

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

Diseño y validación conceptual de un sistema de identificación acústica de vehículos aéreos y terrestres en entornos militares mediante el uso de aprendizaje automático

Andrade Pico Luis Clemente y Gómez Díaz Melissa Malayca

Departamento de Eléctrica, Electrónica y Telecomunicaciones

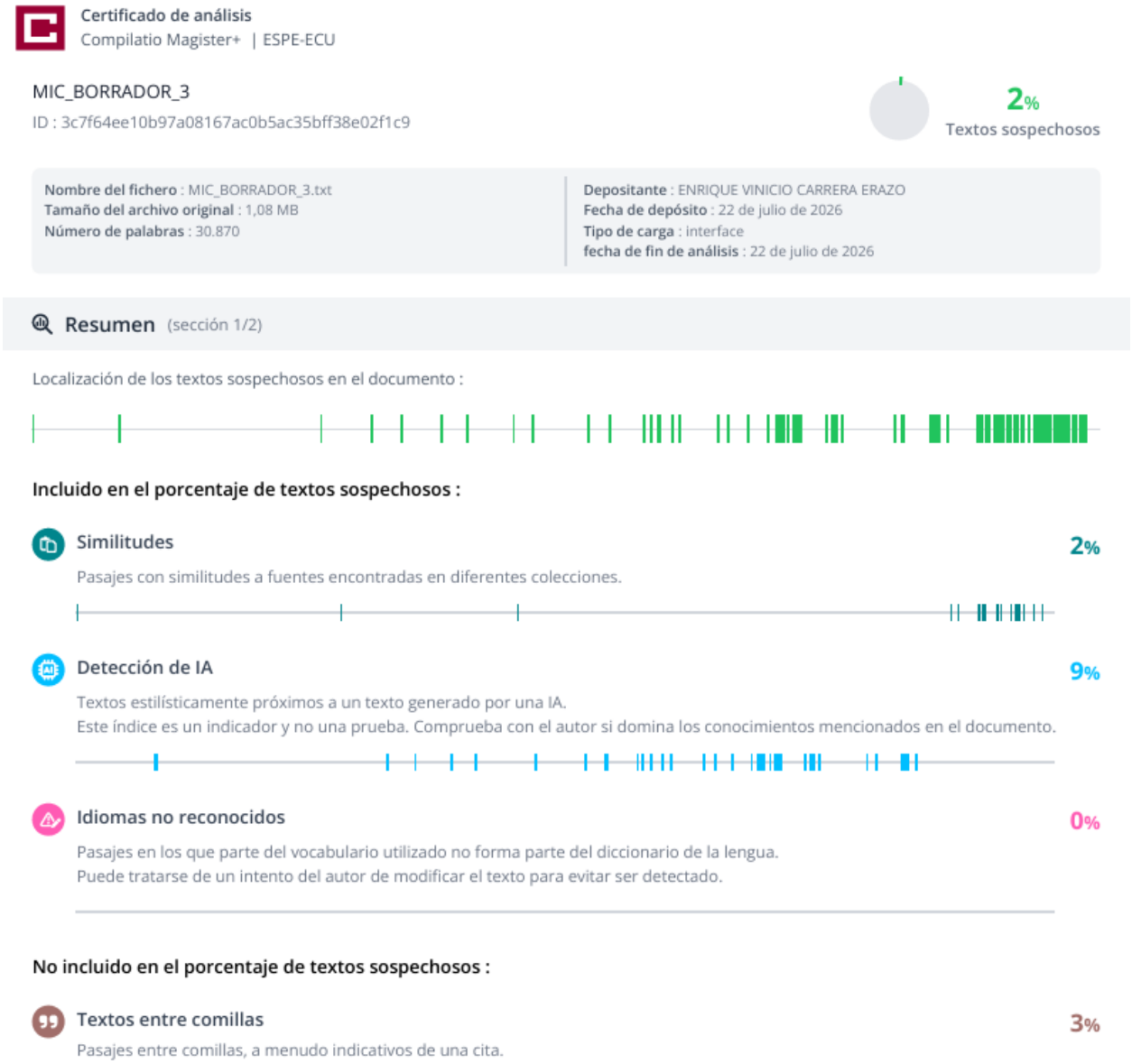
Carrera de Telecomunicaciones

Trabajo de integración curricular, previo a la obtención del título de

Ingeniero en Telecomunicaciones

Ing. Carrera Erazo Enrique Vinicio, PhD

22 de julio de 2026



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
ENRIQUE VINICIO CARRERA ERAZO

Ing. Carrera Erazo, Enrique Vinicio, PhD

Director

C.C.: 1708792104

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Certificación

Certifico que el trabajo de integración curricular: **"Diseño y validación conceptual de un sistema de identificación acústica de vehículos aéreos y terrestres en entornos militares mediante el uso de aprendizaje automático"** fue realizado por los señores Andrade Pico, Luis Clemente y Gómez Díaz, Melissa Malayca, el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE; habiéndose efectuado una revisión integral de forma y de fondo conforme a los lineamientos y formatos institucionales vigentes. Además, fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 22 de julio de 2026

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
ENRIQUE VINICIO CARRERA ERAZO

Ing. Carrera Erazo, Enrique Vinicio, PhD

Director

C.C.: 1708792104

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Responsabilidad de Autoría

Nosotros, Andrade Pico, Luis Clemente y Gómez Díaz, Melissa Malayca, con cédulas de ciudadanía n.º1712158557 y n.º1726856014, declaramos que el contenido, ideas y criterios del trabajo de integración curricular: **"Diseño y validación conceptual de un sistema de identificación acústica de vehículos aéreos y terrestres en entornos militares mediante el uso de aprendizaje automático"** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 22 de julio de 2026

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
LUIS CLEMENTE ANDRADE PICO

Andrade Pico, Luis Clemente

C.C.: 1712158557

Firmado electrónicamente por:
MELISSA MALAYCA GOMEZ DIAZ

Gómez Díaz, Melissa Malayca

C.C.:1726856014

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Autorización de Publicación

Nosotros, Andrade Pico, Luis Clemente y Gómez Díaz, Melissa Malayca, con cédulas de ciudadanía n.º1712158557 y n.º1726856014, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular: **"Diseño y validación conceptual de un sistema de identificación acústica de vehículos aéreos y terrestres en entornos militares mediante el uso de aprendizaje automático"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 22 de julio de 2026

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
LUIS CLEMENTE ANDRADE PICO

Andrade Pico, Luis Clemente

C.C.: 1712158557

Firmado electrónicamente por:
MELISSA MALAYCA GOMEZ DIAZ

Gómez Díaz, Melissa Malayca

C.C.: 1726856014

Agradecimientos

Agradezco en primer lugar a Dios por haberme brindado la oportunidad de formarme en esta casa de estudios tan prestigiosa. Expreso mi más profunda gratitud a mis padres Luis Agustín Andrade Ayala y Yannet Elizabeth Pico Armas por su apoyo incondicional, por permanecer siempre a mi lado a pesar de las dificultades, y a todas aquellas personas que, aun en la distancia y sin poder estar presentes físicamente, fueron el impulso y la motivación detrás de muchas de mis decisiones.

Luis Clemente Andrade Pico

Agradezco a mis padres, Segundo Gómez y Mercedes Díaz, por su apoyo incondicional, por su ayuda y motivación, y por cada consejo que me dieron a lo largo de mi vida; gracias por levantarme cada vez que tropecé en el camino y por recordarme con su ejemplo que siempre se puede seguir adelante. Ustedes son, sin duda, lo más importante en mi vida. A mis hermanos, Jessica y Juan, gracias por ser mis guías y mi compañía incondicional, por esos consejos que me ayudaron a crecer como hija, como estudiante y, sobre todo, como persona. A mis profesores, gracias por compartir con generosidad sus conocimientos y por ser un ejemplo a seguir durante toda mi carrera. Y, finalmente, gracias a cada persona que caminó junto a mí en este proceso de vida estudiantil, porque cada una me dejó un aprendizaje y momentos que llevaré siempre conmigo.

Melissa Malayca Gómez Díaz

Índice general

Certificación **3**

Responsabilidad de Autoría **4**

Autorización de Publicación **5**

Agradecimientos **6**

Lista de Tablas **11**

Lista de Figuras **12**

Resumen **13**

Abstract **14**

Capítulo I: Introducción **15**

Planteamiento del Problema 15

Objetivo 16

Objetivo General 16

Objetivos específicos 16

Justificación y Alcance 16

Capítulo II: Marco Teórico **18**

Estado del Arte 18

Marco Conceptual 19

Inteligencia artificial 19

Aprendizaje Automático 20

Aprendizaje supervisado 20

Clasificación 21

Clasificación multiclase 21

Clasificación jerárquica 22
Redes neuronales artificiales 22
Aprendizaje Profundo 23
Clasificación Acústica de Vehículos 24
Firma acústica vehicular 24
Relación entre configuración mecánica y firma acústica 24
Vehículos militares 24
Clasificación general de vehículos militares 25
Vehículos aéreos tripulados y no tripulados 26
Vehículos terrestres tripulados y no tripulados 27
Procesamiento Digital de Señales Acústicas 27
Señal continua, señal discreta y ventaneo 27
Señales estacionarias y no estacionarias 28
Salto 28
Raíz Media Cuadrática 29
Transformada de Fourier de Tiempo Corto 29
Espectrograma y espectrograma de Mel 30
Representaciones y Características Acústicas 30
Tasa de cruce por cero 31
Centroide espectral 31
Ancho de banda espectral 32
Rolloff espectral 32
Flujo espectral 33
Planitud espectral 33
PLP 34
Log-Mel 35
Filtros Mel 35
Coeficientes cepstrales y MFCC 36
Delta y Delta-Delta 36
Escalamiento de características 37

Estandarización y normalización 37
Desbalance de clases y aumento de datos en audio 38
Ganancia 39
Desplazamiento Circular 40
Inversión de polaridad 40
Estiramiento Temporal 41
SpecAugment 42
MixUp 42
Modelos de aprendizaje automático empleados 43
Fundamentos de la Máquina de Vectores de Soporte 44
Fundamentos de la Red Neuronal Convolucional 47
Métricas de evaluación 51
Entrenamiento, validación, prueba y generalización 51

Capítulo III: Metodología **53**
Obtención de datos 53
Organización de datos 55
Selección de los modelos a emplearse 56
Modelo de aprendizaje automático tradicional 58
Modelo de aprendizaje profundo 59
Preprocesamiento 60
Segmentación del audio 60
Aumento de datos y balanceo jerárquico 63
Flujo de procesamiento de los modelos 66
Flujo de procesamiento del modelo ML 66
Flujo de procesamiento del modelo DL 70

Capítulo IV: Resultados y Discusión **73**
Resultados del modelo de aprendizaje automático tradicional 73
Resultados del modelo de aprendizaje profundo 77
Análisis comparativo: ML frente a DL 82

Capítulo V: Conclusiones y Recomendaciones **85**

Conclusiones 85

Recomendaciones y trabajos futuros 86

Lista de acrónimos **97**

Índice de tablas

1. Medidas empleadas en la normalización de características. 37
2. Métricas de evaluación empleadas en los modelos. 52
3. Especificaciones técnicas de la grabadora empleadas. 54
4. Distribución de audios crudos por tipo de vehículo. 54
5. Arquitectura jerárquica común de clasificación (7 nodos). 55
6. Técnicas de aumento de datos empleadas en el entrenamiento SVM. 65
7. Comparación exploratoria de métricas globales entre configuraciones SVM. 73
8. Métricas generales por nodo – modelo SVM. 74
9. Matriz de confusión – N3 Aéreos/AUV: modelos de dron (SVM). 75
10. Matriz de confusión – N3 Aéreos/Helicópteros: modelos de helicóptero (SVM). . . . 76
11. Matriz de confusión – N3 Terrestres/Oruga: modelos de oruga (SVM). 76
12. Matriz de confusión – N3 Terrestres/Rueda: modelos de rueda (SVM). 77
13. Comparación de métricas globales entre arquitecturas CNN evaluadas. 78
14. Métricas generales por nodo – modelo CNN jerárquico seleccionado. 79
15. Comparación por nodo entre arquitecturas CNN jerárquicas evaluadas. 80
16. Comparación por nodo entre SVM y CNN jerárquico. 82
17. Resumen interpretativo de ventajas por modelo. 84

Índice de figuras

1. Flujograma general del proceso metodológico. 53
2. Árbol jerárquico de clasificación. 56

Resumen

En los escenarios contemporáneos de defensa y seguridad, la vigilancia pasiva basada en señales acústicas representa una alternativa complementaria a sistemas tradicionales como radares, sensores ópticos y sistemas de radiofrecuencia. A diferencia de los sistemas activos o dependientes de condiciones visuales, el análisis acústico permite aprovechar las emisiones sonoras producidas naturalmente por los vehículos durante su operación. Cada plataforma aérea o terrestre genera una firma acústica muy particular asociada a su sistema de propulsión, componentes mecánicos, régimen de funcionamiento y dinámica de desplazamiento, lo que permite plantear su identificación mediante técnicas de procesamiento digital de señales y aprendizaje automático.

El presente trabajo propone el diseño conceptual y validación de un sistema de identificación basado en audio para la clasificación automática de vehículos aéreos, como drones multirrotor y helicópteros, y terrestres, como blindados a oruga y rueda, en entornos militares con alta presencia de ruido e interferencias, como base para un sistema de vigilancia pasivo. El proyecto se desarrolla mediante el empleo de dos modelos un modelo de Aprendizaje de Maquina Tradicional (ML) una línea clásica pero robusta y supervisada que permite trabajar con un conjunto de datos relativamente limitado, y otro de Aprendizaje Profundo (DL), propicios para el aprendizaje de patrones en los que se puede realizar representaciones de audio.

Palabras clave: Identificación Acústica, Vehículos Militares, Aprendizaje de Maquina, Aprendizaje Profundo, Vigilancia Pasiva.

Abstract

In contemporary defense and security scenarios, passive surveillance based on acoustic signals represents a complementary alternative to traditional systems such as radar, optical sensors, and radio-frequency systems. Unlike active systems or those dependent on visual conditions, acoustic analysis makes it possible to exploit the sound emissions naturally produced by vehicles during their operation. Each aerial or ground platform generates a highly distinctive acoustic signature associated with its propulsion system, mechanical components, operating regime, and movement dynamics, which makes it possible to approach its identification through digital signal processing and machine learning techniques.

This work proposes the conceptual design and validation of an audio-based identification system for the automatic classification of aerial vehicles, such as multi-rotor drones and helicopters, and ground vehicles, such as tracked and wheeled armored vehicles, in military environments with a high presence of noise and interference, as a basis for a passive surveillance system. The project is developed through the use of two models: a Traditional Machine Learning (ML) model, a classical yet robust and supervised approach suited to working with a relatively limited dataset, and a Deep Learning (DL) model, well suited to learning patterns from representations that can be derived from audio.

Keywords: Acoustic Identification, Military Vehicles, Machine Learning, Deep Learning, Passive Surveillance.

Capítulo I: Introducción

Planteamiento del Problema

La guerra actual y escenario beligerantes, incluso los que no se enfocan en conflicto armado, presentan una creciente complejidad operacional debido al uso intensivo de vehículos aéreos no tripulados, helicópteros ligeros y diversos vehículos terrestres utilizados en la región. En este contexto, los sistemas tradicionales de vigilancia basados en sensores ópticos, infrarrojos, radar o radiofrecuencia pueden verse limitados por costos, condiciones ambientales, detectabilidad o restricciones frente a plataformas con bajas emisiones. La vigilancia pasiva basada en señales acústicas constituye una alternativa complementaria de bajo costo y alta prospección y escalabilidad [1].

Los vehículos de similares características pueden diferenciarse unos de otros, debido al cambio de patrones espectrales causados por cambios particulares como la carga útil, abriendo una ventana de diferenciación [2]; por tanto, ya que cada vehículo genera una firma sonora particular relacionada con su sistema de propulsión, motores, rotores, sistemas de tracción, vibraciones, rodadura, componentes mecánicos y dinámica de operación; este puede ser aislado y clasificado [3], [4], [5].

El problema central consiste en diseñar y validar conceptualmente un sistema capaz de identificar vehículos aéreos y terrestres en entornos militares, es decir, con presencia de ruido, variabilidad operacional y similitud acústica entre algunas plataformas y condiciones reales en las que fueron recopiladas las muestras para entrenar los clasificadores; en este contexto persiste la necesidad de evaluar las firmas acústicas disponibles para optimizar la obtención de información suficiente para distinguir clases jerárquicas como aéreos y terrestres, AUV y helicópteros, oruga y rueda, así como modelos particulares dentro de cada grupo.

Objetivo

Objetivo General

Diseñar y validar conceptualmente un sistema para la identificación automática de vehículos aéreos y terrestres mediante el análisis de sus señales acústicas, con aplicación directa en sistemas de vigilancia y seguridad en entornos militares.

Objetivos específicos

1. Evaluar si el empleo de una arquitectura jerárquica para clasificación mejora la identificación de los vehículos en distintos niveles de decisión, desde categorías generales hasta modelos específicos.
2. Emplear arquitecturas de aprendizaje automático tradicional y de aprendizaje profundo para la clasificación acústica de vehículos aéreos y terrestres.
3. Evaluar y comparar los modelos empleados mediante métricas de clasificación, con el fin de determinar cuál ofrece el desempeño más óptimo en cada nivel jerárquico.
4. Validar la factibilidad conceptual del sistema propuesto de identificación acústica.

Justificación y Alcance

Después del Conflicto Armado Internacional que sufrió el Ecuador en 1995 donde las capacidades militares en guerra convencional tuvieron su máxima proyección y financiamiento, y debido a una delimitación mucho más formal con los vecinos de nuestras fronteras.

Hasta finales de la década pasada la vigilancia y monitoreo fronterizo ha sido relegada a un papel más que secundario, esto aunado a las nuevas amenazas de conflictos que caen en zonas grises e híbridas han volcado las miradas y los esfuerzos al interior del país, debilitando aún más la frontera y paradójicamente, fortaleciendo las líneas logísticas de estas nuevas amenazas [6], [7], [8].

Debido a las exiguas capacidades económicas del Estado la inversión y recursos en seguridad y defensa se manejan de manera prioritaria, así que es menester, en primer lugar, for-

talecer el control efectivo del estado en las fronteras a fin de contribuir con la misión principal de las Fuerzas Armadas y a su vez coadyuvar con su nueva misión, desarrollando un modelo, ligero, asequible y escalable de identificación basado en aprendizaje automático orientado a los sistemas acústicos siendo estos los más asequibles a fin de poder recuperar las capacidades nacionales en vigilancia pasiva [9], [10].

El presente proyecto se enfoca en diseñar sistemas de identificación basados en dos modelos ampliamente conocidos y validados en tareas de identificación acústica: uno de aprendizaje automático tradicional (ML) y otro de aprendizaje profundo (DL) [11], [12]. No se busca una implementación física ni su integración en una aplicación real, sino validar su correcto funcionamiento y corroborar, mediante métricas de evaluación, la factibilidad de una implementación posterior, considerando las limitaciones tanto en cantidad como en calidad de las muestras obtenidas en entornos reales de operación.

Capítulo II: Marco Teórico

Estado del Arte

La clasificación automática de señales acústicas se ha desarrollado a partir del análisis de características temporales, espectrales y cepstrales, como la energía, el cruce por cero, el centroide espectral y los coeficientes MFCC. Estas características han sido utilizadas como entrada para algoritmos tradicionales de aprendizaje automático, entre ellos las máquinas de vectores de soporte, especialmente en escenarios con conjuntos de datos limitados [3], [4], [11].

Posteriormente, las redes neuronales convolucionales ampliaron este enfoque mediante el procesamiento directo de espectrogramas y representaciones Log-Mel, permitiendo aprender patrones espectro-temporales sin depender exclusivamente de características seleccionadas manualmente. Esta transición se consolidó en tareas de clasificación de sonidos ambientales y de audio a gran escala, en las que las CNN alcanzaron un desempeño competitivo frente a los enfoques clásicos basados en características diseñadas a mano [13], [14], [15], [16], [17].

En el ámbito específico de la clasificación acústica de vehículos, diversos estudios han caracterizado el ruido vehicular como el resultado de la superposición de múltiples fuentes mecánicas y aerodinámicas —motor, transmisión, escape, rodadura— cuya combinación produce firmas espectrales distinguibles entre plataformas [18], [19]. A partir de esta caracterización, se han propuesto tanto descriptores espectrales manuales como arquitecturas convolucionales aplicadas a espectrogramas para diferenciar vehículos y motores, incluyendo casos con modificaciones mecánicas que alteran su firma sonora [20], [21], [22], [23].

Dentro del contexto militar en particular, la identificación acústica de vehículos terrestres mediante sensores distribuidos en el terreno ha demostrado ser una vía viable para la vigilancia pasiva, al permitir reconocer patrones de marcha y de propulsión característicos de plataformas blindadas y mecanizadas incluso bajo condiciones de campo [24]. Este antecedente resulta especialmente relevante para el presente trabajo, ya que constituye una de las pocas referencias que aborda directamente la identificación acústica de vehículos militares terrestres, a diferencia de la mayoría de estudios centrados en entornos civiles o controlados.

En paralelo, la identificación acústica de vehículos aéreos no tripulados ha recibido un impulso considerable a partir del uso de redes convolucionales sobre representaciones espectrales de corto tiempo, empleadas tanto para la detección de la presencia de un dron como para la clasificación de su modelo específico, incluso en condiciones de baja relación señal-ruido [2], [25], [26], [27]. Estos trabajos han evidenciado además la utilidad de combinar información acústica con otras modalidades sensoriales para mejorar la robustez de la detección en aplicaciones de seguridad [28].

No obstante, la revisión de estos antecedentes evidencia que la mayoría de los estudios abordan de forma independiente la clasificación acústica de vehículos terrestres o de plataformas aéreas no tripuladas, evaluados generalmente bajo condiciones controladas y rara vez comparando modelos de aprendizaje automático tradicional y de aprendizaje profundo sobre un mismo conjunto de datos y esquema de preprocesamiento. En este contexto, el presente trabajo aborda la identificación jerárquica de vehículos aéreos y terrestres mediante un modelo de aprendizaje automático tradicional y uno de aprendizaje profundo, evaluados de forma comparable bajo condiciones cercanas a un entorno real de operación. La organización jerárquica permite dividir el problema general en decisiones progresivas y especializar cada nodo en grupos acústicamente relacionados [29].

Marco Conceptual

Los conceptos que sustentan la clasificación de audios se fundamentan tanto en las ramas del aprendizaje automático como en el análisis acústico. Para comprender este proceso, es necesario considerar fundamentos teóricos básicos como la inteligencia artificial, el aprendizaje automático, las señales acústicas, la extracción de características, la normalización de datos, entre otros aspectos que sirven de base para el desarrollo de sistemas de clasificación de audio.

Inteligencia artificial

La inteligencia artificial es el campo de estudio dedicado al diseño y desarrollo de sistemas computacionales capaces de percibir su entorno, representar y procesar información, razonar de manera limitada, aprender a partir de la experiencia y actuar con cierto grado de au-

tonomía para resolver problemas que, tradicionalmente, han requerido de la intervención humana [30]. Esta definición admite al menos dos perspectivas complementarias: por un lado, la que concibe a la inteligencia artificial como el estudio de los sistemas que actúan racionalmente, es decir, que seleccionan la acción más apropiada en función de la información disponible y del objetivo que persiguen; y por otro, la que la entiende como la construcción de entidades capaces de emular procesos cognitivos propios del razonamiento humano [30]. En ambos casos, el propósito es dotar a un sistema artificial de la capacidad de tomar decisiones frente a información incompleta, ruidosa o variable, condición particularmente relevante en el análisis de señales acústicas capturadas en entornos reales.

Aprendizaje Automático

El aprendizaje automático, o *Machine Learning*, es una subárea de la inteligencia artificial en la que los sistemas no se programan únicamente mediante reglas explícitas, sino que ajustan su comportamiento a partir de datos. Este enfoque permite a los computadores aprender automáticamente desde datos y generar algoritmos capaces de realizar una tarea sobre datos del mismo tipo [11]. En este sentido, el aprendizaje no consiste en memorizar ejemplos aislados, sino en construir una función de decisión que permita generalizar hacia nuevas observaciones.

El aprendizaje automático también puede entenderse como el proceso mediante el cual los computadores adquieren la capacidad de aprender desde los datos, especialmente en tareas predictivas como la clasificación [31]. En problemas supervisados, el modelo recibe ejemplos previamente etiquetados y aprende una relación entre las características de entrada y la clase esperada.

Aprendizaje supervisado

El aprendizaje supervisado es una modalidad del aprendizaje automático en la que el modelo se entrena a partir de ejemplos cuya salida esperada es conocida. En este enfoque, cada muestra del conjunto de entrenamiento está formada por una entrada y una etiqueta o valor objetivo, de modo que el algoritmo puede ajustar sus parámetros comparando la predicción generada con la respuesta correcta. El entrenamiento supervisado requiere ejemplos de entrada

asociados a sus salidas esperadas, lo que permite modificar los pesos o parámetros del modelo en función del error observado durante el proceso de aprendizaje [32].

Clasificación

La clasificación es una tarea de análisis en la que se busca asignar una observación a una de varias clases previamente definidas. En este tipo de problemas, se dispone de un conjunto de soluciones posibles o clases, y el objetivo consiste en determinar a cuál de ellas pertenece un conjunto de observables [33]. Desde esta perspectiva, la clasificación puede entenderse como un proceso de decisión que transforma datos medibles en una categoría interpretable.

En el aprendizaje supervisado, el objetivo del clasificador es aprender una frontera de decisión o una regla que permita separar las clases en función de sus características, para la clasificación acústica de vehículos, nos atañe rasgos acústicos como los componentes cepstrales, energía, distribución espectral o variaciones temporales con el fin de poder determinar la clases representada por las categorías como por ejemplo, vehículo aéreo, vehículo terrestre, subgrupo mecánico o modelos específicos.

Clasificación multiclase

La clasificación multiclase es una extensión de la tarea general de clasificación, en la cual una observación debe asignarse a una categoría dentro de un conjunto finito de más de dos clases posibles. En las tareas de análisis, dado un conjunto de observables, se busca determinar a cuál de las clases disponibles pertenece dicho conjunto [33]. A partir de esta idea, la clasificación multiclase amplía el problema de decisión, ya que el modelo no debe elegir entre dos alternativas, sino discriminar entre varias categorías previamente definidas.

En clasificación multiclase, esas salidas esperadas corresponden a etiquetas discretas pertenecientes a un conjunto de clases. Formalmente, si existen K clases posibles, $\hat{y}$ representa la clase predicha, x es el vector de características de entrada, r es el índice de la clase evaluada, K es el número total de clases y $f_r(x)$ representa la puntuación, compatibilidad o función de decisión asociada a la clase r. La clase seleccionada es aquella que obtiene la mayor puntuación para la muestra analizada:

$$\hat{y} = \arg_r \max f_r(x), \qquad r = 1, \ldots, K \tag{1}$$

Clasificación jerárquica

Cuando las clases de un problema de clasificación presentan una taxonomía natural, la tarea puede organizarse de forma jerárquica en lugar de resolverse mediante una única clasificación plana. [29] La clasificación jerárquica resulta ventajosa cuando las categorías se relacionan mediante estructuras de especialización, ya que permite descomponer una decisión compleja en varias decisiones progresivas más simples, cada una restringida a un subconjunto reducido de clases. En la identificación acústica de vehículos, este enfoque permite distinguir primero familias amplias, como vehículos aéreos y terrestres, para posteriormente diferenciar subfamilias, por ejemplo drones frente a helicópteros o vehículos de oruga frente a vehículos de rueda, y finalmente modelos específicos dentro de cada subfamilia.

Redes neuronales artificiales

Una red neuronal artificial es un modelo computacional compuesto por unidades de procesamiento simples, denominadas neuronas artificiales, organizadas en capas y conectadas mediante parámetros ajustables. Cada neurona calcula una suma ponderada de sus entradas, añade un término de sesgo y aplica una función de activación no lineal para producir su salida [34]. Cuando estas unidades se organizan en una capa de entrada, una o varias capas ocultas y una capa de salida, la red puede aproximar relaciones complejas entre los datos de entrada y la categoría o valor esperado [31].

El ajuste de los parámetros de una red neuronal, es decir, sus pesos y sesgos, se realiza mediante un proceso iterativo de optimización basado en el gradiente del error: la red compara su salida con la respuesta esperada y actualiza sus parámetros para reducir dicho error [34]. Las redes neuronales convolucionales empleadas en la clasificación acústica de este trabajo constituyen una arquitectura especializada de red neuronal artificial, en la cual las conexiones entre capas no son completamente densas, sino que se organizan mediante filtros locales compartidos. Esta característica las hace particularmente adecuadas para datos con estructura espacial o tiempo-frecuencial, como los espectrogramas de audio [17], [34].

Aprendizaje Profundo

El aprendizaje profundo, o *Deep Learning*, es una rama del aprendizaje automático que emplea modelos compuestos por múltiples capas de procesamiento no lineal para aprender representaciones jerárquicas de los datos, de manera que las capas más cercanas a la entrada capturan patrones simples y las capas más profundas combinan dichos patrones en representaciones progresivamente más abstractas [35]. A diferencia de los enfoques tradicionales de aprendizaje automático, en los que gran parte del desempeño depende de características diseñadas manualmente por un experto del dominio, los modelos profundos aprenden de forma conjunta tanto la extracción de características como la función de clasificación, optimizando ambas etapas mediante el mismo procedimiento de entrenamiento basado en el gradiente del error [34], [35]. Esta capacidad de aprender transformaciones progresivas desde los datos de entrada hasta representaciones útiles para clasificación constituye la principal diferencia respecto a los métodos clásicos revisados en la sección anterior.

En el caso particular de señales acústicas, esta propiedad permite que una red aprenda directamente patrones espectro-temporales a partir de representaciones como espectrogramas o matrices log-Mel, sin depender exclusivamente de descriptores seleccionados manualmente, como el centroide espectral o los coeficientes cepstrales [17]. Esta ventaja resulta especialmente relevante cuando las diferencias entre clases acústicas no son evidentes a partir de un conjunto reducido de estadísticos, sino que emergen de patrones más complejos distribuidos en el plano tiempo-frecuencia.

Dentro del aprendizaje profundo, las redes neuronales convolucionales son especialmente relevantes cuando los datos presentan estructura local, es decir, cuando existe correlación entre elementos cercanos de la entrada [36]. En audio, dicha estructura local aparece al representar la señal en el plano tiempo-frecuencia, donde regiones cercanas en tiempo y frecuencia suelen contener patrones acústicos relacionados, como armónicos, transitorios o bandas de energía asociadas a un mismo evento sonoro [13]. Por esta razón, las CNN se han utilizado ampliamente en clasificación de sonidos ambientales y en clasificación de audio a gran escala, empleando espectrogramas o representaciones log-Mel como entrada del modelo [14], [15], [16].

Clasificación Acústica de Vehículos

Firma acústica vehicular

La firma acústica vehicular se define como el conjunto de patrones sonoros característicos de un vehículo, determinados por la energía, el contenido frecuencial, la distribución espectral y la variación temporal de la señal acústica que este produce durante su operación [20], [37]. Dichos patrones resultan de la interacción de múltiples fuentes mecánicas y aerodinámicas, como el motor, el sistema de escape, los neumáticos, las orugas, las hélices o los rotores, de modo que el sonido emitido por un vehículo no corresponde a una componente única, sino a la superposición de varias fuentes acústicas cuya contribución relativa varía según la configuración física del vehículo y su condición de funcionamiento [18], [19].

Relación entre configuración mecánica y firma acústica

La firma acústica de un vehículo militar depende de su configuración mecánica, su medio de operación y su régimen dinámico. En plataformas terrestres, los cambios de velocidad, aceleración, tipo de terreno y carga mecánica pueden modificar la energía, periodicidad y distribución espectral del sonido. En plataformas aéreas, las variaciones de régimen del motor, el número de rotores, el tamaño de las hélices y la sustentación influyen en la frecuencia y estabilidad de los componentes acústicos, lo que lo hace mayoritariamente que los subsistemas que producen patrones diferenciables que pueden representarse mediante descriptores temporales, espectrales y cepstrales muy distinguible unos de otros [38], [39], [40].

Vehículos militares

Un vehículo militar puede entenderse como una plataforma móvil diseñada, adaptada o empleada para cumplir funciones operacionales dentro del ámbito de la defensa, tales como transporte, reconocimiento, vigilancia, protección, apoyo de combate, movilidad táctica o empleo de sistemas especializados. Esta categoría puede incluir plataformas terrestres, aéreas o marítimas, tanto tripuladas como no tripuladas, siempre que su operación esté orientada al cumplimiento de una función militar o de seguridad. En el caso de los sistemas no tripulados, la doc-

trina técnica los considera como sistemas integrados, compuestos no solo por el vehículo, sino también por estaciones de control, enlaces de comunicación, equipos de apoyo y subsistemas asociados a la operación [41], [42], [43], [44].

Desde el punto de vista técnico, un vehículo militar no se limita a su estructura externa, sino que integra subsistemas mecánicos, energéticos, electrónicos y de control. En plataformas terrestres intervienen elementos como motor, transmisión, tren de rodadura, suspensión, ejes, frenos y sistema de escape; en plataformas aéreas participan motores, rotores, hélices, superficies de sustentación y sistemas de control de vuelo; mientras que en plataformas marítimas intervienen casco, planta propulsora, hélices, sistemas de gobierno, sensores y equipos de navegación. En el caso de vehículos terrestres blindados y mecanizados, los manuales técnicos describen precisamente estos componentes como parte de la estructura operacional y de mantenimiento del material militar [38], [39], [40].

Acústicamente un vehículo militar se entiende como una fuente sonora compleja, Su firma puede originarse en la combustión del motor, el sistema de escape, la transmisión, el contacto rueda-suelo u oruga-suelo, las vibraciones estructurales, el giro de rotores o hélices, la turbulencia aerodinámica, la cavitación o la interacción del sistema propulsor con el medio, por tanto, también podrían ser clasificados fundamentándose en que la arquitectura mecánica y operacional de cada plataforma genera patrones temporales, espectrales y cepstrales diferenciables, aunque estos puedan verse afectados por ruido ambiental, distancia, terreno, velocidad, régimen de operación o interferencias [5], [45].

Clasificación general de vehículos militares

Los diversos vehículos militares pueden clasificarse de dos maneras, sea por el medio en el que operan o por la presencia o ausencia de tripulación a bordo, estas categorías no son excluyentes pudiendo ser ambas a la vez. En el enfoque según el medio de operación pueden ser vehículos terrestres, aéreos y marítimos; si consideramos el enfoque según su forma de operación, estos pueden ser tripulados, cuando requieren personal a bordo para su conducción o control directo, o no tripulados, cuando operan mediante control remoto, autonomía parcial o autonomía completa manejada por algoritmos.

Los sistemas no tripulados no se reducen únicamente al vehículo físico. En el caso aé-

reo, la doctrina técnica y de interoperabilidad considera que un sistema aéreo no tripulado comprende la aeronave, la estación de control, los enlaces de mando y control, los equipos de comunicaciones y los elementos de apoyo necesarios para la operación [41], [42]. De forma análoga, en el ámbito marítimo, los planes técnicos de la Armada de los Estados Unidos, por ejemplo se distinguen vehículos de superficie no tripulados y vehículos submarinos no tripulados como sistemas destinados a misiones como vigilancia, reconocimiento, contramedidas de minas, inspección, identificación y apoyo a operaciones navales [43], [44]. Para la consecución del presente proyecto no se consideran vehículos marítimos de ninguna clase ni tripulados ni no tripulados.

Vehículos aéreos tripulados y no tripulados

Los vehículos aéreos militares tripulados corresponden a plataformas que transportan personal a bordo para su operación, como helicópteros, aviones de transporte, aeronaves de entrenamiento o aeronaves de combate, de los cuales las Fuerzas Armadas poseen de muy variadas características y capacidades. En el contexto acústico, los helicópteros son particularmente relevantes porque su firma sonora está asociada al rotor principal, rotor de cola, motor, transmisión, flujo aerodinámico y variaciones de régimen durante maniobras de vuelo. Estos elementos producen componentes periódicos y patrones espectro-temporales que pueden diferenciarse de otras fuentes sonoras.

Los vehículos aéreos no tripulados, por su parte, corresponden a plataformas aéreas que no llevan operador humano a bordo y que pueden ser controladas de forma remota o ejecutar funciones con distintos niveles de autonomía. En términos de sistema es un conjunto integrado por vehículo aéreo, estación de control, enlaces de comunicación, navegación, sensores y equipos de apoyo [41], [42]. Desde el punto de vista acústico, los drones multirrotor generan firmas asociadas al giro de múltiples hélices, motores eléctricos, controladores de velocidad, vibración estructural y variaciones rápidas de régimen, bastante distinto al de uno tripulado. Las Fuerzas Armadas del Ecuador poseen también algunas plataformas aéreas no tripuladas muy variadas tanto de uso comercial como militar más orientada a tareas de vigilancia e inteligencia.

Vehículos terrestres tripulados y no tripulados

Los vehículos terrestres militares tripulados comprenden plataformas sobre rueda u oruga empleadas para transporte, movilidad táctica, reconocimiento, apoyo de combate o protección. En esta categoría se incluyen blindados, vehículos mecanizados, transportes tácticos y vehículos de apoyo. Su firma acústica puede estar asociada al motor, sistema de escape, transmisión, suspensión, tren de rodadura, contacto con el terreno, vibraciones del chasis y cambios de régimen durante la marcha .Los manuales técnicos de este material lo clasifican como blindado, mecanizado y táctico y describen estos subsistemas como elementos operacionales y de mantenimiento de vehículos militares [38], [39], [40].

Los vehículos terrestres no tripulados, conocidos como UGV por sus siglas en inglés *Unmanned Ground Vehicle*, corresponden a plataformas terrestres operadas de forma remota o autónoma para tareas como reconocimiento, inspección, transporte de carga, desactivación de explosivos o apoyo a unidades tácticas. Aunque comparten con los vehículos tripulados algunos mecanismos de generación sonora, como motores, ruedas, orugas o actuadores, pueden incorporar además sistemas eléctricos, servomecanismos y patrones de desplazamiento distintos. En términos acústicos, su firma puede diferir por tamaño, tipo de propulsión, peso, velocidad y forma de contacto con el terreno. Ya que el Ecuador y sus países vecinos presentan capacidades similares y unicamente emplea AGV para tareas de desminado y sus unidades son muy limitadas no se ha considerado para el presente trabajo.

Procesamiento Digital de Señales Acústicas

Señal continua, señal discreta y ventaneo

Una señal acústica es una variación de presión en el aire que evoluciona en el tiempo y que, en su forma física original, puede representarse como una señal continua. Sin embargo, para ser procesada mediante un sistema computacional, dicha señal debe convertirse en una secuencia discreta de muestras. Esta transformación se realiza mediante el muestreo, proceso por el cual se toman valores de la señal continua en instantes separados por un intervalo temporal constante, donde se asume una estabilidad local suficiente. El análisis computacional de audio requiere transformar la información perceptual o física en representaciones numéricas

manipulables [11]. La amplitud de la señal en cada instante representa la magnitud de dicha variación de presión respecto al reposo; su evolución a lo largo del tiempo constituye el dominio temporal de la señal, es decir, la forma de onda tal como es capturada por el micrófono [45].

En el dominio continuo se representa como $x(t)$. Para procesarla con un computador debe convertirse en una sucesión de muestras, proceso que da lugar a la señal discreta $x[n]$. El periodo de muestreo T_s indica la separación entre dos muestras consecutivas y la frecuencia de muestreo f_s expresa cuántas muestras se toman por segundo.

$$x[n] = x(nT_s) \tag{2}$$

Señales estacionarias y no estacionarias

Una señal estacionaria es aquella cuyas propiedades estadísticas permanecen constantes en el tiempo, mientras que una señal no estacionaria presenta variaciones temporales en su energía, frecuencia o distribución espectral [45], [46]. El sonido vehicular puede modelarse como una señal no estacionaria, dado que sus características acústicas cambian en función del régimen del motor, la velocidad, la distancia al sensor, la aceleración, la desaceleración y las condiciones del entorno [20], [37]. Debido a esta variación temporal, el análisis global de frecuencia resulta insuficiente para describir completamente la señal, razón por la cual se emplean representaciones tiempo-frecuencia que preservan información local sobre la evolución del sonido en el tiempo [46], [47].

Salto

El salto, también denominado *hop*, es el desplazamiento temporal entre el inicio de una ventana de análisis y el de la siguiente [46]. Cuando una señal se analiza mediante ventanas sucesivas, el salto H determina el número de muestras que se avanza entre dos segmentos consecutivos, controlando así el grado de solapamiento entre ventanas y la densidad temporal de la representación resultante [46], [47]. Para una ventana de longitud N, existe solapamiento entre ventanas cuando $H < N$, mientras que si $H = N$ las ventanas son contiguas y no se superponen [45], [46].

$$\text{Solapamiento} = N - H \tag{3}$$

Raíz Media Cuadrática

El RMS, del inglés *Root Mean Square*, puede traducirse como raíz media cuadrática. En análisis de audio, este descriptor permite estimar la amplitud efectiva o energía promedio de una señal dentro de una ventana temporal. A diferencia de una amplitud instantánea, el RMS resume la intensidad de un tramo completo de la señal, por lo que resulta útil para identificar variaciones de energía asociadas con la actividad sonora, incluyendo análisis y clasificación [5].

Matemáticamente, para una ventana de N muestras, donde $x[n]$ representa la muestra de la señal discreta en el instante n, el RMS se expresa como:

$$RMS = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x[n]^2} \tag{4}$$

Transformada de Fourier de Tiempo Corto

La Transformada de Fourier de Tiempo Corto, conocida como STFT por sus siglas en inglés (*Short-Time Fourier Transform*), permite analizar una señal mediante la aplicación de una ventana temporal y el cálculo de la Transformada de Fourier sobre cada segmento resultante [48]. A diferencia de la Transformada de Fourier aplicada sobre una señal completa, la STFT permite observar la evolución del contenido frecuencial a lo largo del tiempo, ya que calcula espectros locales en ventanas sucesivas de la señal.

En el análisis de señales acústicas y de voz, muchas señales no pueden considerarse estacionarias durante todo el registro, pero sí pueden aproximarse como localmente estacionarias en intervalos cortos [49]. Lo que corrobora el uso de ventanas temporales para analizar señales cuyo contenido cambia con el tiempo. En el caso de sonidos vehiculares, esta condición es relevante porque motores, rotores, vibraciones, rodadura e interferencias pueden variar durante la grabación.

Podemos representar $x[n]$ como la señal discreta, $w[n]$ como la función ventana, m como el índice temporal de la ventana, H como el salto entre ventanas consecutivas, k como el índice

frecuencial y N como el número de puntos usados en la transformada. El resultado $X(m,k)$ indica la contribución espectral de la señal en una ventana temporal específica y en una frecuencia determinada.

$$X(m,k) = \sum_{n=0}^{N-1} x[n+mH]w[n]e^{-j2\pi kn/N} \tag{5}$$

Espectrograma y espectrograma de Mel

El espectrograma es una representación tiempo-frecuencia derivada de la magnitud o la potencia de la STFT, en la cual un eje representa el tiempo, otro eje representa la frecuencia y cada valor indica la energía de la señal en una región específica del plano tiempo-frecuencia [46], [47]. De forma común, el espectrograma de potencia se define como [46]:

$$S(m,k) = |X(m,k)|^2 \tag{6}$$

Por su parte, el espectrograma de Mel transforma el eje de frecuencia lineal en una escala perceptual denominada escala Mel, la cual aproxima la sensibilidad del oído humano ante diferencias de frecuencia [50], [51]. La conversión de frecuencia en Hertz a escala Mel suele expresarse como [51]:

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

Representaciones y Características Acústicas

En los enfoques tradicionales de aprendizaje automático aplicados a señales acústicas, el modelo no aprende directamente desde la forma de onda completa, sino a partir de una representación numérica construida previamente mediante extracción de características. Esta etapa convierte cada segmento de audio en un vector de descriptores que resume propiedades temporales, espectrales, energéticas o cepstrales de la señal. Posteriormente, dicho vector se utiliza como entrada para un clasificador, por ejemplo, una Máquina de Vectores de Soporte, un árbol de decisión, un clasificador bayesiano o un método de vecinos más cercanos. Este enfo-

que ha sido ampliamente utilizado en clasificación de audio antes del predominio de los modelos de aprendizaje profundo [3].

En clasificación general de audio se han empleado características como energía de corto tiempo, tasa de cruces por cero, frecuencia fundamental, ancho de banda, centroide espectral, rolloff espectral y coeficientes cepstrales como base para distinguir diferentes categorías de audio [3]. De forma similar, se ha propuesto la clasificación automática en trabajos similares como en géneros musicales, mediante características de textura tímbrica, incluyendo centroide espectral, rolloff, flujo espectral, cruces por cero, energía y coeficientes MFCC [4].

Tasa de cruce por cero

El término *Zero Crossing Rate* se traduce como tasa de cruces por cero. Este descriptor mide la frecuencia con la que la señal cambia de signo dentro de una ventana de análisis. En rasgos de audio de corto plazo, se presenta como la tasa de cambios de signo de la señal durante una trama determinada [5]. Considerando que $x[n]$ es la señal discreta y N el número de muestras de la ventana, se expresa como:

$$ZCR = \frac{1}{2N} \sum_{n=1}^{N-1} |\mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n-1])| \tag{8}$$

Centroide espectral

El centroide espectral describe el centro de gravedad del espectro de una señal. En el análisis de audio, este descriptor permite ubicar la posición promedio de la energía espectral y se interpreta como una medida de forma espectral [4]. De manera complementaria, dentro del conjunto de características espectrales de corto plazo, también se presenta como el centro de gravedad del espectro [5]. Matemáticamente, se expresa como:

$$C = \frac{\sum_{k=1}^{K} f_k S_k}{\sum_{k=1}^{K} S_k} \tag{9}$$

Donde C representa el centroide espectral, es decir, la frecuencia alrededor de la cual se concentra el centro de gravedad del espectro. El índice k identifica cada componente frecuencial,

mientras que K corresponde al número total de componentes espectrales consideradas. La variable f_k representa la frecuencia asociada a la componente k y S_k indica su magnitud o energía espectral. El producto $f_k S_k$ pondera cada frecuencia de acuerdo con la energía que contiene, mientras que el cociente entre la suma de estas frecuencias ponderadas y la energía espectral total permite obtener la posición promedio de la energía dentro del espectro.

Ancho de banda espectral

El ancho de banda espectral se relaciona con la dispersión de la energía alrededor del centroide espectral. Mientras el centroide indica la posición promedio de la energía en el eje de frecuencia, el ancho de banda indica qué tan extendida está dicha energía. Dentro de los descriptores espectrales de corto plazo, el *spectral spread* se emplea como una característica útil para el análisis y la clasificación acústica [5]. Asimismo, este descriptor forma parte de conjuntos amplios de características orientadas a la descripción y clasificación de sonidos [52]. En clasificación general de audio, el ancho de banda se ha utilizado dentro de los rasgos espectrales para diferenciar señales con energía concentrada en una zona frecuencial estrecha de señales con energía distribuida en un rango más amplio [3], sintetizada en la ecuación:

$$B = \left(\frac{\sum_{k=1}^{K} S_k |f_k - C|^p}{\sum_{k=1}^{K} S_k} \right)^{1/p} \tag{10}$$

Donde k es índice que identifica cada componente frecuencial, mientras que K corresponde al número total de componentes espectrales consideradas. La variable S_k representa la magnitud o energía espectral asociada a la componente k, C representa el centroide espectral, que actúa como frecuencia de referencia para medir la dispersión y p define el orden de la norma utilizada para medir dicha dispersión; generalmente se emplea $p = 2$.

Rolloff espectral

El *rolloff* espectral es la frecuencia por debajo de la cual se concentra un porcentaje determinado de la energía total del espectro, típicamente el 85 %. Cuanto más alto es este valor, mayor es la proporción de energía ubicada en frecuencias altas; cuanto más bajo, más concentrada está la energía en frecuencias bajas [4].

$$\sum_{k:f_k \leq f_r} s_k \geq \rho \sum_{k=1}^{K} s_k \tag{11}$$

Donde f_r es la frecuencia de *rolloff*, es decir, el valor buscado; s_k es la magnitud espectral de la componente k; K es el número total de componentes consideradas; y ρ es la proporción de energía que define el umbral (habitualmente $\rho = 0{,}85$).

Flujo espectral

El flujo espectral, o *spectral flux*, mide la variación del espectro entre ventanas consecutivas. A diferencia de otros descriptores que resumen una sola trama, el flujo espectral captura cambios temporales en la distribución frecuencial de la señal. Este descriptor puede expresarse como la diferencia cuadrática entre las magnitudes normalizadas de dos distribuciones espectrales sucesivas [4].

$$Flux(t) = \sum_{k=1}^{K} \left(S_t[k] - S_{t-1}[k]\right)^2 \tag{12}$$

Donde $S_t[k]$ representa la magnitud espectral normalizada de k en la trama actual t, y $S_{t-1}[k]$ representa la magnitud correspondiente en la trama anterior.

Planitud espectral

La planitud espectral, o *spectral flatness*, es un descriptor que mide el grado de uniformidad de la distribución espectral de una señal. En términos generales, permite estimar si un espectro presenta un comportamiento más cercano al ruido, con energía distribuida de manera relativamente uniforme, o si presenta componentes tonales dominantes concentradas en ciertas frecuencias. Este descriptor forma parte de conjuntos amplios de características de audio orientadas a descripción, similitud y clasificación [52]. Además, la planitud espectral se ha analizado como una medida vinculada al coeficiente de tonalidad, lo que permite distinguir entre señales con estructura tonal y señales con comportamiento más ruidoso [53].

$$Flatness = \frac{\left(\prod_{k=1}^{K} S_k\right)^{1/K}}{\frac{1}{K}\sum_{k=1}^{K} S_k} \tag{13}$$

Donde S_k representa la magnitud o energía espectral correspondiente a la componente de frecuencia identificada por el índice k, mientras que K indica el número total de componentes espectrales consideradas. El término $\prod_{k=1}^{K} S_k$ representa el producto de todas las magnitudes espectrales y, al elevarlo a $1/K$, se obtiene la media geométrica del espectro. Por su parte, el término $\frac{1}{K}\sum_{k=1}^{K} S_k$ corresponde a la media aritmética de dichas magnitudes. La planitud espectral se obtiene mediante la relación entre ambas medias: valores cercanos a uno indican una distribución espectral relativamente uniforme y asociada a señales de comportamiento más ruidoso, mientras que valores cercanos a cero representan una mayor concentración de energía en determinadas componentes frecuenciales, característica de señales más tonales.

PLP

La predicción lineal perceptual, conocida como PLP por sus siglas en inglés Perceptual Linear Prediction, es una técnica de análisis acústico [54], definida como un método que combina principios de percepción auditiva con predicción lineal, incorporando etapas como el análisis espectral, la transformación a escala Bark, la ponderación por igual sonoridad, la compresión intensidad-sonoridad y el modelado autorregresivo mediante predicción lineal.

$$X_n \approx \sum_{p=1}^{P} a_p X_{[}n-p] \tag{14}$$

Donde X_n representa la muestra actual de la señal, estimada a partir de una combinación lineal de muestras anteriores. El término X_{n-p} corresponde a la muestra situada p posiciones antes de la muestra actual, mientras que a_p representa el coeficiente de predicción lineal asociado a dicho retardo. La variable p identifica cada retardo considerado y P determina el orden del modelo, es decir, el número total de muestras anteriores empleadas para realizar la estimación. El símbolo $\approx$ indica que la muestra actual se aproxima mediante la suma ponderada de sus valores pasados.

Log-Mel

La representación Log-Mel puede entenderse como una matriz tiempo-frecuencia obtenida proyectando el espectro de potencia de una señal sobre un banco de filtros distribuidos en Mel y aplicar una compresión logarítmica sobre las energías resultantes. Este procedimiento se fundamenta en el uso de bancos de filtros en escala Mel, energías logarítmicas y representación cepstral empleados en el análisis acústico de corto plazo [50]. En Log-Mel se conserva la matriz de energías logarítmicas por bandas Mel, sin aplicar la transformada discreta del coseno que compacta dicha información en MFCC.

$$\mathrm{LogMel}(t, m) = \log\left(\sum_{k} P(t, k) H_m[k] + \varepsilon\right) \tag{15}$$

Donde $\mathrm{LogMel}(t, m)$ representa la energía logarítmica obtenida en la trama temporal t y en la banda Mel m. El índice t identifica la posición temporal de la trama analizada, mientras que m corresponde al filtro del banco Mel utilizado. Por su parte, k representa cada componente de frecuencia considerada en el espectro y $P(t, k)$ indica la potencia espectral de la señal en la trama t y en la componente frecuencial k. El término $H_m[k]$ representa la respuesta del filtro Mel m aplicada sobre la frecuencia identificada por k. La sumatoria $\sum_k P(t, k) H_m[k]$ acumula la potencia espectral ponderada dentro de la banda cubierta por dicho filtro. Finalmente, ε es un valor positivo pequeño que evita calcular el logaritmo de cero, mientras que la función logarítmica comprime el rango dinámico de las energías obtenidas.

Filtros Mel

Los filtros Mel conforman un banco de filtros, generalmente triangulares, que agrupan la energía del espectro en bandas distribuidas según la escala Mel, reduciendo la dimensionalidad frecuencial y representando el contenido acústico de forma perceptualmente motivada [50], [51]. Si $H_b(k)$ denota el filtro Mel b aplicado sobre el índice de frecuencia k, la energía Mel para una ventana m puede expresarse como [51]:

$$M(m, b) = \sum_{k} S(m, k) H_b(k) \tag{16}$$

Donde $S(m,k)$ corresponde al espectrograma de potencia y $M(m,b)$ a la energía resultante en la banda Mel b [50], [51].

Coeficientes cepstrales y MFCC

Los coeficientes cepstrales son descriptores que representan de forma compacta la envolvente espectral de una señal a partir del logaritmo de su espectro, separando la información asociada a la estructura global del espectro de las variaciones más finas [45], [50]. Los MFCC, o coeficientes cepstrales en frecuencias Mel, se obtienen aplicando el logaritmo sobre las energías calculadas mediante los filtros Mel y, posteriormente, una transformada discreta del coseno, lo que produce una representación compacta y ampliamente utilizada en el reconocimiento de audio [50], [51]. Una formulación común de los MFCC es [51]:

$$c_q(m) = \sum_{b=1}^{B} \log(M(m,b)) \cos\left[\frac{\pi q}{B}\left(b - \frac{1}{2}\right)\right] \tag{17}$$

Donde $c_q(m)$ corresponde al coeficiente cepstral q en la ventana m, B es el número de filtros Mel y $M(m,b)$ es la energía de la banda Mel b [50], [51].

Delta y Delta-Delta

Los coeficientes Delta representan la variación temporal de una característica acústica, mientras que los coeficientes Delta-Delta corresponden a la variación de segundo orden, es decir, a la aceleración de dichos cambios [47]. Si c_t denota una característica acústica en el instante t, el coeficiente Delta puede aproximarse mediante una regresión local de la siguiente forma [47]:

$$\Delta c_t = \frac{\sum_{r=1}^{R} r(c_{t+r} - c_{t-r})}{2\sum_{r=1}^{R} r^2} \tag{18}$$

Los coeficientes Delta-Delta se obtienen aplicando el mismo procedimiento sobre los coeficientes Delta [47]:

$$\Delta^2 c_t = \frac{\sum_{r=1}^{R} r(\Delta c_{t+r} - \Delta c_{t-r})}{2\sum_{r=1}^{R} r^2} \tag{19}$$

Escalamiento de características

Estandarización y normalización

El escalamiento de características es una etapa de preprocesamiento necesaria para colocar variables numéricas en escalas comparables antes de entrenar determinados algoritmos de aprendizaje automático [31]. Muchos modelos y algoritmos de optimización se comportan mejor cuando las características se encuentran en rangos o distribuciones similares, ya que una variable con magnitudes numéricas mayores puede influir de forma desproporcionada sobre el cálculo de distancias, pesos o fronteras de decisión [31]. Por tanto, se distinguen dos procedimientos comunes: la normalización, entendida frecuentemente como el reescalamiento de los valores a un intervalo definido, por ejemplo $[0, 1]$, y la estandarización, que transforma cada característica para que tenga media cercana a cero y desviación estándar unitaria [31]. Ambos procedimientos se detallan en la Tabla 1.

Tabla 1 *Medidas empleadas en la normalización de características.*

Medida	Expresión	Significado	Aplicación en la normalización
Media de una característica	$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$	Representa el valor promedio de la característica j en el conjunto de entrenamiento.	Se usa como referencia central para restar a cada valor de la característica.
Desviación estándar de una característica	$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2}$	Mide la dispersión de los valores de la característica j respecto a su media.	Permite escalar la característica según su variabilidad.
Estandarización de una característica	$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$	Transforma cada valor original x_{ij} en un valor estandarizado Z_{ij}.	Coloca la característica en una escala comparable, con media cercana a cero y desviación estándar cercana a uno.

En redes neuronales, la normalización también es relevante porque entradas con es-

calas muy distintas pueden dificultar la optimización. Cuando la entrada es una matriz tiempo-frecuencia, la estandarización puede aplicarse por canal o por característica para centrar los valores y controlar su dispersión. Esta operación ayuda a que el entrenamiento sea más estable, especialmente cuando se usan optimizadores basados en gradiente [34]. Conceptualmente, la normalización mantiene la misma finalidad que en aprendizaje automático tradicional: evitar que diferencias de escala dominen el proceso de aprendizaje.

Desbalance de clases y aumento de datos en audio

Desbalance de clases

El desbalance de clases ocurre cuando las categorías de un conjunto de datos no están representadas por cantidades similares de muestras, de modo que algunas clases poseen numerosos ejemplos mientras que otras cuentan con una cantidad reducida [55]. En problemas de clasificación supervisada, esta condición puede sesgar el aprendizaje hacia las clases mayoritarias, ya que el modelo puede alcanzar una exactitud aparentemente alta favoreciendo a las categorías con más datos, mientras disminuye su capacidad para reconocer las clases minoritarias [34], [55].

Muestreo balanceado

El muestreo balanceado es una estrategia utilizada para reducir el efecto del desbalance de clases durante el entrenamiento, modificando la frecuencia con la que las muestras de cada clase se presentan al modelo [55]. En lugar de entrenar únicamente con la distribución original de los datos, el muestreo puede aumentar la probabilidad de seleccionar ejemplos de clases minoritarias o reducir la presencia relativa de las clases mayoritarias, con el objetivo de que el clasificador reciba una señal de aprendizaje más equilibrada [34], [55].

Weighted Random Sampler

El *Weighted Random Sampler* es una técnica de muestreo en la que cada muestra recibe un peso que determina su probabilidad de ser seleccionada durante la formación de lotes de entrenamiento [34], [55]. En problemas desbalanceados, los pesos suelen definirse en función

de la frecuencia de cada clase, de modo que las muestras pertenecientes a clases minoritarias tengan mayor probabilidad de aparecer durante el entrenamiento [55]. Si n_k denota el número de muestras de la clase k, un peso de muestreo puede formularse de manera general como:

$$w_k = \frac{1}{n_k} \tag{20}$$

Donde w_k asigna un mayor peso a las clases con menor cantidad de muestras [55].

Aumento de datos en audio

El aumento de datos en audio consiste en generar nuevas muestras de entrenamiento a partir de audios previamente etiquetados mediante transformaciones que conservan la clase original. Esta técnica permite producir datos adicionales a partir de muestras existentes, siempre que las deformaciones aplicadas no modifiquen el significado acústico de la etiqueta [15].

En este sentido, el audio transformado debe seguir representando la misma clase sonora, por lo que el aumento de datos debe aplicarse de forma controlada. En problemas de clasificación acústica, esto resulta especialmente importante, ya que una transformación excesiva podría alterar rasgos relevantes de la firma sonora del vehículo.

Ganancia

La variación de ganancia, también llamada *random gain* o *volume gain*, es una técnica de aumento de datos que modifica la amplitud de la señal de audio sin alterar necesariamente la clase acústica del evento. En revisiones sobre clasificación de audio, esta transformación se presenta como una técnica tradicional para generar nuevas versiones de una señal sonora mediante cambios de volumen o ganancia [56].

Una forma simple de representar esta operación es:

$$y[n] = \alpha x[n] \tag{21}$$

Donde $y[n]$ representa la señal obtenida después de aplicar la variación de ganancia, mientras que $x[n]$ corresponde a la señal original y n identifica el índice de cada muestra. El parámetro α

representa el factor de ganancia aplicado a la amplitud de la señal. Cuando $\alpha > 1$, la amplitud aumenta; cuando $0 < \alpha < 1$, la amplitud disminuye; y cuando $\alpha = 1$, la señal conserva su amplitud original.

Desplazamiento Circular

El desplazamiento circular, o *circular shift*, es una técnica de aumento de datos aplicada sobre señales de audio en la que el contenido temporal de la muestra se desplaza sin eliminar información, ya que la porción desplazada fuera de un extremo se reincorpora por el extremo opuesto. En identificación acústica, este procedimiento puede entenderse como un desplazamiento tipo *wrap-around*, en el cual una parte de la señal o de su representación se traslada conservando la información original de la muestra [57]. Esta operación permite generar variantes temporales del audio sin cambiar la clase sonora.

El desplazamiento circular se expresa como:

$$y[n] = x\left[(n - s) \text{ mód } N\right] \tag{22}$$

Donde $y[n]$ representa la señal obtenida después de aplicar el desplazamiento circular, mientras que $x[n]$ corresponde a la señal original y n identifica el índice de cada muestra. El parámetro s indica el número de muestras desplazadas y N representa el número total de muestras que conforman el segmento de audio. La operación módulo, expresada mediante módN, permite que las muestras que superan uno de los extremos del segmento se reincorporen por el extremo opuesto, conservando así la longitud y la información completa de la señal.

Inversión de polaridad

La inversión de polaridad es una técnica de aumento de datos en el dominio temporal que altera la fase de la señal en 180 grados, produciendo una señal simétrica respecto a la original [58]. En términos prácticos, esta transformación equivale a invertir el signo de las muestras de la forma de onda, por lo que los valores positivos pasan a negativos y los negativos a positivos.

Esta técnica se incluye dentro de las transformaciones aplicadas a clasificación de audio junto con otros procedimientos sobre la señal, como desplazamiento temporal, variación de ganancia y estiramiento temporal [59]. Por ello, puede considerarse una transformación válida dentro de un esquema de aumento de datos, siempre que no modifique el significado acústico de la clase.

$$y[n] = -x[n] \tag{23}$$

Donde $y[n]$ representa la señal obtenida después de aplicar la inversión de polaridad, mientras que $x[n]$ corresponde a la señal original y n identifica el índice de cada muestra. El signo negativo indica que la amplitud de cada muestra se multiplica por -1, de modo que los valores positivos pasan a ser negativos y los negativos pasan a ser positivos. Esta transformación invierte la fase de la señal en 180 grados, pero conserva su duración, contenido frecuencial y magnitud absoluta.

Estiramiento Temporal

El *time stretch* es una técnica de aumento de datos en el dominio temporal que modifica la duración de una señal de audio sin alterar necesariamente su tono. Esta transformación permite obtener versiones más lentas o más rápidas de una muestra original, por lo que puede aumentar la variabilidad temporal del conjunto de entrenamiento manteniendo la clase acústica cuando se aplica de forma moderada [58].

Una forma general de representar el estiramiento temporal es:

$$y(t) = x\left(\frac{t}{a}\right) \tag{24}$$

Donde $y(t)$ representa la señal obtenida después de aplicar el estiramiento temporal, mientras que $x(t)$ corresponde a la señal original y t representa la variable de tiempo. El parámetro a es el factor de estiramiento temporal y determina la modificación de la duración de la señal. Cuando $a > 1$, la señal se alarga y se reproduce más lentamente; cuando $a < 1$, la señal se comprime y se reproduce con mayor rapidez. Para $a = 1$, la duración de la señal permanece sin

cambios.

SpecAugment

SpecAugment es una técnica de aumento de datos aplicada directamente sobre representaciones espectrales del audio, como espectrogramas o espectrogramas de Mel, mediante el enmascaramiento aleatorio de regiones en los ejes de tiempo y frecuencia [60]. A diferencia de los aumentos aplicados sobre la forma de onda, SpecAugment modifica la representación tiempo-frecuencia que recibe el modelo, ocultando parcialmente bandas frecuenciales o intervalos temporales para reducir la dependencia del clasificador respecto a regiones específicas del espectrograma [60]. En el enmascaramiento de frecuencia se selecciona un conjunto de bandas y se reemplazan sus valores por cero o por un valor constante, mientras que en el enmascaramiento temporal se aplica el mismo procedimiento sobre cuadros consecutivos del eje temporal [60]. De forma general, si $S(t, f)$ denota un espectrograma, una máscara de frecuencia puede expresarse como [60]:

$$\tilde{S}(t,f) = \begin{cases} 0, & f_0 \leq f < f_0 + F \\ S(t,f), & \text{en otro caso} \end{cases} \tag{25}$$

y una máscara temporal puede expresarse como [60]:

$$\tilde{S}(t,f) = \begin{cases} 0, & t_0 \leq t < t_0 + T \\ S(t,f), & \text{en otro caso} \end{cases} \tag{26}$$

Donde F corresponde al ancho de la máscara de frecuencia, T al ancho de la máscara temporal, y f_0 y t_0 indican las posiciones iniciales de cada máscara [60].

MixUp

MixUp es una técnica de regularización y aumento de datos que genera nuevas muestras de entrenamiento mediante combinaciones lineales de pares de ejemplos y de sus respectivas etiquetas [61]. Si (x_i, y_i) y (x_j, y_j) son dos muestras de entrenamiento, MixUp construye una

nueva muestra sintética de la siguiente forma [61]:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{27}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{28}$$

Donde λ es un valor obtenido de una distribución Beta que controla la proporción en la que se combinan ambas muestras [61]. En redes neuronales profundas, esta técnica favorece fronteras de decisión más suaves y reduce la tendencia del modelo a memorizar ejemplos específicos [34], [61].

Modelos de aprendizaje automático empleados

La clasificación constituye una de las tareas fundamentales del aprendizaje automático y puede abordarse mediante técnicas de *Machine Learning* (ML) tradicional o de *Deep Learning* (DL) [11]. Entre los clasificadores de ML, las máquinas de vectores de soporte (*Support Vector Machines*, SVM) se caracterizan por construir fronteras de decisión que maximizan el margen de separación entre las clases. Mediante el uso de funciones *kernel* y el ajuste de hiperparámetros como C y γ, este modelo puede adaptarse a problemas lineales y no lineales, incluso cuando se dispone de conjuntos de datos reducidos [3], [12], [22].

Dentro del DL, las redes neuronales convolucionales (*Convolutional Neural Networks*, CNN) permiten aprender patrones jerárquicos a partir de datos organizados en estructuras matriciales [34], [62]. En la clasificación acústica, estas redes pueden utilizar representaciones bidimensionales de las señales, como los espectrogramas en escala Mel y las matrices de coeficientes MFCC, para identificar características espectrales y temporales relevantes [13], [14], [17]. En consecuencia, las SVM y las CNN representan enfoques complementarios para el análisis y la clasificación de señales acústicas, cuyos criterios de selección y configuración se desarrollan en el capítulo metodológico.

Fundamentos de la Máquina de Vectores de Soporte

Las Máquinas de Vectores de Soporte, conocidas como SVM por sus siglas en inglés (Support Vector Machine), son una máquina de aprendizaje para problemas de clasificación supervisada [12] en la que los vectores de entrada pueden ser proyectados, de forma explícita o implícita, hacia un espacio de mayor dimensión donde se construye una superficie de decisión capaz de separar las clases. La idea central del método consiste en encontrar la frontera de decisión con mayor capacidad de generalización, lo que se logra maximizando el margen entre las clases.

En un problema de clasificación, dicha frontera corresponde a un hiperplano que divide el espacio de características en regiones asociadas a cada clase. A diferencia de un clasificador que solo busca separar los datos, el SVM selecciona el hiperplano ubicado a la mayor distancia posible de las muestras más cercanas de cada clase, es decir, el de mayor margen. Estas muestras reciben el nombre de vectores de soporte, debido a que son las observaciones que determinan directamente la posición de la frontera de decisión. Por esta razón, el modelo no depende de todos los puntos de entrenamiento de la misma manera, sino especialmente de aquellos que se ubican cerca del límite entre clases.

SVM de margen suave

Cuando las clases no son linealmente separables, el SVM puede emplear una formulación de margen suave y funciones *kernel*. [12] Este enfoque de separación lineal es en caso de que los datos no puedan separarse perfectamente mediante un hiperplano, permitiendo errores controlados durante el entrenamiento. Esta extensión se conoce como SVM de margen suave (*soft-margin SVM*); esta tolerancia tiene como finalidad evitar que el modelo dependa de una separación rígida cuando los datos presentan ruido, solapamiento entre clases o muestras atípicas. En la formulación de margen máximo ideal, el SVM busca un hiperplano que separe completamente las clases. Sin embargo, en problemas reales esta condición es difícil de cumplir, sobre todo en el contexto militar de la adquisición de las muestras, donde las grabaciones pueden incluir ruido ambiental, interferencias, variaciones de distancia, cambios de régimen mecánico o segmentos poco representativos.

Para resolver este problema, el SVM de margen suave introduce variables de holgura, normalmente representadas como ξ_i, que permiten medir cuánto incumple una muestra la restricción del margen. Si una muestra está correctamente clasificada y fuera del margen, su valor de holgura es cero. Si cae dentro del margen o es clasificada incorrectamente, la variable de holgura toma un valor positivo.

$$\begin{aligned} \min_{w,b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{sujeto a} \quad & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \qquad i = 1, \ldots, N, \\ & \xi_i \geq 0 \end{aligned} \tag{29}$$

Donde w representa el vector que define la orientación del hiperplano de separación y b corresponde al término independiente o sesgo que permite desplazar dicho hiperplano. La variable x_i representa el vector de características de la muestra i, mientras que y_i corresponde a su etiqueta de clase. El término ξ_i es la variable de holgura asociada a cada muestra y cuantifica el grado en que esta incumple la restricción del margen. El parámetro C controla el equilibrio entre la maximización del margen y la penalización de los errores de clasificación, mientras que N indica el número total de muestras de entrenamiento. El término $\frac{1}{2}\|w\|^2$ favorece un margen de separación amplio y la suma $C\sum_{i=1}^{N}\xi_i$ penaliza las muestras situadas dentro del margen o clasificadas incorrectamente.

Kernel RBF

Las funciones kernel son mecanismos que permiten a los modelos lineales operar de forma implícita en espacios de características de mayor dimensión [63]. El RBF, por sus siglas en inglés Radial Base Function, se define como una función de similitud que depende de la distancia entre dos vectores de entrada. Su valor es alto cuando las muestras comparadas se encuentran cercanas en el espacio de características y disminuye a medida que la distancia entre ellas aumenta. [64] también se puede describir el uso de kernels en SVM como una forma de construir fronteras de decisión no lineales, lo cual resulta útil cuando las clases no pueden separarse adecuadamente mediante una frontera lineal en el espacio original. El kernel RBF se expresa

mediante la siguiente ecuación:

$$k(X_i, X_j) = \exp\left(-\gamma \|X_i - X_j\|^2\right) \tag{30}$$

Donde $k(X_i, X_j)$ representa la similitud entre dos muestras, X_i y X_j son vectores de características, $\exp$ corresponde a la función exponencial, γ es el parámetro que controla el alcance de influencia de las muestras y $\|X_i - X_j\|^2$ representa la distancia euclidiana cuadrática entre ambos vectores.

Parámetros C y gamma

En un SVM con kernel RBF, el comportamiento del clasificador depende en gran medida de dos hiperparámetros, C y γ, que deben seleccionarse cuidadosamente por su influencia directa en la capacidad de generalización del modelo [65]. El parámetro C pertenece a la formulación de margen suave y controla el costo asignado a los errores de clasificación, mientras que γ define el alcance de influencia de cada muestra dentro del kernel.

El parámetro C regula el equilibrio entre mantener un margen amplio y reducir los errores de entrenamiento. Un valor alto penaliza fuertemente las muestras mal clasificadas, ajustando la frontera para corregir la mayor cantidad posible de errores; esto puede mejorar el desempeño en entrenamiento, pero también aumenta el riesgo de sobreajuste. Un valor bajo, en cambio, permite más violaciones del margen y favorece una frontera más simple, con el riesgo de producir subajuste si resulta demasiado flexible.

El parámetro γ, por su parte, controla qué tan lejos llega la influencia de una muestra dentro del kernel RBF y se relaciona con el ancho de la función radial mediante:

$$\gamma = \frac{1}{2\sigma^2} \tag{31}$$

Donde σ es el ancho de la función radial. Dado que ambos parámetros son inversamente proporcionales, un valor pequeño de σ produce un γ alto, concentrando la influencia de cada muestra y generando fronteras de decisión más complejas; un valor grande de σ produce un γ bajo, ampliando dicha influencia y generando fronteras más suaves.

Fundamentos de la Red Neuronal Convolucional

Una red neuronal convolucional, o CNN, es un modelo de aprendizaje profundo que utiliza operaciones de convolución para extraer patrones locales a partir de datos con estructura espacial, como imágenes o representaciones bidimensionales de audio [34], [36]. En clasificación acústica, el audio puede transformarse en espectrogramas para que la CNN aprenda patrones espectro-temporales asociados a la distribución de energía en frecuencia y tiempo [21], [47]. Si X denota la entrada, K un filtro convolucional, b un sesgo y σ una función de activación, una operación convolucional puede expresarse de manera general como [34]:

$$Y = \sigma(K * X + b) \tag{32}$$

En una CNN jerárquica, cada nodo puede emplear un clasificador convolucional especializado que distingue únicamente las clases hijas de ese nodo, reduciendo la complejidad local de la decisión en comparación con una clasificación plana entre todas las clases [29], [36].

Convolución y filtros convolucionales

La convolución es la operación fundamental de una red neuronal convolucional, ya que permite extraer patrones locales de una entrada mediante filtros entrenables, también denominados *kernels* [34], [36]. En el caso de señales acústicas representadas como espectrogramas, la entrada puede interpretarse como una imagen bidimensional en la cual un eje representa el tiempo y el otro la frecuencia; por ello, los filtros convolucionales pueden aprender estructuras locales asociadas a concentraciones de energía, bandas armónicas, variaciones temporales o patrones espectro-temporales característicos de las fuentes vehiculares [21], [47]. De forma general, una convolución bidimensional aplicada sobre una entrada X con un filtro K puede expresarse como [34]:

$$Y(i,j) = \sum_{u} \sum_{v} X(i+u, j+v) K(u,v) + b \tag{33}$$

Donde $Y(i,j)$ denota la activación resultante en la posición espacial (i,j), X es la entrada, K es

el filtro convolucional y b es el sesgo aprendido por el modelo [34].

Feature map

Un *feature map*, o mapa de características, es la salida generada por un filtro convolucional al aplicarse sobre una entrada; cada mapa resalta la presencia de un patrón aprendido, como bordes en imágenes o estructuras de energía en espectrogramas [34], [36].

Función de activación ReLU

La función de activación ReLU, del inglés *Rectified Linear Unit*, es una función no lineal ampliamente utilizada en redes neuronales profundas debido a su simplicidad computacional y a su capacidad para favorecer el entrenamiento de modelos con múltiples capas [34], [66]. Esta función conserva los valores positivos de la activación y anula los negativos, por lo que se define como [34]:

$$\mathrm{ReLU}(x) = \mathrm{máx}(0, x) \tag{34}$$

En una CNN, ReLU permite introducir no linealidad después de las operaciones convolucionales o de las capas densas, lo cual es necesario para que la red pueda aproximar relaciones complejas entre las representaciones acústicas de entrada y las clases de salida [34].

Pooling: Max Pooling y Average Pooling

El *pooling* es una operación utilizada en las CNN para reducir la dimensión espacial de los mapas de características, disminuir el costo computacional y favorecer cierto grado de invariancia local frente a pequeñas variaciones de posición en la entrada [34]. En el Max Pooling se conserva el valor máximo dentro de una región local $\mathcal{R}$, lo que resalta la activación dominante de un patrón aprendido [34]:

$$y = \underset{i \in \mathcal{R}}{\mathrm{máx}}\, x_i \tag{35}$$

En el Average Pooling, en cambio, se calcula el promedio de los valores dentro de una

región local, generando una síntesis de la activación media en esa zona [34]:

$$y = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} x_i \tag{36}$$

Batch Normalization

Batch Normalization es una técnica incorporada en redes neuronales profundas para normalizar las activaciones de una capa a partir de la media y la varianza de un mini-lote, con el objetivo de estabilizar y acelerar el entrenamiento [67]. Para un mini-lote de m activaciones, la media y la varianza se calculan como [67]:

$$\mu_B = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{37}$$

$$\sigma_B^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 \tag{38}$$

A continuación, cada activación se normaliza y se transforma mediante los parámetros aprendibles γ y β [67]:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{39}$$

$$y_i = \gamma \hat{x}_i + \beta \tag{40}$$

Dropout y Dropout2d

Dropout es una técnica de regularización utilizada en redes neuronales profundas que consiste en desactivar aleatoriamente un subconjunto de unidades durante el entrenamiento, con el objetivo de reducir la dependencia excesiva entre neuronas y disminuir el sobreajuste [34], [68]. Si h denota un vector de activaciones y r una máscara aleatoria binaria, la operación de Dropout puede expresarse como [68]:

$$\tilde{h} = r \odot h \tag{41}$$

Donde $\odot$ denota el producto elemento a elemento [68]. En redes convolucionales, Dropout2d extiende este principio desactivando mapas de características completos en lugar de unidades individuales, lo cual resulta adecuado cuando las activaciones vecinas dentro de un mismo mapa se encuentran espacialmente correlacionadas [34].

ResNet y bloques residuales

ResNet es una arquitectura de redes neuronales profundas basada en bloques residuales, propuesta para facilitar el entrenamiento de modelos con muchas capas y reducir el problema de degradación del desempeño al aumentar la profundidad [62]. En un bloque residual, la red aprende una función residual $\mathcal{F}(x)$ y la suma a la entrada original x, de modo que la salida puede expresarse como [62]:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{42}$$

Donde $\{W_i\}$ denota los pesos aprendidos dentro del bloque [62].

Squeeze-and-Excitation

El bloque Squeeze-and-Excitation, o SEBlock, es un mecanismo de atención por canal que recalibra de forma adaptativa la importancia de los mapas de características producidos por una red convolucional [69]. La operación *squeeze* resume cada canal mediante un promedio global espacial, definido como [69]:

$$z_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{43}$$

La operación *excitation*, por su parte, aprende los pesos de importancia de cada canal mediante capas completamente conectadas y funciones no lineales [69]:

$$s = \sigma(W_2\delta(W_1 z)) \tag{44}$$

Finalmente, cada canal se recalibra multiplicándolo por su peso correspondiente [69]:

$$\tilde{u}_c = s_c u_c \tag{45}$$

Métricas de evaluación

Las métricas de evaluación permiten medir el desempeño de un modelo de clasificación a partir de la comparación entre las etiquetas reales y las etiquetas predichas. En un sistema de clasificación acústica, estas métricas son necesarias porque no basta con conocer el porcentaje global de aciertos; también es importante identificar qué clases se reconocen correctamente, cuáles presentan errores y entre qué categorías existe mayor confusión. Las métricas empleadas son generales y pueden resumirse y especificarse conforme a la Tabla 2.

Entrenamiento, validación, prueba y generalización

Para evaluar un modelo de clasificación de forma confiable, el conjunto de datos disponible suele dividirse en tres subconjuntos: entrenamiento, validación y prueba. El subconjunto de entrenamiento se utiliza para ajustar los parámetros del modelo; el de validación permite ajustar hiperparámetros y decidir cuándo detener el entrenamiento; y el de prueba se reserva para estimar el desempeño final sobre datos no utilizados durante el ajuste del modelo [31].

Esta separación permite aproximar la capacidad de generalización del modelo, es decir, su desempeño esperado sobre datos nuevos. Cuando el modelo se ajusta demasiado a las particularidades del conjunto de entrenamiento, incluyendo ruido o variaciones específicas, puede perder capacidad de respuesta frente a muestras no vistas; este fenómeno se conoce como sobreajuste. El caso contrario ocurre cuando el modelo no logra capturar adecuadamente los patrones presentes en el conjunto de entrenamiento, lo que se denomina subajuste o *underfitting* [34].

Tabla 2 *Métricas de evaluación empleadas en los modelos.*

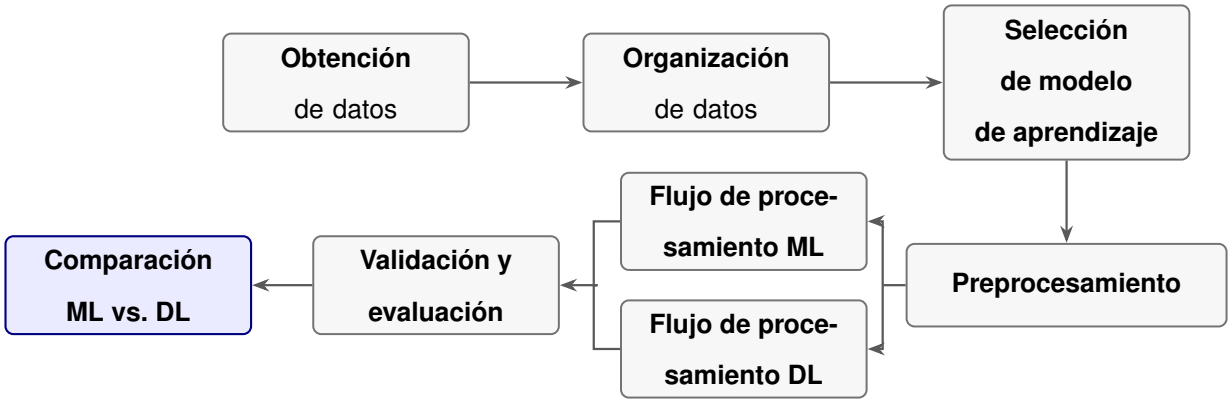
Métrica	Ecuación	Qué mide	Interpretación
Exactitud	$\frac{TP+TN}{TP+TN+FP+FN}$	Proporción total de predicciones correctas respecto al total de muestras evaluadas [70].	Permite observar el rendimiento global del clasificador, tanto en modelos tradicionales como SVM como en redes neuronales convolucionales.
Precision	$\frac{TP}{TP+FP}$	Proporción de muestras predichas como una clase que realmente pertenecen a esa clase [70].	Indica la confiabilidad de las predicciones positivas del modelo.
Recall	$\frac{TP}{TP+FN}$	Proporción de muestras reales de una clase que fueron reconocidas correctamente [70].	Mide la capacidad del modelo para detectar todos los segmentos que realmente pertenecen a una clase.
F1	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Equilibrio entre precision y recall mediante una media armónica [70].	Resume el desempeño de una clase considerando simultáneamente falsas alarmas y omisiones.
F1-Macro	$\frac{1}{K}\sum_{k=1}^{K} F1_k$	Promedio simple del F1 calculado para todas las clases [70].	Evalúa el desempeño dando el mismo peso a cada clase, sin importar cuántas muestras tenga. Es relevante cuando existe desbalance de clases.
F1 ponderado	$\frac{\sum_{k=1}^{K} n_k F1_k}{\sum_{k=1}^{K} n_k}$	Promedio del F1 ponderado por la cantidad de muestras de cada clase [70].	Resume el rendimiento global considerando el peso real de cada clase en el conjunto de evaluación.
Exactitud balanceada	$\frac{1}{K}\sum_{k=1}^{K} \text{Recall}_k$	Promedio del recall obtenido por cada clase, otorgando el mismo peso a cada categoría [70].	Permite evaluar si el clasificador reconoce adecuadamente todas las clases, especialmente cuando existe desbalance.
Pérdida o Loss	Depende de la función de pérdida usada por el modelo.	Mide el error que el modelo intenta minimizar durante el entrenamiento [34].	En una CNN, permite observar si el modelo está aprendiendo durante las épocas de entrenamiento.
Cross-Entropy Loss	$-\sum_{k=1}^{K} y_k \log(\hat{y}_k)$	Mide la diferencia entre la distribución real de clases y_k y la distribución predicha $\hat{y}_k$ por el modelo [34].	Es una función de pérdida habitual en clasificación multiclase con redes neuronales, incluida una CNN jerárquica por nodos.
Loss de validación	$\mathcal{L}_{val}$	Error calculado sobre el conjunto de validación durante el entrenamiento [31], [34].	Permite monitorear generalización y detectar sobreajuste; en CNN puede usarse para seleccionar el mejor modelo o activar parada temprana.
Support	n_k	Cantidad de muestras reales pertenecientes a la clase k en el conjunto evaluado [70].	Ayuda a interpretar precision, recall y F1, especialmente cuando algunas clases tienen pocas muestras.
Matriz de confusión	$C_{i,j}$	Tabla que contabiliza cuántas muestras de la clase real i fueron predichas como clase j [70].	Permite identificar confusiones específicas entre clases o ramas del modelo jerárquico.
Evaluación end-to-end	$\frac{\text{rutas completas correctas}}{\text{total de muestras}}$	Evalúa si la ruta completa desde la raíz hasta la clase final fue predicha correctamente [29].	Mide el desempeño real del sistema jerárquico completo, incluyendo la propagación de errores entre nodos.

Nota: Los acrónimos TP, TN, FP y FN se encuentran definidos en la sección de Acrónimos.

Capítulo III: Metodología

El desarrollo metodológico se organiza en siete etapas secuenciales, desde la captura de audio en campo hasta el análisis comparativo de los modelos entrenados. La Figura 1 resume gráficamente el proceso completo; las secciones siguientes detallan cada etapa, comenzando por la toma de datos y la organización jerárquica de las clases, para luego describir por separado la metodología específica de los modelos.

Figura 1 *Flujograma general del proceso metodológico.*



Obtención de datos

Todos los audios empleados en este trabajo fueron capturados en campo, en condiciones acústicas reales y no simuladas, preservando el ruido ambiental, la reverberación y la variabilidad de distancia propias de un escenario operativo. La grabación se realizó empleando una grabadora portátil Zoom Handy Recorder H4n, equipada con un par de micrófonos condensadores en configuración estéreo XY con ángulo de apertura ajustable entre 90 y 120 grados, dependiendo del objetivo, lo que permitió capturar la escena sonora sin introducir diferencias de fase entre canales. Sus especificaciones técnicas se detallan en la Tabla 3.

Tabla 3 *Especificaciones técnicas de la grabadora empleadas.*

Parámetro	Valor
Modelo	Zoom Handy Recorder H4n
Tipo de micrófono	Condensador estéreo, configuración XY
Ángulo de captura	Ajustable entre 90 y 120 grados
Conversión A/D	24 bits
Formato de grabación	WAV (PCM, sin compresión)
Frecuencia de muestreo	96,000 Hz
Cuantización	24 bits

Como se observa en la Tabla 3, la grabadora se configuró para maximizar la fidelidad espectral de las señales de interés: la conversión analógico-digital se realizó a 24 bits y los archivos se almacenaron en formato WAV sin compresión, a una frecuencia de muestreo de 96,000 Hz. lo que excede ampliamente el límite impuesto por el criterio de Nyquist para el rango audible relevante en motores, rotores y sistemas de tracción, preservando los componentes de alta frecuencia que después aprovechan tanto el vector de descriptores manuales del modelo SVM como la representación espectro-temporal de la CNN. En total durante el período de recolección de datos, el cual recorrió tres unidades de caballería del ejército ecuatoriano, una de aviación y una de inteligencia durante un periodo de cinco semanas, se pudo, considerando tanto el acceso a los vehículos como algunos audios que se obtuvieron por oportunidad alcanzando poco mas de cinco horas de audio totales distribuidos de manera heterogénea conforme a la Tabla 4.

Tabla 4 *Distribución de audios crudos por tipo de vehículo.*

Tipo de vehículo	Audios crudos	Minutos totales	Horas
AUV	26	36.97	0.62
Helicópteros	23	32.61	0.54
Oruga	34	88.89	1.48
Rueda	74	144.29	2.40
Total	**157**	**302.76**	**5.05**

Organización de datos

Debido a que las clases del problema (vehículos aéreos y terrestres, sus subfamilias y sus modelos específicos) presentan una taxonomía natural jerárquica, se organizó en tres niveles que posteriormente generaría siete nodos de clasificación válido para el entrenamiento de los clasificadores de tal manera que los audios se almacenan conforme a la Taba 5.

Tabla 5 *Arquitectura jerárquica común de clasificación (7 nodos).*

Nivel / nodo	**Decisión realizada**	**N.º de clases**
Nivel 1 – raíz	Aéreos vs. Terrestres	2
Nivel 2 – rama aérea	AUV vs. Helicópteros	2
Nivel 2 – rama terrestre	Oruga vs. Rueda	2
Nivel 3 – Aéreos / AUV	Modelos específicos de dron	8
Nivel 3 – Aéreos / Helicópteros	Modelos específicos de helicóptero	4
Nivel 3 – Terrestres / Oruga	Modelos específicos de oruga	4
Nivel 3 – Terrestres / Rueda	Modelos específicos de rueda	6

La Figura 2 ilustra esta misma organización en forma de árbol, mostrando las ocho clases de nivel 3 en la rama de AUV, las cuatro de la rama de helicópteros, las cuatro de la rama de oruga y las seis de la rama de rueda que finalmente distinguen ambos modelos que se forman considerando todas las subclases. Es preciso recordar que estos no son todos los vehículos militares disponibles en el ejército ecuatoriano en su respectivas categorías en existencia, estos corresponden a muchos de los que se encuentran operativos al momento de realizar el presente trabajo.

Figura 2 *Árbol jerárquico de clasificación.*



Selección de los modelos a emplearse

La clasificación constituye una de las tareas fundamentales del aprendizaje automático y puede abordarse mediante modelos de *Machine Learning* (ML) tradicional o de *Deep Learning* (DL) [11]. Dentro del ML existen alternativas como los clasificadores bayesianos o los árboles de decisión, valorados por su simplicidad y rapidez, pero que carecen de un principio de margen que garantice una buena generalización, lo que los vuelve menos confiables cuando el número de muestras es limitado y las fronteras entre clases son complejas. Las máquinas de vectores de soporte (*Support Vector Machines*, SVM) se sustentan, en cambio, en el principio de minimización del riesgo estructural: en lugar de ajustarse únicamente a minimizar el error sobre los datos de entrenamiento, buscan la frontera de decisión que maximiza el margen de separación entre clases, lo que se traduce en cotas de generalización más favorables cuando el número de muestras disponibles es reducido [12]. Esta propiedad resulta determinante para el presen-

te trabajo, cuyo conjunto de datos es de construcción propia, de tamaño limitado y fue obtenido en condiciones de captura cercanas a un entorno real de operación, antes que en un ambiente controlado.

Una segunda ventaja de la SVM es que, mediante el uso de funciones *kernel*, puede representar relaciones no lineales entre las características sin necesidad de aumentar explícitamente la dimensionalidad del espacio de entrada ni de rediseñar el modelo; el problema de optimización subyacente depende del margen alcanzado y no directamente del número de variables consideradas, lo que reduce el riesgo de sobreajuste frente a otros clasificadores cuando la cantidad de muestras es limitada [11], [12]. Esta base teórica se ve respaldada por un uso extendido en la literatura de clasificación acústica y de audio bajo condiciones de datos reducidos, en la que la SVM se ha aplicado con buenos resultados a la discriminación de eventos sonoros y señales mecánicas [3], [22], [23], lo que refuerza su pertinencia como modelo de referencia clásico para este trabajo.

Como contraparte, las redes neuronales convolucionales (*Convolutional Neural Networks*, CNN) se fundamentan en dos principios que las distinguen de una red completamente conectada: la conectividad local, que asume que la información relevante de una señal está contenida en la relación entre elementos cercanos, y el uso de pesos compartidos entre filtros, que reduce drásticamente el número de parámetros y otorga invariancia ante desplazamientos locales del patrón [34], [36]. Al apilar varias capas convolucionales, estos filtros locales se combinan progresivamente hasta formar representaciones jerárquicas, en las que las primeras capas capturan patrones simples y las capas más profundas los combinan en estructuras acústicas más complejas [34], [36]. Esta base teórica es directamente aplicable a la clasificación acústica cuando la señal se representa mediante estructuras bidimensionales tiempo-frecuencia, como los espectrogramas en escala Mel o las matrices de coeficientes cepstrales (MFCC), en las que regiones cercanas suelen contener patrones acústicos relacionados entre sí [13], [14], [15], [16].

La segunda propiedad relevante de la CNN es que aprende de forma automática, durante el propio entrenamiento, las transformaciones que convierten la representación de entrada en características discriminativas, en lugar de depender de descriptores diseñados manualmente [13], [17]. Esta capacidad de aprendizaje de representaciones ha sido validada en tareas de clasificación de sonidos ambientales y de audio a gran escala, con resultados competitivos frente

a enfoques basados en características manuales [14], [15], [16], lo que respalda su pertinencia como segundo modelo de referencia para este trabajo.

La principal diferencia entre ambos modelos radica en el origen de las características que utilizan para clasificar: mientras que la SVM opera sobre un conjunto de descriptores diseñados manualmente a partir del conocimiento del dominio, la CNN aprende dichas características directamente de los datos. Esta distinción convierte a ambos modelos en un contraste metodológico apropiado: no se trata simplemente de comparar dos algoritmos, sino de comparar dos filosofías distintas de representación del conocimiento acústico —una basada en el diseño experto de características, y otra basada en el aprendizaje de representaciones— frente a un mismo problema y bajo las mismas condiciones desfavorables de adquisición y disponibilidad limitada de muestras.

Por esta razón, ambos modelos se integran en el mismo esquema jerárquico de clasificación planteado en la sección anterior, lo que permite evaluarlos bajo condiciones comparables y mantener coherencia metodológica a lo largo del procedimiento experimental [11], [29].

Modelo de aprendizaje automático tradicional

El empleo de un modelo de aprendizaje automático clásico basado en una Máquina de Vectores de Soporte como primer clasificador responde a su adaptabilidad de frontera, adaptabilidad en categorías de clasificación y derivado de esto, su compatibilidad con la arquitectura jerárquica de la clasificación natural de los vehículos militares. Funciona muy bien en un solo nivel como clasificador multinomial separando directamente todas las clases finales; así como especificación binomial donde el sistema divide el problema en decisiones sucesivas y progresivamente más específicas, es decir, puede: primero distingue entre vehículos aéreos y terrestres; posteriormente, entre sus principales subcategorías; y, finalmente, entre modelos particulares. Esta organización reduce la complejidad de cada decisión local y se adapta a un conjunto de clases agrupadas de manera natural, como AUV, helicópteros, vehículos a rueda y vehículos a oruga [29].

Al decantar por la capacidad binomial se forman naturalemnte nodos y dentro de cada uno, la SVM construye una frontera de decisión basada en el principio de maximización del margen, con el propósito de separar las clases que compiten en ese nivel [12]. Asimismo, el empleo

de un núcleo de función de base radial (RBF) permite representar relaciones no lineales entre las características acústicas. Esto resulta pertinente cuando las firmas sonoras de vehículos similares comparten componentes energéticos, espectrales o cepstrales. Esta capacidad adquiere especial relevancia en el contexto militar, donde las muestras disponibles son limitadas, contienen ruido y son obtenidas bajo condiciones operativas variables.

Otra ventaja relevante es que opera sobre características acústicas previamente calculadas y normalizadas, lo que permite controlar la información que ingresa al modelo y mantener la trazabilidad metodológica del proceso de clasificación. Además, puede ofrecer un desempeño competitivo en conjuntos de datos reducidos y en espacios de características de dimensión moderada, siempre y cuando el procedimiento de validación se diseñe cuidadosamente [11], [71].

Modelo de aprendizaje profundo

Dentro del aprendizaje profundo, la elección de la CNN frente a otras arquitecturas como las redes recurrentes (RNN/LSTM) o los modelos basados en mecanismos de atención (*Transformers*) responde a la limitación de datos disponibles para este trabajo. Las redes recurrentes procesan la información de forma secuencial, lo que dificulta su paralelización y las hace más sensibles a la pérdida de información en secuencias largas; los modelos de atención, por su parte, carecen de un sesgo inductivo de localidad y requieren volúmenes de datos considerablemente mayores para alcanzar un desempeño competitivo [17], [72]. La CNN, en cambio, aprovecha directamente la localidad y la invariancia ante desplazamientos propias de las representaciones tiempo-frecuencia, lo que la vuelve más eficiente en número de parámetros y más adecuada cuando el conjunto de datos disponible es limitado, como ocurre en este trabajo.

La selección del modelo CNN jerárquico se fundamenta, además, en la naturaleza estructurada del problema de clasificación acústica de vehículos militares. Las categorías vehiculares pueden organizarse de forma progresiva, distinguiendo primero entre vehículos aéreos y terrestres, y luego entre subgrupos como aeronaves no tripuladas, helicópteros, vehículos de rueda o de oruga. Esta organización se ajusta al principio de la clasificación jerárquica, en la cual un problema multiclase se divide en decisiones locales dentro de una estructura de árbol, donde cada nodo resuelve una clasificación más específica que la anterior. El modelo jerárquico resulta así más coherente que un clasificador plano, pues incorpora la relación natural entre

categorías y permite que cada clasificador local se especialice en un subconjunto de clases relacionadas [29].

Frente a un enfoque plano, en el que un único modelo debe clasificar directamente todas las clases finales en una sola salida, la descomposición jerárquica reduce la complejidad de cada decisión individual: en lugar de aprender simultáneamente fronteras entre categorías acústicamente heterogéneas, cada nodo se especializa en distinguir un subconjunto reducido y relacionado de clases, lo que favorece un aprendizaje más estable y limita las confusiones entre familias que deberían separarse en niveles previos [29], [34]. Las primeras etapas del árbol separan así grupos generales, mientras que los nodos posteriores se concentran en distinciones más específicas dentro de una misma rama.

Esta organización jerárquica exige, además, que la partición de los datos y la evaluación del modelo sean metodológicamente consistentes con la estructura del problema. Dado que varios segmentos acústicos pueden provenir de una misma grabación y compartir condiciones de captura, distribuirlos de forma independiente entre entrenamiento y validación puede sobrestimar el desempeño real del modelo, un riesgo ampliamente documentado en la evaluación de modelos de aprendizaje automático [31], [73]. Por ello, la validación de cada nodo debe realizarse de manera agrupada por fuente de origen y, en la medida de lo posible, estratificada por claro riesgo de fuga de datos y mejorar la representatividad del conjunto de validación.

Preprocesamiento

Segmentación del audio

La segmentación del audio constituye la primera etapa operativa del preprocesamiento. Su finalidad es convertir audios originales de duración, formato y condiciones de captura variables en fragmentos homogéneos, auditables y aptos para las etapas posteriores de extracción de características, entrenamiento y evaluación. Este procedimiento permite que las señales acústicas ingresen al sistema bajo una unidad temporal común, evitando que el modelo compare directamente grabaciones de longitudes distintas o con estructuras temporales no equivalentes.

El proceso inicia con la lectura recursiva del directorio que contiene el conjunto de datos

original. Se consideran archivos de audio en formatos comunes, tales como WAV, FLAC, MP3, OGG, M4A y AAC. La ubicación de cada archivo dentro de la estructura de carpetas permite conservar su etiqueta jerárquica, de modo que la ruta del archivo se transforma posteriormente en niveles de clase. Por ejemplo, una grabación ubicada dentro de una rama aérea, una subrama de AUV y una carpeta de modelo específico conserva esa estructura durante todo el procesamiento.

Para cada audio original se genera un identificador único de fuente. Este identificador se obtiene a partir de la ruta relativa del archivo dentro del conjunto de datos y permite reconocer todos los segmentos derivados de una misma grabación. Su uso es fundamental porque, en señales acústicas segmentadas, varios fragmentos de un mismo audio pueden compartir el mismo ambiente, ruido de fondo, distancia de captura, equipo de grabación, condición operacional y características propias de la fuente sonora. Por tanto, el sistema conserva esa relación de origen para que las etapas posteriores puedan evitar que segmentos hermanos de una misma grabación se mezclen indebidamente entre entrenamiento y evaluación.

Una vez identificado el archivo, la señal se convierte a una representación monofónica cuando posee más de un canal. Esta conversión se realiza mediante la combinación de los canales disponibles, con el propósito de trabajar con una sola serie temporal por audio. Posteriormente, la frecuencia de muestreo se unifica a 96,000 Hz. Esta decisión permite conservar una mayor resolución temporal y espectral respecto a configuraciones de menor frecuencia, manteniendo coherencia con la naturaleza del material acústico utilizado en el proyecto. Si una grabación fue capturada con una frecuencia distinta, se realiza un remuestreo hacia la frecuencia objetivo antes de dividirla en fragmentos.

La señal resultante se divide en segmentos de 4.0 segundos respondiendo a un criterio intermedio entre conservar suficiente información acústica y evitar fragmentos demasiado largos que mezclen eventos o condiciones de operación distintas así como conservar una cantidad media de muestras que no proporcionen un desvalance posterior al momento de clasificar, técnicas observadas en trabajos similares de clasificación acústica con variación de unidades temporales de varios sonidos ambientales o en el ámbito de detección acústica de drones [15] [25].

El desplazamiento entre el inicio de un segmento y el siguiente también es de 4.0 segundos, por lo que los fragmentos se generan de forma contigua y sin solapamiento. Esta con-

figuración evita que dos segmentos consecutivos compartan muestras, reduciendo la repetición directa de información entre fragmentos vecinos. En términos de análisis de señales, el uso de ventanas temporales fijas permite estudiar porciones locales de una señal no necesariamente estacionaria, manteniendo una unidad de análisis comparable entre grabaciones [45].

Cuando un audio tiene una duración menor que la ventana definida, el sistema no descarta automáticamente la grabación. En ese caso se genera un único fragmento desde el inicio de la señal y se completa la duración faltante con valores nulos hasta alcanzar los 4.0 segundos requeridos. Esta operación permite conservar audios cortos sin alterar su etiqueta, aunque deja registrado que el segmento fue llevado a la longitud estándar necesaria para el procesamiento posterior.

Después de formar cada segmento se calcula su energía mediante la raíz cuadrática media. Este valor permite estimar la actividad acústica del fragmento. Solo se conservan aquellos segmentos cuya energía es igual o superior a 0.003. Los fragmentos por debajo de ese umbral se consideran de baja actividad o prácticamente silenciosos, por lo que no se escriben como muestras útiles para el entrenamiento. Esta regla evita alimentar el modelo con fragmentos que contienen poca información discriminante y que podrían debilitar la representación acústica de las clases.

Cada segmento válido se guarda como archivo WAV con codificación PCM de 16 bits. Aunque el audio se analiza internamente con la frecuencia objetivo de 96,000 Hz, la escritura en PCM de 16 bits permite almacenar los fragmentos en un formato ampliamente compatible y manejable para las etapas posteriores. Además, las rutas de salida se organizan de forma compacta mediante el identificador de fuente, lo que reduce la longitud de los nombres de archivo y evita problemas del sistema operativo asociados a rutas excesivamente largas. Esta decisión fue incorporada como medida práctica después de detectar fallos por longitud de ruta durante entrenamientos anteriores.

Junto con los archivos segmentados se genera un manifiesto general del conjunto procesado. Este manifiesto registra, para cada fragmento, la ruta del audio original, la ruta del segmento generado, el identificador de fuente, el identificador del segmento, la ruta de clase, el modo de almacenamiento, el estado del procesamiento, el tipo de muestra, el tiempo inicial y final del fragmento, la duración, el valor RMS, la frecuencia de muestreo original y la frecuencia de

muestreo objetivo. Además, se agregan columnas por nivel jerárquico, de modo que la estructura de carpetas del dataset queda transformada en información tabular utilizable por los clasificadores.

El estado de cada segmento también queda registrado. Los fragmentos válidos pueden aparecer como escritos o ya existentes, mientras que los segmentos que no alcanzan el umbral mínimo de energía se registran como descartados por bajo RMS. Si ocurre un problema durante la lectura o procesamiento de un archivo, el manifiesto conserva una fila de error asociada a la fuente correspondiente. Esta trazabilidad permite auditar posteriormente cuántos audios fueron procesados, cuántos segmentos útiles se obtuvieron, qué fragmentos fueron descartados y si existieron fallos de lectura o escritura.

En conjunto, la etapa de segmentación transforma el conjunto de audios originales en una base de fragmentos normalizada en duración, frecuencia de muestreo, estructura de salida y trazabilidad. El resultado no es únicamente una colección de segmentos, sino un conjunto auditado mediante un manifiesto que conserva la relación entre cada fragmento, su audio raíz y su posición dentro de la jerarquía de clases. Esta información es esencial para que las etapas posteriores puedan entrenar y evaluar los modelos sin perder el control sobre la procedencia de las muestras ni introducir fuga de datos.

Aumento de datos y balanceo jerárquico

El aumento de datos sobre la forma de onda se aplica exclusivamente al conjunto de entrenamiento de cada partición. Los conjuntos de validación y prueba no reciben segmentos aumentados, con el fin de que las métricas reportadas reflejen el desempeño del modelo frente a datos reales no observados durante el entrenamiento. Esta restricción es metodológicamente importante, ya que evita que variaciones artificiales de una misma señal influyan en la evaluación del clasificador. En tareas de clasificación acústica, el aumento de datos se emplea para generar variaciones plausibles de señales etiquetadas, incrementando la diversidad del conjunto de entrenamiento sin modificar la clase acústica original [15], [58], [74].

En el presente trabajo, el aumento de datos se utiliza con un propósito principal: reducir el desbalance entre clases dentro de cada decisión jerárquica. Para ello, el sistema analiza las clases que compiten dentro de un mismo nodo y determina cuántas muestras posee cada una

en el conjunto de entrenamiento. A partir de ese conteo, identifica la clase con mayor número de segmentos y la utiliza como referencia de balance. Las clases con menor cantidad de muestras reciben segmentos aumentados hasta aproximarse al tamaño de la clase mayoritaria. De esta manera, la cantidad de datos generados no se define mediante un valor fijo arbitrario, sino en función del déficit real observado entre clases hermanas.

El criterio de balance se aplica de forma ascendente dentro de la jerarquía. Primero se consideran las clases finales o más profundas, como los modelos específicos de vehículos, y posteriormente se revisan los niveles superiores. Esta estrategia permite que el aumento aplicado a las clases hija también contribuya al equilibrio de sus categorías padre. Por ejemplo, si dentro de una rama aérea una clase de AUV presenta menos muestras que otra, el sistema puede generar nuevos segmentos para esa clase final; al hacerlo, no solo se reduce el desbalance en el nodo de modelos, sino que también se modifica el conteo agregado de la rama a la que pertenece. Así, el balance se construye desde las clases más específicas hacia las agrupaciones superiores.

Para generar una muestra aumentada se selecciona un segmento real del conjunto de entrenamiento de la clase correspondiente y se aplica una sola transformación acústica. La selección del segmento base se realiza con reemplazo, de modo que una misma muestra real puede originar más de una variación cuando el déficit de su clase lo requiere. No obstante, cada segmento aumentado conserva la referencia al fragmento original del cual proviene, lo que permite auditar posteriormente la trazabilidad del proceso y verificar qué muestra real dio origen a cada variación artificial.

Las transformaciones empleadas fueron seleccionadas por su carácter conservador. Se descartaron modificaciones más agresivas, como la adición de ruido externo o cambios amplios de tono, debido a que las grabaciones del proyecto ya presentan condiciones ruidosas y variables. En cambio, se priorizaron operaciones que alteran la posición temporal, la polaridad, la intensidad o la duración de forma leve, procurando preservar la identidad acústica de la clase. La distribución de técnicas se definió con mayor peso para las transformaciones menos invasivas y menor peso para aquellas que pueden modificar con mayor sensibilidad la forma temporal de la señal.

La Tabla 6 resume las técnicas de aumento empleadas, la transformación aplicada y la

proporción con la que son seleccionadas durante la generación de nuevas muestras.

Tabla 6 *Técnicas de aumento de datos empleadas en el entrenamiento SVM.*

Técnica	Transformación aplicada	Proporción
Desplazamiento circular	Reubica circularmente la señal dentro de la misma ventana temporal, con un desplazamiento aleatorio de hasta el 10 % de la longitud del segmento. La duración total del audio no cambia.	40 %
Inversión de polaridad	Multiplica la amplitud de la señal por -1, invirtiendo su polaridad sin modificar la duración, la energía absoluta ni la distribución frecuencial principal.	25 %
Variación de ganancia	Aumenta o reduce la amplitud de la señal mediante una variación aleatoria entre 2 y 3 dB, con signo positivo o negativo. Esta transformación simula diferencias moderadas de nivel de captura.	25 %
Estiramiento temporal leve	Comprime o expande ligeramente la señal con un factor aleatorio entre 0.95 y 1.05. El resultado se ajusta nuevamente a la duración original del segmento para conservar la longitud de entrada del modelo.	10 %

La proporción 40–25–25–10 responde a un criterio de conservación de la firma acústica. El desplazamiento circular recibe el mayor peso porque modifica la posición temporal de la señal sin alterar su contenido energético o espectral. La inversión de polaridad y la variación de ganancia se aplican con pesos intermedios, ya que introducen cambios simples y controlados en la forma de onda o en su nivel de amplitud. El estiramiento temporal recibe el menor peso porque, aunque se mantiene en un rango leve, puede modificar con mayor sensibilidad la dinámica temporal de ciertos sonidos mecánicos.

Cada segmento aumentado se almacena en una carpeta compacta destinada exclusivamente a las muestras generadas durante el entrenamiento. Además, se registra en un manifiesto de aumento de datos que documenta la técnica aplicada, los parámetros utilizados, la clase correspondiente y el segmento real del cual proviene. Junto con este manifiesto, el sistema genera una auditoría de balance por clase, donde se reporta el número de segmentos reales, el número de segmentos aumentados, el total resultante, el déficit respecto a la clase mayoritaria y el porcentaje de aumento aplicado. Esta auditoría permite verificar que el balanceo se haya realizado de manera controlada, trazable y limitada al conjunto de entrenamiento.

Flujo de procesamiento de los modelos

Flujo de procesamiento del modelo ML

El flujo de procesamiento de este ML está basado en SVM, este se organiza como un esquema de clasificación supervisada de carácter jerárquico. En lugar de resolver todas las clases del problema en una sola decisión, el sistema divide la tarea en decisiones locales sucesivas. Primero se distingue entre categorías generales, como vehículos aéreos y terrestres; posteriormente, cada rama se especializa en subcategorías, por ejemplo AUV y helicópteros en la rama aérea, u oruga y rueda en la rama terrestre; finalmente, se evalúan los modelos específicos dentro de cada subgrupo. Este tipo de organización es coherente con los enfoques de clasificación jerárquica, en los cuales un problema multiclase se descompone en decisiones más simples y relacionadas con la estructura natural de las clases [29]. En este contexto, cada clasificador local aprende únicamente a diferenciar las clases hijas que compiten dentro de un mismo nivel de la jerarquía.

El proceso inicia con el ingreso de los audios segmentados y su respectiva información de procedencia. Cada fragmento conserva la referencia al audio original del cual fue obtenido, lo que permite reconocer qué segmentos provienen de una misma grabación raíz. Esta identificación es importante porque, en señales acústicas segmentadas, varios fragmentos de un mismo audio pueden compartir condiciones de captura, ruido de fondo, distancia al sensor, ambiente, tipo de operación y características propias de la fuente sonora.Por ello, el sistema no trata cada segmento como si fuera completamente independiente, sino que conserva la relación entre los fragmentos derivados de una misma grabación. Esta precaución reduce el riesgo de fuga de datos, ya que evita que segmentos muy similares de un mismo audio aparezcan simultáneamente en entrenamiento y evaluación.

Para cada nodo de la jerarquía se seleccionan solo las muestras que pertenecen a las clases que deben competir en esa decisión. Por ejemplo, en el nodo superior se consideran las muestras aéreas y terrestres; en el nodo aéreo se consideran únicamente las muestras correspondientes a AUV y helicópteros; en la rama terrestre se consideran oruga y rueda; y en los nodos finales se consideran los modelos específicos de cada subgrupo. De esta manera, cada clasificador local se entrena con un problema más acotado, evitando que tenga que separar clases

que no forman parte de la decisión evaluada en ese nivel.

La validación del sistema se realiza mediante particiones basadas en grupos de origen. Esto significa que los segmentos derivados de una misma grabación madre permanecen juntos dentro de un mismo subconjunto de evaluación. Con ello se busca estimar de forma más realista la capacidad de generalización del modelo frente a audios no vistos, en lugar de medir únicamente su capacidad para reconocer fragmentos acústicamente muy próximos a los ya observados durante el entrenamiento. Esta decisión metodológica es especialmente relevante en bases de datos acústicas reducidas, donde una evaluación inadecuada puede producir métricas artificialmente elevadas.

Una vez definida la partición de datos para cada decisión local, el aumento de datos se aplica únicamente sobre el subconjunto de entrenamiento. Esta restricción evita que las transformaciones artificiales contaminen la validación o la prueba, preservando el carácter independiente de la evaluación. El aumento se genera con el objetivo de reducir el desbalance entre clases dentro de cada nodo. Para ello, el sistema identifica la clase con mayor cantidad de muestras de entrenamiento dentro de la decisión local y calcula cuántas muestras adicionales requiere cada clase minoritaria para aproximarse a ese tamaño de referencia. De esta manera, la generación de muestras aumentadas no se realiza mediante una cantidad fija arbitraria, sino en función de la diferencia observada entre clases hermanas del mismo nodo.

Las transformaciones empleadas se distribuyen de forma controlada entre las nuevas muestras generadas. Del total de muestras aumentadas requeridas para una clase, aproximadamente el 40 % se produce mediante desplazamiento circular de la señal, el 25 % mediante inversión de polaridad, el 25 % mediante variación de ganancia y el 10 % mediante estiramiento temporal leve. Esta distribución responde a un criterio conservador: se priorizan las transformaciones que modifican la posición temporal o la fase de la señal sin alterar de forma sustancial su contenido espectral, mientras que las modificaciones de ganancia y duración se aplican en proporciones menores para evitar deformaciones excesivas de la firma acústica. En este sentido, el aumento de datos busca generar variaciones plausibles de una muestra etiquetada sin modificar su significado de clase, con el propósito de incrementar la diversidad del entrenamiento y mejorar la robustez del modelo [15], [74].

Después de la selección de muestras y del aumento aplicado al entrenamiento, cada

segmento de audio se transforma en una representación numérica mediante la extracción de características acústicas. El modelo no trabaja directamente con la forma de onda cruda, sino con descriptores que resumen propiedades temporales, energéticas, espectrales y cepstrales de la señal. En la configuración empleada se extraen 92 características por cada segmento de audio. Esta cantidad resulta de combinar coeficientes cepstrales, derivadas temporales, descriptores espectrales, descriptores energéticos, medidas temporales y representaciones perceptuales. La finalidad no es aumentar el número de variables de forma indiscriminada, sino representar la señal desde perspectivas complementarias: la energía, la variación temporal, la distribución frecuencial, la forma espectral y la evolución cepstral.

Dentro de este conjunto se incluyen los coeficientes cepstrales en las frecuencias de Mel, junto con sus variaciones de primer y segundo orden. Los MFCC describen la envolvente espectral de la señal en una escala relacionada con la percepción auditiva, mientras que las derivadas permiten incorporar información sobre la evolución temporal de esos coeficientes. También se incluyen medidas energéticas como la raíz cuadrática media, medidas temporales como la tasa de cruces por cero, y descriptores espectrales como centroide, ancho de banda, frecuencia de caída espectral, planitud y contraste espectral. Este tipo de representación ha sido ampliamente empleado en análisis y clasificación de audio, ya que permite convertir señales acústicas complejas en vectores de características comparables entre clases [4], [5], [11], [50], [52]. Adicionalmente, el flujo incorpora representaciones Log-Mel y PLP, con el fin de capturar información relacionada con la distribución perceptual de la energía y con propiedades cepstrales de la señal, sumando 92 características, cantidad mesurada a la necesidad de mantener un equilibrio entre riqueza descriptiva y complejidad del modelo ya que un número demasiado reducido de descriptores podría omitir diferencias acústicas relevantes entre vehículos similares, mientras que un número excesivamente alto incrementaría el riesgo de sobreajuste, especialmente en un conjunto de datos limitado [11], [71].

Posteriormente, las características extraídas son normalizadas de manera individual por descriptor. Esta normalización por característica transforma cada variable para que tenga una escala comparable, evitando que aquellas con valores numéricos más grandes dominen el proceso de aprendizaje. En este punto, la normalización se ajusta únicamente con los datos de entrenamiento de cada partición y luego se aplica a los datos de evaluación. Esta forma de proce-

der es importante porque impide que la información estadística del conjunto de prueba influya en el entrenamiento del modelo, lo cual también contribuye a evitar fuga de datos.

El clasificador empleado en cada nodo corresponde a una Máquina de Vectores de Soporte. Este método busca construir una frontera de decisión que separe las clases maximizando el margen entre las muestras más cercanas de cada grupo. Dichas muestras, conocidas como vectores de soporte, son las que definen de manera más directa la posición de la frontera de separación [12]. En problemas acústicos, donde cada segmento se representa mediante un conjunto de descriptores numéricos, la SVM resulta adecuada como modelo tradicional de aprendizaje supervisado, especialmente cuando se trabaja con bases de datos limitadas y espacios de características de dimensión relativamente alta [11], [71].

Cuando la separación entre clases no puede describirse adecuadamente mediante una frontera lineal, el modelo utiliza una función núcleo de base radial. Esta función permite representar relaciones no lineales entre las muestras, proyectando implícitamente los datos a un espacio donde la separación entre clases puede resultar más efectiva. En este esquema, el parámetro C controla el compromiso entre permitir errores de clasificación y mantener un margen amplio, mientras que γ regula el alcance de influencia de cada muestra dentro de la función de base radial. Un valor inadecuado de estos parámetros puede conducir a un modelo demasiado rígido o demasiado ajustado a los datos de entrenamiento; por ello, su selección forma parte del proceso de ajuste del clasificador.

La búsqueda de configuraciones se realiza dentro del conjunto de entrenamiento de cada partición, evaluando distintas combinaciones de los parámetros del modelo sin utilizar la información reservada para la prueba. Esto permite seleccionar una configuración adecuada para cada nodo sin comprometer la independencia de la evaluación. En consecuencia, cada clasificador local se entrena y se valida en función de las características propias del problema que debe resolver, ya sea una separación general entre ramas superiores o una diferenciación fina entre modelos de vehículos.

Finalmente, el desempeño se evalúa de manera independiente en cada nodo de la jerarquía. Para ello se emplean métricas como exactitud, exactitud balanceada, F1-Macro, F1 ponderado, reporte de clasificación y matriz de confusión. La exactitud permite observar la proporción general de aciertos; sin embargo, cuando las clases están desbalanceadas, puede resultar

insuficiente por sí sola. Por esta razón se incorporan métricas como la exactitud balanceada y F1-Macro, que permiten valorar mejor el comportamiento del modelo entre clases con diferente cantidad de muestras [70], [75], [76]. La matriz de confusión, por su parte, permite identificar qué clases se confunden entre sí y en qué nivel de la jerarquía ocurre dicha confusión.

En conjunto, el flujo de procesamiento SVM jerárquico recibe audios segmentados, conserva la relación de cada segmento con su grabación de origen, construye particiones de evaluación por fuente, aplica aumento de datos únicamente sobre el entrenamiento, genera muestras aumentadas en función del desbalance observado entre clases hermanas, extrae un conjunto de 92 características acústicas, normaliza los descriptores, entrena clasificadores locales mediante SVM y evalúa cada decisión mediante métricas específicas. De esta forma, el sistema busca transformar señales acústicas complejas en decisiones progresivas, controladas y auditables, manteniendo trazabilidad desde el ingreso del audio hasta la predicción obtenida en el nodo correspondiente.

Flujo de procesamiento del modelo DL

El flujo de procesamiento CNN jerárquico se estructura como una secuencia de etapas que transforma los audios originales en una predicción final organizada por niveles. La arquitectura inicia con el conjunto de audios del dataset y la ruta jerárquica de clase de cada muestra, la cual permite representarla dentro de una estructura de categorías. Esta organización es necesaria porque el modelo no realiza una clasificación plana directa, sino que utiliza una jerarquía de decisiones en la que cada nodo resuelve una tarea local hasta llegar a una clase final.

La primera etapa corresponde al preprocesamiento de audio. En esta fase, las grabaciones se convierten a una tasa de muestreo uniforme de 96 kHz, se transforman a mono y se dividen en segmentos de 4 segundos con un salto de 4 segundos. Esta segmentación permite obtener muestras acústicas homogéneas para el entrenamiento y la evaluación. Además, se aplica un filtro basado en RMS para descartar segmentos con baja energía, evitando que fragmentos de silencio o ruido débil sean utilizados como entradas poco representativas para el modelo. Como resultado, se genera un registro de segmentos que conserva información esencial de cada fragmento, como la grabación de origen a la que pertenece, su ubicación dentro de la jerarquía de clases y su valor RMS.

Posteriormente, cada segmento se transforma en una representación acústica de cuatro canales. Esta representación está compuesta por el espectrograma de Mel en decibelios, los coeficientes Delta, los coeficientes Delta-Delta y los MFCC. El espectrograma de Mel permite representar la distribución de energía del sonido en el dominio tiempo-frecuencia, mientras que los coeficientes Delta y Delta-Delta incorporan información sobre la evolución temporal de las características acústicas. Por su parte, los MFCC aportan una descripción compacta de la envolvente espectral de la señal. Estos cuatro canales forman un tensor acústico que sirve como entrada para la CNN.

Luego de la extracción de características, el flujo de procesamiento aplica normalización Z-Score por canal. Esta normalización ajusta cada representación para que tenga una escala comparable, reduciendo el efecto de diferencias de magnitud entre canales y favoreciendo un entrenamiento más estable. Las características normalizadas se almacenan en caché, lo que permite reutilizarlas durante el entrenamiento sin recalcular nuevamente las transformaciones espectrales.

La siguiente etapa corresponde a la división de datos y al balanceo. El conjunto se separa mediante una partición global agrupada por grabación de origen y estratificada por clase. El agrupamiento por grabación de origen evita que segmentos provenientes de un mismo audio aparezcan simultáneamente en entrenamiento y validación, reduciendo el riesgo de fuga de datos. La estratificación busca conservar representación de las clases en el conjunto de validación, lo cual es importante en problemas con desbalance de datos. Durante el entrenamiento se aplican técnicas de aumento de datos, como cambios de ganancia, desplazamiento circular, inversión de polaridad y estiramiento temporal. Además, se utiliza enmascaramiento aleatorio del espectrograma (*SpecAugment*) y un muestreo ponderado de las clases para disminuir el sesgo hacia aquellas con mayor cantidad de muestras.

El núcleo del flujo de procesamiento está formado por el entrenamiento de una CNN local en cada nodo de la jerarquía. La estructura jerárquica parte de un nodo raíz que separa categorías generales y continúa con nodos intermedios que discriminan subgrupos más específicos. Cada nodo con más de una clase hija entrena su propio clasificador CNN, de manera que el modelo no aprende todas las clases finales en una sola salida, sino que descompone el problema en decisiones progresivas. Esta estrategia permite que cada CNN se especialice en

patrones acústicos propios de su rama.

La arquitectura CNN utilizada por cada nodo puede adoptar una versión profunda, empleada cuando existe una unidad de procesamiento gráfico disponible, o una versión ligera, empleada cuando el entrenamiento se realiza únicamente sobre el procesador central. Ambas variantes emplean capas convolucionales, normalización por lotes, función de activación ReLU, operaciones de *pooling*, bloques residuales, módulos de atención por canal (*Squeeze-and-Excitation*) y *Dropout*. Las convoluciones permiten aprender patrones locales sobre el tensor acústico, los bloques residuales facilitan el entrenamiento de redes más profundas y los módulos de atención por canal recalibran la importancia de los canales aprendidos por la red. El clasificador denso final de cada nodo produce la predicción entre las clases hijas correspondientes.

Durante el entrenamiento se emplea una función de pérdida de entropía cruzada con suavizado de etiquetas, el optimizador AdamW, una reducción dinámica de la tasa de aprendizaje cuando el desempeño en validación se estanca, la técnica de combinación lineal de ejemplos (*MixUp*) y parada temprana. Estas técnicas buscan mejorar la estabilidad del entrenamiento, reducir el sobreajuste y conservar el mejor modelo según el comportamiento observado en validación. Al finalizar, se guardan los artefactos correspondientes a cada nodo, como los pesos del modelo operativo, sus metadatos, el historial de entrenamiento, el reporte de clasificación, la matriz de confusión y los resultados individuales del nodo.

Finalmente, el flujo de procesamiento realiza la inferencia jerárquica. En esta etapa, un segmento de audio procesado desciende por el árbol de decisión: primero se evalúa en el nodo raíz, luego en el nodo correspondiente a la rama predicha y así sucesivamente hasta alcanzar una hoja final. La salida del sistema es una ruta completa de clase, por ejemplo, una secuencia del tipo raíz, subcategoría y modelo específico. La evaluación se realiza en dos niveles: por nodo, para medir el desempeño aislado de cada clasificador local, y de forma *end-to-end*, para verificar si toda la ruta jerárquica fue predicha correctamente. Esta doble evaluación es necesaria porque en modelos jerárquicos los errores pueden propagarse desde nodos superiores hacia decisiones posteriores.

Capítulo IV: Resultados y Discusión

Este capítulo presenta los resultados obtenidos en los nodos jerárquicos de los modelos SVM y CNN, evaluados sobre sus respectivos conjuntos de validación independientes. Los resultados se analizan de manera particular, no solo como valores brutos, sino también considerando el contexto, las características y la capacidad individual de cada nodo. Posteriormente, se realiza un análisis comparativo directo entre ambos modelos, dado que comparten la misma arquitectura jerárquica y el mismo conjunto de datos de origen.

Resultados del modelo de aprendizaje automático tradicional

Si bien el modelo SVM presenta por naturaleza una capacidad de clasificación aceptable, como se puede apreciar en la Tabla 7, se optó por una metodología jerárquica al demostrar una ligera mejora en la precisión y en la exactitud balanceada, métricas que permiten evaluar tanto el desempeño general como el comportamiento del clasificador frente al desbalance entre clases.

Tabla 7 *Comparación exploratoria de métricas globales entre configuraciones SVM.*

Modelo SVM	Alcance de evaluacion	Accuracy	Balanced Acc.	F1-Macro	F1-Weighted
SVM jeráquico	Clasificación por nodos, helicópteros y terrestres	91.35 %	85.97 %	84.91 %	91.24 %
SVM NO jerárquico	Clasificación directa por clases	89.22 %	83.99 %	88.37 %	88.39 %

Como se observa en la Tabla 8, los nodos superiores de la arquitectura jerárquica presentan un desempeño alto y estable. N1–Raíz alcanza una exactitud del 99.64 %, mientras que N2–Aéreos y N2–Terrestres registran valores del 95.10 % y 96.42 %, respectivamente. Este comportamiento se relaciona con la estructura jerárquica, ya que dichos nodos clasifican familias generales con diferencias acústicas más marcadas y utilizan el conjunto agregado de las muestras pertenecientes a sus nodos descendientes. En cambio, en N3 los datos se distribuyen

entre modelos específicos, lo que reduce la cantidad y diversidad de audios disponibles por clase e incrementa la similitud acústica entre ellas. Por tanto, la disminución observada en algunos nodos inferiores no contradice la efectividad del enfoque jerárquico, sino que evidencia la mayor complejidad de la clasificación a medida que aumenta el nivel de especificidad.

Tabla 8 *Métricas generales por nodo – modelo SVM.*

Nodo	Clases evaluadas	N.º muestras	Accuracy	Balanced Acc.	F1-Macro	F1-Weighted
N1 – Raíz	AEREOS vs TERRESTRES	4,460	99.64 %	99.63 %	99.49 %	99.64 %
N2 – Aéreos	AUV vs HELICOPTEROS	1,020	95.10 %	94.77 %	95.04 %	95.07 %
N2 – Terrestres	ORUGA vs RUEDA	3,440	96.42 %	96.76 %	96.26 %	96.44 %
N3 – AUV	8 modelos de dron	542	*86.72 %	86.03 %	80.37 %	81.83 %
N3 – Helicópteros	3 modelos defendibles	462	*74.03 %	67.56 %	69.09 %	72.57 %
N3 – Oruga	4 modelos de oruga	1,322	99.39 %	99.63 %	99.28 %	99.40 %
N3 – Rueda	6 modelos de rueda	2,118	*86.83 %	85.97 %	86.69 %	86.65 %

Los resultados marcados con asterisco deben interpretarse con cautela debido a la limitada independencia de las muestras. En N3–Helicópteros, las 462 muestras proceden de pocos audios raíz; además, la clase H145_COBRA fue excluida por disponer de un único audio raíz. Una situación similar se presenta en N3–AUV, cuyas clases cuentan con un máximo de dos audios raíz. Por tanto, aunque ambos nodos poseen numerosas muestras, su escasa diversidad limita la solidez de la evaluación. En contraste, N3–Rueda dispone de casi cuatro veces más muestras que N3–AUV y de una mayor cantidad de audios raíz, pero alcanza una exactitud similar. Este comportamiento se relaciona principalmente con las condiciones de captura, pues los vehículos de rueda son los más utilizados actualmente por el Ejército en operaciones militares y sus registros fueron obtenidos en dicho contexto. En consecuencia, varios audios presentan ruido ambiental, solapamiento con otras fuentes sonoras e interferencias propias de escenarios reales de operación.

Por su parte, los nodos superiores —Raíz, Aéreos y Terrestres— alcanzan exactitudes superiores al 95 % y valores de F1 macro mayores al 94 %, lo que evidencia una adecuada dife-

renciación entre las familias generales de vehículos.

Un análisis más detallado del tercer nivel jerárquico (N3) permite identificar particularidades que no se reflejan completamente en las métricas generales. Como se observa en la Tabla Como se observa en la Tabla 9, la mayoría de las clases presenta una diagonal bien definida. La principal excepción es MATRICE_350, que no registra predicciones correctas y tiende a confundirse con todas las clases, especialmente con MATRICE_400, a la cual se asignan erróneamente 48 muestras. A excepción de DRAGONEYE2, las clases evaluadas corresponden a drones comerciales DJI utilizados por el Ejército. En particular, la similitud mecánica y electrónica entre MATRICE_350 y MATRICE_400, junto con la limitada diversidad de sus audios raíz, podría contribuir a esta confusión. No obstante, las demás clases alcanzan resultados aceptables o casi perfectos pese a presentar limitaciones similares. Por tanto, el error de MATRICE_350 no se explicaría únicamente por la cantidad de datos, sino también por su proximidad acústica con MATRICE_400 y la baja representatividad de los registros disponibles.

Tabla 9 *Matriz de confusión – N3 Aéreos/AUV: modelos de dron (SVM).*

Real \ Predicho	AIR3S	DRAGONEYE2	MATRICE_30	MATRICE_350	MATRICE_400	MAVIC 3T	MAVIC2_PRO_ENT.	MINI4PRO
AIR3S	30	0	2	0	0	2	0	0
DRAGONEYE2	0	142	0	0	0	0	0	0
MATRICE_30	0	0	36	0	0	0	0	0
MATRICE_350	2	2	2	*0	48	2	10	2
MATRICE_400	0	0	0	0	58	0	0	0
MAVIC 3T	0	0	0	0	0	114	0	0
MAVIC2_PRO_ENT.	0	0	0	0	0	0	58	0
MINI4PRO	0	0	0	0	0	0	0	32

Por su parte, la matriz de confusión de N3–Helicópteros presenta un desempeño heterogéneo, como se observa en la Tabla 10. La principal dificultad corresponde a AS332B_SUPERPUMA: de sus 114 muestras, únicamente 49 son clasificadas correctamente, mientras que 56 se asignan a ECUREUIL_AS350. En contraste, esta última clase identifica correctamente 212 de sus 221 muestras. Asimismo, FENNEC_AS550-C3 registra 81 aciertos, aunque mantiene confusiones con las otras dos clases. Pese a las diferencias mecánicas entre estos helicópteros, el reducido número de audios raíz independientes limita la diversidad acústica disponible para el entrenamiento. A ello se añade que todos los registros fueron obtenidos en condiciones operativas reales, donde la inspección auditiva evidencia ruido y solapamiento

con otras fuentes sonoras. Ambos factores podrían enmascarar las características propias de cada modelo y contribuir a las confusiones observadas. Además, H145_COBRA fue excluido de la evaluación independiente por disponer de un único audio raíz, insuficiente para valorar su capacidad de generalización.

Tabla 10 *Matriz de confusión – N3 Aéreos/Helicópteros: modelos de helicóptero (SVM).*

Real \ Predicho	AS332B_SUPERP.	ECUREUIL_AS350	FENNEC_AS550-C3E
AS332B_SUPERPUMA	49	*56	9
ECUREUIL_AS350	8	212	1
FENNEC_AS550-C3E	27	19	81

El mayor contraste dentro del tercer nivel jerárquico se observa en los vehículos de oruga, cuyo desempeño es prácticamente perfecto, como muestra la Tabla 11. La única confusión corresponde a ocho muestras de AMX_155mm clasificadas como AMX_105mm, mientras que las demás clases no presentan errores. Este resultado es destacable porque tres de los vehículos comparten una base mecánica similar y se diferencian principalmente por su armamento, cuya variación modifica la carga útil y puede introducir componentes espectrales distintivos [2]. A diferencia de los vehículos aéreos, estos audios fueron obtenidos en condiciones cercanas a las ideales, con menor presencia de ruido y solapamiento acústico. Debido al uso limitado de estos vehículos en operaciones urbanas, las grabaciones se realizaron principalmente en hangares o entornos controlados. Por tanto, la calidad de los registros habría favorecido la identificación de sus características acústicas, incluso entre modelos mecánicamente similares.

Tabla 11 *Matriz de confusión – N3 Terrestres/Oruga: modelos de oruga (SVM).*

Real \ Predicho	AMX 13	AMX_105mm	AMX_155mm	APC
AMX 13	308	0	0	0
AMX_105mm	0	182	0	0
AMX_155mm	0	8	538	0
APC	0	0	0	286

Como se observa en la Tabla 12, MBOMBE es la clase más numerosa de N3–Rueda,

con 722 muestras, de las cuales 636 se clasifican correctamente, equivalentes al 88.09 %. No obstante, también es la única clase que presenta confusiones con todos los demás vehículos, principalmente con COBRA II, con 50 asignaciones erróneas. Este comportamiento debe analizarse considerando que sus registros contienen una elevada presencia de ruido, interferencias y solapamiento con otras fuentes acústicas. Pese a estas condiciones y a la similitud sonora existente entre vehículos automotores de rueda, la diagonal continúa siendo predominante. La amplitud y diversidad del conjunto de datos de este nodo contribuyen a mantener una capacidad de discriminación aceptable, al representar con mayor fidelidad las variaciones acústicas propias de escenarios operativos reales.

Tabla 12 *Matriz de confusión – N3 Terrestres/Rueda: modelos de rueda (SVM).*

Real \ Predicho	CASCABEL	COBRA II	DAVID	MBOMBE	URAL	URUTÚ
CASCABEL	286	0	0	2	0	2
COBRA II	8	522	0	76	0	4
DAVID	0	14	124	14	10	0
MBOMBE	16	50	2	636	14	4
URAL	0	1	0	0	141	0
URUTÚ	30	0	0	32	0	130

Resultados del modelo de aprendizaje profundo

Los resultados obtenidos con el modelo CNN jerárquico seleccionado se deben a la corrección progresiva de dos aspectos observados en las arquitecturas evaluadas previamente: la forma de organizar la clasificación y la forma de validar el desempeño. La arquitectura plana obligaba a una sola red a clasificar directamente las 22 clases finales, haciendo competir simultáneamente categorías acústicamente distintas e incrementando la complejidad del aprendizaje. La arquitectura jerárquica seleccionada, en cambio, divide el problema en decisiones locales por nodo, lo cual es más adecuado cuando las clases poseen una organización natural por niveles, permitiendo que cada clasificador aprenda patrones específicos de su rama [29], [34], tal como se observa en la comparación de métricas globales de la Tabla 13.

Tabla 13 *Comparación de métricas globales entre arquitecturas CNN evaluadas.*

Arquitectura CNN	Tipo de clasificación	Accuracy	Balanced Acc.	F1-Macro	F1-Weighted
CNN plana	Clasificación directa de 22 clases	67.34 %	65.43 %	44.22 %	69.90 %
CNN jerárquica seleccionada	Clasificación por nodos jerárquicos	81.65 %	87.80 %	80.25 %	82.81 %

Como se observa en la Tabla 13, el incremento de 67.34 % a 81.65 % en Accuracy y de 44.22 % a 80.25 % en F1-Macro evidencia que la clasificación jerárquica resultó más adecuada que la plana. La mejora en F1-Macro es especialmente relevante, ya que esta métrica otorga igual peso a cada clase y permite evaluar si el modelo reconoce adecuadamente categorías con menor cantidad de muestras [70]. Por esta razón, el F1-Macro se considera una métrica central para comparar las arquitecturas, debido a que evita que el desempeño de clases con mayor soporte o ramas más numerosas oculte fallas en clases minoritarias. En este caso, el aumento observado indica que el modelo no solo mejora en promedio, sino que también logra un reconocimiento más equilibrado entre clases.

En el caso del modelo jerárquico, las métricas globales de la Tabla 13 corresponden a una evaluación *end-to-end*, en la que una predicción se considera correcta únicamente si toda la ruta desde la raíz hasta la clase final coincide con la etiqueta real. Por ello, estas métricas son más estrictas que las métricas aisladas por nodo, ya que incorporan la propagación de errores entre niveles de la jerarquía. Esta característica explica por qué el modelo puede alcanzar valores muy altos en algunos nodos individuales, como se observa en la Tabla 14, y aun así presentar una métrica global menor cuando una rama específica concentra errores relevantes.

Tabla 14 *Métricas generales por nodo – modelo CNN jerárquico seleccionado.*

Nodo	Clases evaluadas	N. º muestras	Accuracy	Balanced Acc.	F1- Macro	F1- Weighted
N1 – Raíz	AEREOS vs TERRESTRES	1,079	99.17 %	98.98 %	99.10 %	99.16 %
N2 – Aéreos	AUV vs HELICOPTEROS	401	95.51 %	95.41 %	95.49	95.50 %
N2 – Terrestres	ORUGA vs RUEDA	678	99.41 %	99.36 %	99.36 %	99.41 %
N3 – AUV	8 modelos de dron	205	100.00 %	100.00 %	100.00 %	100.00 %
N3 – Helicópteros	4 modelos de helicóptero	196	42.86 %	55.02	37.18 %	47.50 %
N3 – Oruga	4 modelos de oruga	243	88.48 %	93.64 %	86.05 %	88.82 %
N3 – Rueda	6 modelos de rueda	435	90.34 %	93.33 %	87.84 %	91.12 %

Como se observa en la Tabla 14, el modelo funciona especialmente bien en las decisiones jerárquicas superiores: el nodo raíz alcanza 99.10 % de F1-Macro y la separación entre vehículos de oruga y rueda llega a 99.36 %, lo que indica que las diferencias acústicas entre categorías generales son suficientemente marcadas para ser aprendidas por la red. Estas diferencias responden a la configuración mecánica, el sistema de propulsión y la distribución de la energía sonora en tiempo y frecuencia de cada categoría vehicular.

El desempeño del modelo también se relaciona con la arquitectura convolucional empleada en cada nodo. El uso de bloques residuales facilita el entrenamiento de redes más profundas al permitir el flujo directo de información entre capas, mientras que los bloques Squeeze-and-Excitation recalibran la importancia de los canales aprendidos por la red. Esto resulta relevante porque la entrada del modelo está compuesta por cuatro canales acústicos, por lo que no toda la información tiene la misma importancia en cada nodo de decisión. Así, cada clasificador local puede aprender representaciones espectro- temporales específicas para diferenciar las clases de su propia rama.

Como muestra la Tabla 14, el nodo AUV obtiene el mejor desempeño, con 100 % en todas las métricas, debido a que los modelos de dron presentan patrones espectro-temporales muy diferenciables, asociados a rotores, modulación temporal y distribución de energía en bandas frecuenciales, que las CNN capturan mediante filtros locales sobre espectrogramas. El uso

conjunto de espectrograma de Mel, MFCC, Delta y Delta-Delta refuerza esta capacidad al aportar información complementaria: el espectrograma de Mel describe la distribución de energía en bandas perceptuales a lo largo del tiempo, revelando concentraciones de energía, armónicos y modulaciones producidas por rotores y motores; los MFCC resumen la envolvente espectral en coeficientes compactos, facilitando la diferenciación entre fuentes con estructuras frecuenciales distintas [50]; y los coeficientes Delta y Delta-Delta aportan información dinámica al capturar la variación temporal de estas características, relevante dado que el régimen del motor y la modulación de rotores no permanecen estáticos durante el segmento de audio [47]. En conjunto, estas representaciones proporcionan a la CNN información sobre energía, forma espectral y evolución temporal, ampliando su capacidad para aprender patrones discriminativos en cada nodo de la jerarquía.

Tabla 15 *Comparación por nodo entre arquitecturas CNN jerárquicas evaluadas.*

Nodo	Clases evaluadas	F1-Macro CNN preliminar	F1-Macro CNN seleccionada
N1	AEREOS vs TERRESTRES	96.02 %	99.10 %
N2	AUV vs HELICOPTEROS	100.00 %	95.49 %
N2	ORUGA vs RUEDA	98.84 %	99.36 %
N3	8 modelos de dron	60.28 %	100.00 %
N3	4 modelos de helicóptero	52.86 %	37.18 %
N3	4 modelos de oruga	72.97 %	86.05 %
N3	6 modelos de rueda	94.06 %	87.84 %

Aunque, según la Tabla 15, algunas métricas de la CNN preliminar parecen altas, su evaluación presentaba clases ausentes en validación, lo que podía inflar o distorsionar los resultados. La arquitectura seleccionada corrige este problema mediante una división global agrupada por grabación de origen y estratificada por clase, reduciendo el riesgo de fuga de datos y mejorando la confiabilidad de la evaluación. La mejora del nodo AUV, visible en la misma tabla, debe interpretarse considerando que la evaluación preliminar estaba afectada por clases ausentes en validación, lo que distorsionaba el F1-Macro. En la arquitectura seleccionada, la validación estratificada permite evaluar de manera más representativa las clases de drones, mostrando que esta rama posee patrones acústicos altamente discriminativos para la CNN.

Según la Tabla 14, el bajo desempeño del nodo de helicópteros se debe principalmente a la escasez de muestras disponibles para esta rama: 196 muestras de validación y 624 muestras de entrenamiento antes del aumento de datos. Esta cantidad resulta limitada para diferenciar modelos con firmas acústicas cercanas, especialmente cuando las grabaciones disponibles no cubren suficiente variabilidad de condiciones de operación, distancia, entorno o régimen de funcionamiento. Una cantidad limitada de datos dificulta la estimación de patrones generalizables y aumenta la sensibilidad del modelo a variaciones específicas de las grabaciones; además, el desbalance o la baja variabilidad acústica entre clases favorece el aprendizaje de las clases dominantes, reduciendo el recall y el F1 en las minoritarias.

El F1-Macro de 37.18 % en este nodo, reportado en la Tabla 14, refleja una dificultad significativa para reconocer de forma equilibrada todas las clases de helicópteros, mientras que la Balanced Accuracy de 55.02 % y el F1-Weighted de 47.50 % confirman que el problema afecta tanto el reconocimiento uniforme entre clases como el rendimiento ponderado por soporte. Este nodo constituye, por tanto, el principal cuello de botella del modelo jerárquico seleccionado. Cabe destacar además que su entrenamiento se detuvo tempranamente, en la época 16, frente a los nodos con mejor desempeño, ya que el modelo no encontró una mejora sostenida en validación debido a la falta de datos representativos o alta similitud acústica entre clases. El aumento de datos puede mitigar este efecto, pero no sustituye la necesidad de mayor diversidad real de grabaciones.

Como respaldan las Tablas 13, 14 y 15, el modelo CNN jerárquico seleccionado constituye la mejor opción entre las arquitecturas evaluadas por tres razones: supera claramente a la CNN plana en métricas globales, en particular en F1-Macro, confirmando que la estructura jerárquica es más adecuada para categorías organizadas por niveles; mejora la evaluación jerárquica preliminar mediante una partición más controlada, agrupada por fuente de audio y estratificada por clase; e incorpora una arquitectura más robusta, con bloques residuales y mecanismos Squeeze-and-Excitation, para aprender representaciones espectro-temporales. No obstante, su desempeño global depende de la calidad de cada nodo: la rama de helicópteros, en particular, requiere mayor cantidad y diversidad de muestras para alcanzar una generalización comparable al resto de la jerarquía.

Análisis comparativo: ML frente a DL

La comparación entre el modelo SVM jerárquico y el modelo CNN jerárquico debe realizarse por nodo y no mediante métricas generales, ya que una comparación global podría mostrar la superioridad de uno de los modelos sin considerar el contexto particular de cada nivel y rama de clasificación. Además, el modelo CNN dispone de una evaluación *end-to-end*, donde una predicción solo se considera correcta si toda la ruta jerárquica es acertada, mientras que el SVM reporta principalmente resultados por nodo mediante validación cruzada agrupada por grabación de origen. Por ello, la comparación más equilibrada consiste en observar el comportamiento local de cada nodo y en qué ramas cada modelo generaliza mejor [29], [70], tal como se muestra en la Tabla 16.

Tabla 16 *Comparación por nodo entre SVM y CNN jerárquico.*

Nodo	Clases evaluadas	F1-Macro CNN	F1-Macro SVM
N1	AEREOS vs TERRESTRES	99.10 %	99.49 %
N2	AUV vs HELICOPTEROS	95.49 %	95.04 %
N2	ORUGA vs RUEDA	99.36 %	96.26 %
N3	8 modelos de dron	100.00 %	80.37 %
N3	Helicópteros	37.18 %	69.09 %
N3	4 modelos de oruga	86.05 %	99.28 %
N3	6 modelos de rueda	87.84 %	86.69 %

Como se observa en la Tabla 16, ningún modelo domina todos los nodos. En las decisiones superiores, ambos presentan un desempeño alto: el SVM obtiene una ligera ventaja en la separación entre AEREOS y TERRESTRES (99.49 % frente a 99.10 % de F1-Macro), diferencia marginal que sugiere que ambos capturan adecuadamente las diferencias acústicas generales, aunque el SVM cuenta con una ventaja metodológica al evaluarse con mayor cantidad de muestras y validación agrupada por fuente de audio, lo que reduce la varianza de la estimación.

La CNN obtiene su mejor resultado en el nodo AUV (100 % frente a 80.37 % del SVM, según la Tabla 16), explicado por que los drones generan patrones espectro-temporales altamente distinguibles —asociados a rotores, modulación temporal y concentración de energía en

bandas específicas— que la CNN captura al aprender filtros directamente sobre espectrogramas, sin depender de características manuales. En esta rama, por tanto, el aprendizaje automático de características favorece claramente a la CNN.

El SVM, en cambio, supera a la CNN en helicópteros (69.09 % frente a 37.18 %) y en oruga (99.28 % frente a 86.05 %), como se observa en la Tabla 16. En helicópteros, la diferencia se explica por la escasez de muestras y la alta similitud acústica entre clases: con datos reducidos o poco diversos, la CNN tiene dificultades para aprender representaciones profundas generalizables, mientras que el SVM funciona mejor sobre características manuales previamente diseñadas; el desbalance de clases además afecta con mayor fuerza a la CNN, reduciendo el recall de las clases menos representadas. En oruga, el resultado indica que las características manuales del SVM (MFCC, Delta, Delta-Delta, PLP, RMS y descriptores espectrales) son suficientemente discriminativas, mientras que la CNN presenta confusiones entre variantes acústicamente cercanas.

La CNN, por su parte, supera al SVM en la separación entre ORUGA y RUEDA (99.36 % frente a 96.26 %) y obtiene una ligera ventaja en el nodo de rueda (87.84 % frente a 86.69 %), de acuerdo con la Tabla 16, lo que sugiere que, con diferencias acústicas suficientemente marcadas y datos más representativos, la CNN aprende patrones espectro-temporales útiles para separar categorías. En el nodo de rueda, sin embargo, la ventaja es moderada y ambos modelos se comportan de forma similar.

La Tabla 17 sintetiza estas observaciones, identificando el modelo con ventaja en cada aspecto comparado y su justificación correspondiente.

Tabla 17 *Resumen interpretativo de ventajas por modelo.*

Aspecto comparado	Modelo con ventaja	Justificación
Clasificación de drones AUV	CNN	Aprende patrones espectro-temporales directamente desde representaciones acústicas multicanal.
Separación AEREOS vs TERRESTRES	SVM marginal	Ambos modelos son muy altos; el SVM obtiene una ligera ventaja y fue evaluado con más cobertura por grupos.
Separación ORUGA vs RUEDA	CNN	La CNN aprende diferencias acústicas generales asociadas a mecanismos de rodadura y oruga.
Clasificación de helicópteros	SVM	Es más estable con pocas muestras y clases acústicamente similares.
Clasificación de vehículos de oruga	SVM	Las características manuales resultan altamente discriminativas para esta rama.
Clasificación de vehículos de rueda	CNN marginal	Ambos modelos son cercanos; la CNN obtiene una ligera ventaja en F1-Macro.
Robustez del peor nodo	SVM	Su peor F1-Macro es superior al peor nodo de la CNN.
Aprendizaje automático de características	CNN	No depende exclusivamente de descriptores manuales; aprende filtros desde espectrogramas.

CNN jerárquica es superior en ramas con patrones espectro-temporales claramente diferenciables, especialmente en drones AUV y en la separación ORUGA-RUEDA, gracias a su capacidad de aprender automáticamente representaciones acústicas desde entradas multicanal (espectrograma de Mel, MFCC, Delta y Delta-Delta). Su desempeño disminuye, en cambio, en ramas con menor cantidad de datos, menor diversidad de grabaciones o mayor similitud acústica entre clases, como en helicópteros. El SVM presenta mayor robustez y estabilidad en estos nodos difíciles, al utilizar características acústicas manuales y un clasificador de margen máximo que generaliza mejor con datos limitados. Por tanto, la CNN resulta más potente para ramas con patrones espectrales diferenciables, mientras que el SVM conserva ventaja donde la disponibilidad y diversidad de datos no permite explotar plenamente una arquitectura profunda.

Capítulo V: Conclusiones y Recomendaciones

Conclusiones

Si bien no contar con una jerarquía natural en los clasificadores ML y DL también produce resultados aceptables, el empleo de la organización jerárquica permitió abordar la identificación acústica de los vehículos de acuerdo con las relaciones existentes entre sus clases y simultáneamente se resuelve el problema de la separación de todos los modelos facilitando una toma de decisiones sucesivas, desde las categorías generales hasta las clases más específicas. Esta descomposición resultó especialmente favorable para el modelo de Aprendizaje Profundo cuyo desempeño mejoró de manera clara respecto de la clasificación plana y del modelo de aprendizaje de máquina el cual fue menos pronunciado, además la estructura jerárquica facilitó la interpretación de los errores y permitió identificar las ramas en las que aumenta la dificultad de clasificación.

La comparación directa entre los modelos mostró que ambos reconocen adecuadamente las diferencias acústicas generales entre vehículos aéreos y terrestres. Las principales diferencias aparecieron en los niveles inferiores de la jerarquía. La DL presentó un mejor comportamiento en la identificación de los modelos AUV y en la separación entre vehículos de oruga y de rueda, mientras que la ML fue más consistente en las ramas correspondientes a helicópteros y modelos específicos de vehículos de oruga. En los vehículos de rueda, el desempeño de ambos modelos fue similar, con una ventaja reducida en el modelo de DL.

Este comportamiento responde a la forma en que cada modelo representa la información acústica. El modelo de DL aprovecha las relaciones espectro-temporales presentes en las entradas multicanal y obtiene mejores resultados cuando las clases exhiben patrones suficientemente diferenciables. EL modelo de ML, en cambio, opera sobre descriptores acústicos previamente calculados y mostró mayor estabilidad en determinadas ramas con menor diversidad de grabaciones o mayor similitud entre clases. Por tanto, los resultados no permiten establecer la superioridad general de un modelo, sino reconocer que su conveniencia depende de las características acústicas y de la disponibilidad de datos en cada nodo.

Las mayores dificultades se concentraron en el tercer nivel jerárquico, donde se distinguen modelos pertenecientes a una misma familia. En estas ramas, las semejanzas mecánicas y acústicas reducen la separación entre las clases, mientras que la limitada cantidad de grabaciones independientes restringe la diversidad disponible para el entrenamiento. Esta situación fue particularmente visible en los helicópteros. En sentido contrario, el resultado obtenido en los modelos AUV demuestra una elevada capacidad de discriminación pese a su reducido número de muestras.

En conjunto, el procesamiento desarrollado demuestra que las señales acústicas contienen información suficiente para diferenciar categorías generales y determinados modelos específicos de vehículos. La integración de la adquisición, el acondicionamiento de las señales, la extracción de características y la clasificación jerárquica respalda la factibilidad conceptual del sistema como medio complementario de vigilancia pasiva abriendo el camino a su factibilidad operativa con algunos matices, pues esta requiere ampliar el conjunto de grabaciones, evaluar el sistema en nuevos escenarios y comprobar su funcionamiento en tiempo real sobre una plataforma física.

La complementariedad observada entre los modelos ML y DL constituye uno de los principales hallazgos del trabajo. En lugar de imponer un único modelo en toda la jerarquía, los resultados plantean la posibilidad de utilizar clasificadores especializados según las condiciones de cada nodo. Una arquitectura híbrida permitiría aprovechar la capacidad de la CNN para aprender patrones espectro-temporales y la estabilidad de la SVM cuando los datos son limitados o las clases presentan mayor similitud, estableciendo una base metodológica para continuar el desarrollo del sistema sin atribuirle un nivel de madurez superior al demostrado experimentalmente.

Recomendaciones y trabajos futuros

Debido a la naturaleza jerárquica del problema y a que los modelos DL y ML demostraron capacidades complementarias en diferentes nodos, se recomienda estudiar la implementación de un clasificador híbrido. Esta propuesta permitiría combinar los descriptores acústicos manuales y la capacidad de clasificación del SVM con las representaciones aprendidas auto-

máticamente por la CNN. En lugar de emplear un único modelo en toda la arquitectura, podría asignarse a cada nodo el clasificador que haya demostrado mayor estabilidad y capacidad de generalización, aprovechando el desempeño superior de cada modelo en las diferentes ramas jerárquicas.

Una vez desarrollado el clasificador híbrido, sería necesario embeberlo en sistemas en chip (*System on Chip*, SoC) de diferentes capacidades de procesamiento. Estas pruebas permitirían determinar los requerimientos mínimos de memoria, potencia de cómputo, consumo energético y tiempo de respuesta necesarios para ejecutar el modelo en condiciones reales. La comparación entre plataformas SoC de distintas capacidades también permitiría comprobar la utilidad práctica del sistema y extrapolar su factibilidad económica para un despliegue a gran escala.

Posteriormente, se recomienda incorporar algoritmos de localización y seguimiento basados en la diferencia de tiempo de llegada (*Time Difference of Arrival*, TDOA). En este caso, la identificación de la plataforma por parte del clasificador establecería el evento de referencia o *clock* a partir del cual cada sensor registraría la marca temporal de llegada de la señal. La diferencia entre estas marcas temporales permitiría calcular el TDOA entre los receptores. Con dos sensores y observaciones sucesivas podría apoyarse la estimación de la dirección de desplazamiento y de la velocidad relativa de la plataforma, mientras que con tres o más sensores correctamente sincronizados y distribuidos sería posible estimar posiciones sucesivas y reconstruir su trayectoria. De esta manera, el sistema integraría la identificación de la plataforma con información sobre su desplazamiento y ubicación.

Referencias

[1] P. K. Atrey, N. C. Maddage y M. S. Kankanhalli, «Audio-based event detection for multimedia surveillance,» en *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2006, págs. 813-816. DOI: 10.1109/ICASSP.2006.1661400

[2] Y. Wang, «A Large-Scale UAV Audio Dataset and Audio-Based UAV Classification Using CNN,» Tesis doct., Purdue University, West Lafayette, IN, USA, 2023. DOI: 10.25394/PGS.23696391

[3] D. Li, I. K. Sethi, N. Dimitrova y T. McGee, «Classification of general audio data for content-based retrieval,» *Pattern Recognition Letters*, vol. 22, n.º 5, págs. 533-544, 2001. DOI: 10.1016/S0167-8655(00)00119-7

[4] G. Tzanetakis y P. Cook, «Musical genre classification of audio signals,» *IEEE Transactions on Speech and Audio Processing*, vol. 10, n.º 5, págs. 293-302, 2002. DOI: 10.1109/TSA.2002.800560

[5] T. Giannakopoulos, «pyAudioAnalysis: An open-source Python library for audio signal analysis,» *PLOS ONE*, vol. 10, n.º 12, e0144610, 2015. DOI: 10.1371/journal.pone.0144610

[6] E. C. Morejón Quito, «La Seguridad Nacional y las Amenazas Híbridas: Los Desafíos Estratégicos de las Fuerzas Armadas en las Fronteras Norte y Sur (2004–2010),» Tesis de mtría., FLACSO Ecuador, Quito, Ecuador, sep. de 2023. dirección: https://hdl.handle.net/10469/20129

[7] J. F. Martínez Bucheli y J. P. Podesta, «Análisis de la Capacidad de las Fuerzas Armadas de Ecuador para Enfrentar Amenazas Híbridas en el Marco de la Doctrina y Leyes Vigentes,» Tesis de mtría., Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador, 2023. dirección: https://repositorio.espe.edu.ec/handle/21000/36494

[8] X. V. Fonseca Meneses y F. E. Sani Cando, «Las amenazas híbridas como amenaza al Estado,» *Revista Academia de Guerra del Ejército Ecuatoriano*, vol. 16, n.º 1, págs. 148-154, 2023. DOI: 10.24133/RCSD.VOL16.N01.2023.11

[9] M. Andrade-Vásquez, «Ampliando el enfoque de la soberanía: Fuerzas Armadas de Ecuador frente al crimen organizado transnacional,» *Estado & Comunes*, vol. 1, n.º 18, págs. 101-120, 2024. DOI: 10.37228/estado_comunes.v1.n18.2024.313

[10] Observatorio Ecuatoriano de Crimen Organizado, «Caracterización del Crimen Organizado en Ecuador,» Pan American Development Foundation, Quito, Ecuador, inf. téc., sep. de 2023. dirección: https://oeco.padf.org/wp-content/uploads/2023/09/Caracterizacion-Crimen-Organizado-Version-corta-V2.pdf

[11] F. Camastra y A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. London, U.K.: Springer, 2008. DOI: 10.1007/978-1-84800-007-0

[12] C. Cortes y V. Vapnik, «Support-vector networks,» *Machine Learning*, vol. 20, n.º 3, págs. 273-297, 1995. DOI: 10.1007/BF00994018

[13] S. Hershey et al., «CNN architectures for large-scale audio classification,» en *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, págs. 131-135. DOI: 10.1109/ICASSP.2017.7952132

[14] K. J. Piczak, «Environmental sound classification with convolutional neural networks,» en *Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, págs. 1-6. DOI: 10.1109/MLSP.2015.7324337

[15] J. Salamon y J. P. Bello, «Deep convolutional neural networks and data augmentation for environmental sound classification,» *IEEE Signal Processing Letters*, vol. 24, n.º 3, págs. 279-283, 2017. DOI: 10.1109/LSP.2017.2657381

[16] 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, págs. 2880-2894, 2020. DOI: 10.1109/TASLP.2020.3030497

[17] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang y T. Sainath, «Deep learning for audio signal processing,» *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, n.º 2, págs. 206-219, 2019. DOI: 10.1109/JSTSP.2019.2908700

[18] D. A. Bies, C. H. Hansen y C. Q. Howard, *Engineering Noise Control*, 5.ª ed. Boca Raton, FL, USA: CRC Press, 2017.

[19] Y.-J. Xing, Z. Qu y W.-B. Shangguan, «Identification of vehicle noise based on transfer path and condition power spectrum analysis,» en *SAE Technical Paper Series*, 2022. DOI: 10.4271/2022-01-0306

[20] A. Wieczorkowska, E. Kubera, T. Słowik y K. Skrzypiec, «Spectral features for audio based vehicle and engine classification,» *Journal of Intelligent Information Systems*, vol. 50, págs. 265-290, 2018. DOI: 10.1007/s10844-017-0459-2

[21] K. W. Cheng et al., «Spectrogram-based classification on vehicles with modified loud exhausts via convolutional neural networks,» *Applied Acoustics*, vol. 205, pág. 109 254, 2023. DOI: 10.1016/j.apacoust.2023.109254

[22] C.-H. Lin et al., «SVM-based sound classification based on MPEG-7 audio LLDs and related enhanced features,» en *Proceedings of the 6th International Conference on Convergence and Hybrid Information Technology (ICHIT)*, ép. Communications in Computer and Information Science, vol. 310, Springer, 2012, págs. 536-543. DOI: 10.1007/978-3-642-32692-9_67

[23] S. Ntalampiras, I. Potamitis y N. Fakotakis, «A multidomain approach for automatic home environmental sound classification,» en *Proceedings of Interspeech 2010*, 2010, págs. 2210-2213. DOI: 10.21437/Interspeech.2010-608

[24] S. Astapov, A. Riid y J.-S. Preden, «Military vehicle acoustic pattern identification by distributed ground sensors,» en *Proceedings of the 15th Biennial Baltic Electronics Conference (BEC)*, 2016, págs. 167-170. DOI: 10.1109/BEC.2016.7743755

[25] Y. Seo, B. Jang y S. Im, «Drone detection using convolutional neural networks with acoustic STFT features,» en *Proceedings of the 15th IEEE International Conference on Advan-*

ced Video and Signal Based Surveillance (AVSS), 2018, págs. 1-6. DOI: 10.1109/AVSS.2018.8639425

[26] S. Al-Emadi, A. Al-Ali, A. Mohammad y A. Al-Ali, «Audio-based drone detection and identification using deep learning,» en *Proceedings of the 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 2019, págs. 459-464. DOI: 10.1109/IWCMC.2019.8766732

[27] S. Al-Emadi, A. Al-Ali y A. Al-Ali, «Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks,» *Sensors*, vol. 21, n.º 15, pág. 4953, 2021. DOI: 10.3390/s21154953

[28] S. Jamil et al., «Malicious UAV detection using integrated audio and visual features for public safety applications,» *Sensors*, vol. 20, n.º 14, pág. 3923, 2020. DOI: 10.3390/s20143923

[29] C. N. Silla y A. A. Freitas, «A survey of hierarchical classification across different application domains,» *Data Mining and Knowledge Discovery*, vol. 22, n.º 1–2, págs. 31-72, 2011. DOI: 10.1007/s10618-010-0175-9

[30] S. J. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 3.ª ed. Upper Saddle River, NJ, USA: Pearson, 2010.

[31] S. Raschka y V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*, 3.ª ed. Birmingham, U.K.: Packt Publishing, 2019.

[32] P. Ponce Cruz, *Inteligencia Artificial con Aplicaciones a la Ingeniería*. Mexico City, Mexico: Alfaomega, 2010.

[33] R. L. Marín Morales y J. T. Palma Méndez, *Inteligencia Artificial: Técnicas, Métodos y Aplicaciones*. Madrid, Spain: McGraw-Hill Interamericana de España, 2008.

[34] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. dirección: https://www.deeplearningbook.org/

[35] Y. LeCun, Y. Bengio y G. Hinton, «Deep Learning,» *Nature*, vol. 521, n.º 7553, págs. 436-444, 2015. DOI: 10.1038/nature14539

[36] Y. LeCun, L. Bottou, Y. Bengio y P. Haffner, «Gradient-based learning applied to document recognition,» *Proceedings of the IEEE*, vol. 86, n.º 11, págs. 2278-2324, 1998. DOI: 10.1109/5.726791

[37] B. Gold, N. Morgan y D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 2.ª ed. Hoboken, NJ, USA: Wiley, 2011. DOI: 10.1002/9781118142882

[38] *Codificación Básica: Manual de Mantenimiento del Material Blindado y Mecanizado*, Comando de Educación y Doctrina del Ejército, Ejército Ecuatoriano, Quito, Ecuador, 2019.

[39] *Illustrated Parts Catalogue: Otokar COBRA II 4x4 Tactical Armored Vehicle*, Otokar Otomotiv ve Savunma Sanayi A.Ş., 2023.

[40] *Manual de Mantenimiento para Transporte Blindado de Personal COBRA II 4x4*, Otokar Otomotiv ve Savunma Sanayi A.Ş., 2023.

[41] *Standard Interfaces of Unmanned Aircraft Control System (UCS) for NATO Unmanned Aircraft Interoperability*, Brussels, Belgium: NATO Standardization Office, abr. de 2017.

[42] *Unmanned Aircraft Systems Airworthiness Requirements (USAR)*, Brussels, Belgium: NATO Standardization Office, abr. de 2019.

[43] U.S. Department of the Navy, «The Navy Unmanned Undersea Vehicle (UUV) Master Plan,» United States Navy, Washington, DC, USA, inf. téc. ADA511748, nov. de 2004. dirección: https://apps.dtic.mil/sti/pdfs/ADA511748.pdf

[44] U.S. Department of the Navy, «The Navy Unmanned Surface Vehicle (USV) Master Plan,» United States Navy, Washington, DC, USA, inf. téc. ADA504867, jul. de 2007.

[45] A. V. Oppenheim y R. W. Schafer, *Discrete-Time Signal Processing*, 3.ª ed. Upper Saddle River, NJ, USA: Pearson, 2010.

[46] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Cham, Switzerland: Springer, 2015. DOI: 10.1007/978-3-319-21945-5

[47] T. Giannakopoulos y A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*. Oxford, U.K.: Academic Press, 2014.

[48] J. B. Allen, «Short-term spectral analysis, synthesis, and modification by discrete Fourier transform,» *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, n.º 3, págs. 235-238, 1977. DOI: 10.1109/TASSP.1977.1162950

[49] L. R. Rabiner y R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.

[50] S. B. Davis y P. Mermelstein, «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,» *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n.º 4, págs. 357-366, 1980. DOI: 10.1109/TASSP.1980.1163420

[51] B. Logan, «Mel frequency cepstral coefficients for music modeling,» en *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[52] G. Peeters, «A Large Set of Audio Features for Sound Description: Similarity and Classification in the CUIDADO Project,» IRCAM, Paris, France, Project Report, 2004. dirección: https://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf

[53] S. Dubnov, «Generalization of spectral flatness measure for non-Gaussian linear processes,» *IEEE Signal Processing Letters*, vol. 11, n.º 8, págs. 698-701, 2004. DOI: 10.1109/LSP.2004.831663

[54] H. Hermansky, «Perceptual linear predictive (PLP) analysis of speech,» *The Journal of the Acoustical Society of America*, vol. 87, n.º 4, págs. 1738-1752, 1990. DOI: 10.1121/1.399423

[55] H. He y E. A. Garcia, «Learning from imbalanced data,» *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, n.º 9, págs. 1263-1284, 2009. DOI: 10.1109/TKDE.2008.239

[56] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe y S. Misra, «Data augmentation and deep learning methods in sound classification: A systematic review,» *Electronics*, vol. 11, n.º 22, pág. 3795, 2022. DOI: 10.3390/electronics11223795

[57] E. Sprengel, M. Jaggi, Y. Kilcher y T. Hofmann, «Audio-based bird species identification using deep learning techniques,» en *Working Notes of CLEF 2016—Conference and Labs of the Evaluation Forum*, ép. CEUR Workshop Proceedings, vol. 1609, 2016, págs. 547-559. dirección: https://ceur-ws.org/Vol-1609/16090547.pdf

[58] A. L. Sarris, N. Vryzas, L. Vrysis y C. Dimoulas, «Investigation of data augmentation techniques in environmental sound recognition,» *Electronics*, vol. 13, n.º 23, pág. 4719, 2024. DOI: 10.3390/electronics13234719

[59] L. Ferreira-Paiva, E. Alfaro-Espinoza, V. M. Almeida, L. B. Felix y R. V. A. Neves, «A survey of data augmentation for audio classification,» *Anais do Congresso Brasileiro de Automática*, vol. 3, n.º 1, 2022. DOI: 10.20906/CBA2022/3469

[60] D. S. Park et al., «SpecAugment: A simple data augmentation method for automatic speech recognition,» en *Proceedings of Interspeech 2019*, 2019, págs. 2613-2617. DOI: 10.21437/Interspeech.2019-2680

[61] H. Zhang, M. Cissé, Y. N. Dauphin y D. Lopez-Paz, «mixup: Beyond empirical risk minimization,» en *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[62] K. He, X. Zhang, S. Ren y J. Sun, «Deep residual learning for image recognition,» en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, págs. 770-778. DOI: 10.1109/CVPR.2016.90

[63] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[64] T. Hastie, R. Tibshirani y J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2.ª ed. New York, NY, USA: Springer, 2009. DOI: 10.1007/978-0-387-84858-7

[65] C.-W. Hsu, C.-C. Chang y C.-J. Lin, «A Practical Guide to Support Vector Classification,» Department of Computer Science, National Taiwan University, Taipei, Taiwan, inf. téc., 2016. dirección: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[66] V. Nair y G. E. Hinton, «Rectified linear units improve restricted Boltzmann machines,» en *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, págs. 807-814.

[67] S. Ioffe y C. Szegedy, «Batch normalization: Accelerating deep network training by reducing internal covariate shift,» en *Proceedings of the 32nd International Conference on Machine Learning*, ép. Proceedings of Machine Learning Research, vol. 37, PMLR, 2015, págs. 448-456.

[68] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever y R. Salakhutdinov, «Dropout: A simple way to prevent neural networks from overfitting,» *Journal of Machine Learning Research*, vol. 15, n.º 56, págs. 1929-1958, 2014.

[69] J. Hu, L. Shen y G. Sun, «Squeeze-and-excitation networks,» en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, págs. 7132-7141. DOI: 10.1109/CVPR.2018.00745

[70] M. Sokolova y G. Lapalme, «A systematic analysis of performance measures for classification tasks,» *Information Processing & Management*, vol. 45, n.º 4, págs. 427-437, 2009. DOI: 10.1016/j.ipm.2009.03.002

[71] F. Pedregosa et al., «Scikit-learn: Machine learning in Python,» *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011. dirección: https://www.jmlr.org/papers/v12/pedregosa11a.html

[72] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea y S. Poria, «A review of deep learning techniques for speech processing,» *Information Fusion*, vol. 99, pág. 101 869, 2023. DOI: 10.1016/j.inffus.2023.101869

[73] R. Kohavi, «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,» en *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, 1995, págs. 1137-1143.

[74] S. Wei, S. Zou, F. Liao y W. Lang, «A comparison on data augmentation methods based on deep learning for audio classification,» *Journal of Physics: Conference Series*, vol. 1453, n.º 1, pág. 012 085, 2020. DOI: 10.1088/1742-6596/1453/1/012085

[75] D. M. W. Powers, «Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,» *Journal of Machine Learning Technologies*, vol. 2, n.º 1, págs. 37-63, 2011.

[76] A. Tharwat, «Classification assessment methods,» *Applied Computing and Informatics*, vol. 17, n.º 1, págs. 168-192, 2021. DOI: 10.1016/j.aci.2018.08.003

[77] C. P. Bernal Oñate, «Time-Frequency Analysis, Audio-Based Features, and Low-Dimensional Manifold Latent Spaces in Cardiac and Seismographic Signals,» Tesis doct., Universidad Rey Juan Carlos, Madrid, Spain, 2024. dirección: https://hdl.handle.net/10115/32955

Lista de acrónimos

A/D: Analog-to-Digital (Conversión Analógico-Digital)

AAC: Advanced Audio Coding (Codificación de Audio Avanzada)

AGV: Autonomous Ground Vehicle (Vehículo Terrestre Autónomo)

AUV: Aerial Unmanned Vehicle (Vehículo Aéreo No Tripulado)

CNN: Convolutional Neural Network (Red Neuronal Convolucional)

dB: Decibel (Decibelio)

DL: Deep Learning (Aprendizaje Profundo)

F1: Medida F1, media armónica entre precisión y recall

FLAC: Free Lossless Audio Codec (Códec de Audio Libre sin Pérdida)

FN: False Negative (Falso Negativo)

FP: False Positive (Falso Positivo)

Hz: Hertz (Hercio)

IA: Inteligencia Artificial

M4A: MPEG-4 Audio (Audio MPEG-4)

MFCC: Mel-Frequency Cepstral Coefficients (Coeficientes Cepstrales en Frecuencia Mel)

ML: Machine Learning (Aprendizaje Automático)

MP3: MPEG-1 Audio Layer III (Capa de Audio 3 de MPEG-1)

PCM: Pulse Code Modulation (Modulación por Codificación de Pulsos)

PLP: Perceptual Linear Prediction (Predicción Lineal Perceptual)

RBF: Radial Basis Function (Función de Base Radial)

ReLU: Rectified Linear Unit (Unidad Lineal Rectificada)

ResNet: Residual Network (Red Residual)

RMS: Root Mean Square (Raíz Media Cuadrática)

SE: Squeeze-and-Excitation (Compresión y Excitación)

SoC: System on Chip (Sistema en Chip)

STFT: Short-Time Fourier Transform (Transformada de Fourier de Tiempo Corto)

SVM: Support Vector Machine (Máquina de Vectores de Soporte)

TDOA: Time Difference of Arrival (Diferencia de Tiempo de Llegada)

TN: True Negative (Verdadero Negativo)

TP: True Positive (Verdadero Positivo)

UGV: Unmanned Ground Vehicle (Vehículo Terrestre No Tripulado)

WAV: Waveform Audio File Format (Formato de Archivo de Audio de Onda)

ZCR: Zero Crossing Rate (Tasa de Cruces por Cero)