# Clasificador de sonidos que indiquen una alerta o amenaza para las personas con discapacidad auditiva

J. Villalobos Tenorio[1], J. Gil González[2]

29 de julio del 2024

[1]*Pontificia Universidad Javeriana Cali, Colombia*
[2]*Pontificia Universidad Javeriana Cali, Colombia*

## Resumen

Este proyecto se centra en entrenar modelos de aprendizaje automático para clasificar algunos sonidos del conjunto de datos *AudioSet* de Google, seleccionados por su cantidad de muestras y relevancia para indicar alertas o amenazas. Se documenta el proceso de entrenamiento y los obstáculos encontrados, con el objetivo de, en un trabajo futuro, implementar estos modelos en dispositivos móviles para ayudar a personas con discapacidad auditiva a identificar sonidos de su entorno. Para el desarrollo, se generaron espectrogramas de los sonidos y se entrenaron varios modelos utilizando *transfer learning*, comparando su desempeño con diversas métricas.

## 1. Introducción

El uso de los sentidos es fundamental para la interacción diaria con el entorno, y entre ellos, la audición juega un papel crucial al permitir la detección de sonidos que pueden indicar situaciones de riesgo o amenaza. Sin embargo, las personas con pérdida auditiva enfrentan una desventaja significativa, aumentando el riesgo de sufrir accidentes, generando mayores costos económicos, dificultando el aprendizaje, reduciendo las oportunidades laborales y, en última instancia, deteriorando su calidad de vida. Ante esta problemática, surge la necesidad de desarrollar soluciones alternativas y accesibles. Una opción viable es un sistema computacional basado en aprendizaje de máquina que pueda describir textualmente eventos o señales sonoras indicativas de riesgo, utilizando el micrófono de un smartphone. Este sistema no solo reduciría la vulnerabilidad de las personas con discapacidad auditiva, sino que también podría ser una herramienta educativa para aquellos que están en proceso de aprender a interpretar sonidos. Por lo tanto, este proyecto se centra en entrenar modelos de aprendizaje automático que aporten a la detección de los sonidos que indiquen una alerta o amenaza para las personas con esta discapacidad.

## 2. Objetivos

El objetivo general de esta investigación es desarrollar un modelo de clasificación de sonidos a partir de técnicas de aprendizaje automático orientado a apoyar a las personas con discapacidad auditiva. En cuanto a los objetivos específicos se incluye:

- Desarrollar un módulo de recolección y procesamiento de datos que permita extraer y preparar sonidos desde una base de datos existente para su clasificación.
- Implementar un modelo de clasificación a partir de técnicas de aprendizaje automático que permita la identificación de sonidos.
- Desarrollar una estrategia de búsqueda de hiperparámetros con el fin de maximizar el rendimiento del modelo entrenado.

- Realizar la validación de los modelos a través de métricas de desempeño según los algoritmos empleados.

# 3. Fundamentación Teórica

La pérdida auditiva se refiere a las personas que no logran escuchar al igual que quienes no tienen esta condición, es decir, que los umbrales de audición están por encima de 20 dB en ambos oídos [1], por lo cual no escuchan sonidos que se encuentren por debajo de 20 decibelios, cuando lo normal es tener el umbral auditivo en 0 dB, siendo este el nivel mínimo de presión sonora que el oído humano es capaz de percibir.

Según el *Servicio Nacional de Salud del Reino Unido* [2], entre los tratamientos existentes para la pérdida auditiva permanente, se encuentran los audífonos e implantes, estos primeros no restablecen por completo la audición, pero pueden ayudar a amplificar los sonidos para que sean más fuertes y claros. En el caso de los implantes, se realiza una cirugía para adherir el dispositivo al cráneo o colocarlo en lo profundo del oído. Es importante aclarar que en algunos casos de pérdida auditiva, como en los que se tiene daño del nervio auditivo después del oído medio, los implantes o audífonos no benefician a las personas con esta condición [3].

El impacto de la perdida auditiva puede ser muy profundo en la vida de las personas, según la OMS [1] puede causar la pérdida de la habilidad para comunicarse con los demás. En el caso de los niños retrasa el desarrollo del lenguaje, ocasionando diversas dificultades como problemas de salud mental y problemas económicos. También es importante saber que para el 2050 se estima más de 700 millones de personas con discapacidad auditiva o pérdida auditiva discapacitante, la cual se define cuando el umbral de la audición inicia después de los 35 decibelios en el mejor oído. Cerca del 80 % de las personas con pérdida auditiva discapacitante se encuentra en países de ingresos medios y bajos. La pérdida auditiva por lo general aumenta con la edad, más del 25 % de las personas mayores de 60 años son afectadas por la pérdida auditiva discapacitante.

Ahora, en cuanto a los espectrogramas, son una herramienta utilizada para visualizar cómo cambian las frecuencias en una señal de audio a lo largo del tiempo [4]. Para generar un espectrograma, se toman múltiples Transformadas Discretas de Fourier (DFT) de pequeños fragmentos de la señal de audio, y los espectros resultantes se organizan uno tras otro. Esto permite que el espectrograma muestre las frecuencias del audio a través del tiempo, mostrando en una sola gráfica la información de tiempo, frecuencia y amplitud. El algoritmo que permite hacer este representación se conoce como STFT o Transformada de Fourier de tiempo reducido.

Un espectrograma en escala de mel, o espectrograma mel es una variante de los espectrogramas que se utiliza comúnmente en tareas de aprendizaje automático. Es similar al espectrograma en la información que contiene, pero su eje de frecuencia es diferente. A diferencia del espectrograma estándar, se añade un paso más en su generación, pues cada espectro de frecuencia generado con la STFT pasa luego por unos filtros para transformar las frecuencias a la escala mel. La escala mel es una escala no lineal que permite acercar las frecuencias a las que puede percibir el sistema auditivo humano, pues este es más sensible a los cambios en frecuencias bajas que a los cambios en las frecuencias altas.

Un enfoque común y altamente efectivo para el aprendizaje profundo en pequeños conjuntos de datos de imágenes es utilizar una red pre-entrenada. Una red pre-entrenada es una red guardada que se entrenó previamente en un conjunto de datos grande, típicamente en una tarea de clasificación de imágenes a gran escala. Si este conjunto de datos original es lo suficientemente grande y general, el modelo o red pre-entrenada puede considerarse un modelo genérico del mundo visual, y por lo tanto, sus características pueden ser útiles para muchos problemas diferentes de clasificación de imágenes, aunque estos nuevos problemas puedan involucrar clases completamente diferentes a las de la tarea original. [5]

El uso de modelos pre-entrenados para otras tareas de clasificación en el mismo dominio, se conoce como *transfer learning*, y la base de datos comúnmente utilizada para entrenar estos grandes modelos en la clasificación de imágenes, se conoce como *ImageNet*.

# 4. Resultados

Se obtuvieron las muestras de 9 clases del conjunto de datos *AudioSet* [6], se generaron los espectrogramas mel por cada una de las muestras, se extrajeron las características según el modelo de *transfer learning* para cada muestra, y se entrenaron diversos modelos que fueron evaluados y optimizados en dos fases. A continuación, se presentan los detalles de estos procedimientos.

## 4.1. Obtención del conjunto de datos

Para poder cumplir con la tarea de entrenar los modelos de aprendizaje automático, fue necesario obtener el conjunto de datos, este se obtuvo a través de una herramienta [7] que permite la descarga de cada una de las muestras por clase, con esta, se lograron obtener las siguientes muestras relevantes para identificar una alerta o amenaza:

| **Clase** | **Número de Muestras** |
|---|---|
| Emergency Vehicle | 1000 |
| Explosion | 974 |
| Gunshot | 1000 |
| Power Tool | 1000 |
| Music | 200 |
| Bell | 200 |
| Speech | 200 |
| Silence | 200 |
| Sneeze | 200 |
| **Total** | **4974** |

Cuadro 1: Número de muestras obtenidas por clase

Cada una de las muestras se encuentra representada como un fragmento de audio en formato *.wav* de 10 segundos conteniendo el sonido según la clase. Las clases principales son «Emergency Vehicle», «Explosion», «Gunshot» y «Power Tool». La quinta clase es la clase «Other» que se compone de las clases «Music», «Bell», «Speech», «Silence» y «Sneeze». Esta clase compuesta se construye para ayudar a la generalización del modelo en la clasificación.

## 4.2. Generación de espectrogramas mel

Luego de tener cada una de las clases organizadas por carpetas según su clase y cantidad de muestras descargadas, se desarrolla una herramienta [9] para transformar los fragmentos de audio en los espectrogramas mel. Esto se logró utilizando el módulo *Librosa* [8] de *Python 3*, el cual se especializa en análisis de música y audio.

A continuación se presentan una muestra aleatoria en espectrograma mel de la clase «Emergency Vehicle».

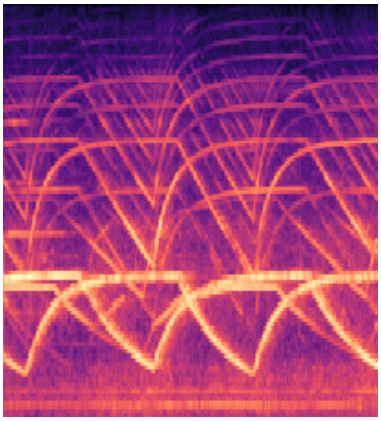

Figura 1: Muestra de espectrograma mel para la clase «Emergency Vehicle».

## 4.3. Resultados primera fase de optimización

Tras entrenar un modelo base con la extracción de características de cada uno de los modelos, se procede a la primera fase de optimización, donde se utiliza la técnica *Random Search* para encontrar los mejores hiperparámetros en cada modelo. En la siguiente figura se puede observar cómo se realizó la extracción de características.

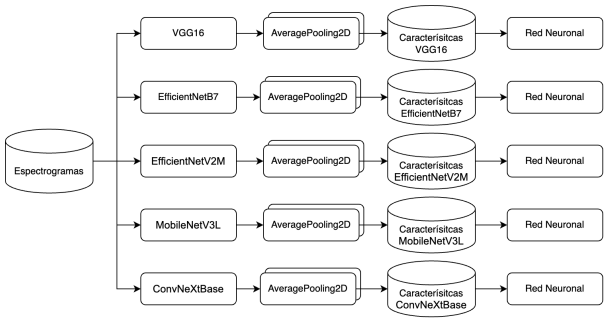


Figura 2: Extracción de características de los espectrogramas.

En esta primera fase se obtienen los siguientes resultados para la métrica de *validation categorical accuracy* que mide la precisión del modelo en el conjunto de datos de validación:

| **Modelo Base** | **Val. Categ. Accuraccy** |
|---|---|
| VGG16 | 70.65 % |
| EfficientNetV2M | 73.27 % |
| EfficientNetB7 | 75.08 % |
| ConvNeXtBase | 75.98 % |
| MobileNetV3L | 76.68 % |

Cuadro 2: Resultados de los modelos primera iteración

### 4.4. Resultados segunda fase de optimización

Debido a los resultados positivos en la métrica de validación de los modelos entrenados con los pesos de *EfficientNetB7*, *ConvNeXtBase* y *MobileNetV3L* se decide utilizar estos mismos para obtener una mejor versión donde mejore el validation categorical accuracy y la estabilidad de los modelos. A continuación se presentan los resultados obtenidos para las métricas *Validation Categorical Accuracy* (renombrado V.C.A. en la tabla 3) y *Macro ROC-AUC* donde esta última mide la capacidad del modelo para distinguir entre clases.

| **Modelo Base** | **V.C.A.** | **ROC-AUC** |
|---|---|---|
| EfficientNetB7 | 75.48 % | 0.9318 |
| ConvNeXtBase | 76.78 % | 0.9424 |
| MobileNetV3L | 77.09 % | 0.9432 |

Cuadro 3: Resultados de los modelos segunda iteración

Como se observa en la tabla 3, se obtuvo una mejora en el rendimiento de la métrica *Validation Categorical Accuracy* con respecto a la primera fase. En cuanto a la métrica *ROC-AUC*, se observan buenos resultados, ya que cuanto más cercano sea el valor a 1 en esta métrica, mejor es la capacidad del modelo para distinguir entre las clases.

## 5. Discusión y Conclusiones

El objetivo final de este trabajo fue entrenar modelos de aprendizaje automático capaces de clasificar clases específicas del conjunto de datos AudioSet de Google. Estas clases fueron elegidas para asistir a personas con discapacidad auditiva, con la idea de identificar alertas y amenazas en su entorno.

Es importante resaltar que uno de los principales retos, fue adaptar el conjunto de datos a un formato de imágenes, pues el proceso es costoso computacionalmente. Aunque se logró encontrar y utilizar una herramienta para la descarga de las muestras de audio [7], fue necesario desarrollar herramientas adicionales para la transformación de esas muestras a imágenes [9].

Uno de los problemas que se logró identificar durante el desarrollo, fue la dificultad para utilizar *data augmentation*, esto debido a que aunque aumentar los datos de un conjunto de imágenes no es complicado si se realizan rotaciones, reflejos, entre otras, en este caso no era válido realizar este tipo de operaciones sobre las imágenes porque representan una gráfica en donde importa la dirección. Es decir, si tenemos una imagen de un gato y la reflejamos sobre uno de sus ejes, sigue siendo la imagen de un gato, pero si reflejamos un espectrograma que representa el sonido de una de nuestras clases, ya no se tendría una representación válida de estos sonidos sino de un sonido

completamente diferente, lo que puede influir negativamente en el modelo. Una posible solución a este problema, es aplicar transformaciones a las muestras de audio originales antes de transformarlas en imágenes, por ejemplo añadir ruido, cambiar la frecuencia, etc., generando así, muestras adicionales válidas.

En cuanto al rendimiento de los modelos, se podría decir que puede ser mejorable. Este fue un trabajo de exploración y como la idea es entrenar un modelo capaz de identificar los sonidos sin ser este muy costoso computacionalmente, funcionaría mejor centrarse en entrenar el modelo *MobileNetV3L*, que según la revisión de literatura es el mejor para este tipo de tareas con recursos limitados [10]. Además, este fue el modelo con el que se alcanzó mejor rendimiento general de los entrenados en esta investigación.

# Referencias


[1] Deafness and hearing loss, en. dirección: https://www.who.int/health-topics/hearing-loss (visitado 02-04-2023).

[2] Hearing loss, en, oct. de 2017. dirección: https://www.nhs.uk/conditions/hearing-loss/ (visitado 02-04-2023)

[3] A. Pavlidou y B. Lo, «Artificial ear - a wearable device for the hearing impaired,» en 2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN), Athens, Greece: IEEE, jul. de 2021, págs. 1-4, isbn: 9781665403627. doi: 10.1109/BSN51625.2021.9507021. dirección: https://ieeexplore.ieee.org/document/9507021/ (visitado 22-03-2023).

[4] Introducción a los datos de audio - Hugging Face Audio Course — huggingface.co, https://huggingface.co/learn/audio-course/es/chapter1/audio_data, (visitado 15-06-2024).

[5] F. Chollet, Deep learning with Python, Second edition. Shelter Island: Manning Publications, 2021, OCLC: on1289290141, isbn: 9781617296864.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman et al., «Audio Set: An ontology and human-labeled dataset for audio events,» en 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA: IEEE, mar. de 2017, págs. 776-780, isbn: 9781509041176. doi:10.1109 / ICASSP.2017.7952261. dirección: http://ieeexplore.ieee.org/document/7952261/ (visitado 22-03-2023)

[7] A. McDonagh y J. Villalobos, Jeremias-V/audioset-processing. 22 de oct. de 2023. dirección: https://github.com/Jeremias-V/audioset-processing.

[8] B. McFee, M. McVicar, D. Faronbi et al., librosa/librosa: 0.10.2.post1, ver. 0.10.2.post1, mayo de 2024. doi: 10.5281/zenodo.11192913. dirección: https://doi.org/10.5281/zenodo.11192913.

[9] J. Villalobos, Jeremias-V/audioset_thesis. 22 de oct. de 2023. dirección: https://github.com/Jeremias-V/audioset_thesis.

[10] A. Howard, M. Sandler, B. Chen et al., «Searching for MobileNetV3,» en 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, págs. 1314-1324. doi: 10.1109/ICCV.2019.00140.