TRABAJO FIN DE GRADO

ESCUELA UNIVERSITARIA POLITÉCNICA

Grado en Ingeniería Informática

CLASIFICADOR AUTOMÁTICO DE GÉNEROS MUSICALES

BASADO EN REDES NEURONALES

Autor: Nombre Apellido

Director: Nombre Director

Murcia, mes de 2025

Nombre Apellido

RESUMEN

Este Trabajo Fin de Grado presenta el diseño y desarrollo de un **clasificador automático de géneros musicales** basado en **redes neuronales**. El sistema sigue una canalización en **tres partes**: (1) **entrada** de un archivo de audio; (2) **análisis** mediante un modelo de IA que infiere el **género musical**; y (3) **salida** en pantalla con la **predicción**. El desarrollo se estructura en **tres fases**: **preprocesamiento** de datasets de audio (normalización, remuestreo y **extracción de características** como **MFCC** y **espectrogramas log-mel**), **entrenamiento** del modelo (selección de hiperparámetros, regularización y validación) y **inferencia** sobre **audios externos** replicando el mismo preprocesado.

Se emplean **datasets públicos de referencia** (p. ej., GTZAN) y se evalúa el rendimiento con métricas estándar (**accuracy**, **F1 macro**) bajo particiones de entrenamiento/validación/prueba. Las contribuciones principales son: (i) una **canalización reproducible** de preprocesamiento–entrenamiento–inferencia; (ii) un **clasificador neuronal** para audio musical; y (iii) una **herramienta de uso** que, dada una pista de audio, devuelve el **género** estimado. Se discuten limitaciones (sesgo de datos, ambigüedad de etiquetas) y vías futuras (mejoras de datos, arquitectura y robustez).

Palabras clave: clasificación musical; géneros musicales; aprendizaje profundo; redes neuronales; MFCC; espectrograma log-mel; audio digital.

Nombre Apellido

ABSTRACT

This Bachelor's Thesis presents the design and development of an **automatic music-genre classifier** based on **neural networks**. The system implements a three-part pipeline: (1) **input** of an audio file; (2) **analysis**, where an AI model infers the **music genre**; and (3) **output** of the predicted label. The project is carried out in **three phases**: **audio dataset preprocessing** (normalization, resampling, and **feature extraction** such as **MFCCs** and **log-mel spectrograms**), **model training** (hyperparameter tuning, regularization, and validation), and **inference** on **external audio** with identical preprocessing.

We rely on **public benchmark datasets** (e.g., GTZAN) and report performance using standard metrics (**accuracy**, **macro-F1**) under train/validation/test splits. The main contributions are: (i) a **reproducible pipeline** spanning preprocessing–training–inference; (ii) a **neural-network-based audio classifier**; and (iii) a **user-facing tool** that returns the estimated **music genre** for a given track. Limitations (data bias, label ambiguity) and future work (data, architecture, robustness) are discussed.

Keywords: music classification; music genres; deep learning; neural networks; MFCC; log-mel spectrogram; digital audio.

Nombre Apellido

1. INTRODUCCIÓN

La clasificación automática de **géneros musicales** es un problema con impacto directo en la **organización**, la **búsqueda** y la **recomendación** de contenidos sonoros. La industria musical y el consumo digital han crecido hasta niveles en los que el etiquetado manual resulta **costoso**, **lento** y propenso a **inconsistencias**: distintos catalogadores pueden asignar géneros diferentes a una misma pista, y mantener esa coherencia a gran escala es complejo. Al mismo tiempo, los avances en *aprendizaje profundo* han demostrado que es posible aprender patrones sonoros que ayudan a **automatizar** esta tarea con criterios reproducibles y evaluables.

Este Trabajo Fin de Grado (TFG) presenta el diseño, desarrollo y evaluación de un **clasificador de géneros musicales** basado en **redes neuronales**. La idea central es sencilla de enunciar: dada una pista de audio, la herramienta **estima a qué género pertenece**. Para lograrlo, se parte de una representación del sonido que resulte informativa para el modelo y se comprueba su rendimiento con métricas claras y comprensibles. El enfoque busca un equilibrio entre **rigor técnico** y **simplicidad de uso**, de manera que el sistema sea útil como punto de partida para futuros desarrollos y, a la vez, suficientemente transparente para entender cómo y por qué ofrece una respuesta.

1.1. Motivación

Este trabajo nace de una combinación de necesidades prácticas y oportunidades tecnológicas:

- **Escalabilidad y eficiencia.** El volumen actual de catálogos musicales impide mantener un etiquetado manual con tiempos razonables. Un sistema automático reduce el esfuerzo y acelera la organización.

- **Consistencia.** Un procedimiento sistemático ayuda a disminuir la variabilidad entre catalogaciones, facilita la **auditoría** del proceso y permite repetir evaluaciones en el tiempo.

- **Utilidad directa.** La etiqueta de género es un atributo básico para **filtrado**, **descubrimiento** y generación de **listas**; disponer de ella de forma automática mejora la experiencia de consulta y análisis.

- **Madurez del ecosistema.** Existen librerías de **procesamiento de audio** y marcos de **deep learning** suficientemente robustos como para construir soluciones con buen balance entre rendimiento y mantenibilidad.

1.2. Contexto y definición del problema

El problema que se aborda es el de **asignar a cada pista de audio una etiqueta de género** perteneciente a un conjunto finito definido por los datos de referencia. Se trata de un caso clásico de **clasificación supervisada**: a partir de ejemplos anotados, el sistema aprende a relacionar **patrones sonoros** con **categorías musicales**.

La resolución de la tarea descansa en tres ideas prácticas. Primero, el audio debe transformarse en **representaciones** que capturen la información relevante para distinguir estilos (por ejemplo, descriptores espectrales habituales en la literatura). Segundo, sobre esas representaciones se entrena un **modelo neuronal** capaz de identificar regularidades y proponer una etiqueta. Y tercero, la calidad de la solución debe **medirse** con datos que el modelo no ha visto durante el entrenamiento y con **métricas estándar**, de forma que el resultado sea comparable y entendible.

A lo largo del trabajo se han tenido en cuenta aspectos que influyen de manera directa en el rendimiento y la validez de las conclusiones: la **homogeneidad** en el tratamiento del audio, la **separación estricta** de los conjuntos de entrenamiento, validación y prueba, y la **documentación** de las decisiones adoptadas. También se reconocen las **limitaciones** propias del dominio, como la posible **ambigüedad** entre géneros cercanos o el **desequilibrio** de clases en los conjuntos de datos disponibles. Estas consideraciones no impiden avanzar; ayudan a interpretar mejor los resultados y a orientar mejoras futuras.

1.3. Objetivos

Objetivo general

Construir y comprobar el funcionamiento de una herramienta que, a partir de una canción, **indique automáticamente su género musical**, utilizando **redes neuronales** y un tratamiento del audio coherente con la literatura actual.

Objetivos específicos

1. **Preparar** un conjunto de datos de trabajo **coherente y limpio**, aplicando el mismo criterio de tratamiento del audio para todas las pistas (niveles, muestreo y duración mínima).

2. **Elegir y justificar** una **representación del audio** sencilla y eficaz para esta tarea, dejando claros los parámetros para poder ajustarlos con facilidad cuando sea necesario.

3. **Diseñar y entrenar** un **modelo neuronal** que aprenda a distinguir los géneros definidos, incorporando buenas prácticas para evitar sobreajuste y validar el progreso.

4. **Evaluar** el comportamiento con un **conjunto de prueba independiente** y con **métricas comprensibles** (por ejemplo, precisión y F1 macro), **explicando** los errores más frecuentes y sus posibles causas.

5. **Presentar** una **aplicación de uso sencillo** que reciba una pista y devuelva el género estimado, aplicando el mismo tratamiento del audio que se usó en el entrenamiento.

6. **Documentar** las decisiones y dependencias para que el trabajo sea **reproducible** y fácil de mantener, de modo que pueda **crecer** con más datos o con nuevas configuraciones sin empezar de cero.

7. **Dejar una base escalable**: que resulte **fácil incorporar más datos** y, si se desea, **ampliar el número de géneros** sin rehacer el proyecto desde cero.

1.4. Estructura del documento

El documento se organiza del siguiente modo:

- **Capítulo 2 — Estado del arte.** Se revisan los conceptos básicos del dominio musical y de IA, las **representaciones** de audio más utilizadas, los principales **enfoques de clasificación** y los **conjuntos de datos** de referencia, justificando las elecciones realizadas.

- **Capítulo 3 — Metodología.** Se describen las metodologías de desarrollo consideradas y se argumenta la elección aplicada al proyecto.

- **Capítulo 4 — Tecnologías y herramientas.** Se detallan los **entornos**, **librerías** y **recursos** utilizados para el desarrollo y las pruebas.

- **Capítulo 5 — Estimación de recursos y planificación.** Se presenta la **planificación temporal** y la estimación del **esfuerzo**.

- **Capítulo 6 — Desarrollo del proyecto.** Se documenta el proceso completo: preparación de datos, configuración del **modelo**, **entrenamiento** y resultados intermedios.

- **Capítulo 7 — Despliegue y pruebas.** Se recoge el **plan de pruebas**, la verificación del funcionamiento y consideraciones de **mantenimiento**.

- **Capítulo 8 — Conclusiones.** Se resumen los **resultados**, se discuten **limitaciones** y se proponen **líneas futuras**.

2. ESTUDIO DE MERCADO

Este capítulo sitúa el proyecto en el **contexto actual** de la clasificación automática de géneros musicales. Se establecen los **conceptos clave** del dominio, se revisan las **representaciones de audio** y los **enfoques de clasificación** más utilizados, y se enmarca el **uso de datos y métricas** habituales para valorar soluciones. Con ello se ofrece una base objetiva para entender *qué se conoce*, *qué limitaciones persisten* y *por qué* el enfoque adoptado resulta adecuado en relación con el estado del arte.

2.1. Conceptos relevantes del dominio de aplicación

La clasificación automática de género musical parte de una premisa sencilla: en el sonido grabado hay huellas del estilo —en cómo se distribuye la energía en frecuencia, en los patrones rítmicos, en la instrumentación y en la producción— que pueden hacerse explícitas mediante representaciones adecuadas y aprenderse con modelos supervisados. Para enmarcar correctamente la tarea, conviene fijar primero cómo se describe el audio digital, qué rasgos musicales se relacionan con los géneros, qué papel juegan los metadatos y bajo qué supuestos se formula el problema de clasificación.

2.1.1. Música digital y señal de audio

El audio musical que se procesa en este trabajo es el resultado de **muestrear** y **cuantizar** una señal analógica. La **frecuencia de muestreo** (p. ej., 44,1 o 48 kHz) establece el techo de frecuencias que podemos representar con fidelidad; la **profundidad de bits** (16–24) condiciona el rango dinámico y el nivel de ruido de cuantización. En la práctica, los conjuntos de datos llegan con tasas y formatos heterogéneos; por coherencia, se convierte todo a una **tasa objetivo** única y se documenta el proceso para que sea repetible.

La organización del material sonoro a lo largo del tiempo exige una unidad de análisis que equilibre detalle y contexto. Por ello se trabaja con **ventanas** breves y solapadas (del orden de decenas de milisegundos) que permiten asumir una cuasiestacionariedad local y describir la señal con transformadas o descriptores bien definidos. A partir de esas ventanas es habitual construir representaciones de mayor alcance (espectrogramas, secuencias de coeficientes) que capturen la evolución temporal.

Antes de cualquier extracción de información, se aplican operaciones que reducen variabilidad ajena al contenido musical: **normalización de nivel** (evitando saturación), **remuestreo** a la tasa objetivo y, cuando procede, **mezcla a mono** para eliminar sesgos debido a panoramizaciones o efectos estéreo que no son relevantes para distinguir géneros. Estas decisiones no persiguen “mejorar” el audio, sino **hacerlo comparable** entre piezas y sesiones de grabación distintas.

2.1.2. Propiedades musicales relevantes para la clasificación

Los géneros no se definen por un único rasgo, sino por **familias de regularidades** que suelen combinarse:

- **Timbre.** Es el "color" del sonido y está íntimamente ligado al **contenido espectral**: envolventes con más energía en agudos, presencia de formantes vocales, armónicos fuertes o saturación propia de determinadas guitarras o sintetizadores. Esta dimensión es la que con más claridad separa instrumentaciones y estéticas de producción.
- **Ritmo y tempo.** La periodicidad, la acentuación y ciertos **patrones rítmicos** caracterizan estilos: baterías programadas con golpes regulares frente a *grooves* sincopados, líneas de bombo marcadas o patrones de caja particulares. El tempo por sí solo rara vez decide, pero refuerza otras pistas.
- **Armonía y tonalidad.** Las relaciones entre alturas y la **distribución de clases de nota** a lo largo del tiempo aportan contexto sobre progresiones y cadencias. Representaciones compactas del contenido armónico (p. ej., **cromas**) ayudan a detectar regularidades de ciertos estilos.

Estas dimensiones se vuelven medibles cuando se proyectan en **dominios de representación** adecuados: la energía distribuida en bandas perceptuales, la evolución temporal de esa energía y resúmenes robustos a pequeñas variaciones de interpretación o mezcla (véase § 2.2).

2.1.3. Metadatos y taxonomías de género

Cualquier sistema supervisado necesita **etiquetas**. En música, esas etiquetas pueden venir de campos **ID3**, anotaciones editoriales o contribuciones comunitarias. Esto introduce particularidades: (i) los géneros son **categorías con fronteras difusas**; (ii) las taxonomías suelen ser **jerárquicas** (género $\rightarrow$ subgénero), y una granularidad excesiva añade ruido; (iii) la **calidad de las etiquetas** condiciona el aprendizaje. En etapas iniciales es sensato trabajar con un **conjunto finito y estable** de géneros y verificar la coherencia mínima de las anotaciones.

2.1.4. Clasificación supervisada en el contexto musical

La tarea se formula como **clasificación supervisada**: a partir de pares (x, y) donde x es una representación del audio y y la etiqueta de género, se aprende un modelo $f(x)$ que asigna a una pista nueva la categoría más plausible. Dos cuestiones prácticas marcan la diferencia entre una evaluación fiable y una ilusoria: (i) la **definición de la instancia** y la posible **agregación** por segmentos; (ii) la **validación honesta** con particiones de *entrenamiento/validación/prueba* sin fugas (por artista, máster,

etc.). Métricas como la **precisión (accuracy)** ofrecen una visión global; la **F1 macro** equilibra el peso de cada clase cuando hay desequilibrios y la **matriz de confusión** ayuda a entender qué estilos se confunden y por qué.

2.1.5. Consideraciones y supuestos de uso

Se asume que las pistas de entrada poseen **duración suficiente**; que los ficheros se pueden convertir a la **tasa objetivo** y formato de trabajo sin pérdida incompatible con la tarea; y que las **transformaciones y parámetros** aplicados en la preparación del audio se mantienen **idénticos** entre entrenamiento e inferencia. Bajo estas condiciones, el problema queda acotado y las comparaciones entre métodos tienen sentido.

2.2. Representaciones y características de audio

Elegir una buena representación es, en la práctica, decidir **qué** del sonido queremos que el modelo vea y **qué** preferimos ocultar. La onda tal cual contiene todo, pero también contingencias de grabación y mezcla. El objetivo aquí es fijar un **punto de vista estable** (invariante a pequeños cambios de nivel, microdesplazamientos o coloraciones poco relevantes) que conserve aquello que diferencia estilos: **timbre**, **patrones rítmicos** y, cuando aplica, **contenido armónico**.

2.2.1. Tiempo–frecuencia: de la STFT a "imágenes" útiles

El paso natural desde la onda es mirar cómo se reparte la **energía por frecuencias** a lo largo del tiempo. La **STFT** divide la señal en ventanas breves y, en cada una, calcula su espectro. La elección de **tamaño de ventana** y **salto** no es menor: ventanas largas separan bien frecuencias (útil para ver armónicos y envolventes tímbricas), pero difuminan eventos rápidos; ventanas cortas capturan transitorios y microgestos rítmicos, a costa de perder resolución espectral. Con la STFT obtenemos **espectrogramas**; conviene **comprimir amplitudes** (log) para evitar que unos pocos picos dominen. Esta representación se comporta como una **imagen**: texturas verticales (transitorios), horizontales (sostenidos), diagonales (glissandi), peines regulares (armónicos), "nubes" de energía de determinadas producciones. Esta "visualidad" abre la puerta a modelos que explotan **regularidades locales en 2D**.

2.2.2. Perceptuales: mel/log-mel, MFCC y cromas (y por qué se usan)

(Log-)mel: el banco de filtros **mel** redistribuye la energía en bandas que crecen en anchura según aumenta la frecuencia, aproximando la sensibilidad auditiva. El **mel-espectrograma** filtra detalles poco relevantes; su versión **log-mel** estabiliza el rango dinámico. Resultado: una **matriz 2D** muy alineada con lo que distinguimos como oyentes; funciona especialmente bien con **convolucionales**.

MFCC: al aplicar una transformada coseno a log-mel se obtienen los **coeficientes**

cepstrales en Mel, que condensan la forma global del espectro en un vector corto. Capturan **timbre promedio** con bajo coste, aunque pierden detalle fino; útiles como entrada ligera o complemento.

Cromas: cuando la **tonalidad** y las **progresiones** discriminan estilos, las **cromas** proyectan la energía en 12 clases de nota, parcialmente independientes del timbre. Aportan una vista armónica que complementa al mel/log-mel.

Resumen operativo: para un clasificador generalista, **log-mel** como base 2D es un punto de partida sólido; **MFCC** aportan ligereza y timbre promedio; **cromas** suman discriminación armónica. Combinar dos suele equilibrar mejor que apostar por uno solo.

2.2.3. Preparación, normalización y longitud: lo pequeño que decide lo grande

Proyectos idénticos en arquitectura pueden divergir por **detalles de preparación**. Por eso se unifican **tasa de muestreo** y formato, se controla el **nivel** y se decide cómo gestionar la **longitud**: recortes a duración fija, **padding** controlado o agregación de fragmentos por pista. Tras extraer mel/log-mel o cromas, conviene una **normalización estadística** consistente (media/desviación del conjunto de entrenamiento). Cambiar una ventana de 25 a 40 ms, pasar de 64 a 128 bandas mel o normalizar por pista en lugar de por conjunto puede mover varios puntos de métrica: no son afinaciones menores, son **decisiones de diseño**.

Aspectos puntuales

- **Señales en el dominio temporal.** Indicadores como **RMS** o **ZCR** son útiles para control de calidad, detección de actividad o segmentación inicial. Como única base para el género, suelen quedarse cortos por su escaso poder tímbrico y estructural.
- **Aumento de datos (data augmentation).** Conjuntos desbalanceados se benefician de variaciones **realistas**: ligeros cambios de ganancia, pequeños estiramientos temporales, enmascaramientos de tiempo/frecuencia o añadidos de ruido suave. La regla: respetar las **invariancias** del género.

2.2.4. Criterios de elección (qué usar y cuándo)

- Equilibrio precisión/coste y modelos 2D $\Rightarrow$ **log-mel**.
- Entrada ligera o capa adicional de timbre promedio $\Rightarrow$ **MFCC**.
- Corpus con rasgos armónicos distintivos $\Rightarrow$ añadir **cromas**.
- Con más presupuesto, explorar **combinaciones** (p. ej., log-mel + cromas) aporta robustez.

- En cualquier caso, documentar **parámetros** (ventana, *hop*, n_{mel}, escalados) y mantener **consistencia** entre entrenamiento e inferencia garantiza resultados **repetibles**.

2.3. Relación con proyectos con la misma funcionalidad

La clasificación de género musical está presente en **prototipos y bibliotecas abiertas**, en **corpora públicos** y en **servicios industriales**. A continuación se destacan referencias concretas y lo que aportan para un sistema que extrae características, entrena una red y evalúa/infiere sobre audio nuevo.

2.3.1. Prototipos y bibliotecas abiertas

musicnn (MTG/UPF). Redes **convolucionales** preentrenadas para *music tagging* sobre **(log-)mel**, con modelos listos para inferencia y ejemplos de uso. Útil como referencia de **pipeline reproducible** y como extractor de *embeddings* o punto de partida para *fine-tuning*.

Essentia. Librería C++/Python de MIR que cubre desde **procesado básico** (MFCC, cromas, mel) hasta **inferencia de modelos** con pesos publicados. Acelera la **extracción de características** y aporta guías de **normalización** y empaquetado de **inferencia**.

Qué nos llevamos: buenas prácticas de **preprocesado** (resampleo, ventana, *hop*, escalado), uso de **representaciones perceptuales** y ejemplos de **convolucionales 2D** para audio.

2.3.2. Corpora y recursos públicos

MagnaTagATune (MTAT). Clips de $\sim$30 s con **etiquetas humanas** recogidas vía juego; clásico para *tagging* y comparativas (útil para estudiar **ruido** de etiquetas y **equilibrio** de clases).

Million Song Dataset (MSD) + Last.fm tags. Un millón de pistas con **metadatos** y etiquetas a nivel de canción (incluidos géneros), habitual en trabajos de caracterización y recomendación.

AudioSet. Ontología de cientos de clases y millones de clips etiquetados; abarca **instrumentos**, **eventos** y **géneros**. Se emplea para **preentrenar** modelos que luego se especializan en tareas musicales.

2.3.3. Servicios industriales y productos reales

Spotify (Web API). Expone **audio features** (tempo, energy, valence, etc.) y **"genre seeds"** para recomendaciones. Aunque los clasificadores internos no son públicos, estas interfaces evidencian que **género** y **rasgos de audio** son atributos operativos de primer nivel y que existen **taxonomías mantenidas** a gran escala.

Google/YouTube. La investigación en **AudioSet** ha permeado productos y servicios: el patrón **preentrenamiento masivo + especialización** es frecuente con audio heterogéneo. Para un clasificador de géneros, esto sugiere opciones de *transfer learning* y la conveniencia de **datasets amplios y diversos** cuando se busque robustez.

Lecciones de operación: en producción se priorizan **ingestión a escala**, **actualización continua** de modelos, **monitorización** de sesgos y **trazabilidad**. Incluso en un TFG conviene separar **extracción** $\rightarrow$ **entrenamiento** $\rightarrow$ **inferencia**, fijar **semillas** y documentar **parámetros**.

2.3.4. Convergencia técnica en la literatura

Los últimos años muestran una convergencia: **CNN/CRNN** sobre **(log-)mel** como línea base sólida para etiquetado musical (género incluido), con protocolos de evaluación que evitan **fugas** entre particiones y reportan **accuracy** junto a **F1 macro**. Esta combinación equilibra **expresividad**, **robustez** y **coste**.

2.3.5. Decisiones aplicables al presente trabajo

- **Extracción: log-mel** como base; **MFCC** y **cromas** como complementos según el corpus. Implementación apoyada en bibliotecas consolidadas para asegurar **reproducibilidad**.
- **Entrenamiento:** arquitecturas **convolucionales** con validación honesta y control de sobreajuste; registro de **semillas**, **versiones** y **parámetros**.
- **Evaluación e inferencia:** particiones **train/val/test** sin fugas (idealmente por **artista**), métricas **accuracy** y **F1 macro**, y **matriz de confusión** para interpretar errores; diseño de inferencia compatible con pipelines de recomendación (género como atributo).

3. ESTUDIO DE VIABILIDAD

Este apartado comprueba que la solución propuesta es **pertinente**, **factible** con recursos realistas y **sostenible** a futuro. Primero se acota el **alcance** (qué entra y qué no), después se **valoran alternativas** tecnológicas con criterios objetivos (calidad, coste, esfuerzo de mantenimiento, escalabilidad) y, finalmente, se **selecciona** la combinación más coherente con los objetivos del proyecto.

3.1. Alcance del proyecto

Propósito. Desarrollar un sistema que, a partir del contenido de una pista de audio, **estime su género musical** dentro de un conjunto finito de clases, ofreciendo resultados **reproducibles** y **comparables**.

Entradas y salidas.

- **Entrada:** ficheros WAV/MP3 convertibles a una **tasa de muestreo objetivo**; duración suficiente para que exista evidencia musical (evitando silencios prolongados o tramos ajenos a la música).
- **Salida: etiqueta de género** y, opcionalmente, **probabilidades** por clase para análisis y diagnóstico.

Hipótesis de trabajo.

- El preprocesado (normalización, remuestreo, segmentación) **no destruye** información discriminativa.
- Las **etiquetas** de los datasets de referencia son razonablemente coherentes para entrenar y evaluar.
- Se dispone de **recursos académicos** (CPU/GPU moderada) y de un entorno Python estándar.

Dentro del alcance (fase actual).

- Representaciones **perceptuales** del audio (mel/log-mel) como base; **MFCC** y **cromas** como complementos evaluables.
- Clasificador **neuronal** (CNN 2D ligera; variante **CRNN** si los datos lo justifican).
- **Evaluación** con *accuracy*, **F1 macro** y **matriz de confusión**; **inferencia** que replica exactamente el preprocesado del entrenamiento.

Fuera del alcance (fase actual).

- Etiquetado **multi-género** o taxonomías jerárquicas completas.

- Requisitos estrictos de **tiempo real** o despliegues móviles/embebidos.
- Curación legal de catálogos masivos o armonización editorial de metadatos.

Riesgos principales y mitigación.

- **Desequilibrio de clases:** particiones estratificadas, uso de **F1 macro** y *augmentation* prudente.
- **Ruido de etiquetas** / **ambigüedad de género:** muestreo manual de calidad, análisis de confusión y revisión de casos frontera.
- **Sobreajuste:** *early stopping*, regularización y separación estricta **train**/**val**/**test** (idealmente evitando cruce de artistas).
- **Fugas entre particiones:** control por artista/álbum cuando sea posible y verificación automática de duplicados.

3.2. Estudio y valoración de las alternativas de solución

A continuación se comparan, de forma separada, las decisiones clave del sistema. Cada bloque incluye definiciones breves, una tabla comparativa y una conclusión operativa.

Representación del audio *Qué resuelve.* Convertir la onda en una **vista estable e informativa** que conserve señales de **timbre**, **patrones rítmicos** y, si procede, **armonía**. *Conceptos rápidos:* **mel**/**log-mel** (energía por bandas en escala perceptual; con **log** se comprime el rango dinámico), **MFCC** (transformada coseno sobre log-mel; "timbre promedio" en pocos coeficientes) y **cromas** (energía proyectada en 12 clases de nota; resumen armónico/tonal).

Representación	Ventajas	Limitaciones	Coste	Cuándo usarla
Mel / **log-mel**	Alineada con la **percepción**; eficaz con **CNN**/**CRNN**; captura timbre + gestos rítmicos	Requiere ajustar ventana/*hop*; mayor coste que MFCC	Medio	**Base principal** en clasificación de género
MFCC	**Compactos** y rápidos; buen "timbre promedio"	Pierden **detalle 2D** y contexto temporal fino	Bajo	**Canal ligero** o *baseline*; complemento a mel
Cromas	Aportan **armonía**/**tonalidad**; complementan mel/log-mel	Poco útiles cuando el género es más **textural**/**production-driven**	Bajo–Medio	**Segundo canal** si el corpus muestra señales armónicas claras

Conclusión. Adoptar **log-mel** como **base**; **MFCC** y **cromas** se incorporarán sólo si aportan **mejora consistente** en el corpus elegido.

Arquitectura del modelo *Qué resuelve.* Aprender funciones que separen géneros a partir de la representación.

Alternativa	Fortalezas	Debilidades	Idoneidad
SVM/k-NN + MFCC	Simplicidad, rapidez, reproducibles	Techo de rendimiento inferior; ignoran estructura 2D/temporal	**Línea base** comparativa
CNN 2D sobre log-mel	Aprende **texturas 2D**; excelente relación **precisión/coste**	Memoria temporal limitada (según ventana)	**Opción principal**
CRNN (CNN + recurrente)	Añade **memoria temporal** para dependencias largas	Más compleja; riesgo de sobreajuste en datasets pequeños	**Variante** si hay datos/tiempo

Conclusión. Empezar con **CNN 2D ligera** y evaluar una **CRNN** si la diversidad del dataset lo justifica y se observa ganancia tangible.

Segmentación y estrategia de decisión *Qué resuelve.* Reducir sensibilidad a intros/silencios y secciones atípicas.

Estrategia	Pros	Contras	Decisión
Pista completa	Implementación muy simple	Susceptible a segmentos no representativos	—
Por segmentos + agregación	**Robustez** a variaciones locales; decisiones más estables	Más cómputo; requiere coherencia de preprocesado	**Elegida** (promedio / voto)

Conclusión. Procesar ventanas/segmentos homogéneos y **agregar** (promedio o voto), manteniendo el mismo *pipeline* en entrenamiento e inferencia.

Datasets candidatos *Criterio.* Iterar primero con un conjunto **equilibrado y claro**; ampliar después para robustez.

Conclusión. Empezar con **FMA-small** por limpieza y equilibrio; contrastar con **GTZAN** bajo particiones controladas; plan de ampliación a **FMA-medium**/**large** y, a futuro, **MTAT**/**MSD** para generalización.

Protocolo de evaluación y métricas *Qué resuelve.* Comparar de forma honesta y **diagnosticar** errores.

Nota operativa. Todos los parámetros de extracción (ventana, *hop*, nº de bandas mel, escalado), semillas y versiones deben quedar **documentados** para que los resultados puedan **reproducirse**.

Dataset	Tamaño/Clases	Ventajas	Inconvenientes	Uso recomendado
FMA-small	8 000 clips/30 s; **8** géneros equilibrados	Licencia clara; comparabilidad; ideal para prototipado	Menos diverso que *medium/large*	**Principal** (primera iteración)
GTZAN	1 000 clips/30 s; 10 géneros	Clásico; fácil de usar	Fallos documentados (duplicados, misetiquetado)	**Benchmark** secundario con *splits* cuidados
MagnaTagATune	∼30 s por clip; etiquetas humanas	Tamaño razonable; útil para *tagging*	Etiquetas ruidosas; no centrado sólo en género	**Complemento** (robustez)
MSD + Last.fm	1M pistas con *tags*	**Escala** y diversidad	Audio no siempre accesible; etiquetas heterogéneas	**Futuro** (transfer/robustez)

Elemento	Elección	Motivo
Particionado	**Train/Val/Test** estratificado; si es posible, separación por **artista**	Evitar **fugas** y medir **generalización**
Métricas	*Accuracy* + **F1 macro**	*Accuracy* resume; **F1 macro** equilibra clases desbalanceadas
Análisis	**Matriz de confusión**	Localiza **fronteras** entre géneros y guía mejoras

3.3. Selección de la solución

Representación adoptada. *Base:* **log-mel** (64–128 bandas, compresión logarítmica, ventana/*hop* documentados), por su equilibrio entre **información perceptual**, **robustez** y **coste**. *Complementos evaluables:* **MFCC** (timbre promedio) y **cromas** (armonía) como canales adicionales si aportan mejora **consistente** en validación.

Modelo y decisión. *Arquitectura principal:* **CNN 2D ligera** sobre log-mel, diseñada para operar con **segmentos** y **agregación** de probabilidades (promedio o voto). *Variante opcional:* **CRNN** si el tamaño/diversidad del conjunto final justifica capturar **dependencias temporales** más largas con ganancia medible.

Dataset y plan de crecimiento. *Primera iteración:* **FMA-small** (8 géneros equilibrados) para iterar rápido y fijar una línea base sólida. *Ampliación:* **FMA-medium/large** y validación cruzada con **GTZAN** bajo *splits* cuidadosos; a medio plazo, exploración de **MagnaTagATune**/**MSD** para robustez y *transfer learning*.

Criterios de aceptación.

- **Reproducibilidad:** scripts de **preparación**, **entrenamiento** e **inferencia** con parámetros y semillas fijadas.
- **Rendimiento: F1 macro** superior a la línea base clásica (MFCC + SVM/k-NN) y

accuracy acorde al estado del arte del dataset empleado.

- **Trazabilidad:** registro de versiones y configuraciones para repetir resultados y auditar cambios.

Justificación final. La combinación **log-mel + CNN 2D** con decisión por **segmentos** proporciona el mejor **equilibrio** entre calidad, coste y mantenibilidad. Permite **crecer** en datos y clases sin rehacer el sistema y mantiene una ruta clara para mejorar (variantes CRNN, canales adicionales, ampliación de datasets) cuando se disponga de mayor volumen y diversidad de información.