el error no proviene solo de señales ruidosas, sino de la interacción entre distribución desigual de datos, etiquetas ambiguas, similitud acústica entre clases y límites de generalización del modelo (Assarzadeh et al., 2025; Knight et al., 2017b).

El análisis de errores también puede apoyarse en herramientas de interpretación del espacio de representación, como el análisis de componentes principales (*Principal Component Analysis*, PCA) y la incrustación estocástica de vecinos distribuidos en (t) (*t-distributed Stochastic Neighbor Embedding*, t-SNE) (Assarzadeh et al., 2025). PCA proyecta datos de alta dimensión en componentes principales que resumen la variación lineal dominante, mientras que t-SNE facilita la visualización de agrupamientos locales en espacios de menor dimensión. En reconocimiento bioacústico, estas técnicas permiten explorar si las representaciones aprendidas separan adecuadamente las clases o si ciertos sonidos aparecen mezclados en el espacio latente. Esta lectura ayuda a identificar patrones de confusión asociados con clases acústicamente similares, muestras ambiguas, desbalance de datos o presencia de etiquetas ruidosas (Assarzadeh et al., 2025; Kershenbaum et al., 2025).

De manera complementaria, las explicaciones locales interpretables agnósticas al modelo (*Local Interpretable Model-Agnostic Explanations*, LIME) permiten examinar qué regiones de

una entrada influyen en una predicción específica (Assarzadeh et al., 2025). En el estudio de Assarzadeh et al. (2025), LIME se aplica a espectrogramas clasificados incorrectamente por BirdNET para observar qué zonas orientan la decisión del modelo. Esta técnica no corrige la predicción, pero permite revisar si el modelo se apoya en regiones informativas del espectrograma o si la clasificación parece estar afectada por problemas previos del dato, como etiquetas inconsistentes o ejemplos acústicos ambiguos (Assarzadeh et al., 2025; Kershenbaum et al., 2025).

Una aproximación complementaria para examinar representaciones complejas proviene del análisis topológico de datos (*Topological Data Analysis*, TDA), en particular de la homología persistente. Este enfoque permite estudiar la forma de una nube de puntos a distintas escalas y describir rasgos estructurales como componentes conectados, ciclos o vacíos mediante diagramas de persistencia (Fernández et al., 2023). En el análisis de señales y series temporales, esta perspectiva resulta útil para explorar patrones geométricos de los datos, detectar comportamientos atípicos y reconocer estructuras que pueden quedar ocultas en proyecciones convencionales.

Fernández et al. (2023) proponen una formulación de homología persistente intrínseca basada en una métrica de Fermat dependiente de la densidad, orientada a recuperar características topológicas con menor sensibilidad a outliers y a la forma en que los datos están embebidos en el espacio ambiente. Dentro del reconocimiento bioacústico, este tipo de herramientas puede aportar una lectura exploratoria de espacios de características, *embeddings* o trayectorias derivadas de señales acústicas. Su aporte no reemplaza las métricas de clasificación ni la validación ecológica, pero puede complementar el análisis de errores al mostrar cómo se organizan las representaciones aprendidas y dónde aparecen posibles separaciones, solapamientos o anomalías entre clases.

Desde esta perspectiva, el desarrollo de sistemas bioacústicos robustos en entornos reales no depende únicamente de incorporar arquitecturas más complejas. Requiere examinar cómo se distribuyen los datos, cuánta confiabilidad tienen las etiquetas, qué patrones acústicos aprende el modelo, cómo se organiza el espacio de representación y en qué condiciones aparecen los errores de clasificación (Assarzadeh et al., 2025; Fernández et al., 2023; Knight et al., 2017b). Este enfoque fortalece la explicabilidad del sistema, porque desplaza la evaluación desde el valor final de una métrica hacia el análisis de los factores que pueden originar una predicción incorrecta (Assarzadeh et al., 2025; Kershenbaum et al., 2025). Así, el análisis de errores deja de ocupar un

lugar secundario y pasa a formar parte del diseño metodológico necesario para construir modelos más robustos, transferibles y ecológicamente interpretables (Assarzadeh et al., 2025; Lostanlen et al., 2019).

En conjunto, la evidencia sugiere que la robustez del reconocimiento bioacústico depende de la interacción entre arquitectura, distribución de datos, calidad del etiquetado, representación acústica y naturaleza del error (Assarzadeh et al., 2025; Kershenbaum et al., 2025). Por tanto, el desarrollo futuro del campo debería combinar el aumento de la complejidad de los modelos con el fortalecimiento de la calidad de los datos, el diseño de protocolos de validación más exigentes y el uso de herramientas de interpretación que permitan comprender las condiciones bajo las cuales el sistema falla (Knight et al., 2017b; Lostanlen et al., 2019).

4. METODOLOGÍA

La presente investigación se desarrolló mediante una revisión sistemática de literatura con enfoque descriptivo y comparativo. El propósito metodológico fue identificar, seleccionar, evaluar y sintetizar estudios sobre reconocimiento bioacústico de especies en entornos naturales mediante redes neuronales profundas y métodos comparables. El proceso se estructuró con base en los lineamientos PRISMA 2020, que proporcionan una guía para reportar de forma transparente las fases de identificación, cribado, elegibilidad e inclusión de estudios en revisiones sistemáticas (Page et al., 2021).

La metodología se organizó en cuatro momentos principales. Primero, se definieron palabras clave mediante una adaptación de la lógica PICO al contexto de revisión tecnológica y metodológica. Luego, se aplicó una búsqueda estructurada en Scopus y se depuraron los registros recuperados. Después, se realizó el cribado por metadatos, combinando asistencia computacional y revisión manual de los casos no concluyentes. Finalmente, se evaluó la elegibilidad por texto completo mediante una matriz de calidad y se extrajeron variables metodológicas para la síntesis descriptiva del corpus final.

Esta organización permitió mantener trazabilidad entre la pregunta de investigación, la estrategia de búsqueda, los criterios de selección y las categorías de análisis utilizadas en los resultados. La metodología se orientó a caracterizar el estado de la literatura mediante la comparación de arquitecturas, tareas, representaciones acústicas, métricas de desempeño, estrategias de validación y factores que afectan la aplicabilidad de los modelos en entornos naturales. Debido a la heterogeneidad metodológica del corpus, la estimación de efectos agregados mediante metaanálisis quedó fuera del alcance de la revisión.

4.1 DISEÑO METODOLÓGICO Y CRITERIOS DE SELECCIÓN

La revisión se planteó como un estudio documental, descriptivo y comparativo. Su unidad de análisis fueron artículos científicos que abordaran tareas de reconocimiento, detección, clasificación, identificación o estimación de especies a partir de señales acústicas. El interés principal estuvo en estudios aplicados a contextos ecológicos o bioacústicos, especialmente aquellos

que utilizaran redes neuronales profundas, modelos CNN, arquitecturas no-CNN comparables o estrategias híbridas de aprendizaje automático.

La selección de estudios priorizó artículos con evidencia sustantiva sobre al menos una de las siguientes dimensiones: arquitecturas y preprocesamiento acústico, variación del desempeño ante ruido o calidad del conjunto de datos, aplicaciones ecológicas más allá de la clasificación simple y métricas o enfoques de validación utilizados en la literatura reciente. La sola mención de aprendizaje profundo en el texto no fue considerada criterio suficiente de pertinencia.

Los criterios de inclusión y exclusión se definieron para asegurar que el corpus final respondiera directamente al objetivo de este trabajo. Se incluyeron estudios que utilizaran audio como fuente principal de información, abordaran especies silvestres o escenarios ecológicos, aplicaran redes neuronales profundas o métodos no-CNN comparables, reportaran métricas cuantitativas de desempeño y superaran la evaluación de elegibilidad por texto completo.

Se excluyeron estudios centrados en reconocimiento de voz humana, música, audio médico o clasificación general de sonidos sin relación explícita con especies o monitoreo ecológico. También se excluyeron trabajos sin evaluación cuantitativa, sin relación con especies silvestres, fuera del tipo documental definido o con puntaje inferior al umbral mínimo de elegibilidad.

La Tabla 4.1 sintetiza los criterios utilizados.

Tabla 4.1

Criterios de inclusión y exclusión aplicados en la revisión

Criterios de inclusión	**Criterios de exclusión**
Uso de audio como fuente principal de información	Estudios centrados en voz humana, música, audio médico o audio no ecológico
Reconocimiento, detección, clasificación o identificación de especies silvestres	Estudios sin relación explícita con especies silvestres o monitoreo ecológico
Aplicación de redes neuronales profundas o métodos no-CNN comparables	Estudios puramente descriptivos sin método computacional de reconocimiento
Reporte de métricas cuantitativas de desempeño	Estudios sin evaluación cuantitativa del desempeño
Correspondencia con las preguntas de investigación	Estudios con relación débil, indirecta o insuficiente con el objetivo de este trabajo
Puntaje de elegibilidad igual o superior a 3 sobre 4	Puntaje de elegibilidad inferior a 3 sobre 4

Nota. Elaboración propia. Los criterios se aplicaron durante la selección de estudios para asegurar correspondencia temática, metodológica y evaluativa con el objetivo de la revisión.

La Tabla 4.1 permite explicitar que la selección se basó en criterios metodológicos y temáticos más amplios que la presencia de palabras clave. Un artículo podía mencionar términos como *deep learning* o *bioacoustics*, pero ser excluido si no abordaba reconocimiento de especies, si no reportaba métricas o si su aplicación no correspondía a un contexto ecológico.

4.2 ESTRATEGIA DE BÚSQUEDA

Antes de construir la cadena de búsqueda, se definieron grupos de palabras clave mediante una adaptación de la lógica PICO. En revisiones clínicas, PICO suele organizar la búsqueda en población, intervención, comparación y resultado. En esta monografía, dicha lógica se adaptó al análisis de métodos computacionales: la población corresponde al dominio acústico y bioacústico, la intervención corresponde a las tareas de reconocimiento automático, la comparación corresponde a familias de modelos de aprendizaje profundo y métodos relacionados y el resultado corresponde a métricas de desempeño, validación y robustez.

La Tabla 4.2 presenta los términos utilizados para estructurar la búsqueda. Esta organización permitió evitar una consulta demasiado amplia que recuperara trabajos de audio no ecológico, pero también redujo el riesgo de excluir estudios relevantes por usar sinónimos como *detection*, *classification*, *recognition*, *bioacoustics*, *PAM* o *deep learning*.

Tabla 4.2

Palabras clave y sinónimos utilizados en la estrategia de búsqueda

Elemento adaptado	**Término base**	**Sinónimos y términos relacionados**
Población	acoustic	PAM; acoustics; audio; bioacoustics; environmental recordings; passive acoustic monitoring; grabaciones ambientales; señales acústicas
Intervención	recognition	classification; detection; identification; reconocimiento; clasificación; detección; identificación; etiquetado acústico
Comparación	CNN	ANN; CNN; CRNN; deep learning; embedded models; neural networks; transfer learning; redes neuronales; modelos embebidos
Resultado	validation	accuracy; precision; recall; F1; ROC-AUC; AUPRC; calibration; class imbalance; generalization; robustness; metrics; noise; ruido; precisión; robustez

Nota. Elaboración propia. La tabla organiza los términos empleados para construir la estrategia de búsqueda a partir de una adaptación del esquema PICO al contexto del reconocimiento acústico automático de especies. PAM = monitoreo acústico pasivo; CNN = red neuronal convolucional; ANN = red neuronal artificial; CRNN = red neuronal convolucional recurrente; ROC-AUC = área bajo la curva ROC; AUPRC = área bajo la curva precisión-recall.

La fase de identificación se realizó mediante una búsqueda estructurada en la base de datos recopilada, utilizando el campo `TITLE-ABS-KEY`. La búsqueda se concentró en título, resumen y palabras clave porque estos campos permiten recuperar estudios cuya contribución principal se relaciona explícitamente con el dominio temático de interés. La estrategia combinó términos asociados con el dominio acústico, tareas de reconocimiento, aprendizaje profundo y especies silvestres. También se excluyeron trabajos centrados en reconocimiento de voz humana para reducir contaminación temática con literatura de procesamiento de habla.

La consulta empleada fue:

```
TITLE-ABS-KEY ( ( acoustic OR acoustics OR audio OR bioacoustics OR
"acoustic monitoring" OR "passive acoustic monitoring" OR PAM ) AND
( classification OR detection OR identification OR recognition ) AND
( CNN OR "convolutional neural network" OR CRNN OR "deep learning" OR
"neural networks" ) AND ( species OR wildlife OR fauna OR biodiversity ) )
AND NOT TITLE-ABS-KEY ( speech OR voice OR spoken )
AND PUBYEAR > 2017
AND OA
AND ( LIMIT-TO ( DOCTYPE , "ar" ) )
```

La consulta identificó 481 registros iniciales. Este número corresponde al conjunto recuperado antes de la depuración de duplicados y antes de aplicar los criterios de cribado. El uso de acceso abierto y tipo documental artículo permitió acotar el corpus a documentos revisados y disponibles para evaluación completa, mientras que la exclusión de voz humana evitó incluir estudios de reconocimiento de habla que no formaban parte del problema de investigación.

4.3 IDENTIFICACIÓN Y CRIBADO DE REGISTROS

Los resultados recuperados fueron exportados en un archivo tabular con metadatos bibliográficos. La matriz inicial incluyó campos como título, autores, año de publicación, fuente, resumen, palabras clave, DOI y tipo documental. Esta estructura permitió organizar el flujo de selección y conservar trazabilidad entre cada registro y su decisión metodológica posterior.

Después de eliminar duplicados, quedaron 466 registros únicos. El cribado se realizó a partir de título, resumen y palabras clave, con apoyo de evaluaciones asistidas computacionalmente y revisión manual de los casos no concluyentes. La revisión manual permitió resolver registros ambiguos sin tratarlos como rechazos automáticos.

Como resultado del cribado, 264 registros fueron aceptados para recuperación de informe

completo y 202 fueron rechazados por no ajustarse al enfoque temático o metodológico de la revisión. La Tabla 4.3 resume esta etapa.

Tabla 4.3

Resultados del cribado de registros

Resultado del cribado	**Número de registros**
Aceptados para recuperación de informe completo	264
Rechazados durante el cribado	202
Total de registros únicos cribados	466

Nota. Elaboración propia. El cribado combinó asistencia computacional y revisión manual de registros no concluyentes. Los registros aceptados incluyen aquellos aprobados directamente por metadatos y los incorporados después de la revisión manual.

La Tabla 4.3 presenta el resultado consolidado del cribado. Esta forma de reporte evita separar innecesariamente los registros preliminares y los casos manuales, manteniendo la lógica PRISMA: registros cribados, registros excluidos e informes buscados para recuperación.

4.4 ELEGIBILIDAD POR TEXTO COMPLETO

La evaluación de elegibilidad se aplicó sobre 262 informes evaluados a texto completo. La diferencia entre los 264 informes buscados para recuperación y los 262 informes evaluados se debió a que dos informes no fueron recuperados o no contaban con texto completo procesable dentro del corpus.

Sobre cada artículo se aplicó un proceso de *Quality Assessment*, entendido como una evaluación estructurada de pertinencia metodológica. Esta evaluación se basó en cuatro criterios directamente relacionados con las preguntas de investigación:

1. aplicación de arquitecturas CNN al reconocimiento bioacústico;
2. comparación entre modelos CNN y enfoques no-CNN mediante métricas cuantitativas;
3. análisis del efecto del ruido, la calidad del conjunto de datos o las condiciones de grabación;

4. inclusión de tareas ecológicas como presencia, ocupación, abundancia o cuantificación.

Cada criterio fue calificado mediante una escala gradual. La respuesta afirmativa recibió 1 punto, la respuesta parcial recibió 0.5 puntos y la respuesta negativa recibió 0 puntos. La puntuación global de cada artículo se calculó sumando los cuatro criterios, por lo que el resultado posible varió entre 0 y 4 puntos. Se estableció un umbral de aceptación de 3 puntos sobre 4. En términos operativos, un artículo fue aceptado cuando alcanzó una puntuación global igual o superior a 3 y fue rechazado cuando obtuvo una puntuación inferior.

La Tabla 4.4 resume la lógica de calificación utilizada.

Tabla 4.4

Escala de evaluación utilizada en el proceso de Quality Assessment

Valor	**Interpretación**	**Puntaje**
Sí	El criterio se cumple de forma clara y explícita en el artículo	1.0
Parcial	El criterio se menciona o se cumple de forma incompleta, indirecta o limitada	0.5
No	El criterio no se aborda o no existe evidencia suficiente en el texto	0.0

Nota. Elaboración propia. La escala fue aplicada a cuatro criterios de pertinencia metodológica, con una puntuación máxima de 4 puntos por artículo.

La Tabla 4.4 muestra que la evaluación dependió de la combinación de cuatro dimensiones. Esta lógica fue necesaria porque la revisión buscaba investigaciones con evidencia sobre arquitectura, comparación, condiciones de datos y aplicación ecológica. Así, un estudio podía ser relevante aunque no cubriera todos los ejes con la misma profundidad, siempre que su puntuación global alcanzara el umbral definido.

Con el umbral de aceptación igual o superior a 3 puntos, 67 estudios fueron aceptados y 195 fueron rechazados. En consecuencia, el corpus final de la revisión quedó conformado por 67

estudios. La Tabla 4.5 presenta el resultado de esta etapa.

Tabla 4.5

Resultados de la evaluación de elegibilidad por texto completo

Resultado de elegibilidad	**Número de artículos**
Aceptados con puntaje igual o superior a 3	67
Rechazados con puntaje inferior a 3	195
Total de artículos evaluados a texto completo	262

Nota. Elaboración propia. La inclusión final se definió mediante el umbral mínimo de 3 puntos sobre 4 en la evaluación de calidad metodológica.

La Tabla 4.5 documenta el tránsito desde los informes evaluados para elegibilidad hasta la conformación del corpus final. La inclusión de estudios se definió mediante una regla explícita de puntuación, no por disponibilidad, conveniencia o selección discrecional. Este criterio permitió conservar coherencia entre el proceso de selección, los criterios metodológicos definidos y las preguntas de investigación que orientan la revisión.

4.5 EXTRACCIÓN, NORMALIZACIÓN Y SÍNTESIS DE LA INFORMACIÓN

Una vez definido el corpus final de 67 estudios, se extrajo información mediante una matriz común diseñada para comparar los artículos bajo criterios homogéneos. En esta matriz se registraron variables relacionadas con el tipo de arquitectura empleada, la representación acústica utilizada, la tarea abordada, las métricas de desempeño reportadas, las características generales del conjunto de datos y las estrategias complementarias aplicadas, como transferencia de aprendizaje, ajuste fino, aumento de datos, modelos híbridos o enfoques de fusión.

La extracción se diseñó para responder a las preguntas de investigación. Las arquitecturas y representaciones acústicas permitieron abordar la RQ1; las condiciones de ruido, calidad del conjunto de datos y heterogeneidad acústica permitieron abordar la RQ2; las tareas ecológicas como presencia, ocupación, abundancia o cuantificación permitieron abordar la RQ3; y las métricas, validaciones y tendencias metodológicas permitieron abordar la RQ4.

La información fue organizada en categorías comunes, aun cuando cada artículo usara terminología distinta. Por ejemplo, expresiones como *convolutional neural network*, CNN o redes convolucionales fueron agrupadas bajo una misma familia metodológica. Lo mismo se aplicó a métricas, tareas y representaciones acústicas que podían aparecer con nombres equivalentes o variantes terminológicas.

Se distinguieron dos tipos de conteo. El conteo excluyente se utilizó para asignar un método principal a cada artículo, lo que permitió construir una síntesis comparativa general del corpus. En cambio, el conteo no excluyente registró todas las técnicas, tareas, representaciones o métricas mencionadas dentro de un mismo estudio. Esta segunda lógica fue necesaria porque varios artículos combinan enfoques, por ejemplo CNN con transferencia de aprendizaje, modelos híbridos, transformadores o métodos tradicionales de aprendizaje automático.

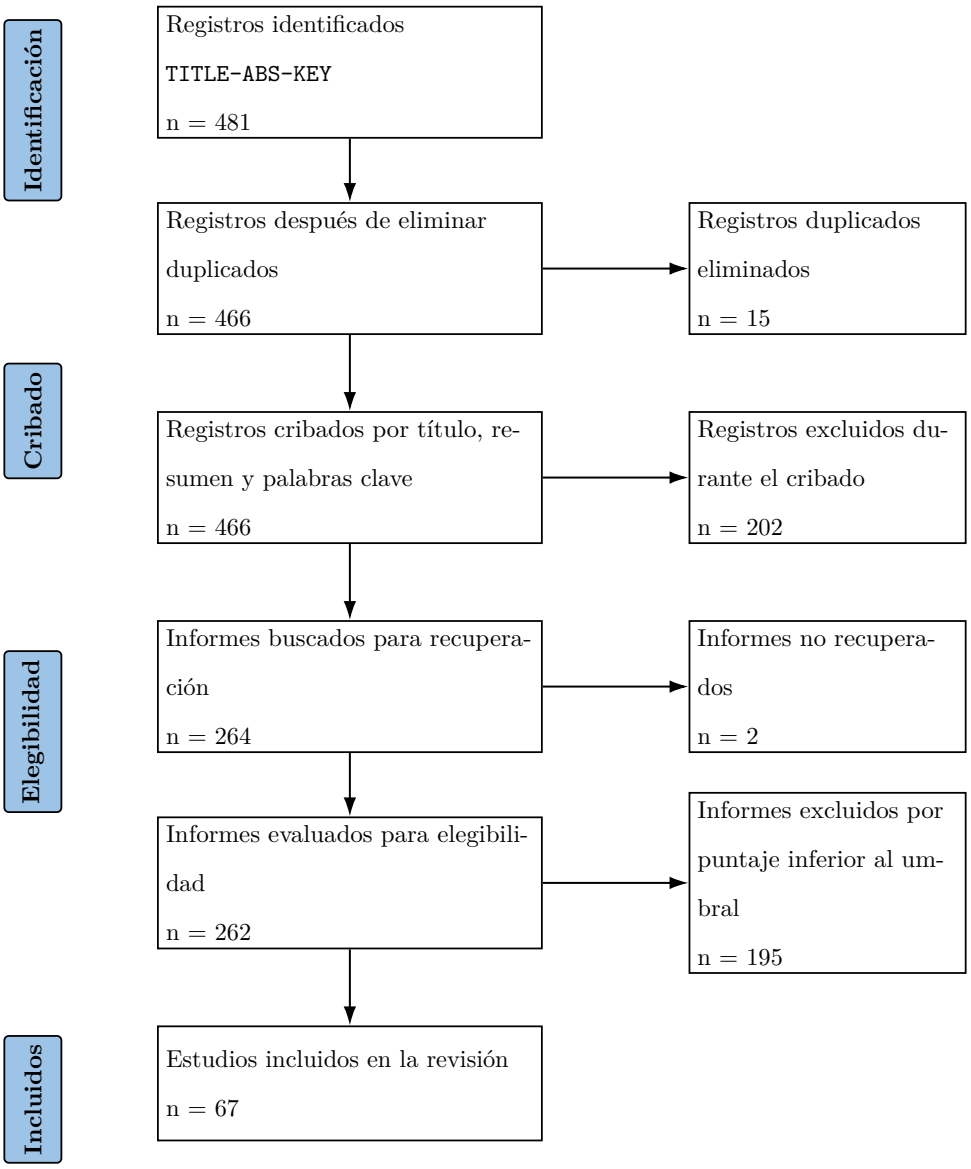
La síntesis de resultados combinó lectura crítica de los estudios incluidos con análisis descriptivo de frecuencias. La lectura crítica permitió interpretar el enfoque metodológico de cada artículo, las condiciones de aplicación, las limitaciones reportadas y el alcance ecológico de sus resultados. El análisis descriptivo permitió organizar el corpus en categorías comparables y calcular frecuencias absolutas sobre los principales elementos metodológicos detectados.

La síntesis no se planteó como un metaanálisis. Esta decisión se tomó porque los estudios incluidos presentan heterogeneidad en especies, sensores, hábitats, duración de grabaciones, representaciones acústicas, arquitecturas, métricas y esquemas de validación. En estas condiciones, combinar numéricamente los resultados como si fueran directamente equivalentes podría producir una interpretación engañosa. Por ello, se priorizó una síntesis narrativa y comparativa, apoyada en frecuencias descriptivas y lectura contextual de los resultados.

El flujo completo de identificación, cribado, elegibilidad e inclusión final se resume en la Figura 4.1.

Figura 4.1

Diagrama PRISMA 2020 del proceso de identificación, cribado, elegibilidad e inclusión de estudios



Nota. Elaboración propia con base en el proceso de revisión sistemática y la nomenclatura PRISMA 2020.

La Figura 4.1 representa la trazabilidad completa del proceso de selección. La fase de identificación muestra los registros recuperados y la eliminación de duplicados. La fase de cribado resume la evaluación por título, resumen y palabras clave, junto con la exclusión de registros no pertinentes. La fase de elegibilidad documenta los informes buscados para recuperación, los informes no recuperados y los informes evaluados a texto completo. La fase de inclusión presenta

el corpus final de 67 estudios. Esta división permite verificar que la selección respondió a una secuencia documentada de filtros y criterios de aceptación alineada con la lógica de reporte PRISMA 2020.

5. RESULTADOS Y DISCUSIÓN

A partir del proceso de revisión sistemática se conformó un corpus final de 67 estudios, seleccionados tras la evaluación de elegibilidad por texto completo. Sobre este conjunto se desarrolló una síntesis descriptiva y comparativa orientada a identificar arquitecturas empleadas, representaciones acústicas, tareas abordadas, métricas de evaluación y factores metodológicos asociados al rendimiento de los modelos. El propósito de este capítulo es interpretar, a partir de las frecuencias observadas, qué revelan estos patrones sobre el estado actual del reconocimiento bioacústico mediante redes neuronales profundas.

Los resultados se organizan en función de las preguntas de investigación. La primera sección analiza arquitecturas y métodos de preprocesamiento; la segunda examina la variación del desempeño ante ruido, calidad del conjunto de datos y heterogeneidad acústica; la tercera evalúa el alcance ecológico de las tareas reportadas; y la cuarta revisa métricas, protocolos de validación y tendencias metodológicas. La lista completa de estudios incluidos se presenta en la matriz resumida de identificación metodológica del corpus que se muestra en el Apéndice A.1.

La lectura crítica de los artículos se acompañó de una extracción estructurada de información sobre los 67 estudios incluidos. Este procedimiento permitió describir patrones internos del corpus en relación con familias metodológicas, representaciones acústicas, tareas de reconocimiento y métricas de desempeño. Los conteos obtenidos tienen un alcance descriptivo, no inferencial; por ello, deben interpretarse como una síntesis de tendencias observadas dentro del conjunto revisado y no como estimadores generalizables a toda la literatura disponible. Esta delimitación resulta necesaria porque los estudios difieren en especies analizadas, hábitats, sensores, escalas de análisis, métricas reportadas y protocolos de validación.

Los resultados evidencian tres patrones que atraviesan el corpus revisado. El primero corresponde al predominio de arquitecturas convolucionales y representaciones tiempo–frecuencia como base metodológica del reconocimiento acústico automatizado. El segundo refleja una diversificación progresiva hacia estrategias de transferencia de aprendizaje, ajuste fino, modelos híbridos, fusión de clasificadores y arquitecturas recientes, entre ellas los transformadores. El

tercero se relaciona con la persistente heterogeneidad en métricas, protocolos de validación y formas de definir las tareas ecológicas, lo que restringe la comparación directa entre estudios. Estos patrones orientan la discusión posterior y permiten interpretar el campo como un espacio metodológicamente consolidado en torno a las CNN, aunque todavía marcado por diferencias relevantes en evaluación, diseño experimental y alcance ecológico.

La Tabla 5.1 resume los resultados agregados de la extracción. Estos valores se interpretan como indicadores descriptivos del corpus y no como resultados inferenciales.

Tabla 5.1

Variables metodológicas y métricas detectadas en el corpus final

Indicador descriptivo	**n**	**%**
Corpus final analizado	67	100.0
Textos disponibles para extracción estructurada	67	100.0
Artículos con uso de CNN detectado	67	100.0
Artículos con uso de transformadores detectado	18	26.9
Artículos con métodos tradicionales de aprendizaje automático detectados	43	64.2
Artículos con enfoques híbridos o de fusión detectados	45	67.2
Artículos con transferencia de aprendizaje detectada	27	40.3
Artículos que reportan F1-score	44	65.7
Artículos que reportan accuracy	61	91.0
Artículos que reportan AUC	17	25.4
Artículos que reportan AUPRC	17	25.4
Artículos que reportan MAE	6	9.0
Artículos sin métricas detectables	0	0.0

Nota. Elaboración propia a partir de la matriz de extracción estructurada del corpus final. Los porcentajes se calcularon sobre el total de estudios incluidos en la revisión ($n = 67$). Las categorías no son excluyentes, por lo que un mismo artículo puede registrar más de una arquitectura, estrategia metodológica o métrica de desempeño.

La Tabla 5.1 documenta el alcance del procesamiento aplicado al corpus final. La disponibilidad de los 67 textos permitió realizar una extracción homogénea de variables metodológicas y métricas de desempeño. La presencia de CNN en todos los estudios confirma la correspondencia del corpus con el eje central de esta revisión, mientras que la detección de transformadores, métodos tradicionales, enfoques híbridos y transferencia de aprendizaje evidencia que la literatura reciente combina distintas familias metodológicas. La presencia de métricas detectables en todos los artículos respalda la pertinencia del corpus para desarrollar una síntesis comparativa del desempeño.

5.1 RQ1. ¿QUÉ ARQUITECTURAS Y MÉTODOS DE PREPROCESAMIENTO PREDOMINAN O REPORTAN MEJOR DESEMPEÑO EN EL RECONOCIMIENTO ACÚSTICO DE ESPECIES?

La evidencia del corpus muestra que el reconocimiento acústico automatizado de especies se estructura principalmente alrededor del aprendizaje profundo y, dentro de este, de arquitecturas convolucionales. Esta predominancia aparece en la frecuencia de uso de los modelos, en la elección de representaciones acústicas tiempo–frecuencia y en la forma en que los estudios reportan el desempeño. La respuesta a esta pregunta requiere analizar las arquitecturas junto con las etapas de preprocesamiento, porque el rendimiento depende tanto del clasificador como de la forma en que el audio se segmenta, transforma y presenta al modelo.

En el conteo excluyente, cada artículo fue asignado a una sola categoría principal según el enfoque dominante reportado. Esta clasificación permite identificar qué método organiza técnicamente cada estudio, aunque no agota todos los componentes empleados dentro del sistema. La Tabla 5.2 muestra que las CNN constituyen el método principal más frecuente, con 43 artículos, equivalentes al 64,2 % del corpus. Les siguen los enfoques híbridos, con 18 artículos, equivalentes al 26,9 %. Las demás categorías aparecen con una frecuencia menor: workflow, herramienta o revisión técnica con 2 artículos; aprendizaje automático tradicional, self-supervised learning, transformer y GAN con 1 artículo cada una.

Tabla 5.2

Distribución del método principal asignado a cada artículo

Método principal	**n**	**%**
CNN	43	64,2
Híbrido	18	26,9
Workflow, herramienta o revisión técnica	2	3,0
Aprendizaje automático tradicional	1	1,5
Self-supervised learning	1	1,5
Transformer	1	1,5
GAN	1	1,5

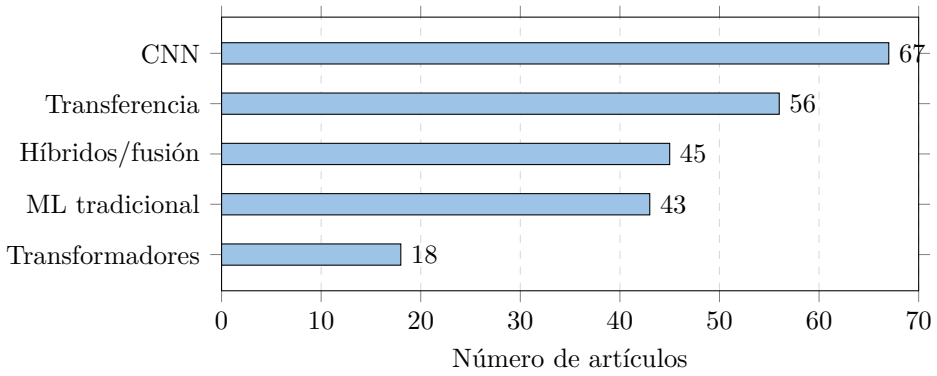
Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los porcentajes se calcularon sobre el total de estudios incluidos en la revisión ($n = 67$). La clasificación corresponde a un conteo excluyente, por lo que cada artículo fue asignado a un único método principal según el enfoque predominante reportado.

La Tabla 5.2 confirma que las CNN ocupan el lugar central dentro del corpus, aunque su presencia no implica que operen como único componente del sistema. LeBien et al. (2020) proponen un flujo de reconocimiento para aves y ranas en paisajes sonoros tropicales donde la CNN se alimenta de mel-espectrogramas construidos a partir de segmentos validados manualmente, lo que evidencia una cadena técnica compuesta por detección, validación, segmentación y representación acústica. Kahl et al. (2021) presentan BirdNET como una solución profunda para monitoreo de diversidad aviar, mientras que estudios posteriores reutilizan o ajustan sus representaciones para contextos locales, como bosques alpinos italianos o especies árticas (Kahl et al., 2021; Schiavo et al., 2025; Ziegenhorn et al., 2025). En una línea complementaria, Xie y Zhu (2019) combinan características diseñadas manualmente con aprendizaje profundo mediante *late fusion*, mostrando que la mejora puede provenir de integrar descriptores acústicos y modelos neuronales dentro de un mismo sistema. Estos casos explican por qué el conteo excluyente identifica el enfoque predominante, pero no resume toda la estructura técnica de los sistemas revisados.

Para representar esa complejidad, la Figura 5.1 presenta un conteo no excluyente de familias metodológicas. En este nivel, un mismo artículo puede aparecer en más de una categoría, ya que varios estudios combinan arquitecturas, representaciones de entrada, estrategias de entrenamiento y métodos de comparación dentro de un mismo flujo de análisis. En el corpus, las CNN estuvieron presentes en los 67 artículos, lo que representa el 100,0 % de los estudios incluidos. También se identificaron 56 estudios con transferencia de aprendizaje o ajuste fino, equivalentes al 83,6 %; 45 estudios con enfoques híbridos o de fusión, equivalentes al 67,2 %; 43 estudios con algún método tradicional de aprendizaje automático, equivalentes al 64,2 %; y 18 estudios con transformadores, equivalentes al 26,9 %.

Figura 5.1

Frecuencia de familias metodológicas detectadas en el corpus



Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los valores corresponden a un conteo no excluyente calculado sobre los 67 estudios incluidos en la revisión; por tanto, un mismo artículo puede registrar más de una familia metodológica. ML tradicional = aprendizaje automático tradicional.

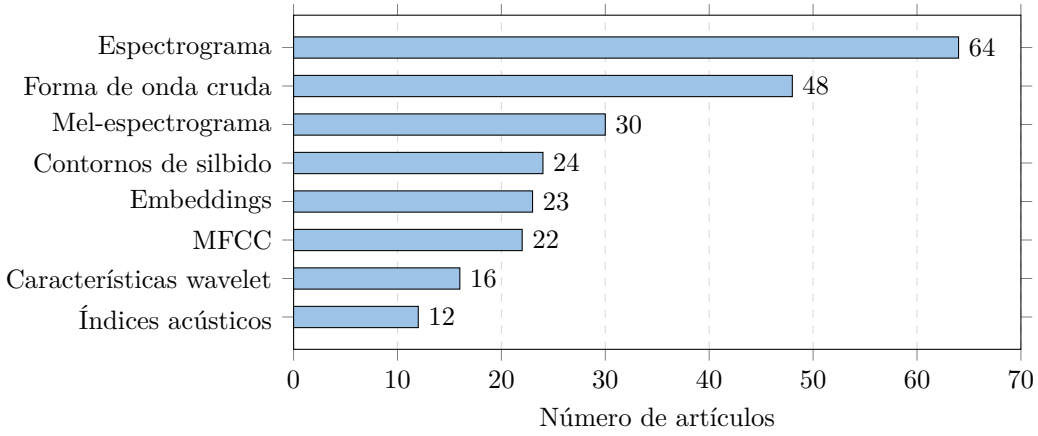
La Figura 5.1 permite observar que el campo combina distintas familias metodológicas dentro de un mismo sistema de reconocimiento. La alta presencia de transferencia de aprendizaje y ajuste fino indica que gran parte de la literatura reutiliza representaciones previamente aprendidas y las adapta a dominios acústicos específicos. Ghani et al. (2023) emplean *embeddings* globales de cantos de aves para fortalecer la transferencia en clasificación bioacústica, mientras que Schiavo et al. (2025) ajustan BirdNET para monitoreo ecoacústico en bosques alpinos italianos. Ziegenhorn et al. (2025) desarrollan ArcticSoundsNET a partir de *embeddings* de BirdNET y los aplican a especies árticas, lo que muestra el valor de las representaciones preentrenadas cuando el dominio local difiere del conjunto de entrenamiento original. La frecuencia de enfoques híbridos o de fusión

coincide con una tendencia metodológica reciente: mejorar la robustez mediante la combinación de entradas, clasificadores o estrategias de decisión, en lugar de depender únicamente de una arquitectura (Dias et al., 2025; Xie y Zhu, 2019).

La evidencia sobre representaciones acústicas refuerza esta interpretación. En el corpus, el espectrograma apareció en 64 artículos, equivalente al 95,5 %; la forma de onda cruda en 48 artículos, equivalente al 71,6 %; el mel-espectrograma en 30 artículos, equivalente al 44,8 %; el contorno de silbido en 24 artículos, equivalente al 35,8 %; los *embeddings* en 23 artículos, equivalente al 34,3 %; los MFCC en 22 artículos, equivalente al 32,8 %; las características basadas en *wavelet* en 16 artículos, equivalente al 23,9 %; y los índices acústicos en 12 artículos, equivalente al 17,9 %. La Figura 5.2 resume estas representaciones bajo un conteo no excluyente.

Figura 5.2

Representaciones acústicas más frecuentes detectadas en el corpus



Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los valores corresponden a un conteo no excluyente calculado sobre los 67 estudios incluidos en la revisión; por tanto, un mismo artículo puede registrar más de una representación acústica de entrada. MFCC = coeficientes cepstrales en las frecuencias de Mel.

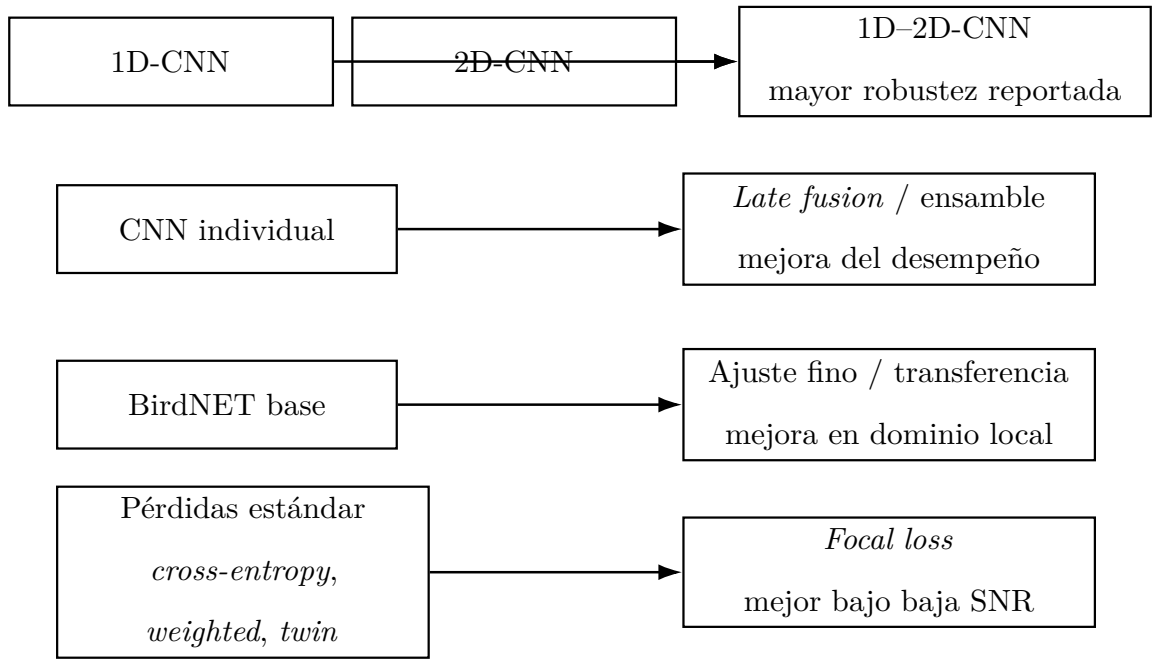
La Figura 5.2 muestra que el reconocimiento bioacústico reciente se apoya principalmente en representaciones tiempo–frecuencia. Este patrón es coherente con el predominio de las CNN, ya que los espectrogramas y mel-espectrogramas transforman el audio en matrices donde las vocalizaciones aparecen como patrones de energía distribuidos en tiempo y frecuencia (Kiskin et al., 2020; Stowell et al., 2019). LeBien et al. (2020) emplean mel-espectrogramas para entrenar una

CNN multietiqueta en paisajes sonoros tropicales, mientras que Kiskin et al. (2020) exploran CNN condicionadas por transformadas *wavelet* para mejorar la detección bioacústica en condiciones de baja relación señal–ruido. La presencia de MFCC mantiene el vínculo con el procesamiento clásico de audio, basado en descriptores compactos de la envolvente espectral (Davis y Mermelstein, 1980). Los contornos de silbido aparecen en estudios centrados en vocalizaciones de cetáceos, como el trabajo de Jin et al. (2022) sobre extracción de silbidos de delfines mediante segmentación semántica. Esta diversidad confirma que el preprocesamiento acústico define qué rasgos quedan disponibles para el modelo y qué patrones podrán ser aprendidos durante el entrenamiento (Kershenbaum et al., 2025; Knight et al., 2017b).

La pregunta de investigación también exige considerar el desempeño reportado. El corpus no permite afirmar que exista una arquitectura universalmente superior para todos los contextos, especies o condiciones acústicas. Los mejores desempeños tienden a reportarse cuando las arquitecturas profundas se integran con representaciones informativas, estrategias de adaptación al dominio, aumento de datos, fusión de modelos o ajuste fino de modelos preentrenados. Esta relación aparece de manera clara en estudios representativos: Xie et al. (2019) comparan distintos modelos CNN para clasificación de cantos de aves; Xie et al. (2021) evalúan funciones de pérdida para mejorar la clasificación bioacústica; Xie y Zhu (2019) integran características manuales con aprendizaje profundo mediante *late fusion*; y los trabajos derivados de BirdNET muestran que la adaptación local puede mejorar la utilidad de modelos preentrenados (Kahl et al., 2021; Schiavo et al., 2025; Ziegenhorn et al., 2025).

Figura 5.3

Síntesis cualitativa de comparaciones de desempeño reportadas en estudios representativos del corpus



Nota. Elaboración propia con base en estudios representativos del corpus sobre comparación de arquitecturas CNN, fusión de representaciones, ajuste fino de modelos preentrenados y funciones de pérdida para clasificación bioacústica (Kahl et al., 2021; Schiavo et al., 2025; Xie y Zhu, 2019; Xie et al., 2019, 2021; Ziegenhorn et al., 2025). La figura resume relaciones cualitativas observadas en la literatura revisada y no constituye una comparación estadística directa entre todos los artículos incluidos. SNR = relación señal–ruido.

La Figura 5.3 sintetiza patrones cualitativos de desempeño identificados en estudios representativos del corpus. La figura debe interpretarse como una organización conceptual de relaciones observadas en la literatura revisada, no como una comparación estadística directa entre todos los artículos incluidos. Xie et al. (2019) comparan modelos CNN para clasificación de cantos de aves, lo que respalda la línea de contraste entre arquitecturas 1D, 2D y combinaciones convolucionales orientadas a mejorar el reconocimiento bioacústico. En una dirección complementaria, Xie et al. (2021) analizan funciones de pérdida aplicadas a CNN y muestran que la elección del criterio de entrenamiento modifica el desempeño bajo condiciones acústicas difíciles, especialmente cuando la relación señal–ruido es baja. La fusión de representaciones aparece ejemplificada por Xie y Zhu (2019), quien combina características diseñadas manualmente con aprendizaje profundo mediante *late fusion* para clasificación de sonidos de aves. La adaptación al dominio se observa en trabajos basados en BirdNET: Kahl et al. (2021) presentan el modelo base

para monitoreo de diversidad aviar, mientras que Schiavo et al. (2025) y Ziegenhorn et al. (2025) reutilizan o ajustan sus representaciones para contextos locales, como bosques alpinos italianos o especies árticas. Estos ejemplos muestran que las mejoras de desempeño pueden originarse en la arquitectura, la representación acústica, la función de pérdida, la fusión de información o la adaptación del modelo a un dominio específico de aplicación.

La respuesta a la RQ1 evidencia que la literatura reciente se concentra principalmente en arquitecturas CNN y en representaciones tiempo–frecuencia, con predominio de los espectrogramas como entrada acústica. El corpus revisado no permite sostener la superioridad universal de una arquitectura específica. Los desempeños más sólidos aparecen cuando el modelo se inserta dentro de un *pipeline* metodológicamente consistente, compuesto por segmentación adecuada, representación acústica pertinente, adaptación al dominio y protocolos de evaluación alineados con las condiciones reales de aplicación (Kershenbaum et al., 2025; Knight et al., 2017b). La complejidad arquitectónica tiene un valor condicionado: puede mejorar el reconocimiento cuando opera sobre datos bien preparados y representaciones informativas, pero pierde capacidad explicativa si se analiza de forma aislada respecto al preprocesamiento, la calidad del conjunto de datos y la rigurosidad de la validación (LeBien et al., 2020; Lostanlen et al., 2019).

5.2 RQ2. ¿CÓMO VARÍA LA PRECISIÓN DE LOS MODELOS ANTE RUIDO AMBIENTAL, CALIDAD DEL DATASET Y HETEROGENEIDAD ACÚSTICA?

La evidencia revisada muestra que la precisión de los modelos bioacústicos no puede entenderse como una propiedad fija de la arquitectura. El rendimiento reportado depende de las condiciones acústicas, la calidad del conjunto de datos, la distribución de clases y la similitud entre el contexto de entrenamiento y el contexto de evaluación (Kershenbaum et al., 2025; Knight et al., 2017b). Un modelo puede obtener valores altos en un conjunto controlado y perder estabilidad cuando se aplica a grabaciones reales con ruido meteorológico, variación entre sensores, fondos acústicos no estacionarios o especies poco representadas (Lostanlen et al., 2019; Stowell et al., 2019). Esta relación permite interpretar la RQ2 desde una perspectiva de robustez: la precisión no expresa únicamente capacidad de clasificación, también refleja la resistencia del sistema frente a cambios ambientales y metodológicos.

Entre los factores que más afectan el desempeño se identifican el ruido meteorológico, la baja relación señal–ruido, el solapamiento entre fuentes acústicas, la variabilidad del fondo sonoro, las diferencias entre sensores y la distancia entre condiciones de entrenamiento y despliegue (Lostanlen et al., 2019; Teixeira et al., 2024). Estos factores modifican la información disponible para el modelo, ya que pueden ocultar la vocalización objetivo, introducir patrones espurios o alterar la forma en que una especie aparece representada en el espectrograma (Kiskin et al., 2020; Knight et al., 2017b). Lostanlen et al. (2019) muestran este problema en redes de sensores bioacústicos, donde el ruido de fondo varía en el tiempo y en el espacio, por lo que un detector entrenado bajo ciertas condiciones puede perder confiabilidad cuando se aplica en otros sitios o periodos de muestreo. En el Bird Audio Detection Challenge, Stowell et al. (2019) también evidencian que la detección automática de aves mediante aprendizaje profundo enfrenta una variabilidad considerable asociada con grabaciones reales y condiciones acústicas heterogéneas.

La calidad del conjunto de datos condiciona la precisión alcanzable. El desbalance de clases puede favorecer el reconocimiento de especies frecuentes y reducir el desempeño en especies raras o con pocas observaciones; el ruido de etiquetas introduce inconsistencias entre el contenido real de la grabación y la clase asignada; la escasez de ejemplos limita la capacidad del modelo para aprender patrones estables (Assarzadeh et al., 2025; Knight et al., 2017b). Assarzadeh et al. (2025). analizan este problema en BirdNET y muestran que la misclasificación puede relacionarse con ruido de etiquetas, distribución desigual de datos, sonidos no objetivo y ambigüedad acústica entre clases. Por ello, parte de la variación en el desempeño reportado por la literatura se explica por la estructura estadística del dataset, no únicamente por diferencias entre arquitecturas (Assarzadeh et al., 2025; Kershenbaum et al., 2025).

La distribución de métricas reportadas refuerza esta lectura. En el corpus, *recall* apareció en 62 artículos, equivalente al 92,5 %; *accuracy* en 61 artículos, equivalente al 91,0 %; *precision* en 59 artículos, equivalente al 88,1 %; F1-score en 44 artículos, equivalente al 65,7 %; y mAP en 43 artículos, equivalente al 64,2 %. En cambio, AUPRC apareció en 17 estudios, equivalente al 25,4 %. Esta distribución muestra una tensión metodológica: aunque la literatura reconoce que el desbalance, el ruido y la generalización afectan el rendimiento, la evaluación práctica sigue apoyándose con alta frecuencia en métricas generales. En escenarios con clases desbalanceadas, una *accuracy* elevada puede ocultar bajo desempeño en especies minoritarias, mientras que

métricas como F1-score, AUPRC, matrices de confusión y evaluación por clase permiten observar mejor los errores de comisión y omisión (Fawcett, 2006; Knight et al., 2017b).

La revisión también identificó estrategias orientadas a mitigar la pérdida de desempeño ante ruido o baja calidad de datos. Entre ellas se encuentran el aumento de datos, funciones de pérdida más robustas, ajuste fino de modelos preentrenados, curación del conjunto de datos, normalización energética, técnicas adaptativas frente al ruido de fondo y validaciones más exigentes en condiciones no vistas (Kiskin et al., 2020; Lostanlen et al., 2019; Xie et al., 2021). Kiskin et al. (2020) proponen CNN condicionadas por wavelets para detección bioacústica, una estrategia dirigida a mejorar la representación de señales bajo baja relación señal–ruido. Xie et al. (2021) analizan distintas funciones de pérdida aplicadas a CNN y muestran que el criterio de entrenamiento puede modificar el desempeño del modelo ante condiciones acústicas difíciles. En modelos preentrenados, trabajos basados en BirdNET muestran que el ajuste fino y el uso de *embeddings* pueden mejorar la adaptación a dominios locales, aunque esta ventaja depende de las características del conjunto de datos y de la validación aplicada (Ghani et al., 2025; Kahl et al., 2021; Schiavo et al., 2025; Ziegenhorn et al., 2025).

El contraste entre estudios permite distinguir limitaciones asociadas al entorno acústico y limitaciones originadas en la estructura del conjunto de datos. Algunas investigaciones reportan restricciones vinculadas con ruido meteorológico, fondos sonoros no estacionarios o variación entre sensores (Lostanlen et al., 2019; Stowell et al., 2019). En estos casos, la pérdida de desempeño suele explicarse por la distancia entre las condiciones acústicas representadas durante el entrenamiento y aquellas encontradas durante el despliegue, más que por una incapacidad intrínseca del modelo. Otros estudios, centrados en análisis de errores, muestran que parte del bajo rendimiento puede surgir del propio dataset, especialmente cuando existen etiquetas ruidosas, clases desbalanceadas o sonidos no objetivo (Assarzadeh et al., 2025). Esta distinción orienta la elección de respuestas metodológicas: los problemas de dominio requieren validación externa, adaptación o preprocesamiento robusto; los problemas del dataset exigen curación de etiquetas, revisión de clases minoritarias y métricas sensibles al desbalance (Kershenbaum et al., 2025; Knight et al., 2017b).

Las estrategias de mejora deben seleccionarse según el origen del error. Si la pérdida de

desempeño proviene del ambiente acústico, resultan pertinentes procedimientos como aumento de datos, normalización energética, preprocesamiento robusto y validación con condiciones no observadas durante el entrenamiento (Kiskin et al., 2020; Lostanlen et al., 2019). Si el problema se relaciona con la estructura del conjunto de datos, adquieren mayor peso la curación de etiquetas, el análisis de errores de clasificación, la revisión de clases minoritarias y el uso de métricas sensibles al desbalance (Assarzadeh et al., 2025; Knight et al., 2017b). Reportar que un modelo pierde precisión ante ruido o heterogeneidad acústica ofrece una descripción incompleta; el análisis debe precisar si esa pérdida responde a una representación insuficiente, a un cambio de dominio, a errores de anotación o a una distribución desigual de los datos.

La respuesta a la RQ2 es que la precisión disminuye o se vuelve menos confiable cuando aumentan el ruido, la heterogeneidad acústica y las limitaciones del conjunto de datos. La variación del rendimiento responde a causas metodológicamente identificables: datos desbalanceados, etiquetas imperfectas, diferencias entre sitios, variabilidad de sensores, fondos acústicos no estacionarios y cambios entre el dominio de entrenamiento y el dominio de aplicación (Assarzadeh et al., 2025; Knight et al., 2017b; Lostanlen et al., 2019). Por esta razón, evaluar el rendimiento implica evaluar también la robustez ecológica del sistema. Comparar modelos sin considerar las condiciones acústicas, la estructura del dataset y el protocolo de validación puede producir conclusiones metodológicamente débiles, aunque las métricas reportadas parezcan altas (Kershenbaum et al., 2025; Knight et al., 2017b).

5.3 RQ3. ¿QUÉ ALCANCES Y LIMITACIONES PRESENTAN LOS MÉTODOS DE RECONOCIMIENTO ACÚSTICO PARA ESTIMAR PRESENCIA, OCUPACIÓN O ABUNDANCIA?

Los resultados muestran que la literatura reciente desplaza el reconocimiento acústico más allá de la clasificación taxonómica de vocalizaciones. Una parte del corpus busca vincular las salidas de los modelos con variables ecológicas como presencia, ocupación, actividad relativa, abundancia o cuantificación de eventos acústicos (Saoud et al., 2024; Teixeira et al., 2024). Esta ampliación conecta el reconocimiento automático con preguntas de monitoreo y conservación, pero exige una interpretación metodológica cuidadosa: detectar o clasificar una vocalización no equivale, por sí mismo, a estimar una variable ecológica compleja (Knight et al., 2017b;

Runge, 2023). Para sostener inferencias de ese tipo, las predicciones acústicas deben relacionarse con detectabilidad, diseño de muestreo, estructura temporal de las observaciones y supuestos ecológicos explícitos (Runge, 2023; Teixeira et al., 2024).

En el conteo no excluyente, detección y clasificación aparecieron en los 67 artículos del corpus, lo que representa el 100,0 %. Segmentación apareció en 62 artículos, equivalente al 92,5 %; abundancia o cuantificación en 49 artículos, equivalente al 73,1 %; denoising en 25 artículos, equivalente al 37,3 %; aprendizaje de embeddings en 23 artículos, equivalente al 34,3 %; presencia-ausencia en 18 artículos, equivalente al 26,9 %; ocupación en 17 artículos, equivalente al 25,4 %; y separación de fuentes en 8 artículos, equivalente al 11,9 %. La Tabla 5.3 presenta esta distribución.

Tabla 5.3

Tareas bioacústicas detectadas en el corpus

Tarea detectada	**n**	**%**
Detección	67	100,0
Clasificación	67	100,0
Segmentación	62	92,5
Abundancia o cuantificación	49	73,1
Reducción de ruido (*denoising*)	25	37,3
Aprendizaje de *embeddings*	23	34,3
Presencia-ausencia	18	26,9
Ocupación	17	25,4
Separación de fuentes	8	11,9

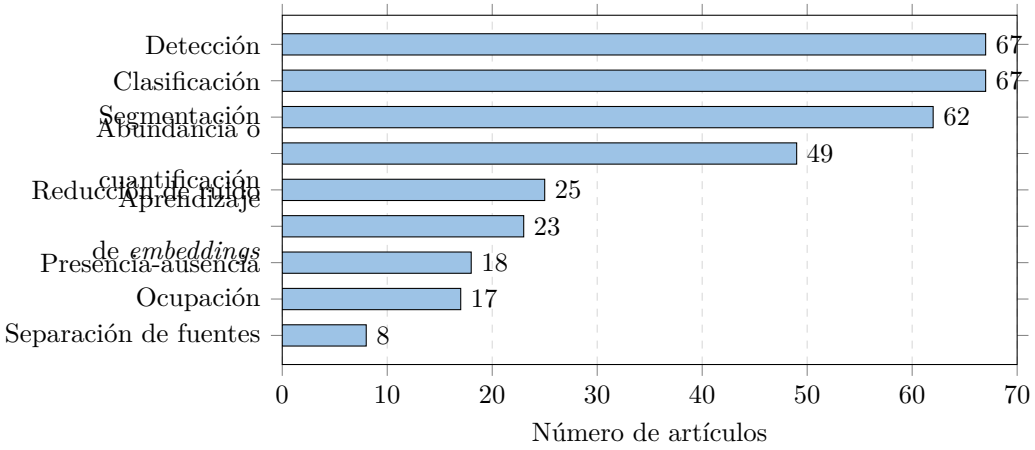
Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los porcentajes se calcularon sobre el total de estudios incluidos en la revisión ($n = 67$). Las categorías corresponden a un conteo no excluyente, por lo que un mismo artículo puede estar asociado con más de una tarea bioacústica.

La Tabla 5.3 debe leerse como un conteo no excluyente. Un mismo artículo puede estar asociado con varias tareas, por ejemplo detección, clasificación y presencia-ausencia, porque

muchos estudios bioacústicos operan mediante flujos compuestos: primero localizan eventos acústicos, luego asignan clases y después interpretan esas salidas en relación con una pregunta ecológica. LeBien et al. (2020) desarrollan un *pipeline* para aves y ranas en paisajes sonoros tropicales donde las detecciones se transforman en probabilidades de presencia por especie y por ventana temporal. En ese tipo de flujo, la tarea computacional y la interpretación ecológica están conectadas, aunque no son equivalentes. La tabla permite observar qué objetivos aparecen en el corpus, sin asumir que todos tienen el mismo grado de madurez metodológica.

Figura 5.4

Distribución de tareas bioacústicas detectadas en el corpus



Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los valores corresponden a un conteo no excluyente calculado sobre los 67 estudios incluidos en la revisión; por tanto, un mismo artículo puede estar asociado con más de una tarea bioacústica.

La Figura 5.4 permite diferenciar tareas computacionales de base y tareas ecológicas derivadas. Detección, clasificación y segmentación aparecen con frecuencias muy altas, lo que confirma que el núcleo técnico del campo sigue orientado a localizar eventos acústicos y asignarlos a clases. Presencia-ausencia y ocupación aparecen con menor frecuencia, lo que sugiere que las inferencias ecológicas explícitas ocupan todavía una posición menos desarrollada dentro del corpus. Stowell et al. (2019) ilustran este puente metodológico al comparar salidas de clasificación automática de aves en grabaciones de larga duración con datos anotados manualmente para producir resultados de modelos de ocupación similares. Ese caso muestra que la ocupación puede derivarse de clasificadores acústicos, siempre que exista un diseño analítico que conecte detecciones, repetición temporal, probabilidad de detección y modelo ecológico.

La categoría abundancia o cuantificación requiere una lectura especialmente cautelosa. Su frecuencia elevada no significa que 49 estudios estimen abundancia real con validación ecológica completa. En varios casos, la categoría puede referirse al conteo de eventos acústicos, la cuantificación de detecciones, la actividad vocal o medidas indirectas de intensidad acústica. Dias et al. (2022) proponen un enfoque de clasificación y cuantificación para generar características en ecología del paisaje sonoro mediante redes neuronales, lo que ejemplifica una aproximación orientada a cuantificar componentes acústicos del ambiente, no necesariamente a estimar individuos de forma directa. Esta distinción es metodológicamente necesaria porque una mayor cantidad de vocalizaciones detectadas puede reflejar mayor actividad vocal, diferencias de detectabilidad, comportamiento temporal o repetición de llamados por un mismo individuo, antes que un aumento real de abundancia (Knight et al., 2017b; Runge, 2023).

Las tareas de *denoising* y separación de fuentes muestran otra línea de desarrollo relacionada con la calidad de la señal antes de la inferencia ecológica. En grabaciones naturales, el ruido ambiental, el solapamiento entre especies y la presencia de sonidos no objetivo pueden afectar tanto la detección como la clasificación (Kershenbaum et al., 2025; Lostanlen et al., 2019). M. Wang et al. (2025) abordan este problema mediante redes generativas adversariales aplicadas a identificación acústica, reducción de ruido y separación de fuentes animales, lo que muestra que algunas estrategias recientes buscan mejorar la entrada acústica antes de producir una predicción ecológicamente interpretable. Estos procedimientos pueden fortalecer la detección en paisajes sonoros complejos, pero no reemplazan la necesidad de validar si la salida corregida conserva información biológica suficiente para sostener inferencias sobre presencia, ocupación o abundancia (Kershenbaum et al., 2025; Knight et al., 2017b).

Los métodos de reconocimiento acústico ofrecen un potencial claro para apoyar inferencias ecológicas, aunque su validez depende de la cadena que conecta la salida del modelo con la variable de interés. Para que una predicción sostenga una inferencia sobre presencia, ocupación o abundancia, se requiere definir la unidad de análisis, controlar la detectabilidad, justificar la relación entre vocalización y variable ecológica y validar los resultados frente a un criterio independiente cuando sea posible (Runge, 2023; Teixeira et al., 2024). Sin esa cadena de validación, la salida del modelo debe interpretarse como evidencia acústica procesada, no como una estimación ecológica directa. Esta distinción evita tratar una probabilidad de clasificación, un

conteo de llamadas o una detección positiva como si fueran equivalentes a presencia confirmada, uso del sitio o número de individuos (Kershenbaum et al., 2025; Knight et al., 2017b).

La respuesta a la RQ3 indica que el reconocimiento acústico automatizado muestra mayor madurez para tareas de detección, clasificación y presencia-ausencia que para estimaciones directas de ocupación o abundancia. La literatura avanza hacia aplicaciones ecológicas más ambiciosas, pero persiste una brecha entre la salida algorítmica y la inferencia ecológica. Esa brecha no invalida el uso de los modelos; exige interpretar sus resultados según el diseño de muestreo, la detectabilidad, el comportamiento vocal de las especies y la incertidumbre asociada al clasificador (Knight et al., 2017b; Teixeira et al., 2024). Una mayor cantidad de vocalizaciones detectadas no puede asumirse como evidencia directa de mayor número de individuos, ya que la relación entre detección acústica, actividad vocal, detectabilidad y abundancia depende del contexto ecológico y del protocolo de validación (Kershenbaum et al., 2025; Runge, 2023). Por tanto, las salidas del modelo deben entenderse como evidencia acústica procesada y no como equivalentes directos de variables ecológicas complejas.

5.4 RQ4. ¿QUÉ MÉTRICAS, PROTOCOLOS DE VALIDACIÓN Y TENDENCIAS METODOLÓGICAS PREDOMINAN EN LA LITERATURA RECIENTE?

El análisis del corpus muestra que la evaluación del rendimiento en reconocimiento acústico automatizado se apoya principalmente en métricas clásicas de clasificación y detección. Esta tendencia facilita una comparación inicial entre estudios, aunque también revela una limitación metodológica: las métricas más frecuentes no siempre son las más informativas cuando existen clases desbalanceadas, ruido ambiental, especies raras o cambios de dominio (Fawcett, 2006; Knight et al., 2017b). La RQ4 examina, por ello, qué indicadores predominan, cómo se usan los protocolos de validación y qué líneas de desarrollo metodológico aparecen con mayor fuerza en la literatura reciente.

En términos cuantitativos, las métricas más frecuentes fueron *recall*, reportada en 62 artículos, equivalente al 92,5 %; *accuracy*, reportada en 61 artículos, equivalente al 91,0 %; *precision*, reportada en 59 artículos, equivalente al 88,1 %; F1-score, reportada en 44 artículos,

equivalente al 65,7 %; y mAP, reportada en 43 artículos, equivalente al 64,2 %. En un segundo nivel aparecen *sensitivity* en 32 artículos, equivalente al 47,8 %; AUPRC y AUC-ROC en 17 artículos cada una, equivalentes al 25,4 %; matriz de confusión en 16 artículos, equivalente al 23,9 %; *specificity* en 9 artículos, equivalente al 13,4 %; y MAE e IoU en 6 artículos cada una, equivalentes al 9,0 %. La Tabla 5.4 presenta esta distribución.

Tabla 5.4

Métricas y recursos de evaluación más frecuentes en el corpus

Métrica o recurso de evaluación	**n**	**%**
Recall	62	92,5
Accuracy	61	91,0
Precision	59	88,1
F1-score	44	65,7
mAP	43	64,2
Sensitivity	32	47,8
AUPRC	17	25,4
AUC-ROC	17	25,4
Matriz de confusión	16	23,9
Specificity	9	13,4
MAE	6	9,0
IoU	6	9,0

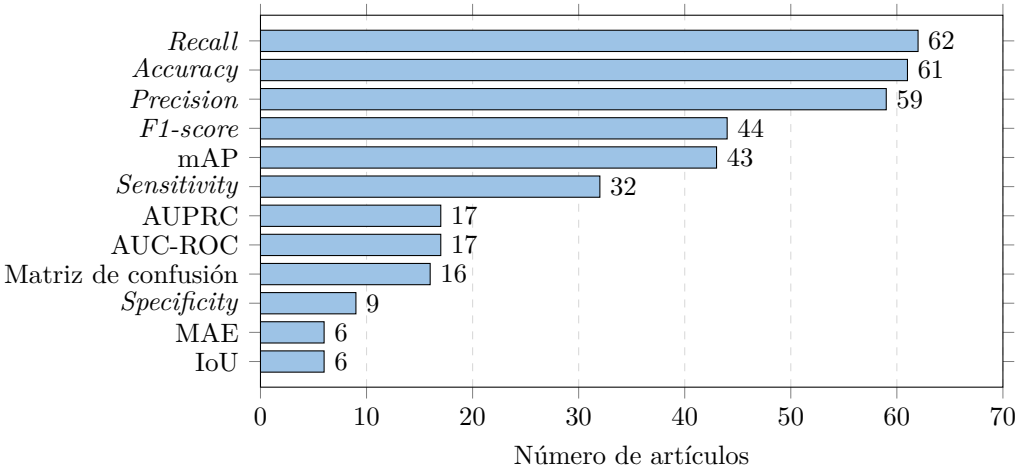
Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los porcentajes se calcularon sobre el total de estudios incluidos en la revisión ($n = 67$). Las categorías corresponden a un conteo no excluyente, por lo que un mismo artículo puede reportar más de una métrica o recurso de evaluación. mAP = precisión media promedio; AUPRC = área bajo la curva precisión-recall; AUC-ROC = área bajo la curva ROC; MAE = error absoluto medio; IoU = intersección sobre unión.

La Tabla 5.4 muestra que el corpus utiliza una base común de indicadores de desempeño, aunque con distintos niveles de profundidad evaluativa. La frecuencia de *recall*, *accuracy* y *precision* permite describir aspectos generales del rendimiento, pero estos valores agregados no

siempre revelan cómo se distribuyen los errores entre especies, eventos acústicos o condiciones de grabación. Knight et al. (2017b) advierten que la evaluación de reconocedores acústicos debe considerar falsos positivos, falsos negativos, sensibilidad, especificidad y contexto de aplicación, debido a que una métrica global puede ocultar fallas relevantes para el monitoreo ecológico. En escenarios con clases desbalanceadas, indicadores como AUPRC, matrices de confusión, *specificity* y evaluaciones por clase aportan información más precisa sobre errores concentrados en especies poco representadas (Fawcett, 2006; Knight et al., 2017b). La baja presencia relativa de estos indicadores sugiere que la evaluación bioacústica aún presenta criterios heterogéneos para escenarios acústicos complejos.

Figura 5.5

Frecuencia de métricas y recursos de evaluación detectados en el corpus



Nota. Elaboración propia a partir de la matriz de extracción del corpus final. Los valores corresponden a un conteo no excluyente calculado sobre los 67 estudios incluidos en la revisión; por tanto, un mismo artículo puede reportar más de una métrica o recurso de evaluación. mAP = precisión media promedio; AUPRC = área bajo la curva precisión-recall; AUC-ROC = área bajo la curva ROC; MAE = error absoluto medio; IoU = intersección sobre unión.

La Figura 5.5 permite diferenciar métricas de uso frecuente y métricas con menor presencia dentro del corpus revisado. Esta distribución muestra que la evaluación del desempeño no puede depender de una única medida, especialmente cuando los estudios trabajan con especies raras, grabaciones ruidosas o clases desbalanceadas. Una métrica global puede ofrecer una lectura favorable del modelo y, al mismo tiempo, ocultar fallas concentradas en categorías minoritarias. Una alta *accuracy*, por ejemplo, puede reflejar aciertos sobre especies frecuentes sin revelar errores

sistemáticos en especies escasas o de mayor interés ecológico. En estos escenarios, indicadores como *F1-score*, AUPRC, matrices de confusión y evaluaciones por clase permiten examinar con mayor precisión dónde se producen los errores y qué tan confiable resulta el modelo para los objetivos ecológicos del estudio (Fawcett, 2006; Knight et al., 2017b).

El contraste entre métricas evidencia una diferencia entre evaluación general y evaluación ecológicamente informativa. *Accuracy*, *precision* y *recall* describen dimensiones básicas del desempeño, pero pueden resultar insuficientes cuando las especies presentan frecuencias muy desiguales o cuando los falsos positivos y falsos negativos tienen consecuencias ecológicas distintas (Knight et al., 2017b). En reconocimiento bioacústico, un falso positivo puede inflar la presencia de una especie en un sitio, mientras que un falso negativo puede ocultar actividad vocal real. Esta asimetría exige seleccionar métricas según la pregunta ecológica, la unidad de análisis y el costo interpretativo del error, no por simple disponibilidad o uso habitual (Kershenbaum et al., 2025; Knight et al., 2017b).

La comparación entre estudios también se complica cuando las métricas se calculan sobre unidades de análisis diferentes. Algunos trabajos evalúan segmentos de audio; otros, eventos detectados, clips completos, ventanas temporales o listas de especies presentes en una grabación. Esta variación modifica el significado de una misma métrica: un F1-score calculado sobre eventos acústicos no es directamente equivalente a un F1-score calculado sobre presencia de especies por archivo. La discusión de métricas debe considerar el indicador reportado, la unidad de evaluación, el umbral de decisión y el tipo de tarea, porque esos elementos definen qué error se está midiendo y qué inferencia puede sostenerse a partir del resultado (Kershenbaum et al., 2025; Knight et al., 2017b).

La revisión muestra también heterogeneidad en los protocolos de validación. La comparación entre estudios se vuelve difícil cuando varían simultáneamente la tarea evaluada, la unidad de análisis, el umbral de decisión, la distribución de clases, el conjunto de prueba y las condiciones acústicas. En algunos trabajos, entrenamiento y prueba comparten condiciones similares, lo que puede producir estimaciones optimistas del rendimiento. En otros, el modelo se evalúa bajo cambios de sitio, sensor o ambiente, lo que ofrece una prueba más exigente de generalización. Stowell et al. (2019), mediante el Bird Audio Detection Challenge, muestran que

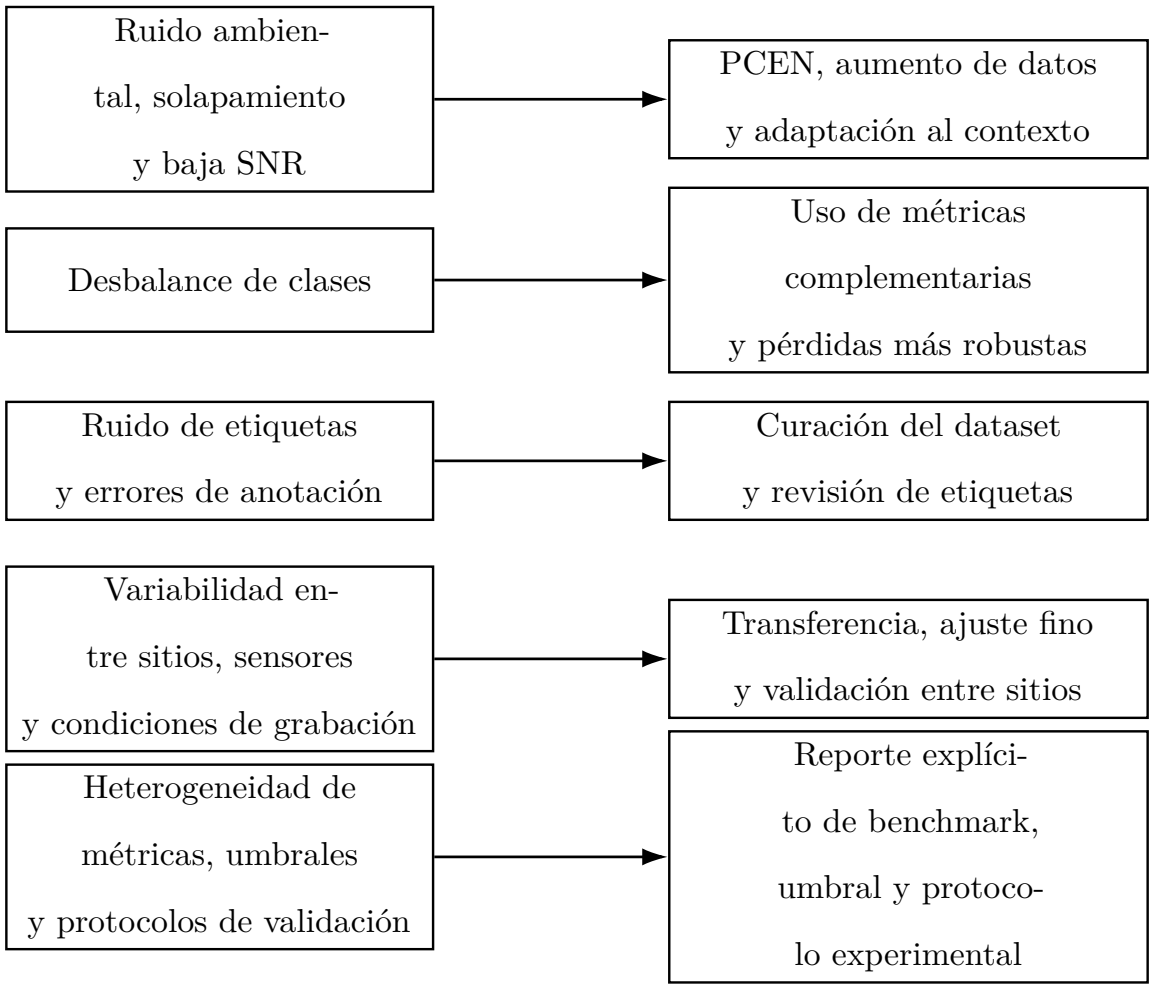
la evaluación comparativa bajo condiciones acústicas reales permite observar diferencias entre modelos que no serían visibles en escenarios más controlados. Lostanlen et al. (2019) refuerzan esta lectura al estudiar detección robusta en redes de sensores bioacústicos, donde la variación espacial y temporal del ruido afecta la confiabilidad del detector.

En este contexto, la validación *cross-site*, *cross-sensor* o bajo condiciones acústicas no observadas durante el entrenamiento constituye una exigencia metodológica para estimar robustez. Su valor radica en que permite distinguir si el modelo aprendió patrones transferibles o si depende de características específicas del conjunto de entrenamiento, como un sensor, un sitio, una temporada o un fondo acústico particular (Kershenbaum et al., 2025; Lostanlen et al., 2019). Sin este tipo de validación, los resultados pueden ser útiles dentro del conjunto evaluado, pero débiles como evidencia de aplicabilidad ecológica general. Los modelos preentrenados ilustran este punto: BirdNET amplía la cobertura taxonómica y facilita el análisis de grandes volúmenes de audio, aunque su aplicación en contextos locales requiere calibración, ajuste de umbrales o validación específica para evitar errores de transferencia (Kahl et al., 2021; Pérez-Granados, 2023; Schiavo et al., 2025; Ziegenhorn et al., 2025).

Respecto a las tendencias metodológicas, el corpus muestra que las CNN continúan como base técnica predominante y se amplían mediante transferencia de aprendizaje, ajuste fino, fusión de modelos, arquitecturas híbridas, aumento de datos y exploración gradual de transformadores. Esta configuración revela una etapa de madurez parcial: existe una base metodológica consolidada, pero el campo sigue buscando mecanismos para mejorar robustez, adaptación y comparabilidad. Kiskin et al. (2020) ejemplifican esta búsqueda al usar CNN condicionadas por wavelets para mejorar la detección bioacústica bajo baja relación señal–ruido. Schiavo et al. (2025) y Ziegenhorn et al. (2025) muestran otra línea de avance al adaptar representaciones de BirdNET a contextos locales mediante ajuste fino o *embeddings*. Assarzadeh et al. (2025) incorporan una perspectiva complementaria al analizar errores de clasificación y ruido de etiquetas en BirdNET, mostrando que el desempeño agregado no basta para comprender por qué falla un sistema.

Figura 5.6

Factores que condicionan el rendimiento en escenarios reales y respuestas metodológicas reportadas en la literatura



Nota. Elaboración propia con base en los factores y respuestas metodológicas identificadas en la literatura revisada (Kahl et al., 2021; Kershenbaum et al., 2025; Kiskin et al., 2020; Knight et al., 2017b; Lostanlen et al., 2019; Schiavo et al., 2025; Ziegenhorn et al., 2025). La figura resume relaciones cualitativas observadas en el corpus y no constituye una comparación estadística directa entre todos los estudios incluidos. SNR = relación señal–ruido; PCEN = normalización energética por canal.

La Figura 5.6 sintetiza factores que condicionan el rendimiento de los sistemas bioacústicos y respuestas metodológicas reportadas en la literatura. Su estructura relaciona problemas recurrentes, como ruido ambiental, desbalance de clases, cambio de dominio, ruido de etiquetas y heterogeneidad de métricas, con estrategias orientadas a mitigarlos. Esta organización permite observar que las soluciones no actúan de forma universal: cada técnica responde a una fuente específica de variación y su efecto depende del contexto acústico, del diseño del dataset y del

protocolo de evaluación. Lostanlen et al. (2019) emplean normalización energética por canal para enfrentar variaciones de ruido en redes de sensores bioacústicos, mientras que Kiskin et al. (2020). recurren a representaciones *wavelet* para mejorar la detección bajo señales difíciles. Los trabajos centrados en BirdNET muestran que el ajuste fino y los *embeddings* pueden apoyar la transferencia a dominios locales, aunque requieren validación específica (Kahl et al., 2021; Schiavo et al., 2025; Ziegenhorn et al., 2025). Procedimientos como PCEN, aumento de datos, funciones de pérdida robustas o *fine-tuning* pueden mejorar el desempeño, pero no reemplazan la validación en condiciones reales ni el reporte transparente de umbrales, unidades de evaluación y protocolos experimentales (Kershenbaum et al., 2025; Knight et al., 2017b).

La respuesta a la RQ4 muestra que la literatura reciente ha ampliado el uso de métricas y estrategias metodológicas para evaluar modelos de reconocimiento acústico, aunque mantiene una heterogeneidad evaluativa marcada. Las métricas clásicas, como *recall*, *accuracy*, *precision* y F1-score, continúan predominando, mientras que medidas sensibles al desbalance de clases, como AUPRC, *specificity* o matrices de confusión detalladas, se reportan con menor frecuencia. Esta asimetría limita la comparación directa entre estudios y puede ocultar errores en especies raras, clases minoritarias o escenarios acústicos complejos. La evidencia revisada indica que uno de los retos del campo consiste en desarrollar modelos más precisos y fortalecer, al mismo tiempo, los criterios de evaluación, reporte e interpretación para estimar con mayor rigor su aplicabilidad en condiciones ecológicas reales (Kershenbaum et al., 2025; Knight et al., 2017b; Lostanlen et al., 2019).

5.4.1 CONTRASTE CRÍTICO ENTRE PROPUESTAS METODOLÓGICAS REPRESENTATIVAS

El contraste entre propuestas metodológicas representativas permite matizar la interpretación de las frecuencias reportadas. Aunque las CNN constituyen la familia metodológica dominante en el corpus, su desempeño no puede interpretarse de manera independiente de la tarea evaluada, la escala del conjunto de datos, la representación acústica, el protocolo de validación y el contexto ecológico de aplicación. La comparación entre estudios debe formularse como una evaluación de configuraciones metodológicas que responden a problemas distintos, evitando la búsqueda de una arquitectura universalmente superior (Kershenbaum et al., 2025;

Knight et al., 2017b).

Una primera diferencia aparece entre enfoques orientados a detección acústica a gran escala y enfoques diseñados para robustez en condiciones de campo. Los sistemas de detección profunda aplicados a grandes colecciones de audio muestran la utilidad de representaciones espectrotemporales y arquitecturas neuronales para procesar datos acústicos extensos (Stowell et al., 2019). Las redes de sensores bioacústicos, en cambio, evidencian que la variación espacial, temporal y ambiental del ruido puede afectar la estabilidad del detector cuando se aplica fuera de las condiciones representadas durante el entrenamiento (Lostanlen et al., 2019). Esta diferencia indica que escalabilidad computacional y robustez ecológica no son propiedades equivalentes: un modelo puede procesar grandes volúmenes de audio y, aun así, requerir evaluación rigurosa frente a cambios de sitio, sensor o fondo acústico.

Un segundo contraste se relaciona con modelos específicos frente a modelos preentrenados de gran escala. Los modelos desarrollados para tareas o conjuntos de datos concretos permiten controlar con mayor precisión el dominio de entrenamiento y evaluación, pero pueden presentar menor transferibilidad cuando se aplican a nuevos contextos acústicos. Sistemas preentrenados como BirdNET amplían la cobertura taxonómica y facilitan el análisis de grandes volúmenes de grabaciones, lo que los vuelve atractivos para monitoreo ecológico amplio (Kahl et al., 2021; Pérez-Granados, 2023). Su desempeño, aun así, depende de la cobertura geográfica, taxonómica y acústica del entrenamiento original; la transferencia hacia nuevos hábitats o comunidades requiere validación local, ajuste de umbrales o estrategias de adaptación (Schiavo et al., 2025; Ziegenhorn et al., 2025).

Un tercer contraste aparece en la estrategia utilizada para mejorar el rendimiento. En algunos estudios, la mejora proviene de reutilizar representaciones aprendidas previamente mediante transferencia de aprendizaje o ajuste fino; en otros, se busca mejorar la representación de entrada mediante transformaciones tiempo–frecuencia alternativas, como enfoques basados en wavelets (Kiskin et al., 2020; Schiavo et al., 2025). Esta diferencia desplaza el punto de intervención: la transferencia de aprendizaje actúa sobre el conocimiento ya codificado en el modelo, mientras que las representaciones alternativas modifican la información acústica entregada al clasificador. Dos sistemas pueden perseguir el mismo objetivo de mejora, pero

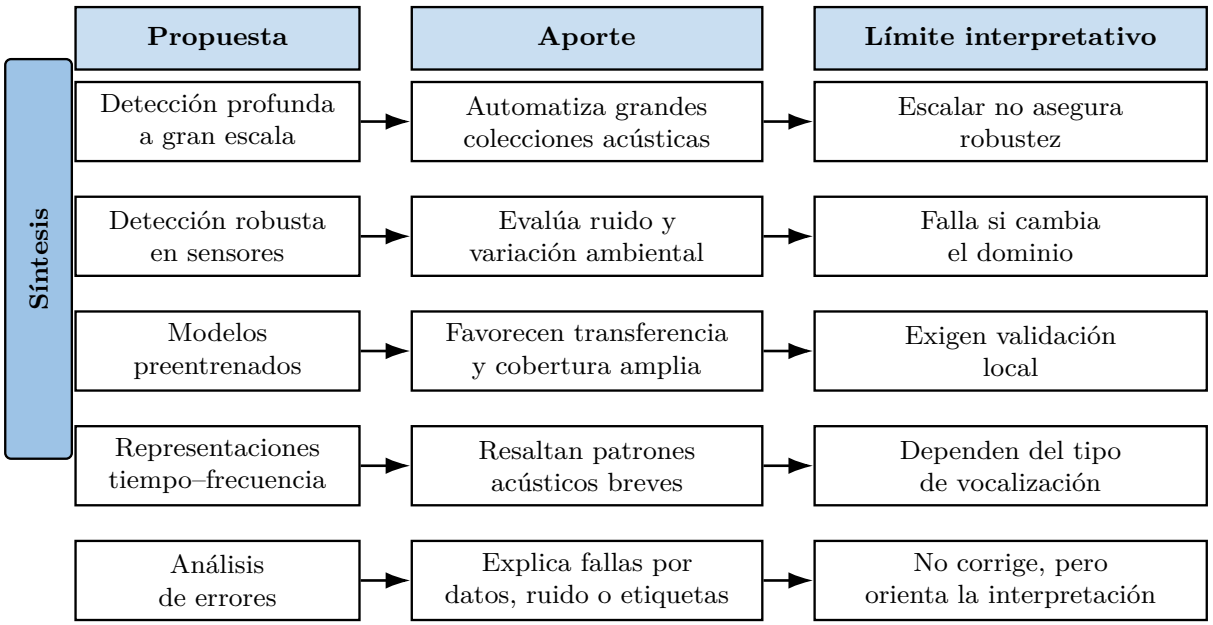
intervenir en etapas distintas del flujo analítico.

La literatura permite distinguir entre estudios orientados a maximizar métricas de desempeño y trabajos que examinan las causas de los errores de clasificación. Los enfoques centrados en el análisis del error muestran que el bajo rendimiento puede estar asociado con ruido de etiquetas, desbalance entre clases, presencia de sonidos no objetivo o ambigüedad acústica entre especies (Assarzadeh et al., 2025). Esta perspectiva amplía las comparaciones basadas en métricas agregadas, ya que permite identificar dónde se concentra el error, qué factores pueden explicarlo y qué propiedades del conjunto de datos condicionan la predicción. En reconocimiento bioacústico, este tipo de análisis aporta una dimensión de explicabilidad necesaria para valorar la confiabilidad ecológica de los sistemas automáticos.

La Figura 5.7 sintetiza estos contrastes mediante tres componentes: propuesta metodológica, aporte principal y límite interpretativo. La figura se organiza de esta manera porque cada línea metodológica responde a un problema distinto dentro del reconocimiento bioacústico. La detección profunda enfatiza la escalabilidad, la detección robusta prioriza la estabilidad frente a variación ambiental, los modelos preentrenados amplían la transferencia, las representaciones tiempo–frecuencia modifican la entrada acústica y el análisis de errores permite interpretar fallas que no son visibles mediante una métrica agregada (Assarzadeh et al., 2025; Kahl et al., 2021; Kiskin et al., 2020; Lostanlen et al., 2019; Stowell et al., 2019).

Figura 5.7

Síntesis gráfica de los aportes y límites interpretativos de propuestas metodológicas representativas del corpus



Nota. Elaboración propia con base en estudios representativos del corpus sobre escalabilidad, robustez, transferencia, representación acústica y análisis de errores en reconocimiento bioacústico (Kershenbaum et al., 2025; Knight et al., 2017b; Lostanlen et al., 2019). La figura resume relaciones cualitativas observadas en la literatura revisada y no constituye una comparación estadística directa entre todos los artículos incluidos.

En conjunto, esta lectura muestra que las propuestas representativas atienden dimensiones diferentes del problema y no se reducen a la búsqueda de una métrica más alta. La pregunta metodológica más sólida se centra en las condiciones bajo las cuales el desempeño puede considerarse robusto, transferible e interpretable, más allá del valor absoluto alcanzado por un modelo. Esta perspectiva permite diferenciar entre precisión dentro de un conjunto de prueba, estabilidad frente a variación ambiental, aplicabilidad en nuevos dominios y confiabilidad ecológica de las predicciones (Kershenbaum et al., 2025; Knight et al., 2017b; Lostanlen et al., 2019).

5.5 SÍNTESIS INTEGRAL DE LAS RQ

En conjunto, las cuatro preguntas de investigación muestran que el reconocimiento acústico automatizado de especies en entornos naturales constituye un campo con una base

técnica consolidada, pero con madurez metodológica desigual. El corpus final de 67 estudios confirma el predominio de arquitecturas convolucionales, representaciones tiempo–frecuencia y métricas clásicas de clasificación. Sin embargo, también evidencia una transición hacia estrategias más complejas, como transferencia de aprendizaje, ajuste fino, enfoques híbridos, fusión de modelos y exploración progresiva de transformadores.

El principal hallazgo integrador indica que el rendimiento surge de una configuración metodológica amplia. La evidencia revisada sugiere que la precisión y la aplicabilidad emergen de la interacción entre representación acústica, preprocesamiento, calidad del conjunto de datos, estrategia de entrenamiento, métricas de evaluación y condiciones ecológicas de aplicación. Por esta razón, no resulta metodológicamente adecuado buscar un modelo universalmente superior sin considerar el contexto experimental en el que fue entrenado y evaluado.

La revisión también identifica una tensión entre desempeño computacional y utilidad ecológica. Los modelos muestran una base sólida para detección, clasificación y presencia-ausencia, pero la inferencia sobre ocupación, abundancia o número de individuos requiere mayor validación. Esta limitación no reduce el valor del reconocimiento acústico automatizado, pero exige interpretar sus salidas como parte de un proceso ecológico más amplio y no como estimaciones directas sin mediación metodológica.

Otro límite estructural del área es la comparabilidad. La heterogeneidad de métricas, umbrales, conjuntos de datos, sensores, hábitats, tareas y protocolos de validación impide acumular evidencia de forma completamente homogénea. En consecuencia, dos estudios pueden reportar valores altos de desempeño y, aun así, no ser directamente comparables. Esta situación refuerza la necesidad de reportes más transparentes, validaciones más exigentes y análisis de error más detallados.

El contraste entre estudios representativos muestra que cada línea metodológica aporta ventajas y limitaciones distintas. Los estudios de detección a gran escala ofrecen escalabilidad, pero enfrentan problemas de robustez; los modelos preentrenados facilitan la transferencia, pero requieren validación local; las representaciones alternativas pueden mejorar la sensibilidad ante ciertos patrones acústicos, pero no eliminan la necesidad de evaluación rigurosa; y los

análisis de error muestran que parte del desempeño depende de la calidad estadística y semántica del conjunto de datos. Esta diversidad confirma que el avance del campo depende de diseños experimentales más transparentes y ecológicamente interpretables, además del desarrollo de modelos más complejos. En este punto, la contribución de la Ciencia de Datos al monitoreo ambiental depende de la capacidad predictiva de los modelos y de la solidez metodológica con la que sus resultados se validan e interpretan.

6. CONCLUSIONES

La presente monografía examinó el estado actual, la precisión y la aplicabilidad de las redes neuronales profundas en el reconocimiento acústico de especies en entornos naturales, a partir de una revisión sistemática de 67 estudios publicados entre 2017 y 2025. La evidencia revisada muestra un campo con avances técnicos relevantes, especialmente por el uso extendido de arquitecturas convolucionales, representaciones tiempo–frecuencia, modelos preentrenados y estrategias de aprendizaje por transferencia. Esta consolidación técnica debe interpretarse junto con límites metodológicos persistentes, relacionados con la calidad de los datos, la heterogeneidad acústica, la validación local, la comparabilidad entre estudios y la traducción de las predicciones computacionales hacia inferencias ecológicas.

La evidencia correspondiente a la RQ1 muestra que las redes neuronales convolucionales constituyen la arquitectura predominante en el reconocimiento acústico automatizado de especies. Esta predominancia se asocia con su compatibilidad con espectrogramas, mel-espectrogramas y otras representaciones bidimensionales del audio, donde las vocalizaciones pueden expresarse como patrones espectrotemporales. La revisión no permite establecer una arquitectura universalmente superior, ya que el desempeño reportado depende de una configuración metodológica completa. En dicha configuración intervienen la segmentación, el preprocesamiento, la normalización, el aumento de datos, la transferencia de aprendizaje, el ajuste fino, la calidad de las etiquetas y el protocolo de validación. La comparación entre modelos requiere analizar el flujo completo de procesamiento y evaluación, dado que una misma arquitectura puede producir resultados distintos bajo condiciones de datos y validación diferentes.

Los resultados asociados con la RQ2 evidencian que la precisión de los modelos se encuentra condicionada por factores propios de los entornos naturales. El ruido ambiental, la baja relación señal–ruido, el solapamiento de vocalizaciones, la variación entre sensores, el desbalance de clases y el ruido de etiquetas afectan la estabilidad del sistema y su capacidad de generalización. Estos factores muestran que la precisión no constituye una propiedad fija del modelo, ya que depende del contexto acústico, de la estructura del conjunto de datos y del protocolo de evaluación. Un resultado elevado en un conjunto controlado puede perder

confiabilidad al aplicarse en nuevos sitios, hábitats, comunidades acústicas o condiciones de grabación. Esta observación resulta central para el monitoreo ecológico, donde los modelos se despliegan en escenarios variables, con señales incompletas, clases poco representadas y fuentes sonoras superpuestas.

La RQ3 permite observar que el reconocimiento acústico automatizado se ha extendido hacia aplicaciones ecológicas como presencia, ocupación, actividad vocal, abundancia relativa y cuantificación de eventos acústicos. La evidencia indica mayor desarrollo metodológico en tareas de detección, clasificación y presencia–ausencia que en estimaciones directas de abundancia o número de individuos. Esta diferencia tiene implicaciones importantes para la interpretación de los resultados, debido a que una vocalización detectada representa evidencia acústica procesada y no una medición directa de individuos o poblaciones. Para que las salidas del modelo sostengan inferencias ecológicas más robustas, se requiere controlar la detectabilidad, definir con precisión la unidad de análisis, considerar la dependencia entre registros y validar la relación entre eventos acústicos detectados y variables ecológicas de interés.

El análisis de la RQ4 muestra que las métricas más frecuentes continúan siendo accuracy, precision, recall, F1-score y mAP. Estas medidas permiten describir el desempeño general de los modelos, con limitaciones importantes cuando existen clases desbalanceadas, especies raras, umbrales variables o consecuencias ecológicas distintas para falsos positivos y falsos negativos. La baja presencia de métricas como AUPRC, specificity, matrices de confusión y evaluaciones por clase evidencia una necesidad de mayor estandarización en la evaluación bioacústica. El predominio de métricas globales puede generar una lectura optimista del rendimiento cuando los errores se concentran en especies poco representadas o en condiciones acústicas difíciles. La heterogeneidad en unidades de análisis, protocolos de validación y condiciones de prueba también limita la comparación directa entre estudios con valores de desempeño aparentemente similares.

El progreso reciente del reconocimiento bioacústico no se reduce a la incorporación de arquitecturas más complejas. Las contribuciones metodológicas actuales se orientan hacia el fortalecimiento de la robustez frente a condiciones acústicas variables, la transferencia entre contextos, la adaptación de modelos preentrenados a escenarios locales, el enriquecimiento de las representaciones acústicas y el análisis sistemático de los errores de clasificación. Estas

líneas desplazan la discusión desde la búsqueda del modelo con mayor valor métrico hacia la evaluación de sistemas confiables bajo condiciones reales de aplicación. En este marco, el ajuste fino, la validación local, las representaciones tiempo–frecuencia alternativas y el análisis de error permiten interpretar el rendimiento de los modelos de acuerdo con las condiciones del sitio de estudio, la estructura del conjunto de entrenamiento y la variabilidad ambiental.

La comparabilidad entre estudios continúa siendo uno de los principales obstáculos del campo. Dos investigaciones pueden reportar desempeños altos y responder a condiciones experimentales muy distintas, por diferencias en especies analizadas, tipo de sensor, hábitat, duración de los segmentos, balance de clases, métricas, umbrales de decisión o protocolos de validación. Esta heterogeneidad limita la identificación de mejoras atribuibles directamente a la arquitectura del modelo, al preprocesamiento, al conjunto de datos o al diseño de evaluación. La precisión debe interpretarse junto con análisis de robustez, estabilidad, transferencia, error y validez ecológica. Una evaluación centrada en una sola métrica ofrece una representación parcial del rendimiento del sistema.

El potencial del reconocimiento acústico automatizado mediante redes neuronales profundas se relaciona con la capacidad de procesar grandes volúmenes de grabaciones, reducir la dependencia exclusiva del análisis manual y apoyar el seguimiento de especies vocales en escalas temporales y espaciales amplias. Su utilidad depende de un diseño metodológico riguroso, con datos de calidad, validación local, documentación de las condiciones de grabación y criterios ecológicos claros para interpretar las predicciones. Las salidas de un modelo deben entenderse como resultados computacionales sujetos a incertidumbre, no como evidencia ecológica definitiva. La automatización amplía la capacidad de observación ambiental y también puede amplificar errores cuando se aplica sin controles adecuados.

Los hallazgos permiten proponer cinco líneas prioritarias para investigaciones futuras. La primera corresponde a la estandarización de métricas, umbrales y protocolos de validación. La segunda apunta a ampliar evaluaciones cross-site, cross-sensor y bajo condiciones acústicas no observadas durante el entrenamiento. La tercera se relaciona con la mejora de los conjuntos de datos mediante curación de etiquetas, balance de clases y documentación de las condiciones de grabación. La cuarta exige profundizar en métodos de explicabilidad y análisis de errores

para identificar las condiciones bajo las cuales fallan los modelos. La quinta requiere establecer vínculos más rigurosos entre las salidas computacionales y las inferencias ecológicas, especialmente cuando se busca estimar ocupación, abundancia o cambios poblacionales. Estas líneas representan condiciones necesarias para avanzar desde resultados experimentales favorables hacia sistemas bioacústicos confiables en escenarios naturales.

En conjunto, las redes neuronales profundas han transformado el reconocimiento acústico de especies y han ampliado las posibilidades del monitoreo bioacústico en entornos naturales. Su principal aporte radica en permitir sistemas escalables capaces de procesar información acústica compleja y apoyar tareas ecológicas de detección, clasificación y seguimiento. La revisión también muestra que su valor aplicado depende de la calidad de los datos, la transparencia metodológica, la validación ecológica y la interpretación crítica de los resultados. Desde la Ciencia de Datos, el reto consiste en desarrollar sistemas precisos, evaluables, transferibles e interpretables, capaces de sostener inferencias ecológicas proporcionales a la evidencia disponible.

7. REFERENCIAS

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., Cattiau, J. y Oleson, E. M. (2021). A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset. *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.607321

Allen-Ankins, S., Mula Laguna, J. y Schwarzkopf, L. (2024). Towards automated detection of the endangered southern black-throated finch (Poephila cincta cincta). *Wildlife Research*, *51*(10). https://doi.org/10.1071/WR23151

Assarzadeh, M., Atufi, P., Eibl, M. y Kahl, S. (2025). Investigating misclassification in bird sounds: The adverse effect of label noise on BirdNET. *Ecological Informatics*, *92*, 103494. https://doi.org/10.1016/j.ecoinf.2025.103494

Badi, A., Ko, K. y Ko, H. (2019). Bird sounds classification by combining PNCC and robust Mel-log filter bank features. *Journal of the Acoustical Society of Korea*, *38*, 39-46. https://cdn.apub.kr/journalsite/sites/ask/2019-038-01/N0660380105/N0660380105.pdf

Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., Hofer, H. y Maier, A. K. (2019). ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-47335-w

Bergler, C., Smeele, S. Q., Tyndel, S. A., Barnhill, A., Torres Ortiz, S., Kalan, A. K., Cheng, R. X., Brinkløv, S. M., Osiecka, A. N., Tougaard, J., Jakobsen, F., Wahlberg, M., Nöth, E., Maier, A. K. y Klump, B. C. (2022). ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-26429-y

Brautaset, O., Waldeland, A. U., Johnsen, E., Malde, K., Eikvil, L., Salberg, A. B. y Handegard, N. O. (2020). Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. *ICES Journal of Marine Science*, *77*(4), 1391-1400. https://doi.org/10.1093/icesjms/fsz235

Brinkløv, S. M., MacAulay, J. D. J., Bergler, C., Tougaard, J., Beedholm, K., Elmeros, M. y Madsen, P. T. (2023). Open-source workflow approaches to passive acoustic monitoring of bats. *Methods in Ecology and Evolution*, *14*(7), 1747-1763. https://doi.org/10.1111/2041-210X.14131

Clark, M. L., Salas, L. A., Baligar, S., Quinn, C. A., Snyder, R. L., Leland, D. F., Schackwitz, W. S., Goetz, S. J. y Newsam, S. (2023). The effect of soundscape composition on bird vocalization classification in a citizen science biodiversity monitoring project. *Ecological Informatics*, *75*. https://doi.org/10.1016/j.ecoinf.2023.102065

Cole, J. S., Michel, N. L., Emerson, S. A. y Siegel, R. B. (2022). Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. *Ornithological Applications*, *124*(2). https://doi.org/10.1093/ornithapp/duac003

Cotillard, T., Sécheresse, X., Aubin, J. A., Mikus, M. A., Vergara, V., Gambs, S., Michaud, R., Martins, C. C., Turgeon, S., Chion, C. y Roca, I. T. (2024). Automatic detection and classification of beluga whale calls in the St. Lawrence estuary. *Journal of the Acoustical Society of America*, *156*(6), 3723-3740. https://doi.org/10.1121/10.0030472

Davis, S. B. y Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357-366. https://doi.org/10.1109/TASSP.1980.1163420

Dias, F. F., Ponti, M. A. y Minghim, R. (2022). A classification and quantification approach to generate features in soundscape ecology using neural networks. *Neural Computing and Applications*, *34*(3), 1923-1937. https://doi.org/10.1007/s00521-021-06501-w

Dias, F. F., Ponti, M. A. y Minghim, R. (2025). Enhancing sound-based classification of birds and anurans with spectrogram representations and acoustic indices in neural network architectures. *Ecological Informatics*, *90*. https://doi.org/10.1016/j.ecoinf.2025.103232

Eichinski, P., Alexander, C., Roe, P., Parsons, S. E. y Fuller, S. J. (2022). A Convolutional Neural Network Bird Species Recognizer Built From Little Data by Iteratively Training, Detecting, and Labeling. *Frontiers in Ecology and Evolution*, *10*. https://doi.org/10.3389/fevo.2022.810330

Fairbrass, A. J., Firman, M., Williams, C. A., Brostow, G. J., Titheridge, H. y Jones, K. E. (2019). CityNet—Deep learning tools for urban ecoacoustic assessment. *Methods in Ecology and Evolution*, *10*(2), 186-197. https://doi.org/10.1111/2041-210X.13114

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández, X., Borghini, E., Mindlin, G. y Groisman, P. (2023). Intrinsic persistent homology via density-based metric learning. *Journal of Machine Learning Research*, *24*, 1-42. https://jmlr.org/papers/volume24/21-1044/21-1044.pdf

Fernandez-Garcia, G., Corpetti, T., Nevoux, M., Beaulaton, L. y Martignac, F. (2023). AcousticIA, a deep neural network for multi-species fish detection using multiple models of acoustic cameras. *Aquatic Ecology*, *57*(4), 881-893. https://doi.org/10.1007/s10452-023-10004-2

Ferreira, A. I. S., da Silva, N. F. F., Mesquita, F. N., Rosa, T. C., Buchmann, S. L. y Mesquita Neto, J. N. (2025). Transformer Models improve the acoustic recognition of buzz-pollinating bee species. *Ecological Informatics*, *86*. https://doi.org/10.1016/j.ecoinf.2025.103010

Florentin, J., Dutoit, T. y Verlinden, O. (2020). Detection and identification of European woodpeckers with deep convolutional neural networks. *Ecological Informatics*, *55*. https://doi.org/10.1016/j.ecoinf.2019.101023

Frainer, G., Dufourq, E., Fearey, J. L., Dines, S., Probert, R., Elwen, S. H. y Gridley, T. (2023). Automatic detection and taxonomic identification of dolphin vocalisations using convolutional neural networks for passive acoustic monitoring. *Ecological Informatics*, *78*. https://doi.org/10.1016/j.ecoinf.2023.102291

Ghani, B., Denton, T., Kahl, S. y Klinck, H. (2023). Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-49989-z

Ghani, B., Kalkman, V. J., Planqué, B., Vellinga, W. P., Gill, L. F. y Stowell, D. (2025). Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-00996-2

Goodfellow, I., Bengio, Y. y Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org

Hamard, Q., Pham, M. T., Cazau, D. y Heerah, K. (2024). A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data. *Ecological Informatics*, *84*. https://doi.org/10.1016/j.ecoinf.2024.102906

Hsu, S., Lee, C., Chang, P., Han, C. y Fan, K. (2018). Local Wavelet Acoustic Pattern: A Novel Time-Frequency Descriptor for Birdsong Recognition. *IEEE Transactions on Multimedia*, *20*(12), 3187-3199. https://doi.org/10.1109/TMM.2018.2834866

Hyer, M. D., Anderson, A. T., Mann, D. A., Mooney, T. A., Aoki, N. y Jensen, F. H. (2025). Robust real-time detection of right whale upcalls using neural networks on the edge. *Ecological Informatics*, *89*. https://doi.org/10.1016/j.ecoinf.2025.103130

Ibrahim, A. K., Zhuang, H., Schärer-Umpierre, M. T., Woodward, C., Erdöl, N. y Chérubin, L. M. (2024). Fish Acoustic Detection Algorithm Research: a deep learning app for Caribbean grouper calls detection and call types classification. *Frontiers in Marine Science*, *11*. https://doi.org/10.3389/fmars.2024.1378159

Jacuzzi, G. y Olden, J. D. (2025). Few-shot transfer learning enables robust acoustic monitoring of wildlife communities at the landscape scale. *Ecological Informatics*, *90*. https://doi.org/10.1016/j.ecoinf.2025.103294

Jeantet, L. y Dufourq, E. (2023). Improving deep learning acoustic classifiers with contextual information for wildlife monitoring. *Ecological Informatics*, *77*, 102256. https://doi.org/10.1016/j.ecoinf.2023.102256

Jin, C., Kim, M., Jang, S. y Paeng, D.-g. (2022). Semantic segmentation-based whistle extraction of Indo-Pacific Bottlenose Dolphin residing at the coast of Jeju island. *Ecological Indicators*, *137*. https://doi.org/10.1016/j.ecolind.2022.108792

Junqueira, B. F., Vieira, R. G., Alves, E. C., Karaziack, B. B. y Dos Santos, M. V. (2025). Environmental Monitoring with Low-Processing Embedded AI through Sound Event Classification. *Journal of the Brazilian Computer Society*, *31*(1), 523-544. https://doi.org/10.5753/jbcs.2025.4210

Kahl, S., Wood, C. M., Eibl, M. y Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*. https://doi.org/10.1016/j.ecoinf.2021.101236

Kershenbaum, A., Akçay, Ç., Babu-Saheer, L., Barnhill, A., Best, P., Cauzinille, J., Clink, D., Dassow, A., Dufourq, E., Growcott, J., Markham, A., Marti-Domken, B., Marxer, R., Muir, J., Reynolds, S., Root-Gutteridge, H., Sadhukhan, S., Schindler, L., Smith, B. R., ... y Dunn, J. C. (2025). Automatic detection for bioacoustic research: A practical guide from and for biologists and computer scientists. *Biological Reviews*, *100*, 620-646. https://doi.org/10.1111/brv.13155

Kiskin, I., Zilli, D., Li, Y., Sinka, M. E., Willis, K. J. y Roberts, S. J. (2020). Bioacoustic detection with wavelet-conditioned convolutional neural networks. *Neural Computing and Applications*, *32*(4), 915-927. https://doi.org/10.1007/s00521-018-3626-7

Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M. y Bayne, E. M. (2017a). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology*, *12*(2), 14. https://doi.org/10.5751/ACE-01114-120214

Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M. y Bayne, E. M. (2017b). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs; Recommandations pour l'évaluation des performances de reconnaissance acoustique et application à cinq programmes courants de reconnaissance automatisée de signaux sonores. *Avian Conservation and Ecology*, *12*(2). https://doi.org/10.5751/ACE-01114-120214

Kumar, A. S., Schlosser, T., Kahl, S. y Kowerko, D. (2024). Improving learning-based birdsong classification by utilizing combined audio augmentation strategies. *Ecological Informatics*, *82*. https://doi.org/10.1016/j.ecoinf.2024.102699

Kumar, R. T., Gupta, M. S., Ahmed, S., Alhumam, A. y Aggarwal, T. (2022). Intelligent audio signal processing for detecting rainforest species using deep learning. *Intelligent Automation and Soft Computing*, *31*(2), 693-706. https://doi.org/10.32604/iasc.2022.019811

Lauha, P., Somervuo, P. J., Lehikoinen, P., Geres, L. S., Richter, T., Seibold, S. y Ovaskainen, O. T. (2022). Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution*, *13*(12), 2799-2810. https://doi.org/10.1111/2041-210X.14003

LeBien, J. G., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Lavista-Ferres, J. M. y Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, *59*. https://doi.org/10.1016/j.ecoinf.2020.101113

Li, X., Luo, N., Yu, F., Li, J., Li, K., Li, Y., Zhao, Z., Liu, Y. y Shi, X. (2026). Human Auditory Representation Learning for cross-dialect bird species recognition. *Ecological Informatics*, *93*. https://doi.org/10.1016/j.ecoinf.2025.103554

Liu, M., Wang, L., Jia, R., Ji, S., Wu, Y., Wu, Y., Xie, L. y Dong, M. (2025). Cross-Modal and Contrastive Optimization for Explainable Multimodal Recognition of Predatory and Parasitic Insects. *Insects*, *16*(12). https://doi.org/10.3390/insects16121187

Lostanlen, V., Cramer, A., Salamon, J., Farnsworth, A. J., Van Doren, B. M., Kelling, S. T. y Bello, J. P. (2024). BirdVoxDetect: Large-Scale Detection and Classification of Flight Calls for Bird Migration Monitoring. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *32*, 4134-4145. https://doi.org/10.1109/TASLP.2024.3444486

Lostanlen, V., Salamon, J., Farnsworth, A. J., Kelling, S. T. y Bello, J. P. (2019). Robust sound event detection in bioacoustic sensor networks. *PLOS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0214168

Maclean, K. y Triguero, I. (2023). Identifying bird species by their calls in Soundscapes. *Applied Intelligence*, *53*(19), 21485-21499. https://doi.org/10.1007/s10489-023-04486-8

Manzano-Rubio, R., Bota, G., Brotons, L. L., Soto-Largo, E. y Pérez-Granados, C. (2022). Low-cost open-source recorders and ready-to-use machine learning approaches provide effective monitoring of threatened species. *Ecological Informatics*, *72*. https://doi.org/10.1016/j.ecoinf.2022.101910

Marchal, J., Fabianek, F. y Aubry, Y. (2022). Software performance for the automated identification of bird vocalisations: the case of two closely related species. *Bioacoustics*, *31*(4), 397-413. https://doi.org/10.1080/09524622.2021.1945952

Marolt, M., Pesek, M., Šturm, R., López Díez, J. J., Rexhepi, B. y Virant-Doberlet, M. (2025). Computational methods for detecting insect vibrational signals in field vibroscape recordings. *Ecological Informatics*, *86*. https://doi.org/10.1016/j.ecoinf.2025.103003

Márquez-Rodríguez, A., Mohedano-Munoz, M. Á., Marín-Jiménez, M. J., Santamaría-García, E., Bastianelli, G., Jordano, P. y Mendoza, I. (2025). A bird song detector for improving bird identification through deep learning: A case study from Doñana. *Ecological Informatics*, *90*. https://doi.org/10.1016/j.ecoinf.2025.103254

Michaud, F., Sueur, J., Le Cesne, M. y Haupert, S. (2023). Unsupervised classification to improve the quality of a bird song recording dataset. *Ecological Informatics*, *74*. https://doi.org/10.1016/j.ecoinf.2022.101952

Mouy, X., Archer, S. K., Dosso, S. E., Dudas, S. E., English, P. A., Foord, C., Halliday, W. D., Juanes, F., Lancaster, D., van Parijs, S. M. y Haggarty, D. R. (2024). Automatic detection of unidentified fish sounds: a comparison of traditional machine learning with deep learning. *Frontiers in Remote Sensing*, *5*. https://doi.org/10.3389/frsen.2024.1439995

Nolan, V., Scott, C., Yeiser, J. M., Wilhite, N. G., Howell, P. E., Ingram, D. R. y Martin, J. A. (2023). The development of a convolutional neural network for the automatic detection

of Northern Bobwhite Colinus virginianus covey calls. *Remote Sensing in Ecology and Conservation*, *9*(1), 46-61. https://doi.org/10.1002/rse2.294

Noumida, A. y Rajan, R. (2023). MULTI-LABEL BIRD SPECIES CLASSIFICATION USING SEQUENTIAL AGGREGATION STRATEGY FROM AUDIO RECORDINGS. *Computing and Informatics*, *42*(5), 1255-1280. https://doi.org/10.31577/cai_2023_5_1255

Okamoto, R. y Oguma, H. (2025). A simulation based approach for enhancing automatic detection of cicada songs in challenging chorus conditions. *Ecological Informatics*, *90*. https://doi.org/10.1016/j.ecoinf.2025.103299

Oppenheim, A. V. y Schafer, R. W. (2009). *Discrete-Time Signal Processing* (3.ª ed.). Prentice Hall.

Owens, A. F., Hockings, K. J., Imron, M. A., Madhusudhana, S. K., Mariaty, n., Setia, T. M., Sharma, M. D., Maimunah, S., van Veen, F. J. y Erb, W. M. (2024). Automated detection of Bornean white-bearded gibbon (Hylobates albibarbis) vocalizations using an open-source framework for deep learning. *Journal of the Acoustical Society of America*, *156*(3), 1623-1632. https://doi.org/10.1121/10.0028268

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., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... y Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pérez-Granados, C. (2023). BirdNET: applications, performance, pitfalls and future opportunities. *Ibis*, *165*(3), 1068-1075. https://doi.org/10.1111/ibi.13193

Prince, P., Hill, A. P., Piña-Covarrubias, E., Doncaster, C. P., Snaddon, J. L. y Rogers, A. C. (2019). Deploying acoustic detection algorithms on low-cost, open-source acoustic sensors for environmental monitoring. *Sensors*, *19*(3). https://doi.org/10.3390/s19030553

Rasmussen, N., Rizk, R., Matoo, O. B. y Santosh, K. C. (2024). DeepWhaleNet: Climate Change-Aware FFT-Based Deep Neural Network for Passive Acoustic Monitoring. *International Journal of Pattern Recognition and Artificial Intelligence*, *38*(14). https://doi.org/10.1142/S0218001424590146

Ravaglia, D., Ferrario, V., De Gregorio, C., Carugati, F., Raimondi, T., Cristiano, W., Torti, V., Hardenberg, A. V., Ratsimbazafy, J. H., Valente, D., Giacoma, C. y Gamba, M.

(2023). There You Are! Automated Detection of Indris' Songs on Features Extracted from Passive Acoustic Recordings. *Animals*, *13*(2). https://doi.org/10.3390/ani13020241

Ruff, Z. J., Lesmeister, D. B., Duchac, L. S., Padmaraju, B. K. y Sullivan, C. M. (2020). Automated identification of avian vocalizations with deep convolutional neural networks. *Remote Sensing in Ecology and Conservation*, *6*(1), 79-92. https://doi.org/10.1002/rse2.125

Runge, J. (2023). Modern causal inference approaches to investigate biodiversity-ecosystem functioning relationships. *Nature Communications*, *14*, 1917. https://doi.org/10.1038/s41467-023-37546-1

Sanderson, G. (2018). But what is the Fourier Transform? A visual introduction [3Blue1Brown YouTube channel].

Saoud, L. S., Sultan, A., Elmezain, M., Heshmat, M., Seneviratne, L. y Hussain, I. (2024). Beyond observation: Deep learning for animal behavior and ecological conservation. *Ecological Informatics*, *84*, 102893. https://doi.org/10.1016/j.ecoinf.2024.102893

Sasek, J., Allison, B., Contina, A. J., Knobles, D. P., Wilson, P. S. y Keitt, T. H. (2024). Semiautomated generation of species-specific training data from large, unlabeled acoustic datasets for deep supervised birdsong isolation. *PeerJ*, *12*(9). https://doi.org/10.7717/peerj.17854

Sattar, F. (2023). A New Acoustical Autonomous Method for Identifying Endangered Whale Calls: A Case Study of Blue Whale and Fin Whale. *Sensors*, *23*(6). https://doi.org/10.3390/s23063048

Schiavo, G., Portaccio, A. y Testolin, A. (2025). Fine-Tuning BirdNET for the Automatic Ecoacoustic Monitoring of Bird Species in the Italian Alpine Forests. *Information*, *16*(8), 628. https://doi.org/10.3390/info16080628

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y. Y. y Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution*, *10*(3), 368-380. https://doi.org/10.1111/2041-210X.13103

Syunkova, A., Lapp, S. M., Basanta, M. D., Lambertini, C., Ruiz Guzman, S., Voyles, J. L., Richards-Zawacki, C. L. y Kitzes, J. A. (2025). Transfer learning outperforms other methods of detecting vocalizations of a critically endangered tropical anuran. *Ecological Informatics*, *92*. https://doi.org/10.1016/j.ecoinf.2025.103427

Teixeira, D., Roe, P., van Rensburg, B. J., Linke, S., McDonald, P. G., Tucker, D. y Fuller, S. (2024). Effective ecological monitoring using passive acoustic sensors: Recommendations for conservation practitioners. *Conservation Science and Practice*, *6*(6), e13132. https://doi.org/10.1111/csp2.13132

Thakur, A., Abrol, V., Sharma, P. y Rajan, P. (2018). Local compressed convex spectral embedding for bird species identification. *The Journal of the Acoustical Society of America*, *143*, 3819-3828. https://doi.org/10.1121/1.5042241

Truong, K., Eidsvik, J., Rorstadbotnen, R. A., Morten, J. P., Andres, L. y Sladen, A. (2025). Automated detection of fin whales with distributed acoustic sensing in the Arctic and Mediterranean. *Journal of the Acoustical Society of America*, *158*(6), 5057-5074. https://doi.org/10.1121/10.0041855

Tyshko, A. A., Krinitskiy, M. A., Shatravin, A. V. y Belikov, R. A. (2023). Automatic Detection of Acoustic Signals of Beluga Whales and Bottlenose Dolphins. *Moscow University Physics Bulletin (English Translation of Vestnik Moskovskogo Universiteta, Fizika)*, *78*, S217-S225. https://doi.org/10.3103/S0027134923070305

Ugarte, J. P. y Arias-Arias, J. M. (2024). Unveiling relevant acoustic features for bird species automatic classification. *Expert Systems with Applications*, *257*, 125046. https://doi.org/10.1016/j.eswa.2024.125046

Wang, H., Xu, Y., Yu, Y., Lin, Y. y Ran, J. (2022). An Efficient Model for a Vast Number of Bird Species Identification Based on Acoustic Features. *Animals*, *12*(18). https://doi.org/10.3390/ani12182434

Wang, M., Darras, K. F. A., Xue, R. y Liu, F. (2025). Animal acoustic identification, denoising and source separation using generative adversarial networks. *Methods in Ecology and Evolution*, *16*(10), 2472-2486. https://doi.org/10.1111/2041-210X.70148

Wang, Y., Ye, J. y Borchers, D. L. (2022). Automated call detection for acoustic surveys with structured calls of varying length. *Methods in Ecology and Evolution*, *13*(7), 1552-1567. https://doi.org/10.1111/2041-210X.13873

Wei, Y., Chen, W., Tuanmu, M. N., Lu, S. y Shiao, M.-T. (2024). Advanced montane bird monitoring using self-supervised learning and transformer on passive acoustic data. *Ecological Informatics*, *84*. https://doi.org/10.1016/j.ecoinf.2024.102927

White, E. L., White, P. R., Bull, J. M., Risch, D., Quer, S. y Beck, S. (2025). Evaluating the performance of automated detection systems for long-term monitoring of delphinids in

diverse marine soundscapes. *PLOS ONE*, *20*(6 June). https://doi.org/10.1371/journal.pone.0323768

Xie, J. y Zhu, M. (2019). Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics*, *52*, 74-81. https://doi.org/10.1016/j.ecoinf.2019.05.007

Xie, J., Hu, K., Guo, Y., Zhu, Q. y Yu, J. (2021). On loss functions and CNNs for improved bioacoustic classification. *Ecological Informatics*, *64*, 101331. https://doi.org/10.1016/j.ecoinf.2021.101331

Xie, J., Hu, K., Zhu, M., Yu, J. y Zhu, Q. (2019). Investigation of different CNN-based models for improved bird sound classification. *IEEE Access*, *7*, 175353-175361. https://doi.org/10.1109/ACCESS.2019.2957572

Xu, W., Zhang, X., Yao, L., Xue, W. y Wei, B. (2020). A multi-view CNN-based acoustic classification system for automatic animal species identification. *Ad Hoc Networks*, *102*. https://doi.org/10.1016/j.adhoc.2020.102115

Zhou, X., Hu, K., Guan, Z., Yu, C., Wang, S., Fan, M., Sun, Y., Cao, Y., Wang, Y. y Miao, G. (2023). Methods for processing and analyzing passive acoustic monitoring data: An example of song recognition in western black-crested gibbons. *Ecological Indicators*, *155*. https://doi.org/10.1016/j.ecolind.2023.110908

Ziegenhorn, M. A., Lanctot, R. B., Brown, S. C., Brengle, M., Schulte, S., Saalfeld, S. T., Latty, C. J., Smith, P. A. y Lecomte, N. (2025). ArcticSoundsNET: BirdNET embeddings facilitate improved bioacoustic classification of Arctic species. *Ecological Informatics*, *90*, 103270. https://doi.org/10.1016/j.ecoinf.2025.103270

Zinemanas, P., Rocamora, M., Miron, M., Font, F. y Serra, X. (2021). An Interpretable Deep Learning Model for Automatic Sound Classification. *Electronics*, *10*(7), 850. https://doi.org/10.3390/electronics10070850

A. MATRIZ DE IDENTIFICACIÓN METODOLÓGICA DEL CORPUS

La siguiente matriz resume la identificación de arquitecturas y enfoques metodológicos en el corpus final de artículos incluidos en la revisión sistemática. Cada fila corresponde a un estudio y cada columna indica la presencia o ausencia de un enfoque detectado mediante la extracción automática estructurada. Las marcas metodológicas son no excluyentes, por lo que un mismo artículo puede presentar múltiples enfoques simultáneamente. En cambio, la columna de método principal corresponde a una clasificación excluyente utilizada para resumir el enfoque dominante de cada estudio.

Tabla A.1

Matriz resumida de identificación de arquitecturas y enfoques metodológicos en el corpus final de artículos.

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN / RNN	Otros
art055	Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs; Recommandations pour l'évaluation des performances de reconnaissance acoustique et application à cinq programmes courants de reconnaissance automatisée de signaux sonores	Workflow / herramienta / revisión	X	-	X	X	X	X	-	X
art063	Local Wavelet Acoustic Pattern: A Novel Time-Frequency Descriptor for Birdsong Recognition	Aprendizaje automático tradicional	X	X	-	X	X	-	-	X
art072	CityNet—Deep learning tools for urban ecoacoustic assessment	CNN	X	X	X	-	-	-	-	X
art073	Deploying acoustic detection algorithms on low-cost, open-source acoustic sensors for environmental monitoring	Híbrido	X	-	-	-	X	-	-	X
art074	Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge	CNN	X	X	-	-	-	X	-	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art083	Robust sound event detection in bioacoustic sensor networks	CNN	X	X	X	X	-	-	X	X
art084	ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning	CNN	X	X	X	X	X	X	X	X
art086	Detection and identification of European woodpeckers with deep convolutional neural networks	CNN	X	X	X	-	X	-	-	X
art090	Bioacoustic detection with wavelet-conditioned convolutional neural networks	CNN	X	X	-	X	X	X	-	X
art091	Automated identification of avian vocalizations with deep convolutional neural networks	CNN	X	-	-	-	X	-	-	X
art093	A multi-view CNN-based acoustic classification system for automatic animal species identification	CNN	X	X	X	X	X	-	X	X
art098	Acoustic classification in multifrequency echosounder data using deep convolutional neural networks	CNN	X	X	-	-	-	X	-	-
art103	A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network	CNN	X	X	-	-	-	-	X	X
art112	BirdNET: A deep learning solution for avian diversity monitoring	CNN	X	X	X	X	-	X	-	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art115	A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset	CNN	X	X	X	-	-	-	-	-
art147	Software performance for the automated identification of bird vocalisations: the case of two closely related species	Híbrido	X	-	X	-	X	-	-	X
art149	Intelligent audio signal processing for detecting rainforest species using deep learning	Híbrido	X	X	-	-	-	-	X	-
art157	A classification and quantification approach to generate features in soundscape ecology using neural networks	CNN	X	X	X	X	-	-	-	X
art163	A Convolutional Neural Network Bird Species Recognizer Built From Little Data by Iteratively Training, Detecting, and Labeling	CNN	X	X	X	-	-	-	-	-
art164	Semantic segmentation-based whistle extraction of Indo-Pacific Bottlenose Dolphin residing at the coast of Jeju island	CNN	X	X	X	X	-	-	-	X
art174	Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data	CNN	X	-	-	-	-	-	-	X
art177	Automated call detection for acoustic surveys with structured calls of varying length	Híbrido	X	X	-	X	X	X	X	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art184	An Efficient Model for a Vast Number of Bird Species Identification Based on Acoustic Features	CNN	X	X	X	X	X	X	X	X
art198	Domain-specific neural networks improve automated bird sound recognition already with small amount of local data	CNN	X	X	X	-	X	-	-	X
art202	Low-cost open-source recorders and ready-to-use machine learning approaches provide effective monitoring of threatened species	Híbrido	X	-	-	-	X	-	-	-
art205	ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning	CNN	X	X	X	X	-	-	X	X
art207	There You Are! Automated Detection of Indris' Songs on Features Extracted from Passive Acoustic Recordings	CNN	X	X	-	-	-	-	-	-
art216	MULTI-LABEL BIRD SPECIES CLASSIFICATION USING SEQUENTIAL AGGREGATION STRATEGY FROM AUDIO RECORDINGS	Híbrido	X	X	X	-	-	-	X	X
art220	The development of a convolutional neural network for the automatic detection of Northern Bobwhite Colinus virginianus covey calls	CNN	X	X	X	-	-	-	-	-

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art224	A New Acoustical Autonomous Method for Identifying Endangered Whale Calls: A Case Study of Blue Whale and Fin Whale	Híbrido	X	-	X	X	-	-	X	X
art229	Unsupervised classification to improve the quality of a bird song recording dataset	Híbrido	X	X	X	-	-	-	-	X
art237	BirdNET: applications, performance, pitfalls and future opportunities	CNN	X	-	-	-	-	-	-	-
art238	The effect of soundscape composition on bird vocalization classification in a citizen science biodiversity monitoring project	CNN	X	X	X	-	-	-	-	-
art239	Open-source workflow approaches to passive acoustic monitoring of bats	Workflow / herramienta / revisión	X	-	X	-	-	X	-	-
art248	Identifying bird species by their calls in Soundscapes	CNN	X	X	X	X	-	-	-	X
art257	Learning to detect an animal sound from five examples	Híbrido	X	X	X	-	-	-	X	X
art258	Methods for processing and analyzing passive acoustic monitoring data: An example of song recognition in western black-crested gibbons	CNN	X	X	X	-	-	-	X	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art261	AcousticIA, a deep neural network for multi-species fish detection using multiple models of acoustic cameras	CNN	X	X	X	-	-	-	X	-
art264	Automatic detection and taxonomic identification of dolphin vocalisations using convolutional neural networks for passive acoustic monitoring	CNN	X	X	X	-	-	-	-	-
art269	Global birdsong embeddings enable superior transfer learning for bioacoustic classification	Híbrido	X	X	X	-	-	-	-	X
art271	Automatic Detection of Acoustic Signals of Beluga Whales and Bottlenose Dolphins	CNN	X	X	-	X	X	X	-	X
art277	Fish Acoustic Detection Algorithm Research: a deep learning app for Caribbean grouper calls detection and call types classification	CNN	X	X	X	X	-	X	X	X
art278	BirdVoxDetect: Large-Scale Detection and Classification of Flight Calls for Bird Migration Monitoring	Híbrido	X	X	X	X	X	X	-	X
art279	Automatic detection of unidentified fish sounds: a comparison of traditional machine learning with deep learning	Híbrido	X	X	-	X	-	X	-	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art280	Semiautomated generation of species-specific training data from large, unlabeled acoustic datasets for deep supervised birdsong isolation	CNN	X	X	X	-	-	X	X	X
art317	Improving learning-based birdsong classification by utilizing combined audio augmentation strategies	CNN	X	X	X	-	X	-	-	X
art318	Automated detection of Bornean white-bearded gibbon (Hylobates albibarbis) vocalizations using an open-source framework for deep learning	CNN	X	-	-	X	X	-	X	X
art319	Towards automated detection of the endangered southern black-throated finch (Poephila cincta cincta)	CNN	X	X	-	-	-	-	-	-
art333	DeepWhaleNet: Climate Change-Aware FFT-Based Deep Neural Network for Passive Acoustic Monitoring	CNN	X	X	-	-	-	-	-	-
art341	A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data	CNN	X	X	X	X	X	X	-	-
art342	Advanced montane bird monitoring using self-supervised learning and transformer on passive acoustic data	Self-supervised	X	X	X	X	-	-	-	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art343	Automatic detection and classification of beluga whale calls in the St. Lawrence estuary	CNN	X	X	X	X	-	X	-	X
art373	Environmental Monitoring with Low-Processing Embedded AI through Sound Event Classification	CNN	X	-	X	X	X	X	-	X
art393	Computational methods for detecting insect vibrational signals in field vibroscape recordings	Híbrido	X	X	-	-	-	-	-	X
art394	Transformer Models improve the acoustic recognition of buzz-pollinating bee species	Transformer	X	X	X	-	-	-	-	X
art397	Evaluating the performance of automated detection systems for long-term monitoring of delphinids in diverse marine soundscapes	Híbrido	X	X	-	-	-	-	-	-
art424	Animal acoustic identification, denoising and source separation using generative adversarial networks	GAN	X	X	X	-	-	-	X	X
art428	Robust real-time detection of right whale upcalls using neural networks on the edge	CNN	X	X	-	-	-	-	-	-
art434	Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification	Híbrido	X	X	-	-	-	-	-	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art437	Enhancing sound-based classification of birds and anurans with spectrogram representations and acoustic indices in neural network architectures	CNN	X	X	X	X	-	-	-	X
art438	Few-shot transfer learning enables robust acoustic monitoring of wildlife communities at the landscape scale	Híbrido	X	X	X	X	-	-	-	-
art440	A simulation based approach for enhancing automatic detection of cicada songs in challenging chorus conditions	CNN	X	X	X	X	X	-	-	X
art443	A bird song detector for improving bird identification through deep learning: A case study from Doñana	CNN	X	X	X	-	-	X	-	X
art448	Transfer learning outperforms other methods of detecting vocalizations of a critically endangered tropical anuran	Híbrido	X	X	X	-	-	-	-	X
art457	Cross-Modal and Contrastive Optimization for Explainable Multimodal Recognition of Predatory and Parasitic Insects	CNN	X	X	X	-	-	X	-	X
art459	Automated detection of fin whales with distributed acoustic sensing in the Arctic and Mediterranean	Híbrido	X	X	X	-	-	-	-	X

Artículo	Título	Método principal	CNN	Transfer	Híbrido	SVM	HMM	RF	CRNN/RNN	Otros
art464	Human Auditory Representation Learning for cross-dialect bird species recognition	CNN	X	X	X	X	-	X	X	X

Nota: La marca **X** indica que el artículo reporta, utiliza, compara o menciona explícitamente ese enfoque metodológico. La marca **-** indica ausencia de detección en la extracción automática estructurada. La columna **Otros** agrupa el resto de enfoques menos frecuentes no mostrados individualmente en la matriz resumida. Las columnas metodológicas son **no excluyentes**, por lo que un mismo artículo puede presentar múltiples marcas simultáneamente.