Santiago de Cali, 25 de Mayo del 2026

Doctor

Diego Luis Linares O.

Director Maestría en Ciencia de Datos

Facultad de Ingeniería y Ciencias

Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado **"Clasificación de espectrogramas de audio con técnicas de aprendizaje profundo para el reconocimiento de sonidos urbanos**", el cual fue realizado por el (los) estudiante (s) Andrés Arroyo Marín con código 0073629, Juan Camilo Domínguez con código 5713099 y Maria Paula Jaramillo Giraldo con código 0066228 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de David Arango Londoño.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Andrés Arroyo Marín

CC. 1.112.228.283 de Pradera, Valle

Juan Camilo Domínguez Benítez

CC. 1.144.078.261 de Cali, Valle

Maria Paula Jaramillo Giraldo

CC. 1.144.056.799 de Cali, Valle

David Arango Londoño

CC. 1.130.586.950 de Cali

FICHA RESUMEN

PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

TÍTULO: CLASIFICACIÓN DE ESPECTROGRAMAS DE AUDIO CON TÉCNICAS DE APRENDIZAJE PROFUNDO PARA EL RECONOCIMIENTO DE SONIDOS URBANOS

1. ÁREA DE TRABAJO: Acústica Ambiental Urbana
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): Andrés Arroyo Marín, Juan Camilo Domínguez Benítez y Maria Paula Jaramillo Giraldo
4. CORREO ELECTRÓNICO: aarroyo82@javerianacali.edu.co, juancamilo9516@javerianacali.edu.con, mpjaramillog@javerianacali.edu.co
5. DIRECCIÓN Y TELÉFONO: Calle 18 # 66-68, 3137459929
6. DIRECTOR: David Arango Londoño
7. VINCULACIÓN DEL DIRECTOR: Docente Catedrático
8. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): No aplica
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): No aplica
11. OTROS GRUPOS O EMPRESAS: No aplica
12. PALABRAS CLAVE (al menos 5): Ciencia de datos, Deep Learning, redes neuronales, paisajes sonoros y espectrogramas
13. FECHA DE INICIO: 23/06/2025
14. FECHA DE FINALIZACIÓN: 25/05/2026

RESUMEN:

El presente proyecto aplicado abordó la clasificación automática de sonidos urbanos mediante técnicas de aprendizaje profundo, utilizando espectrogramas Mel como representación visual de señales acústicas en la ciudad de Popayán, Cauca. Esta iniciativa nace con el propósito de fortalecer el monitoreo inteligente del paisaje sonoro urbano, considerando que la contaminación acústica es una problemática creciente con efectos sobre la salud física y mental, la calidad de vida y sostenimiento ambiental.

Hoy en día, la clasificación manual de sonidos representa un proceso muy operativo, subjetivo y poco escalable debido a la mezcla o superposición constante de sonidos urbanos y a la ausencia de herramientas automatizadas para el reconocimiento de los tipos de sonidos. En este contexto, el objetivo principal consistió en diseñar un modelo basado en redes neuronales convolucionales capaz de clasificar sonidos urbanos según su origen: biofofía (sonido animales), geofonía (sonidos lluvias o viento), antropofonía (transporte motorizado, voces/ música).

Para esto se trabajó con un conjunto de datos de 7.138 fragmentos de audio obtenidos a partir de 250 grabaciones recolectadas en 50 puntos del centro histórico de Popayán. Luego, las señales de audio fueron transformadas en espectrogramas de Mel, convirtiendo el sonido en imágenes para el reconocimiento visual mediante arquitecturas convolucionales de aprendizaje profundo.

Se evaluaron diferentes arquitecturas mediante la búsqueda aleatoria de hiperparámetros, se incluyeron estrategias de transferencia de aprendizaje, aumento de datos y manejo de desbalance de clases. Como resultado, la arquitectura EfficientNetB0 presentó mejores resultados y desempeño, alcanzando un F1 macro de 0,51, un accuracy del 60,9% y un Balanced Accuracy de 0,59 sobre el conjunto de prueba. Con estos resultados, se evidencia la capacidad de un modelo para identificar patrones acústicos complejos en entornos urbanos y demuestran elo potencial de la inteligencia artificial al análisis de sonidos, con aplicaciones en monitoreo ambiental, control de contaminación acústica y apoyo a políticas públicas en las zonas.

CLASIFICACIÓN DE ESPECTROGRAMAS DE AUDIO CON TÉCNICAS DE APRENDIZAJE PROFUNDO PARA EL RECONOCIMIENTO DE SONIDOS URBANOS

Andres Arroyo Marín — Código 0073629

Juan Camilo Domínguez — Código 5713099

Maria Paula Jaramillo — Código 0066228

Proyecto Aplicado para optar al título de

Magíster en Ciencia de Datos

Director(a)

David Arango Londoño — C.C. 113.058.695

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, MAYO 25 DE 2026

TABLA DE CONTENIDO

Contenido

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

El crecimiento constante de las ciudades ha generado nuevos retos ambientales que afectan la calidad de vida de los habitantes y el equilibrio de los ecosistemas urbanos. Según el censo del 2018, Popayán, Cauca contaba con una población de 318.059 habitantes, y se estimó que para el año 2024 la ciudad habría experimentado un crecimiento del 7,8% [1]. Uno de los principales desafíos asociados al desarrollo urbanístico es la contaminación acústica, un fenómeno con efectos negativos tanto en la salud física y mental de los payaneses. Diversos artículos de la Organización Mundial de la Salud (OMS) han evidenciado que la exposición prolongada a altos niveles de ruido urbano puede provocar trastornos del sueño, estrés, enfermedades cardiovasculares y deterioro cognitivo, especialmente en niños y adultos mayores [10].

El monitoreo y control del ruido en la ciudad presenta limitaciones, tanto en frecuencia como en resolución espacial. En este contexto, el presente proyecto propuso el uso de técnicas de aprendizaje profundo aplicadas a espectrogramas de Mel para la clasificación automática de sonidos urbanos registrados en el perímetro urbano de Popayán. El trabajo se enmarca en el proyecto Urbanphony, una iniciativa internacional que analiza paisajes sonoros en Popayán (Colombia), Santiago de Compostela (España) y Venecia (Italia). Para Popayán se dispuso de 250 grabaciones capturadas en 50 puntos estratégicos del centro histórico durante cinco franjas horarias del día, produciendo un total de 7.138 segmentos de audio de dos segundos. Estos segmentos fueron etiquetados manualmente en seis categorías de sonido y transformados en espectrogramas de Mel como representación visual de entrada para los modelos, permitiendo convertir los sonidos en imágenes que se pueden procesar mediante redes neuronales convolucionales.

Se evaluaron diferentes arquitecturas de aprendizaje profundo mediante búsqueda aleatoria de hiperparámetros, utilizando el F1 macro como métrica principal de comparación dado el desbalance de clases presente en el dataset. La arquitectura con mejor desempeño fue EfficientNetB0, la cual fue reentrenada con una estrategia optimizada de aumento de datos, pesos de clase balanceados, regularización L2 y un esquema de entrenamiento bifásico. Como resultado, se alcanzó un F1 macro de 0,51 y una exactitud del 60,9% sobre el conjunto de prueba independiente, evidenciando el potencial de aprendizaje profundo como herramienta de apoyo.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Los altos niveles de ruido y la exposición prolongada han demostrado tener efectos negativos tanto en la salud física como mental de la población. En entornos urbanos, esta problemática puede originarse por múltiples fuentes, como el tráfico vehicular, las actividades industriales, las obras de construcción y algunas interacciones sociales. Según la Organización Mundial de la Salud (OMS,2022) el ruido, al superar ciertos umbrales, representa una amenaza a la salud pública [3], afectando el bienestar, el rendimiento y el equilibrio psicológico de las personas.

Es importante aclarar que no todos los sonidos representan un riesgo para la salud, ni la exposición breve implica consecuencias negativas. Sin embargo, cuando la exposición al ruido es constante y prolongada, se asocia con el desarrollo de enfermedades cardiovasculares, trastornos del sueño, dificultades de aprendizaje en niños y alteraciones en la salud mental [2]. Frente a esta realidad, la Agencia Europea de Medio Ambiente ha impulsado medidas de control y sistemas de monitoreo de la exposición al ruido ambiental, con el fin de mitigar su impacto en zonas urbanas.

En este contexto, surge la necesidad de comprender y clasificar los paisajes sonoros que conforman la acústica de una ciudad. Estos se agrupan comúnmente en tres grandes taxonomías: biofonía, geofonía y antropofonía [4]. La biofonía hace referencia a los sonidos biológicos no humanos, como el canto de los pájaros o el zumbido de los insectos. La geofonía engloba sonidos naturales generados por fenómenos físicos, como el viento, la lluvia, los ríos o los terremotos. Finalmente, la antropofonía corresponde a los sonidos producidos por las actividades humanas, tales como el ruido del tráfico, la música o las construcciones.

No obstante, esta clasificación puede resultar insuficiente en entornos urbanos, donde los sonidos se mezclan constantemente, dificultando la identificación clara de cada categoría. La superposición de fuentes sonoras complica el análisis acústico, especialmente cuando se realiza mediante espectrogramas y métodos manuales. Esta mezcla de sonidos representa un desafío considerable para lograr una clasificación precisa y eficiente de los paisajes sonoros urbanos. Si esta situación persiste sin intervención tecnológica, las ciudades seguirán dependiendo de procesos dispendiosos, lentos y no escalables, lo cual limita el análisis oportuno del entorno acústico y su impacto en la calidad de vida.

Es por eso que las entidades gubernamentales y académicas desarrollaron iniciativas orientadas al estudio y análisis del impacto de los paisajes sonoros. Como resultado de estos esfuerzos, se consolidó un proyecto de investigación denominado Urbanphony, en el cual se recolectaron muestras de tres ciudades: Popayán, Santiago de Compostela y Venecia, con el propósito de

analizar posteriormente el efecto de estos paisajes en las personas. Para llevar a cabo este estudio, fue necesario realizar la clasificación de las grabaciones obtenidas, las cuales se representaron mediante espectrogramas, es decir, representaciones visuales de la variación de las frecuencias de una señal a lo largo del tiempo.

Esta clasificación resultó ser un proceso dispendioso, debido a que requirió recurso humano y tiempo para revisar manualmente todos los espectrogramas generados. En respuesta a esta necesidad, Urbanphony desarrolló un modelo de clasificación basado en aprendizaje automático mediante redes neuronales convolucionales. No obstante, surgió la necesidad de comparar dicho modelo y sus resultados con otras herramientas de aprendizaje profundo aplicables a la clasificación acústica compleja.

El problema abordado en este proyecto estuvo relacionado con la clasificación de sonidos mediante redes neuronales modernas, a partir de la necesidad de identificar diferentes tipos de sonidos que podían impactar significativamente la calidad de vida de las personas, la planificación urbana, la salud pública y la sostenibilidad ambiental. En este contexto, la ausencia de herramientas automatizadas para la identificación precisa y oportuna de sonidos dificultó la toma de decisiones basadas en evidencia en entornos urbanos. Por esta razón, el proyecto se enfocó en desarrollar un sistema automatizado de clasificación de sonidos utilizando técnicas de aprendizaje profundo y espectrogramas, con el propósito de apoyar futuras estrategias de monitoreo acústico, análisis ambiental y toma de decisiones basadas en evidencia en entornos urbanos.

1.2. FORMULACIÓN DEL PROBLEMA

Se desarrolló un modelo basado en técnicas de aprendizaje profundo para la clasificación automática de paisajes sonoros registrados en la ciudad de Popayán. Como insumo para las redes neuronales convolucionales empleadas, se utilizaron espectrogramas, entendidos como representaciones visuales del sonido, como insumo para redes neuronales, con el fin de clasificar los sonidos según su origen: biofonía (sonidos provenientes de seres vivos), geofonía (sonidos asociados a fenómenos naturales) y antropofonía (sonidos generados por la actividad humana).

Por tal razón, la pregunta de investigación que guía este trabajo es: ¿cómo desarrollar un modelo de clasificación automática de sonidos basado en técnicas de aprendizaje profundo que permita caracterizar los paisajes sonoros en zonas urbanas, tomando como caso de estudio la ciudad de Popayán?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Diseñar un modelo basado en técnicas de aprendizaje profundo para la clasificación de espectrogramas de sonidos urbanos, con el fin de identificar y categorizar los sonidos según su origen (biofonía, geofonía y antropofonía) en la ciudad de Popayán, Cauca, facilitando su monitoreo y análisis.

2.2. OBJETIVOS ESPECÍFICOS

- Recolectar espectrogramas de sonidos urbanos en la ciudad de Popayán, Cauca.
- Etiquetar los espectrogramas según su origen sonoro (biofonía, geofonía, antropofonía) garantizando la consistencia del conjunto de datos.
- Estructurar una base de datos funcional con los espectrogramas recolectados, destinada al entrenamiento del modelo de clasificación.
- Diseñar un modelo de clasificación de sonidos a través de técnicas de aprendizaje profundo.
- Evaluar la precisión y eficiencia del modelo.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

La contaminación acústica constituye uno de los principales desafíos ambientales en las ciudades modernas debido al crecimiento de las actividades urbanas, el incremento del tráfico vehicular y la expansión de las zonas comerciales e industriales. En este contexto, el estudio del paisaje sonoro ha adquirido relevancia como una herramienta para comprender la interacción entre las diferentes fuentes acústicas presentes en el entorno y sus posibles efectos sobre la calidad de vida de la población.

El análisis de los paisajes sonoros permite identificar, caracterizar y monitorear los sonidos que conforman un territorio, facilitando la comprensión de fenómenos asociados al ruido ambiental, la percepción sonora y la sostenibilidad de los entornos urbanos. Sin embargo, la creciente disponibilidad de registros acústicos ha generado la necesidad de desarrollar métodos automáticos capaces de procesar grandes volúmenes de información de manera eficiente y escalable.

Los avances recientes en procesamiento digital de señales y aprendizaje profundo han permitido desarrollar sistemas capaces de analizar información acústica de forma automática, facilitando la identificación y clasificación de diferentes tipos de sonidos presentes en ambientes urbanos. Estas tecnologías ofrecen nuevas alternativas para complementar los métodos tradicionales de monitoreo acústico y apoyar procesos de gestión ambiental, planificación urbana y análisis del paisaje sonoro.

En este sentido, el presente marco teórico reúne los conceptos fundamentales que sustentan esta investigación, abordando aspectos relacionados con la contaminación acústica, el paisaje sonoro, el procesamiento de señales de audio, los espectrogramas de Mel, las redes neuronales convolucionales, el aprendizaje por transferencia, las estrategias para el manejo del desbalance de clases y las métricas utilizadas para evaluar el desempeño de los modelos de clasificación desarrollados.

3.1.1 Contaminación acústica urbana

La contaminación acústica urbana se define como la presencia en el ambiente de ruidos o vibraciones que pueden generar molestia, riesgo o daño para las personas, sus actividades o los bienes materiales [20]. A diferencia de otros contaminantes ambientales, el ruido se caracteriza por su naturaleza intangible e inmediata, ya que desaparece cuando cesa la fuente emisora; sin embargo, sus efectos sobre la salud pueden ser acumulativos y manifestarse a largo plazo.

La contaminación acústica constituye uno de los principales problemas ambientales en las ciudades modernas debido al incremento de las actividades urbanas, el crecimiento del parque automotor y la expansión de las zonas comerciales e industriales. De acuerdo con la Organización Mundial de la Salud (OMS), la exposición prolongada a niveles elevados de ruido puede afectar el bienestar de las personas y generar impactos negativos sobre la calidad del sueño, el rendimiento cognitivo y la salud cardiovascular [25]. En consecuencia, el ruido ambiental ha sido reconocido como uno de los factores que más afectan la calidad de vida en los entornos urbanos.

En Colombia, la Resolución 0627 de 2006 establece los estándares máximos permisibles de emisión de ruido y ruido ambiental, definiendo límites diferenciados según el uso del suelo y la franja horaria [20]. No obstante, el monitoreo y caracterización de los sonidos presentes en las ciudades continúa representando un desafío debido a las limitaciones de cobertura espacial y temporal de los métodos tradicionales de medición.

Los avances recientes en procesamiento digital de señales y aprendizaje profundo han permitido desarrollar sistemas capaces de analizar grandes volúmenes de información acústica de forma automática, facilitando la identificación y clasificación de diferentes fuentes sonoras presentes en el entorno urbano. Estas tecnologías representan una alternativa prometedora para complementar los mecanismos tradicionales de monitoreo acústico y contribuir a una mejor comprensión de los paisajes sonoros urbanos.

3.1.2. Paisaje sonoro urbano y ecología acústica

El concepto de paisaje sonoro (soundscape) fue introducido por el compositor canadiense R. Murray Schafer en la década de 1970 en el contexto del World Soundscape Project. Schafer propuso comprender el entorno sonoro como una composición acústica susceptible de ser escuchada, analizada e interpretada, estableciendo las bases de la ecología acústica como disciplina de estudio. Desde esta perspectiva, el paisaje sonoro no se limita únicamente a la medición física del ruido, sino que incorpora la relación entre las fuentes sonoras, el entorno y la percepción humana.

Dentro de este enfoque, Schafer distinguió tres elementos fundamentales: las keynotes, que corresponden a los sonidos de fondo que caracterizan un lugar; las sound signals, entendidas como sonidos que destacan sobre el entorno y reclaman atención; y las soundmarks, que representan sonidos únicos e identitarios de una comunidad o territorio [31]. Estos conceptos permitieron ampliar la comprensión del entorno acústico más allá de los niveles de presión sonora, incorporando aspectos culturales, sociales y perceptuales.

Posteriormente, Bernie Krause desarrolló la hipótesis del nicho acústico (Acoustic Niche Hypothesis), según la cual las especies biológicas tienden a ocupar diferentes bandas de frecuencia para evitar interferencias entre sus señales sonoras [16]. A partir de este planteamiento, propuso una taxonomía del paisaje sonoro compuesta por tres categorías principales: biofonía, geofonía y antropofonía.

La biofonía agrupa los sonidos producidos por organismos vivos no humanos, tales como aves, insectos y otros animales. La geofonía comprende los sonidos generados por fenómenos naturales no biológicos, incluyendo la lluvia, el viento, las corrientes de agua y las tormentas. Finalmente, la antropofonía corresponde a los sonidos originados por las actividades humanas y tecnológicas, como el tráfico vehicular, la maquinaria, las construcciones, las voces y la música [16].

Aunque esta clasificación fue concebida inicialmente para el estudio de ecosistemas naturales, ha sido ampliamente adoptada en investigaciones relacionadas con paisajes sonoros urbanos debido a su capacidad para caracterizar las diferentes fuentes acústicas presentes en una ciudad. En estos entornos, la antropofonía suele constituir la fuente sonora predominante debido a la alta concentración de actividades humanas y sistemas de transporte.

La complejidad del paisaje sonoro urbano radica en la superposición constante de múltiples fuentes acústicas presentes de manera simultánea. Este fenómeno, conocido en la literatura como cocktail party effect, dificulta la identificación y separación de sonidos individuales tanto para los seres humanos como para los sistemas automáticos de clasificación. Como consecuencia, las señales acústicas urbanas suelen presentar mezclas espectrales donde diferentes fuentes comparten bandas de frecuencia y patrones temporales similares, aumentando la dificultad de las tareas de reconocimiento y categorización sonora.

Estas características han motivado el uso de técnicas de procesamiento digital de señales y aprendizaje profundo capaces de extraer patrones complejos a partir de representaciones tiempo-frecuencia, permitiendo abordar problemas de clasificación acústica en entornos con alta superposición de sonidos.

3.1.3 Espectrogramas de Mel como representación de señales de audio

En el análisis de señales acústicas, resulta necesario transformar las ondas de audio en representaciones que permitan observar simultáneamente su comportamiento temporal y frecuencial. Una de las representaciones más utilizadas para este propósito es el espectrograma, el cual muestra cómo se distribuye la energía de una señal en función del tiempo y la frecuencia.

La construcción de un espectrograma se fundamenta en la Transformada de Tiempo Corto de Fourier (STFT, por sus siglas en inglés), una técnica que divide la señal en pequeñas ventanas temporales y aplica la Transformada Discreta de Fourier (DFT) sobre cada una de ellas. Este procedimiento permite analizar cómo evolucionan las componentes frecuenciales a lo largo del tiempo, generando una representación bidimensional donde es posible observar simultáneamente la información temporal y frecuencial de la señal [24].

Como se observa en la Figura 1, el eje horizontal representa el tiempo, mientras que el eje vertical corresponde a la frecuencia medida en hercios (Hz). La intensidad del color refleja la energía presente en cada componente frecuencial, generalmente expresada en decibelios (dB).

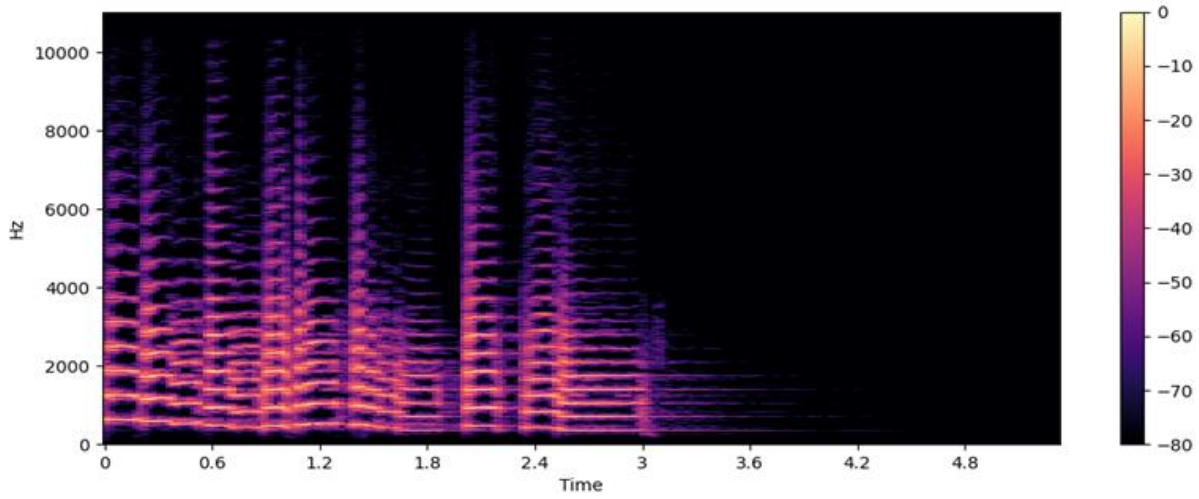

Fig. 1. Espectrograma generado a partir de una señal de audio.[14]

No obstante, la percepción humana del sonido no sigue un comportamiento lineal. El sistema auditivo presenta una mayor sensibilidad a variaciones en frecuencias bajas que en frecuencias altas. Por ejemplo, aunque para un sistema computacional la diferencia entre 100 Hz y 200 Hz es comparable a la diferencia entre 15.000 Hz y 15.100 Hz, perceptualmente el primer cambio resulta mucho más significativo para el oído humano.

Con el fin de incorporar esta característica perceptual al análisis acústico, se propuso la escala Mel, una transformación no lineal de la frecuencia inspirada en la forma como los seres humanos perciben los sonidos [5]. La conversión entre frecuencia lineal y frecuencia Mel puede expresarse mediante la siguiente ecuación:

$$m = 2595(1 + \frac{f}{700})$$

Donde (f) corresponde a la frecuencia expresada en hercios (Hz) y (m) representa la frecuencia equivalente en escala Mel.

Esta transformación asigna una mayor resolución a las frecuencias bajas y medias, donde se concentra gran parte de la información relevante para la percepción humana, y una menor resolución a las frecuencias altas. Como resultado, permite representar de manera más adecuada las características acústicas percibidas por el oído humano.

A partir de esta escala se construyen los espectrogramas de Mel, una representación ampliamente utilizada en tareas de reconocimiento de voz, clasificación de sonidos ambientales y análisis acústico mediante aprendizaje profundo. Su popularidad se debe a que preservan características perceptualmente relevantes del sonido, reducen la dimensionalidad de la información y permiten reformular los problemas de clasificación de audio como tareas de clasificación de imágenes, facilitando el uso de arquitecturas convolucionales y técnicas de aprendizaje por transferencia [13].

Como se observa en la Figura 2, los espectrogramas de Mel enfatizan aquellas regiones frecuenciales que resultan más relevantes para la percepción humana, permitiendo representar de manera más eficiente los patrones acústicos presentes en las señales de audio.

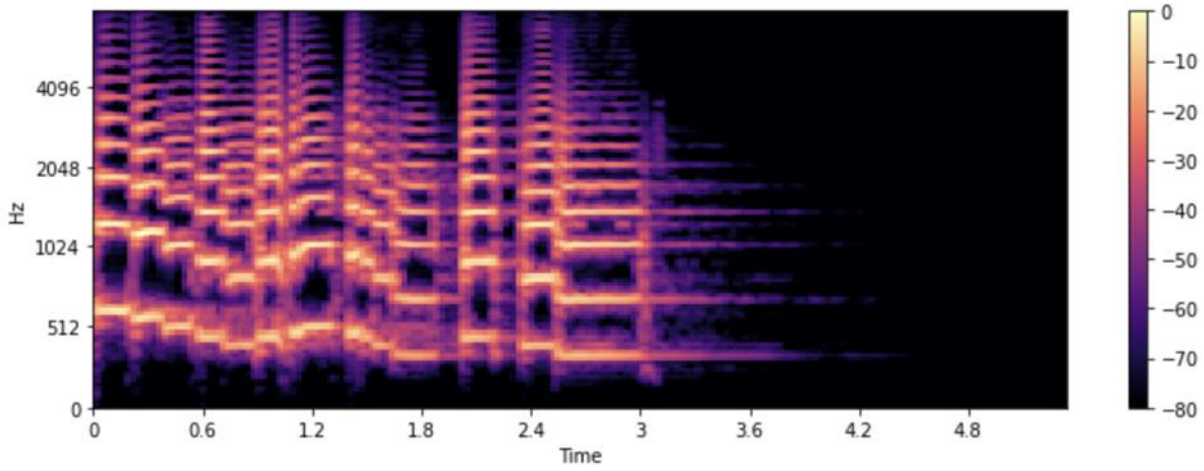

Fig. 2. Ejemplo de espectrograma de Mel.[14]

3.1.4 Fundamentos de Aprendizaje Profundo

El aprendizaje profundo (Deep Learning) es una rama del aprendizaje automático basada en redes neuronales artificiales con múltiples capas de procesamiento, capaces de aprender representaciones jerárquicas de los datos a partir de grandes volúmenes de información [7]. A diferencia de los enfoques tradicionales, donde las características deben ser definidas manualmente por un experto, los modelos de aprendizaje profundo aprenden automáticamente patrones relevantes directamente desde los datos de entrada.

Gracias a esta capacidad, el aprendizaje profundo ha demostrado resultados sobresalientes en tareas de visión por computador, procesamiento del lenguaje natural y análisis de señales de audio. En el caso de la clasificación automática de sonidos, estas técnicas permiten identificar patrones acústicos complejos presentes en representaciones tiempo-frecuencia como los espectrogramas de Mel.

3.1.4.1 Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) constituyen una de las arquitecturas más utilizadas en tareas de clasificación de imágenes debido a su capacidad para extraer automáticamente características relevantes a partir de datos con estructura espacial [17]. A diferencia de las redes neuronales densas tradicionales, las CNN emplean capas convolucionales que procesan regiones locales de la entrada mediante filtros aprendidos durante el entrenamiento.

Esta arquitectura permite detectar patrones jerárquicos de complejidad creciente. Mientras las primeras capas identifican características simples, como bordes, texturas o formas básicas, las capas más profundas combinan dichas características para reconocer patrones de mayor nivel. En el contexto de esta investigación, las CNN permiten identificar patrones acústicos presentes en los espectrogramas de Mel, transformando el problema de clasificación de sonidos en una tarea de reconocimiento de imágenes.

Además de las capas convolucionales, estas arquitecturas incorporan **capas de pooling**, encargadas de reducir la dimensionalidad espacial de los mapas de características, disminuyendo el costo computacional y aumentando la robustez frente a pequeñas variaciones de la entrada. Finalmente, las **capas densas** (fully connected) realizan la clasificación a partir de las características extraídas durante las etapas anteriores.

La función de activación más utilizada en este tipo de arquitecturas es **ReLU** (Rectified Linear Unit), definida como:

$$ReLU(x) = max\ (0, x)$$

Esta función introduce no linealidad al modelo y contribuye a mitigar el problema del gradiente durante el entrenamiento [27].

3.1.4.2 Aprendizaje por Transferencia y Fine-Tuning

El aprendizaje por transferencia (Transfer Learning) consiste en reutilizar modelos previamente entrenados sobre grandes conjuntos de datos para resolver nuevas tareas relacionadas [26]. Esta estrategia permite aprovechar características aprendidas previamente, reduciendo la cantidad de datos y tiempo de entrenamiento requeridos para desarrollar modelos con buen desempeño.

En aplicaciones de visión por computador es común utilizar arquitecturas preentrenadas sobre ImageNet, un conjunto de datos compuesto por millones de imágenes pertenecientes a diferentes categorías. Las primeras capas de estos modelos aprenden características generales, como bordes, formas y texturas, que posteriormente pueden reutilizarse en otros dominios, incluyendo la clasificación de espectrogramas de audio.

Una técnica ampliamente utilizada dentro del aprendizaje por transferencia es el fine-tuning o ajuste fino. Este proceso consiste en congelar inicialmente las capas preentrenadas del modelo y entrenar únicamente las capas de clasificación agregadas para la nueva tarea. Posteriormente, algunas capas del modelo base son descongeladas para adaptar progresivamente las

representaciones aprendidas al dominio específico de interés, mejorando la capacidad de generalización del modelo [39].

El uso de transfer learning resulta especialmente útil cuando se dispone de conjuntos de datos de tamaño moderado, ya que permite aprovechar conocimiento previamente adquirido por modelos entrenados sobre grandes volúmenes de información, reduciendo el riesgo de sobreajuste y mejorando la estabilidad del entrenamiento.

3.1.4.3 Regularización y Prevención del Sobreajuste

El sobreajuste (overfitting) ocurre cuando un modelo aprende patrones específicos del conjunto de entrenamiento en lugar de capturar relaciones generales que puedan aplicarse a nuevos datos. Como consecuencia, el modelo obtiene resultados muy favorables durante el entrenamiento, pero presenta un desempeño inferior al enfrentarse a datos no vistos previamente [7].

Para reducir este problema, se han desarrollado diversas técnicas de regularización. Una de las más utilizadas es **Dropout,** la cual consiste en desactivar aleatoriamente una fracción de las neuronas durante el entrenamiento, obligando a la red a aprender representaciones más robustas y menos dependientes de conexiones específicas [34].

Otra estrategia ampliamente utilizada es la **regularización L2**, que incorpora una penalización sobre los pesos del modelo dentro de la función de pérdida. Esta técnica favorece la obtención de modelos con parámetros más pequeños y distribuidos, contribuyendo a mejorar su capacidad de generalización.

Adicionalmente, mecanismos como **Early Stopping** permiten detener el entrenamiento cuando el desempeño sobre el conjunto de validación deja de mejorar, evitando que el modelo continúe ajustándose excesivamente a los datos de entrenamiento. Estas técnicas constituyen herramientas fundamentales para el desarrollo de modelos de aprendizaje profundo con capacidad de generalizar adecuadamente sobre nuevos datos.

3.1.5 Arquitecturas Profundas para Clasificación de Imágenes

Las arquitecturas profundas de visión por computador han experimentado avances significativos durante la última década, alcanzando resultados sobresalientes en tareas de clasificación, detección y segmentación de imágenes. Debido a que los espectrogramas de Mel constituyen representaciones visuales de las señales de audio, estas arquitecturas pueden emplearse para identificar patrones acústicos complejos mediante técnicas de aprendizaje profundo.

3.1.5.1 MobileNetV2

MobileNetV2 es una arquitectura propuesta por Howard et al. [11] con el objetivo de reducir el costo computacional de las redes convolucionales sin afectar significativamente su desempeño. Su diseño se basa en el uso de convoluciones separables en profundidad (Depthwise Separable Convolutions) y bloques residuales invertidos (Inverted Residual Blocks), mecanismos que disminuyen considerablemente el número de parámetros y operaciones requeridas durante el entrenamiento y la inferencia.

Gracias a estas características, MobileNetV2 se ha consolidado como una alternativa eficiente para aplicaciones donde existen restricciones de memoria o capacidad de procesamiento, manteniendo niveles competitivos de precisión respecto a arquitecturas convolucionales más complejas.

3.1.5.2 EfficientNet

La familia EfficientNet fue propuesta por Tan y Le [36] como una estrategia para mejorar el desempeño de las redes convolucionales mediante un método de escalado compuesto (compound scaling). A diferencia de enfoques anteriores que incrementaban únicamente la profundidad o el ancho de la red, EfficientNet ajusta simultáneamente la profundidad, el ancho y la resolución de entrada utilizando un único coeficiente de escalado.

Esta estrategia permite obtener modelos con una mejor relación entre precisión y costo computacional. EfficientNetB0 constituye la arquitectura base de esta familia y sirve como punto de partida para el desarrollo de variantes más profundas y complejas mediante el proceso de escalado compuesto.

3.1.5.3 EfficientNetV2

EfficientNetV2 representa una evolución de la familia EfficientNet orientada a mejorar la eficiencia computacional y la velocidad de entrenamiento [37]. Para ello incorpora bloques denominados Fused-MBConv, los cuales simplifican ciertas operaciones convolucionales presentes en versiones anteriores y permiten acelerar el proceso de aprendizaje.

Además de mantener las ventajas del escalado compuesto, EfficientNetV2 busca optimizar el equilibrio entre precisión y tiempo de entrenamiento, convirtiéndose en una alternativa ampliamente utilizada en aplicaciones modernas de clasificación visual.

3.1.5.4 Arquitecturas Basadas en Transformers

Los modelos Transformer fueron introducidos originalmente para tareas de procesamiento del lenguaje natural mediante mecanismos de atención capaces de modelar dependencias de largo alcance dentro de una secuencia [38]. Posteriormente, estos mecanismos fueron adaptados al campo de la visión por computador, permitiendo analizar relaciones globales entre diferentes regiones de una imagen.

A diferencia de las redes convolucionales tradicionales, cuya capacidad de aprendizaje depende principalmente de relaciones locales determinadas por el tamaño de los filtros, los Transformers pueden establecer conexiones entre regiones distantes de la entrada mediante mecanismos de atención. Esta característica les permite capturar patrones complejos distribuidos a lo largo de toda la representación visual.

Las arquitecturas híbridas CNN-Transformer combinan las ventajas de ambos enfoques: las capas convolucionales extraen características locales de bajo nivel, mientras que los bloques Transformer modelan relaciones globales de largo alcance. Esta combinación ha demostrado resultados prometedores en problemas de clasificación donde es necesario capturar simultáneamente patrones locales y dependencias globales dentro de los datos [38].

3.1.6 Estrategias de Manejo del Desbalance de Clases

El desbalance de clases es uno de los problemas más frecuentes en tareas de clasificación supervisada. Este fenómeno ocurre cuando algunas categorías poseen una cantidad significativamente mayor de muestras que otras, provocando que los modelos tiendan a favorecer las clases mayoritarias durante el entrenamiento [11].

Una de las estrategias más utilizadas para mitigar este problema consiste en asignar pesos diferenciados a cada clase durante el cálculo de la función de pérdida. De esta manera, los errores cometidos sobre clases minoritarias reciben una penalización mayor que aquellos asociados a clases frecuentes, obligando al modelo a prestar una atención más equilibrada a todas las categorías.

La biblioteca Scikit-Learn implementa este enfoque mediante la estrategia *balanced*, donde el peso asignado a cada clase se calcula de forma inversamente proporcional a su frecuencia dentro del conjunto de entrenamiento [28]:

$$\text{Peso de Clases} = \frac{\text{Total de Muestras}}{\text{(No. Clases x No. Muestras)}}$$

Esta estrategia permite compensar el efecto del desbalance sin alterar la distribución original de los datos ni generar observaciones sintéticas adicionales, favoreciendo una optimización más equilibrada del modelo.

3.1.6.1 Función de Pérdida Focal (Focal Loss)

Otra alternativa ampliamente utilizada para abordar problemas de desbalance es la función de pérdida focal (*Focal Loss*), propuesta por Lin et al. [18]. Esta técnica surge como una modificación de la entropía cruzada tradicional y tiene como objetivo reducir la influencia de las muestras que el modelo clasifica fácilmente, concentrando el proceso de aprendizaje en aquellos ejemplos que presentan mayor dificultad.

A diferencia de la entropía cruzada convencional, la pérdida focal incrementa la importancia relativa de las observaciones mal clasificadas y reduce progresivamente la contribución de aquellas que ya son correctamente reconocidas por el modelo. Como resultado, favorece el aprendizaje de categorías minoritarias o complejas y contribuye a disminuir el sesgo hacia las clases más frecuentes.

Tanto los pesos de clase balanceados como la función de pérdida focal constituyen estrategias ampliamente utilizadas para el tratamiento del desbalance de clases en problemas de clasificación supervisada. La selección de una u otra alternativa depende de las características específicas del conjunto de datos y de los objetivos particulares de cada aplicación.

3.1.7 Métricas de Evaluación para Clasificación Multiclase Desbalanceada

La evaluación de modelos de clasificación requiere métricas capaces de cuantificar objetivamente el desempeño alcanzado por el algoritmo. En problemas multiclase con desbalance entre categorías, la utilización exclusiva de la exactitud (accuracy) puede conducir a interpretaciones erróneas, ya que un modelo podría obtener resultados aparentemente favorables clasificando correctamente únicamente las clases mayoritarias [33].

Por esta razón, en tareas de clasificación multiclase desbalanceada es recomendable complementar la exactitud con métricas que permitan evaluar de manera individual el comportamiento de cada categoría y su contribución al desempeño global del modelo.

3.1.7.1 Exactitud (Accuracy)

La exactitud corresponde a la proporción de predicciones correctas respecto al total de observaciones evaluadas y se calcula mediante:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (33)$$

donde:

- TP: Verdaderos positivos.
- TN: Verdaderos negativos.
- FP: Falsos positivos.
- FN: Falsos negativos.

Aunque esta métrica proporciona una visión general del desempeño del modelo, su interpretación puede resultar limitada cuando existe un desbalance importante entre clases [33].

3.1.7.2 Precisión (Precision)

La precisión mide la proporción de predicciones positivas realizadas por el modelo que realmente pertenecen a la categoría evaluada:

$$Precision = \frac{TP}{(TP+FP)} \quad (33)$$

Valores elevados de precisión indican que el modelo produce pocos falsos positivos para la clase analizada.

3.1.7.3 Sensibilidad (Recall)

La sensibilidad o recall cuantifica la capacidad del modelo para identificar correctamente las observaciones pertenecientes a una determinada clase:

$$Recall = \frac{TP}{(TP+FN)} \quad (33)$$

Esta métrica resulta especialmente relevante cuando el objetivo es minimizar la cantidad de ejemplos que pasan inadvertidos durante el proceso de clasificación.

3.1.7.4 F1-Score

El F1-Score combina precisión y recall mediante su media armónica:

$$F1 = 2\frac{(Precision*Recall)}{(Precision+Recall)} \quad (33)$$

Esta métrica proporciona una medida equilibrada entre ambas dimensiones y resulta especialmente útil cuando existe desbalance entre categorías [33].

3.1.7.5 F1 Macro

El F1 Macro se obtiene calculando el F1-Score de cada clase de forma independiente y posteriormente promediando los resultados sin considerar el número de muestras de cada categoría. Esta métrica otorga la misma importancia a todas las categorías, independientemente de su frecuencia, por lo que es ampliamente utilizada en problemas multiclase desbalanceados [33].

3.1.7.6 Exactitud Balanceada (Balanced Accuracy)

La exactitud balanceada se calcula como el promedio del recall obtenido para cada clase. Esta métrica evita que las clases mayoritarias dominen la evaluación global y proporciona una medida más representativa del desempeño real del modelo cuando existe desbalance entre categorías [40].

3.1.7.7 Matriz de Confusión

La matriz de confusión es una herramienta visual que permite comparar las clases reales con las clases predichas por el modelo. Cada fila representa las observaciones reales y cada columna las predicciones realizadas.

El análisis de esta matriz permite identificar patrones de error, categorías frecuentemente confundidas y clases que presentan mayores dificultades de clasificación. Por esta razón, constituye uno de los mecanismos más utilizados para interpretar el comportamiento de modelos multiclase y complementar la información proporcionada por las métricas globales.

3.2. ANTECEDENTES

El campo de la clasificación automática de sonidos ambientales y urbanos mediante aprendizaje profundo ha experimentado un crecimiento acelerado en la última década, impulsado por la disponibilidad de datasets públicos y el avance de las arquitecturas convolucionales. A continuación se presentan los antecedentes más relevantes para el presente proyecto, organizados según su aporte metodológico y temático.

3.2.1 Clasificación de sonidos ambientales mediante redes neuronales convolucionales

Uno de los trabajos pioneros en este campo fue desarrollado por Piczak (2015), quien propuso el uso de redes neuronales convolucionales para la clasificación de sonidos ambientales a partir de espectrogramas de Mel utilizando los conjuntos de datos ESC-50 y UrbanSound8K [29]. Sus resultados demostraron que las CNN superaban los enfoques tradicionales basados en características manuales y clasificadores convencionales, evidenciando el potencial de estas arquitecturas para el reconocimiento automático de patrones acústicos.

Posteriormente, Salamon y Bello (2017) evaluaron el impacto de diferentes estrategias de aumento de datos sobre modelos convolucionales aplicados a sonidos urbanos [30]. Los autores demostraron que transformaciones como cambios de velocidad, desplazamientos temporales y adición de ruido permitían mejorar significativamente la capacidad de generalización de los modelos, alcanzando resultados superiores a los reportados previamente en UrbanSound8K.

Por su parte, Hershey et al. (2017) realizaron un estudio a gran escala utilizando el conjunto de datos AudioSet de Google [13]. Sus hallazgos mostraron que los modelos preentrenados sobre grandes volúmenes de datos podían transferir eficazmente conocimiento hacia tareas de clasificación acústica, consolidando el aprendizaje por transferencia como una estrategia efectiva para conjuntos de datos de tamaño moderado.

3.2.2 Arquitecturas modernas para clasificación de audio

Con la evolución de las arquitecturas profundas, diferentes investigaciones han explorado modelos más avanzados para el análisis de señales acústicas. Gong et al. (2021) propusieron el Audio Spectrogram Transformer (AST), una arquitectura basada en mecanismos de atención

aplicada directamente sobre espectrogramas de Mel [8]. Sus resultados evidenciaron que los modelos basados en Transformers pueden capturar relaciones globales presentes en las representaciones tiempo-frecuencia, alcanzando desempeños competitivos frente a arquitecturas convolucionales tradicionales.

De forma complementaria, Kong et al. (2020) desarrollaron las Pretrained Audio Neural Networks (PANNs), una familia de modelos entrenados sobre AudioSet que proporcionan representaciones acústicas transferibles para múltiples tareas de clasificación de audio [15]. Este trabajo demostró la importancia del preentrenamiento específico en el dominio acústico para mejorar el desempeño en escenarios con disponibilidad limitada de datos.

3.2.3 Clasificación de paisajes sonoros urbanos

Dentro del contexto específico del paisaje sonoro urbano, Mesaros, Heittola y Virtanen (2016) desarrollaron el TUT Database for Acoustic Scene Classification and Sound Event Detection [21], uno de los conjuntos de datos de referencia más utilizados en los desafíos internacionales DCASE. Este trabajo contribuyó significativamente al desarrollo de metodologías estandarizadas para la clasificación automática de escenas acústicas.

En el contexto latinoamericano, Durán-Paredes et al. (2025) desarrollaron el modelo Urbanphony-3-CNN utilizando grabaciones recolectadas en la ciudad de Popayán como parte del proyecto Urbanphony [6]. Empleando una arquitectura EfficientNetB3 y validación cruzada de cinco pliegues, los autores alcanzaron una exactitud del 75,2 %, demostrando la viabilidad de aplicar arquitecturas profundas para la clasificación de paisajes sonoros urbanos con alta superposición acústica. Este antecedente constituye la referencia más cercana al presente trabajo, dado que utiliza el mismo contexto urbano y un conjunto de datos relacionado.

3.2.4 Trabajos relacionados desarrollados en la Pontificia Universidad Javeriana Cali

En la Pontificia Universidad Javeriana Cali se han desarrollado diversos trabajos que evidencian el potencial de las redes neuronales convolucionales para resolver problemas de clasificación basados en imágenes.

Villalobos Tenorio (2024) desarrolló un clasificador de sonidos de alerta dirigido a personas con discapacidad auditiva utilizando el conjunto de datos AudioSet [19]. El estudio implementó un flujo metodológico basado en la conversión de audio a espectrogramas y el uso de aprendizaje por transferencia, demostrando la aplicabilidad de estas técnicas para el reconocimiento automático de eventos sonoros.

De manera similar, Castro Duarte (2024) empleó modelos convolucionales para el diagnóstico automatizado de Leishmaniasis cutánea a partir de imágenes médicas.[22] Por su parte, Mejía Salgado (2024) comparó distintas arquitecturas CNN para la identificación de actividad de toxoplasmosis ocular mediante imágenes de fondo de ojo [23]. Ambos trabajos resaltan la importancia de la selección de arquitecturas, la optimización de hiperparámetros y el uso de aprendizaje por transferencia en problemas de clasificación visual con conjuntos de datos de tamaño moderado.

En términos generales, los antecedentes revisados evidencian que los espectrogramas de Mel constituyen una representación adecuada para la clasificación automática de sonidos mediante aprendizaje profundo. Asimismo, muestran una tendencia creciente hacia el uso de arquitecturas preentrenadas, estrategias de transferencia de aprendizaje y modelos híbridos que combinan mecanismos convolucionales y de atención. No obstante, se identificó una cantidad limitada de estudios enfocados específicamente en la clasificación de paisajes sonoros urbanos en ciudades latinoamericanas utilizando taxonomías acústicas adaptadas a contextos urbanos complejos. En este sentido, el presente trabajo busca aportar evidencia empírica mediante la comparación de diferentes arquitecturas profundas sobre el conjunto de datos generado en el proyecto Urbanphony para la ciudad de Popayán.

4. CONJUNTO DE DATOS Y PREPARACIÓN DE LOS DATOS

4.1 Contexto y origen de los datos

El conjunto de datos utilizado en este estudio se derivó del proyecto Urbanphony, una iniciativa internacional orientada al análisis del paisaje sonoro urbano en ciudades de Europa y Latinoamérica. Para el caso de Popayán, se dispuso de 250 registros de audio obtenidos en 50 puntos estratégicamente distribuidos en el centro histórico de la ciudad y recolectados en cinco franjas horarias representativas de la actividad urbana.

En cada uno de estos puntos se realizaron mediciones acústicas en cinco periodos de muestreo durante el día: 08:00–10:00, 10:00–12:00, 12:00–14:00, 14:00–16:00 y 16:00–18:00. Este esquema permitió capturar la variabilidad sonora asociada a diferentes momentos de la jornada, incluyendo cambios en la intensidad del tráfico, la actividad comercial, la movilidad peatonal y la presencia de sonidos naturales.

Los registros fueron suministrados al equipo de trabajo por el director del proyecto aplicado y correspondían a grabaciones previamente etiquetadas dentro del proyecto Urbanphony. A partir de este corpus acústico, los autores desarrollaron las etapas de segmentación de audio, generación de espectrogramas Mel, construcción del conjunto de datos para entrenamiento y desarrollo de modelos de aprendizaje profundo orientados a la clasificación automática de sonidos urbanos.

La disponibilidad de grabaciones previamente etiquetadas permitió concentrar el alcance del proyecto en las fases de preparación de datos, construcción del dataset en espectrogramas y diseño de modelos de clasificación automática, manteniendo como referencia las categorías definidas durante la construcción del corpus acústico.

4.2 Esquema de etiquetado del conjunto de datos

Las grabaciones utilizadas en este estudio fueron suministradas como parte del proyecto y contaban con etiquetas previamente definidas. Estas etiquetas fueron asignadas mediante un proceso de clasificación manual realizado durante la construcción del corpus acústico, permitiendo identificar el tipo de sonido predominante presente en cada grabación.

El etiquetado manual consistía en escuchar cada grabación y asignar una etiqueta de sonido. Al realizar este proceso, se presentan grandes retos debido al tiempo que se requiere, el esfuerzo humano y el conocimiento especializado, especialmente en entornos donde múltiples fuentes sonoras pueden mezclarse unos con otras.

Una de las motivaciones de este trabajo surge de la necesidad de apoyar la automatización de este proceso. Actualmente, la clasificación de paisajes sonoros depende en gran medida de la intervención de expertos para identificar y categorizar los sonidos presentes en las grabaciones, lo que limita la escalabilidad de los estudios acústicos. En este contexto, los modelos de aprendizaje profundo desarrollados buscan contribuir a la clasificación automática de nuevos registros sonoros mediante el reconocimiento de patrones acústicos característicos.

Inicialmente, el esquema de clasificación del paisaje sonoro urbano se fundamentó en la taxonomía propuesta por Krause, la cual contempla tres categorías principales: biofonía, geofonía y antropofonía. No obstante, durante la construcción del corpus acústico se evidenció que la categoría de antropofonía agrupaba fuentes sonoras con características acústicas y comportamientos diferenciados, tales como el tránsito vehicular, las voces o la música, y los sonidos asociados al movimiento de personas.

Con el propósito de representar de manera más precisa la diversidad sonora, la categoría de antropofonía fue subdividida en tres clases específicas: transporte motorizado (TM), voces o música (VM) y movimiento humano (MH). Adicionalmente, se incorporó la categoría de silencio (SI) para aquellos segmentos en los que no se identificó un evento sonoro predominante relevante para el análisis. Como resultado, el conjunto de datos final quedó conformado por seis categorías: biofonía (BI), geofonía (GE), silencio (SI), movimiento humano (MH), transporte motorizado (TM) y voces o música (VM), las cuales se presentan en la Figura 3.

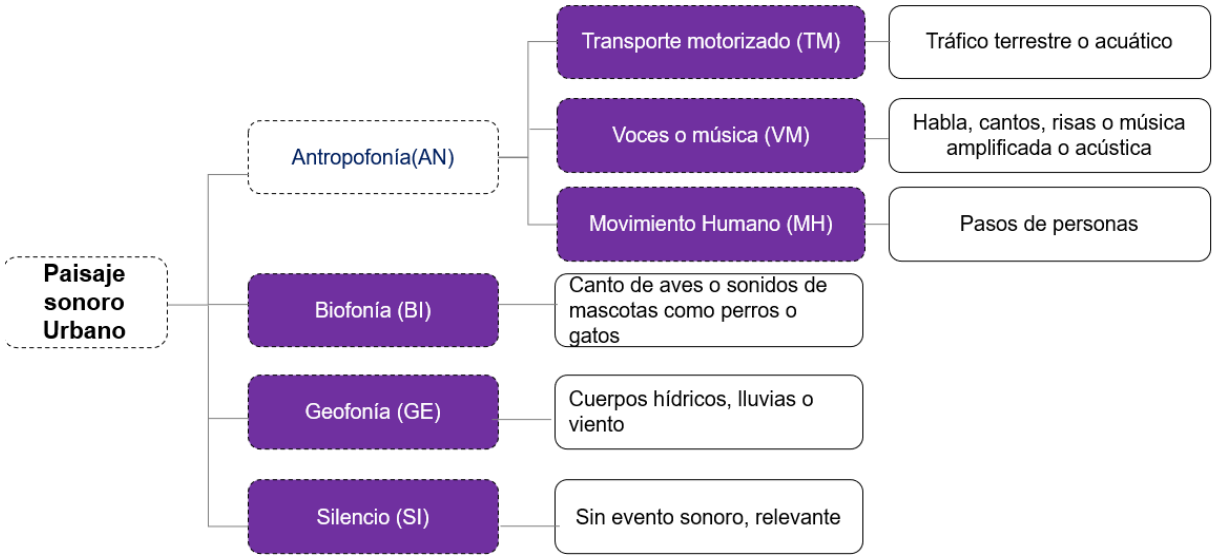

Fig. 3. Categorías acústicas empleadas para el etiquetado del conjunto de datos.

Como se observa en la Figura 3, el esquema de clasificación adoptado permitió representar con mayor nivel de detalle la diversidad acústica presente en los entornos urbanos. La desagregación de la categoría de antropofonía facilitó la diferenciación de fuentes sonoras de origen humano, mientras que las categorías de biofonía, geofonía y silencio complementaron la caracterización del paisaje sonoro utilizada para el entrenamiento y evaluación de los modelos desarrollados.

4.3 Segmentación y preprocesamiento de audio

A partir de las 250 grabaciones originales previamente etiquetadas, se llevó a cabo un proceso de segmentación mediante scripts desarrollados en Python utilizando las bibliotecas Librosa y FFmpeg. Cada grabación fue dividida en fragmentos de dos segundos de duración con el propósito de obtener muestras acústicas más homogéneas para el proceso de clasificación.

La utilización de ventanas temporales cortas permite capturar eventos sonoros específicos presentes en el entorno urbano, reduciendo el efecto de la superposición de múltiples fuentes acústicas que suele presentarse en grabaciones de mayor duración. Esto facilita la identificación de patrones característicos asociados a cada categoría de sonido y mejora la consistencia de las muestras utilizadas durante el entrenamiento de los modelos.

Como resultado de este procedimiento se obtuvo un total de 7.138 archivos de audio en formato WAV, los cuales constituyeron la base para la generación de espectrogramas Mel y la posterior construcción del conjunto de datos utilizado en el entrenamiento y evaluación de los modelos de aprendizaje profundo.

4.4 Generación de espectrogramas de Mel

En esta sección se describe la transformación de los 7.138 fragmentos de audio en espectrogramas de Mel, utilizados como datos de entrada para los modelos de clasificación. Los espectrogramas permiten representar las señales de audio en el dominio tiempo-frecuencia, proporcionando una representación visual de los patrones acústicos presentes en cada muestra y facilitando su procesamiento mediante técnicas de aprendizaje profundo.

La decisión de transformar las señales de audio a espectrogramas de Mel responde a varias ventajas técnicas para la clasificación automática de sonidos urbanos. En primer lugar, esta representación permite visualizar simultáneamente la evolución temporal y frecuencial de la señal, facilitando la identificación de patrones acústicos característicos que no son fácilmente observables en la forma de onda original.

En segundo lugar, la representación bidimensional de los espectrogramas permite reformular el problema como una tarea de clasificación de imágenes, posibilitando el uso de arquitecturas convolucionales preentrenadas mediante técnicas de transfer learning. Esta estrategia resulta especialmente útil cuando se dispone de conjuntos de datos de tamaño moderado, ya que aprovecha características visuales previamente aprendidas por los modelos y reduce la necesidad de entrenarlos completamente desde cero.

Finalmente, la escala Mel incorpora principios de la percepción auditiva humana, otorgando una mayor resolución a las frecuencias bajas y medias donde se concentra gran parte de la información presente en sonidos urbanos como voces, tráfico vehicular y actividad humana. Estas características han convertido a los espectrogramas Mel en una de las representaciones más utilizadas en tareas de clasificación automática de audio.

4.4.1 Carga y estandarización de las señales

Cada archivo de audio fue cargado a una frecuencia de 22.050 Hz. Este proceso fue necesario debido a que las grabaciones originales presentaban diferentes frecuencias de muestreo (44.100 Hz y 48.000 Hz), lo que podría generar inconsistencias en la representación espectral de las

señales.

4.4.2 Transformación a espectrogramas Mel

Cada señal fue transformada en un espectrograma de Mel mediante la Transformada de Tiempo Corto de Fourier (STFT), técnica que permite analizar cómo se distribuye la energía frecuencial de la señal a lo largo del tiempo. A partir de esta transformación se construyó el espectrograma de Mel, el cual aplica una escala no lineal basada en la percepción auditiva humana. Esta escala genera mayor sensibilidad a las frecuencias bajas y menor sensibilidad a las frecuencias altas, asemejando al oído humano.

Los parámetros utilizados para la generación de los espectrogramas se definieron con el propósito de lograr un equilibrio entre resolución espectral, resolución temporal y costo computacional. La tabla 1 resume los parámetros implementados durante el procesamiento.

Tabla 1. Parámetros utilizados para la generación de los espectrogramas de Mel.

Parámetro	Valor	Justificación
sr (frecuencia de muestreo)	22.050 Hz	Frecuencia estandarizada utilizada para garantizar uniformidad entre señales
n_mels (filtros Mel)	128	Balance entre resolución frecuencial y costo computacional
n_fft (tamaño de ventana FFT)	2.048 muestras	Resolución espectral adecuada para distinguir patrones acústicos
hop_length (salto de trama)	512 muestras	Resolución temporal aproximada de 23 ms por frame

Duración del segmento	2 segundos	Reduce la mezcla excesiva de múltiples fuentes sonoras
Formato de salida	PNG escala a grises	Compatible con arquitecturas convolucionales

La selección de estos parámetros corresponde a criterios técnicos. El valor de n_fft = 2.048 permite una resolución frecuencial para diferenciar los sonidos urbanos como voces o motores sin incurrir en un costo computacional. El hop_length de 512 muestras, permite representar variaciones de sonido a lo largo del tiempo con buena granularidad.

Una vez obtenido los espectrogramas de Mel, los valores de energía fueron convertidos a escala logarítmica en decibelios (dB) y posteriormente normalizados para su representación como imágenes en escala de grises. Una vez obtenidos los espectrogramas de Mel, los valores de energía fueron convertidos a escala logarítmica en decibelios (dB) y posteriormente normalizados para su representación como imágenes en escala de grises. Esta representación conserva la información espectral necesaria para el proceso de clasificación y reduce la complejidad computacional asociada al procesamiento de imágenes multicanal.

Por último, cada espectrograma fue almacenado en formato PNG para su uso durante el entrenamiento de los modelos.

4.5 Construcción del dataset y partición estratificada

Una vez obtenido el conjunto completo de espectrogramas, se realizó la división del dataset en subconjunto de entrenamiento, validación y prueba. Para ello se empleó una partición estratificada, con el fin de preservar la proporción original de cada categoría en todos los subconjuntos y evitar sesgos durante el proceso de aprendizaje.

La distribución adoptada fue de 70% para entrenamiento, 15% para validación y 15% para prueba. El subconjunto de entrenamiento fue utilizado para el ajuste de los parámetros de los modelos, mientras que el conjunto de validación permitió monitorear el desempeño durante el entrenamiento y apoyar la selección de hiperparámetros. Finalmente, el conjunto de prueba fue reservado para la evaluación final de los modelos, garantizando una medición objetiva de su capacidad de generalización. La Tabla 2 presenta la distribución final de las muestras utilizadas en cada subconjunto.

Tabla 2. Distribución de las muestras del conjunto de datos en los subconjuntos de entrenamiento, validación y prueba

Conjunto	Número de muestras	Porcentaje
Entrenamiento	4.994	70%
Validación	1.071	15%
Prueba	1.073	15%

Con el propósito de garantizar la reproducibilidad de los experimentos y asegurar que todas las arquitecturas fueran evaluadas bajo las mismas condiciones, se definió una semilla fija (SEED = 2026), permitiendo conservar exactamente la misma partición de datos durante todas las fases de entrenamiento, validación y prueba.

4.6 Análisis exploratorio del conjunto de datos

4.6.1 Distribución de clases y desbalance

El análisis exploratorio del conjunto de datos de los 7.138 fragmentos etiquetados evidenció un desbalance significativo entre las clases. La clase TM (transporte motorizado) concentra el 43% de las muestras, seguido de VM (voces o música) con el 27% y MH (movimiento humano) con el 15,7%. Mientras que las clases SI, BI y GE presentaron entre un 4,5% y 4,7% del total de las muestras. Este desbalance es importante, ya que puede influir en el desempeño del modelo favoreciendo las clases mayoritarias, por lo que se consideró en la etapa de entrenamiento del modelo para evitar sesgos.

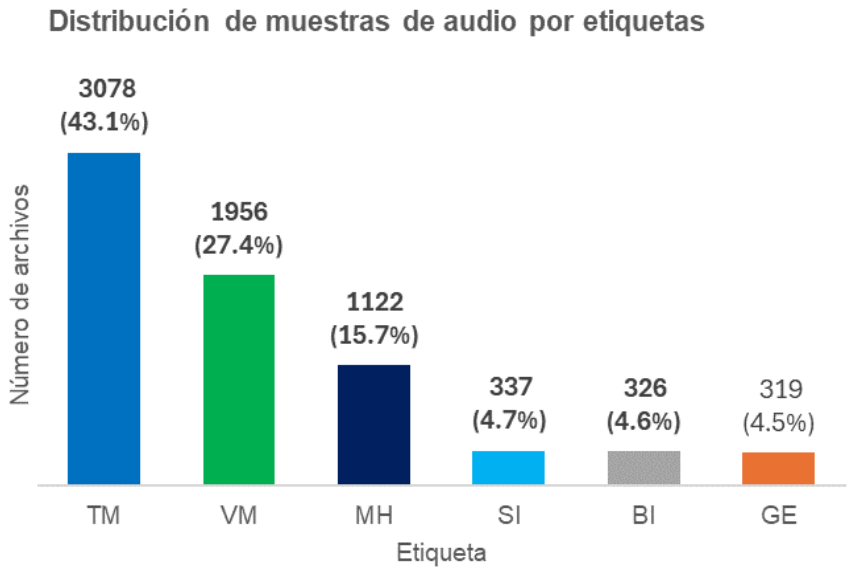

Fig. 4. Distribución de muestras por categoría acústica en el conjunto de datos

Se evidencia un desbalance significativo entre clases, con TM representando el 43,1% del total de muestras.

4.6.2 Comportamiento temporal de los registros

En relación con la duración de los fragmentos, éstos presentaron una homogeneidad cercana a 2 segundos, lo que garantiza consistencia para la entrada del modelo. Este análisis permitió verificar la consistencia de los segmentos generados durante el proceso de preprocesamiento, garantizando que las muestras utilizadas para el entrenamiento presentaran una duración homogénea y adecuada para la entrada de los modelos de clasificación.

Tabla 3. Estadísticas descriptivas de la duración de los segmentos de audio por categoría acústica

Clase	N muestras	Media (s)	Desv. Std (s)	Mínimo (s)	Máximo (s)

BI	326	2.0000	0.0000	2.000	2.000
GE	319	1.9998	0.0037	1.935	2.000
MH	1122	1.9994	0.0131	1.632	2.000
SI	337	2.0000	0.0000	2.000	2.000
TM	3078	1.9994	0.00183	1.224	2.000
VM	1956	1.9996	0.0137	1.488	2.000

La Tabla 3 presenta las estadísticas descriptivas de la duración de los segmentos de audio para cada una de las categorías del conjunto de datos. Se observa que la duración promedio se mantiene cercana a los 2 segundos en todas las clases, con desviaciones estándar mínimas, lo que evidencia la uniformidad alcanzada durante el proceso de segmentación.

Los resultados obtenidos durante el análisis exploratorio permitieron identificar dos características relevantes del conjunto de datos: la presencia de un desbalance significativo entre las categorías y la homogeneidad en la duración de los segmentos de audio. Estas condiciones fueron consideradas posteriormente durante las etapas de construcción del dataset y entrenamiento de los modelos de clasificación.

4.6.3 Exploración geoespacial preliminar de los registros acústicos

Como parte de la caracterización exploratoria del conjunto de datos, se realizó una representación geográfica de los registros acústicos recolectados en el centro histórico de Popayán. Para ello se utilizaron las coordenadas geográficas asociadas a los puntos de muestreo, permitiendo visualizar la distribución espacial de las categorías acústicas presentes en el estudio. La Figura 5 muestra la ubicación de los registros sobre el mapa del área de estudio, donde cada color representa la categoría dominante identificada en el correspondiente punto de observación. Se observa una mayor presencia de registros asociados a transporte motorizado (TM) y voces o música (VM), categorías que también presentaron la mayor representación dentro del conjunto de datos.

Aunque este análisis tiene un carácter exploratorio y no constituye el objetivo principal del trabajo, permitió obtener una comprensión inicial de la distribución espacial de las diferentes categorías acústicas y contextualizar las condiciones bajo las cuales fueron recolectadas las grabaciones utilizadas para el entrenamiento de los modelos.

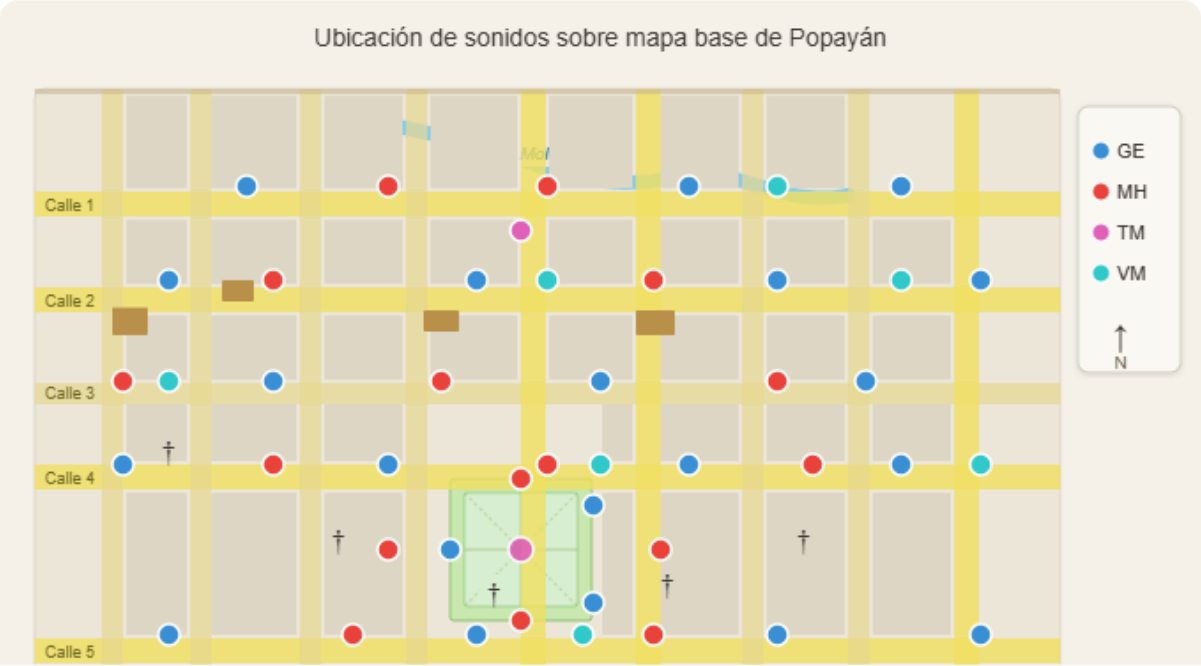

Fig.5. Distribución espacial de las categorías acústicas registradas en los puntos de muestreo de Popayán

Fuente: Elaboración propia a partir de los registros georreferenciados del proyecto Urbanphony.

Nota: La representación cartográfica presentada corresponde a una visualización exploratoria de los puntos de muestreo y no a un análisis espacial inferencial.

4.6.4 Exploración en el dominio frecuencial

Con fines exploratorios, se analizaron las señales en el dominio de la frecuencia mediante la generación de espectrogramas tipo Mel para muestras representativas de cada clase.

Este análisis permitió observar cómo se distribuye la energía del sonido en diferentes bandas de frecuencia a lo largo del tiempo. Algunas categorías presentan mayor concentración en frecuencias bajas, mientras que otras evidencian patrones más dispersos o eventos energéticos puntuales en frecuencias medias.

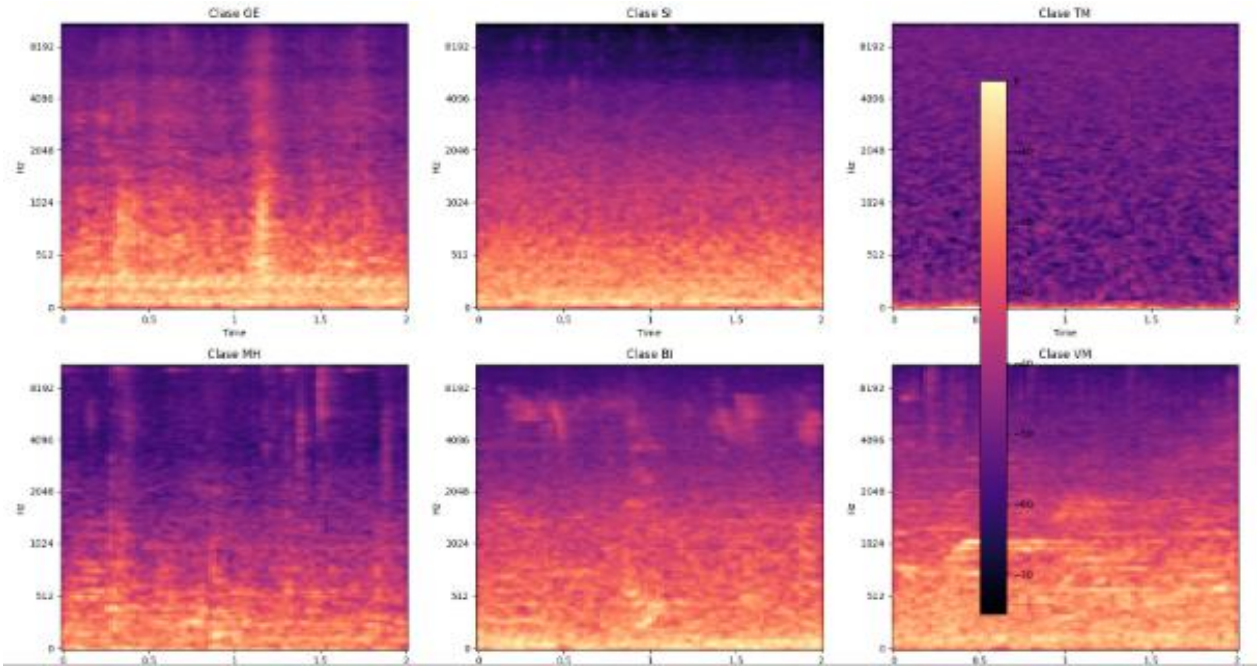

Fig. 6. Espectrogramas de Mel representativos de las categorías acústicas del conjunto de datos.

En la Fig.6, se presentan seis espectrogramas correspondientes a las clases de sonidos analizados. El eje horizontal representa el tiempo (segundos), el eje vertical representa la frecuencia (Hz), y la escala de colores codifica la energía de cada componente frecuencial. Los tonos de color oscuro (azul/morado) indican baja energía del sonido, mientras que los colores cálidos (naranja/ amarillo) representan mayor intensidad.

A partir de estos espectrogramas es posible identificar diferencias entre las clases. Por ejemplo, las clases GE y SI presentan una mayor concentración de energía en frecuencias bajas, lo que se refleja en zonas de mayor intensidad en la parte inferior del espectrograma, indicando sonidos más constantes. Mientras, la clase TM presenta menor variabilidad espectral, evidenciada por una predominancia de tonos de baja intensidad. Las clases MH, BI y VM muestran una mayor variabilidad en la distribución energética, con patrones intermitentes que aparecen en distintas frecuencias y momentos del tiempo. Estas diferencias en la distribución de energía, texturas y patrones constituyen características discriminativas que pueden ser utilizadas por modelos de aprendizaje profundo para distinguir entre las distintas clases de sonidos.

Finalmente, a partir de este análisis integral, se logró comprender la estructura, distribución y comportamiento acústico del conjunto de datos, evidenciando que las diferencias entre clases se manifiestan con mayor claridad en el dominio frecuencial que en el dominio temporal, lo que

justifica la transformación de los fragmentos de audio en espectrogramas como representación de entrada para el modelo de clasificación.

En este sentido, en el siguiente capítulo se describe el proceso de recolección, preprocesamiento y construcción del dataset completo en espectrogramas, así como su estructuración para el entrenamiento del modelo de aprendizaje profundo.

5 DISEÑO DE MODELOS DE CLASIFICACIÓN DE SONIDOS

5.1 Estrategia general del modelado

Dado que los fragmentos de audio fueron representados mediante espectrogramas de Mel , se evaluaron distintas arquitecturas de redes neuronales convolucionales con el propósito de identificar el mejor desempeño.

Sin embargo, el desbalance entre clases representó un reto adicional para el proceso de clasificación. En consecuencia, la estrategia adoptada consistió en evaluar cuatro arquitecturas de visión mediante búsqueda aleatoria de hiperparámetros (random search), 3 trials por arquitectura, SEED= 2026, utilizando el F1 macro como métrica de selección, dada la distribución desbalanceada de las clases. Esta métrica fue seleccionada por otorgar la misma importancia a

todas las clases, evitando que las categorías mayoritarias dominaran la evaluación del desempeño.

El modelo con mejor desempeño fue posteriormente reentrenado bajo una estrategia optimizada que incorporó técnicas de ponderación de clases, un esquema de entrenamiento en dos fases y una regularización L2 para evitar sobreajustes y obtener un mejor resultado.

5.2 Preparación de datos

5.2.1 Pesos de clases balanceados

El desbalance entre clases fue abordado mediante el cálculo de peso de clases usando la estrategia "balanced" de scikit-learn. Con esta decisión, se mitigó el riesgo de que el modelo aprendiera a predecir las clases mayoritarias, obteniendo un accuracy alto y sin clasificar adecuadamente las clases BI, GE y SI.

Por ende, se aplicó la siguiente fórmula para asignar un mayor peso a las clases minoritarias y así obtener resultados más favorables en su clasificación:

$$\text{Peso de Clases} = \frac{\text{Total de Muestras}}{(\text{No. Clases x No. Muestras})}$$

Los pesos calculados se presentan en la Tabla 4.

Tabla 4. Pesos de clase calculados mediante la estrategia "balanced" de Scikit-learn para compensar el desbalance del conjunto de datos.

Clase	N muestras (train)	Peso asignado	Categoría
GE	223	3,7324	Clase minoritaria → peso alto (penaliza más errores)
BI	228	3,6506	Clase minoritaria → peso alto (penaliza más errores)

SI	235	3,5418	Clase minoritaria → peso alto (penaliza más errores)
MH	785	1,0603	Clase balanceada → peso neutro
VM	1.369	0,6080	Clase mayoritaria → peso bajo (penaliza menos errores)
TM	2.154	0,3864	Clase mayoritaria → peso bajo (penaliza menos errores)

Los pesos de clase actúan sobre la función de pérdida durante el entrenamiento: cuando el modelo comete un error en una muestra de clase GE, el gradiente se amplifica 3,73 veces más que si el error fuera sobre una muestra de clase TM. Esto obliga al modelo a prestar mayor atención a las clases minoritarias durante la optimización, mitigando el sesgo hacia las clases mayoritarias sin necesidad de generar muestras sintéticas.

5.3 Estrategia de entrenamiento y validación

Todas las arquitecturas fueron entrenadas utilizando la misma partición estratificada definida en el Capítulo 4. El conjunto de entrenamiento fue empleado para el ajuste de los parámetros del modelo, mientras que el conjunto de validación fue utilizado para monitorear el desempeño durante el proceso de aprendizaje y seleccionar la configuración con mejor capacidad de generalización.

La métrica principal utilizada para comparar arquitecturas fue el F1 Macro, debido al desbalance existente entre las categorías. A diferencia del accuracy, esta métrica otorga la misma importancia a todas las clases y permite evaluar de forma más equilibrada el desempeño global del modelo.

El proceso de entrenamiento se desarrolló mediante una estrategia bifásica. En una primera etapa de calentamiento (warmup), el backbone preentrenado permaneció congelado, permitiendo estabilizar los pesos de las capas de clasificación agregadas sobre las representaciones aprendidas en ImageNet. Posteriormente, se ejecutó una fase de ajuste fino (fine-tuning), donde se

descongelaron progresivamente capas superiores del backbone con el propósito de adaptar las características aprendidas al dominio específico de los espectrogramas acústicos.

Con el fin de controlar el sobreajuste y mejorar la estabilidad del entrenamiento, se implementaron tres mecanismos de regularización. EarlyStopping detuvo el entrenamiento cuando la pérdida de validación dejó de mejorar durante cuatro épocas consecutivas, restaurando automáticamente los mejores pesos obtenidos. ModelCheckpoint almacenó el modelo con mejor desempeño en términos de exactitud de validación (val_accuracy), mientras que ReduceLROnPlateau redujo dinámicamente la tasa de aprendizaje cuando no se observaron mejoras en la función de pérdida.

5.4 Arquitecturas evaluadas

5.4.1 MobileNetV2 con bloque Squeeze-and-Excitation

MobileNetV2 fue seleccionada por ser una arquitectura liviana y eficiente computacionalmente, adecuada para conjuntos de datos de tamaño moderado. Adicionalmente, se incorporó un bloque Squeeze-and-Excitation con el propósito de recalibrar la importancia relativa de los canales de características extraídos del espectrograma, permitiendo al modelo enfatizar patrones acústicos relevantes para la clasificación.

5.4.2 EfficientNetB0

EfficientNetB0 fue seleccionada por su estrategia de escalado compuesto, la cual permite balancear profundidad, ancho y resolución de la red manteniendo una complejidad computacional moderada. Adicionalmente, fue la única arquitectura a la que se aplicó regularización L2 ($\lambda = 0{,}00001$), con el objetivo de reducir el sobreajuste y favorecer una mejor capacidad de generalización.

5.4.3 EfficientNetV2S con Multi-Head Attention Pooling

EfficientNetV2S fue seleccionada por ser una versión optimizada de EfficientNet orientada a mejorar la velocidad de entrenamiento. Sobre esta arquitectura se incorporó un mecanismo de Multi-Head Attention Pooling con el propósito de capturar relaciones globales entre diferentes regiones del espectrograma y complementar la información local extraída por los filtros convolucionales.

5.4.4 CNN transformer

Se evaluó una arquitectura híbrida CNN-Transformer con el objetivo de combinar la capacidad de las redes convolucionales para extraer patrones locales del espectrograma con la habilidad de los mecanismos de atención para modelar dependencias globales. A diferencia de las demás arquitecturas evaluadas, este modelo no contaba con pesos preentrenados, por lo que debió aprender las representaciones acústicas directamente a partir de los datos disponibles.

5.5 Búsqueda de hiperparámetros (Random Search)

La búsqueda aleatoria de hiperparámetros permitió identificar la mejor configuración para cada arquitectura evaluada. Debido a las restricciones computacionales del estudio, se realizaron tres configuraciones por modelo, seleccionando posteriormente aquella que obtuvo el mayor valor de F1 Macro sobre el conjunto de validación.

Tabla 5. Espacio de hiperparámetros explorado durante la búsqueda aleatoria para las arquitecturas evaluadas.

Hiperparámetro	Valores explorados	Justificación
Learning rate	0,001 / 0,0005 / 0,0001	Valores bajos preservan el conocimiento preentrenado y evitan degradar representaciones aprendidas en ImageNet
Dropout	0,2 / 0,3	Regularización moderada acorde al tamaño del dataset disponible
Batch size	16 / 32	Balance entre velocidad de entrenamiento y uso de memoria en el entorno GPU disponible
Dense units	128 / 256 / 384 / 512	Tamaño de capa de clasificación; valores limitados para evitar compensación artificial de errores del backbone
Fine-tune layers	0 / 10 / 20 / 30 / 50 / 60 / 80	Número de capas superiores del backbone a descongelar para ajuste fino del dominio

Num. attention heads	4 / 8	Para arquitecturas con atención; más cabezas capturan más patrones, pero incrementan el costo computacional

Los resultados de la búsqueda aleatoria evidenciaron diferencias importantes entre las arquitecturas evaluadas sobre el conjunto de validación. La configuración asociada a EfficientNetB0 alcanzó los valores más altos de accuracy de validación (0,7254) y F1 Macro (0,645) entre las configuraciones exploradas, por lo que fue seleccionada para su evaluación posterior sobre el conjunto de prueba independiente. Cabe resaltar que esta arquitectura fue la única en la que se incorporó regularización L2 ($\lambda = 0{,}00001$) como parte de la configuración experimental. Por su parte, MobileNetV2 y CNN-Transformer obtuvieron desempeños intermedios, con valores de F1 Macro de 0,5497 y 0,624, respectivamente. En contraste, EfficientNetV2S presentó el menor desempeño sobre el conjunto de validación, alcanzando un F1 Macro de 0,079 y un accuracy de validación de 0,0934, lo que evidencia dificultades para adaptarse al conjunto de datos bajo las configuraciones exploradas durante la búsqueda de hiperparámetros.
En términos generales, los resultados obtenidos en la etapa de validación permitieron identificar las configuraciones más prometedoras para cada arquitectura. Sin embargo, la determinación del desempeño final de los modelos se realizó posteriormente mediante la evaluación sobre el conjunto de prueba independiente, cuyos resultados se presentan en el capítulo siguiente.

6. EVALUACIÓN DEL DESEMPEÑO DEL MODELO SELECCIONADO

A partir de los resultados obtenidos durante la etapa de búsqueda de hiperparámetros, se seleccionó la configuración basada en **EfficientNetB0** para realizar una evaluación más detallada sobre el conjunto de prueba independiente. Esta evaluación permitió analizar el comportamiento del modelo utilizando métricas de clasificación y determinar su capacidad de generalización frente a datos no observados durante el entrenamiento.

6.1. Parámetros del modelo final

El modelo final está compuesto por las siguientes capas (Tabla 6), con un total de 4.379.049 parámetros. Durante la fase de warmup, únicamente los 329.478 parámetros de las capas densas

son entrenables (backbone congelado). En la fase de fine-tuning se descongelan las capas superiores del backbone.

Tabla 6. Estructura de capas y parámetros del modelo EfficientNetB0 seleccionado para la evaluación final

Capa (Tipo)	Forma de salida (Output Shape)	Parámetros
InputLayer	(None, 224, 224, 3)	0
Lambda (efficientnet_preprocess)	(None, 224, 224, 3)	0
EfficientNetB0 (Functional)	(None, 7, 7, 1280)	4049571
GlobalAveragePooling2D	(None, 1280)	0
Dropout	(None, 1280)	0
Dense	(None, 256)	327936
Dense (Salida)	(None, 6)	1542

Total de parámetros: 4.379.049

Parámetros entrenables : 329.478

Parámetros no entrenables: 4.049.571

Nota: En la fase warmup solo se entrenan 329.478 parámetros (cabeza clasificadora).

6.2 Curva de entrenamiento

Las curvas de entrenamiento del modelo final (Figura 11) muestran el comportamiento de la pérdida y la exactitud a lo largo de las épocas, con la línea punteada indicando el fin de la fase warmup (época 3).

La curva de pérdida evidencia una disminución consistente en entrenamiento, con la pérdida de validación siguiendo una tendencia similar aunque con mayor variabilidad. La exactitud de validación presenta un comportamiento creciente durante las primeras épocas tras el warm up, estabilizándose en valores entre 0,55 y 0,65. La brecha entre curvas de entrenamiento y

validación es moderada, indicando un grado controlado de sobreajuste compatible con el tamaño del dataset disponible.

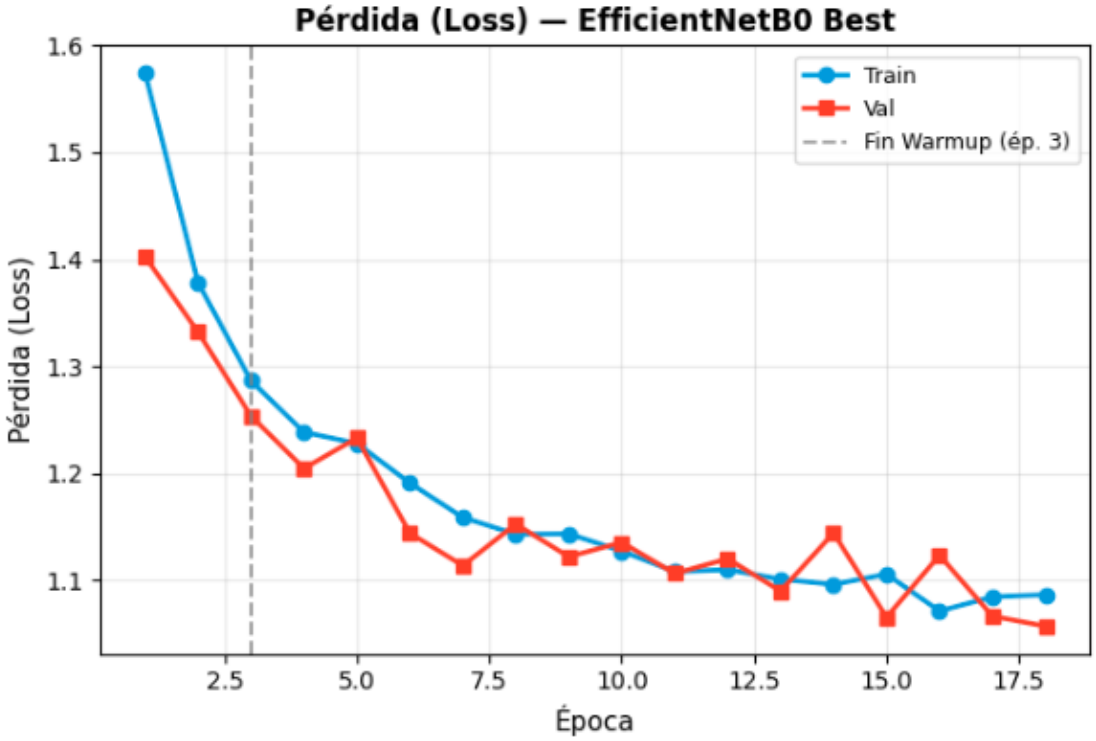

Fig.7. Evolución de la pérdida y la exactitud durante el entrenamiento del modelo EfficientNetB0

6.3 Métricas finales sobre el conjunto de prueba

La Tabla 7 presenta las métricas finales del modelo sobre los conjuntos de validación y prueba independiente.

Tabla 7. Métricas de desempeño del modelo EfficientNetB0 sobre los conjuntos de validación y prueba independiente

Conjunto	Accuracy	Balanced Accuracy	F1 Macro	F1 Weighted
Validación	0,6116	0,6339	0,5288	0,6446
Prueba (Test)	0,6086	0,5873	0,5089	0,6411

La similitud entre las métricas de validación y prueba (F1 macro: 0,52 vs. 0,51; *accuracy*: 0,61 vs. 0,60) indica una adecuada capacidad de generalización sobre datos no vistos, sin evidencia de sobreajuste significativo. La exactitud balanceada en el conjunto de prueba (0,587) es notablemente superior al desempeño esperado por azar en un problema de seis clases (0,167), lo que confirma que el modelo aprendió representaciones discriminatorias relevantes para todas las categorías, no solo para las mayoritarias.

6.4 Reporte de clasificación por clase

La Tabla 8 presenta el reporte de clasificación completo sobre el conjunto de prueba independiente (1.073 muestras).

Tabla 8. Reporte de clasificación del modelo EfficientNetB0 sobre el conjunto de prueba independiente.

Clase	Precisión	Recall	F1-Score	Support
BI (Biofonía)	0,38	0,49	0,43	49
GE (Geofonía)	0,19	0,44	0,26	48
MH (Movimiento Humano)	0,63	0,54	0,58	169
SI (Silencio)	0,25	0,80	0,38	51
TM (Tráfico/Motor)	0,86	0,64	0,74	462

VM (Voces o Música)	0,74	0,61	0,67	294

El análisis por clase revela patrones coherentes con la naturaleza acústica del problema. TM (F1=0,74) y VM (F1=0,67) obtienen el mejor desempeño, consistente con su mayor representación en el dataset y con la distintividad de sus patrones espectrales. MH (F1=0,58) presenta un resultado sólido pese a ser la tercera clase por volumen.

Por otro lado, la clase SI muestra un comportamiento particular: alta precisión del recall (0,80) con baja precisión (0,25), lo que indica que el modelo tiende a clasificar incorrectamente sonidos de otras categorías como silencio. Esto es razonable en el contexto de grabaciones de campo donde el "silencio" no es ausencia total de señal sino baja energía global, característica que puede compartirse con fragmentos de baja intensidad de otras clases.

La clase GE presenta el peor desempeño (F1=0,26), con precisión muy baja (0,19). Con solo 223 muestras de entrenamiento y patrones espectrales difusos el viento y la lluvia producen distribuciones de energía espectralmente similares a algunas grabaciones de tráfico lejano. Esta clase representa el mayor desafío para el modelo y constituye la principal limitación de los resultados.

6.5 Análisis de la matriz de confusión

La Tabla 9 presenta la matriz de confusión del modelo EfficientNetBO sobre el conjunto de prueba, donde los valores en la diagonal representan las clasificaciones correctas por clase.

Tabla 9. Matriz de confusión del modelo EfficientNetB0 sobre el conjunto de prueba independiente

Real/ Predicho	BI	GE	MH	SI	TM	VM
BI (49)	24	4	3	7	6	5
GE (48)	0	21	10	10	5	2
MH (169)	2	31	92	15	6	23
SI (51)	0	4	0	41	5	1

TM (462)	10	34	14	76	297	31
VM (294)	27	19	27	18	25	178

Los errores más frecuentes tienen interpretación acústica clara. La confusión de GE con MH (10 casos) y con SI (10 casos) se explica porque el viento o el agua corriendo con baja energía pueden sonar espectralmente similares a pasos suaves o a segmentos de silencio parcial. La confusión de MH con GE (31 casos) es el error más numeroso fuera de la diagonal y refleja la dificultad de separar pasos de personas en exteriores de sonidos geofónicos de fondo.

La confusión de TM con SI (76 casos) es la más difícil de explicar desde la física acústica: el ruido vehicular tiene energía bien diferenciada del silencio. Una posible explicación es que algunos fragmentos etiquetados como TM corresponden a instantes de baja actividad vehicular (entre vehículos), cuya energía global es similar a la de los fragmentos de silencio. Esta confusión sugiere que el criterio de etiquetado de TM podría refinarse en trabajo futuro.

Consideramos que el problema principal del dataset es la mezcla de sonidos en los segmentos de 2 segundos: al reducir los audios a esta ventana temporal no se elimina la superposición de fuentes, que continúa afectando la calidad del etiquetado y, por extensión, el desempeño del modelo.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 Conclusiones

Los resultados obtenidos evidencian que es posible abordar la clasificación automática del paisaje

sonoro urbano de Popayán mediante técnicas de aprendizaje profundo aplicadas a espectrogramas de Mel. Aunque el desempeño alcanzado aún presenta oportunidades de mejora, los resultados muestran que los modelos son capaces de identificar patrones acústicos relevantes y diferenciar parcialmente las categorías consideradas en el estudio.

En relación con los objetivos específicos planteados, se construyó un conjunto de 7.138 segmentos de audio a partir de 250 grabaciones previamente etiquetadas dentro del proyecto Urbanphony, se generó un dataset de espectrogramas Mel con partición estratificada reproducible, se compararon cuatro arquitecturas mediante búsqueda aleatoria de hiperparámetros y se evaluó el modelo final sobre un conjunto de prueba independiente, alcanzando un F1 Macro de 0,51 y una exactitud de 60,9%.

La reformulación del problema acústico como una tarea de reconocimiento de imágenes, combinada con estrategias de transfer learning y manejo del desbalance, constituyó una aproximación metodológica robusta dado el tamaño moderado del dataset disponible. EfficientNetB0, con regularización L2 y entrenamiento bifásico, logró el mejor balance entre desempeño y generalización entre las cuatro arquitecturas evaluadas.

Entre los principales desafíos identificados, se destaca que el desbalance de clases afecta significativamente el desempeño en las categorías minoritarias GE y BI, las cuales presentaron las métricas F1 más bajas con un 0,26 y 0,43 respectivamente. Asimismo, la mezcla de sonidos en los segmentos de 2 segundos introduce un ruido de etiqueta que limita la capacidad discriminativa del modelo, reflejando que en el entorno urbano de Popayán el silencio puro y la geofonía aislada son eventos acústicos poco frecuentes y difíciles de capturar sin superposición. Finalmente, la ausencia de una métrica de acuerdo inter-anotador, como el coeficiente Kappa de Cohen, surge como una limitación metodológica que impide cuantificar la confiabilidad del etiquetado, constituyendo un punto clave a corregir en trabajos futuros

A pesar de estas limitaciones, la exactitud balanceada de 0,587, notablemente superior al desempeño esperado por azar en un problema de seis clases (0,167), evidencia que el modelo logró aprender patrones discriminatorios asociados a las diferentes categorías acústicas evaluadas. Asimismo, el conjunto de datos utilizado, proveniente del proyecto Urbanphony y compuesto por grabaciones reales del paisaje sonoro urbano de Popayán, constituyó un recurso valioso para el desarrollo y evaluación de estrategias de clasificación automática, aportando evidencia sobre los desafíos y oportunidades que presenta el análisis acústico urbano en contextos latinoamericanos.

Los resultados obtenidos son consistentes con los reportados por el proyecto Urbanphony [6], el cual desarrolló un modelo de clasificación CNN aplicado también a grabaciones del paisaje sonoro

urbano de Popayán, alcanzando una exactitud del 75,2% con la arquitectura EfficientNetB3 bajo validación cruzada de cinco pliegues. La diferencia respecto a la exactitud del 60,9% de este proyecto puede estar asociada a diferencias metodológicas entre ambos trabajos. Este trabajo involucró las seis categorías del paisaje sonoro, entre ellas la geofonía (GE), la clase con mayor dificultad con F1 del 0,26. Se comparó con cuatro arquitecturas mediante la búsqueda aleatoria de hiperparámetros y se incorporó pesos de clase balanceados. Ambos estudios confirman que las arquitecturas de la familia EfficientNet constituyen la aproximación más adecuada para la clasificación de sonidos urbanos en entornos con alta superposición acústica y que el dataset de Popayán fue valioso para el desarrollo y monitoreo del paisaje sonoro en ciudades latinoamericanas.

Este proyecto evidencia el potencial de la inteligencia artificial como herramienta de apoyo para el análisis y monitoreo automático de paisajes sonoros urbanos. Los resultados obtenidos muestran que este tipo de aproximaciones puede servir como base para futuros sistemas de monitoreo acústico, apoyando procesos de gestión ambiental, planificación urbana y evaluación de la calidad acústica en diferentes ciudades. Lo anterior cobra relevancia considerando que la exposición prolongada al ruido se asocia con trastornos del sueño, estrés y enfermedades cardiovasculares. En este sentido, este trabajo constituye una contribución al estudio de herramientas tecnológicas orientadas a comprender y gestionar los entornos sonoros urbanos desde una perspectiva de sostenibilidad y bienestar.

7.2 Trabajos Futuros

A partir de los hallazgos obtenidos, se proponen las siguientes líneas de investigación futura:

- **Ampliar las clases minoritarias con muestreo dirigido:** recolectar muestras adicionales de GE y BI con el propósito de reducir el desbalance existente entre categorías y mejorar la representación de estas clases durante el entrenamiento.
- **Incorporar métricas de acuerdo inter-anotador:** calcular Cohen's Kappa para los segmentos etiquetados por múltiples anotadores, con el fin de cuantificar la confiabilidad del etiquetado y filtrar los segmentos con alta incertidumbre antes del entrenamiento.
- **Explorar arquitecturas especializadas en audio:** evaluar PANNs (Pretrained Audio Neural Networks) [9] y modelos como HTS-AT o BEATs, que fueron preentrenados directamente sobre grandes colecciones de audio y pueden capturar características acústicas no presentes en modelos de visión.
- **Aplicar análisis de interpretabilidad:** generar mapas de activación Grad-CAM para el modelo EfficientNetB0 sobre muestras correctamente e incorrectamente clasificadas, con el objetivo de identificar qué regiones del espectrograma determinan las decisiones del modelo y si estas son acústicamente coherentes.
- **Explorar validación cruzada estratificada:** en este proyecto se optó por una partición fija del dataset (70/15/15) con semilla reproducible, lo que permitió evaluar el modelo sobre un conjunto de prueba completamente independiente. Como trabajo futuro, la incorporación de validación cruzada estratificada permitiría analizar la estabilidad de los resultados frente a diferentes particiones de los datos, especialmente en clases con menor representación como GE y BI.

Anexo A. Disponibilidad del código del Proyecto

Con el propósito de favorecer la reproducibilidad de la investigación, el código fuente desarrollado durante el proyecto fue organizado y publicado en un repositorio GitHub. Este repositorio incluye los scripts utilizados para el preprocesamiento de las señales de audio, la generación de espectrogramas de Mel, la construcción del conjunto de datos, el entrenamiento de los modelos de aprendizaje profundo y la evaluación de resultados.
Asimismo, se documenta la estructura del corpus utilizado y los procedimientos empleados para la generación de las representaciones gráficas presentadas en este trabajo.

Repositorio GitHub:
https://github.com/MariaPaulaJ/Clasificacion-espectrogramas-sonidos-urbanos

La información contenida en el repositorio permite replicar las principales etapas metodológicas desarrolladas durante la investigación, facilitando la validación y reutilización de los resultados obtenidos.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] DANE, "Proyecciones de población municipales por área 2018–2035," Departamento Administrativo Nacional de Estadística, Bogotá, Colombia, 2025. [En línea]. Disponible en: https://www.dane.gov.co

[2] Ministerio de Ambiente, "Resolución 0627 de 2006 — Norma nacional de emisión de ruido y ruido ambiental", Diario Oficial 46.231, 2006.

[3] B. L. Krause, "The niche hypothesis: a virtual symphony of animal sounds, the origins of musical expression and the health of habitats", Soundscape Newsletter, vol. 6, pp. 6-10, 1993.

[4] D. O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley, 1987.

[5] S. Davis and P. Mermelstein,"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,"IEEE Transactions on Acoustics, Speech, and Signal Processing,vol. 28, no. 4,pp. 357–366,1980.

[6] Durán-Paredes, C. A., Grijalba-Obando, J. A., Cajas-Ordóñez, S. A., & Sánchez-Ferreira, C. (2025). Urbanphony-3-CNN: A Convolutional Neural Network for Identifying the Urban Soundscape Taxonomy in Spectrograms Generated from Audios of Historic Cities. Ingeniería e Investigación, 45(2), art. e114942. https://doi.org/10.15446/ing.investig.114942

[7] I. Goodfellow, Y. Bengio and A. Courville,Deep Learning.Cambridge, MA, USA: MIT Press, 2016.

[8] Y. Gong, Y.-A. Chung and J. Glass,"AST: Audio Spectrogram Transformer,"in Proc. Interspeech 2021,pp. 571–575,2021.

[9] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang y M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880-2894, 2020.

[10]] Organización Mundial de la Salud, *Environmental Noise Guidelines for the European Region*. Copenhagen, Denmark: WHO Regional Office for Europe, 2018, ISBN 978-92-890-5356-3.

[11] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861

[12]T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[13] S. Hershey et al., "CNN Architectures for Large-Scale Audio Classification," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131–135.

[14] Hugging Face, "Audio Course — Chapter 1: Working with Audio Data," *Hugging Face*, 2024. [En línea]. Disponible en: https://huggingface.co/learn/audio-course

[15] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition,"IEEE/ACM Transactions on Audio, Speech, and Language Processing,vol. 28, pp. 2880–2894, 2020.

[16]B. L. Krause,"The niche hypothesis: A virtual symphony of animal sounds, the origins of musical expression and the health of habitats,"Soundscape Newsletter,no. 6,pp. 6–10,1993.

[17] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner,"Gradient-based learning applied to document recognition,"Proceedings of the IEEE,vol. 86, no. 11,pp. 2278–2324,1998.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár,"Focal Loss for Dense Object Detection,"in Proc. IEEE Int. Conf. Computer Vision (ICCV),pp. 2980–2988,2017.

[19] J. Villalobos Tenorio, "Clasificador de sonidos que indiquen una alerta o amenaza para las personas con discapacidad auditiva," Trabajo de grado, Facultad de Ingeniería y Ciencias, Pontificia Universidad Javeriana, Cali, Colombia, 2024.

[20] Ministerio de Ambiente, Vivienda y Desarrollo Territorial,"Resolución 0627 de 2006, por la cual se establece la norma nacional de emisión de ruido y ruido ambiental,"Diario Oficial No. 46.231,Bogotá, Colombia,4 abr. 2006.

[21] A. Mesaros, T. Heittola and T. Virtanen,"TUT database for acoustic scene classification and sound event detection,"in Proc. 24th European Signal Processing Conference (EUSIPCO),pp. 1128–1132,2016

[22] M. Castro Duarte, "Aprendizaje automático aplicado al diagnóstico de la ocurrencia de la Leishmaniasis a través de imágenes de lesiones cutáneas," Trabajo de grado, Facultad de Ingeniería y Ciencias, Pontificia Universidad Javeriana, Cali, Colombia, 2024.

[23] D. Mejía Salgado, "Identificación de la actividad de la toxoplasmosis ocular mediante distintas redes neuronales convolucionales," Trabajo de grado, Facultad de Ingeniería y Ciencias, Pontificia Universidad Javeriana, Cali, Colombia, 2024.

[24] D. O'Shaughnessy,Speech Communication: Human and Machine.Reading, MA, USA: Addison-

Wesley, 1987.

[25] World Health Organization (WHO),Environmental Noise Guidelines for the European Region. Copenhagen, Denmark: WHO Regional Office for Europe, 2018.

[26] S. J. Pan and Q. Yang,"A survey on transfer learning,"IEEE Transactions on Knowledge and Data Engineering,vol. 22, no. 10,pp. 1345–1359,2010.

[27] V. Nair and G. E. Hinton,"Rectified linear units improve restricted Boltzmann machines,"in Proc. 27th International Conference on Machine Learning (ICML),pp. 807–814,2010.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel and É. Duchesnay,"Scikit-learn: Machine learning in Python,"Journal of Machine Learning Research,vol. 12,pp. 2825–2830,2011.

[29] K. J. Piczak,"Environmental sound classification with convolutional neural networks,"in Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP),pp. 1–6, 2015

[30] J. Salamon and J. P. Bello,"Deep convolutional neural networks and data augmentation for environmental sound classification,"IEEE Signal Processing Letters,vol. 24,no. 3,pp. 279–283, 2017.

[31] R. M. Schafer,The Tuning of the World.New York, NY, USA: Knopf, 1977.

[32] A. Mesaros, T. Heittola y T. Virtanen, "TUT database for acoustic scene classification and sound event detection", en Proc. European Signal Processing Conference (EUSIPCO), 2016.

[33] M. Sokolova and G. Lapalme,"A systematic analysis of performance measures for classification tasks,"Information Processing & Management,vol. 45,no. 4,pp. 427–437,2009.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov,"Dropout: A simple way to prevent neural networks from overfitting,"Journal of Machine Learning Research,vol. 15,pp. 1929–1958,2014.

[35] G. Gwardys y D. Jankowitz, "Deep image features in music information retrieval", International Journal of Electronics and Telecommunications (IJET), vol. 60, no. 4, pp. 321-326, 2014.

[36] M. Tan and Q. V. Le,"EfficientNet: Rethinking model scaling for convolutional neural networks,"in Proc. International Conference on Machine Learning (ICML),pp. 6105–6114,2019.

[37] M. Tan and Q. V. Le,"EfficientNetV2: Smaller models and faster training,"in Proc. International Conference on Machine Learning (ICML),pp. 10096–10106,2021.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez and I. Polosukhin,"Attention is all you need,"in Advances in Neural Information Processing Systems (NeurIPS),vol. 30,pp. 5998–6008,2017

[39] J. Yosinski, J. Clune, Y. Bengio and H. Lipson,"How transferable are features in deep neural networks,"in Advances in Neural Information Processing Systems (NeurIPS),vol. 27,pp. 3320–3328,2014.

[40]Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. Proceedings of the International Conference on Pattern Recognition (ICPR), 3121–3124.

[41] I. Goodfellow, Y. Bengio, y A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[42]S. Haykin, *Neural Networks and Learning Machines*, 3.ª ed. Upper Saddle River, NJ, USA: Pearson, 2009.

[43] MathWorks, "Convolutional Neural Network (CNN)," *MathWorks Documentation*, 2025. [En línea]. Disponible en: https://www.mathworks.com/discovery/convolutional-neural-network.html

[44] Agencia Europea de Medio Ambiente, "¿Percibe ruidos nocivos a su alrededor?" *Señales de la AEMA 2023*. Copenhagen, Denmark: European Environment Agency, 2023. [En línea]. Disponible en: https://www.eea.europa.eu

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He y P. Dollár, "Focal Loss for Dense Object Detection," en Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980-2988. doi: 10.1109/ICCV.2017.324