Muzca – Rhythm Representation Studies

Daniel Martínez

Alfredo Aponte

Director: PhD Daniel Gómez

Tutor: M.Sc. Jose Giraldo

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI

2025

Trabajo de Grado
Universidad Icesi
Maestría en Ciencia de datos

Resumen

Este trabajo presenta un enfoque basado en aprendizaje profundo para la predicción de representaciones rítmicas explicables directamente desde archivos de audio (.WAV). La representación utilizada, denominada Frequency-Weighted Onset Density (FWOD), permite sintetizar la densidad rítmica de un compás en un vector unidimensional de 16 valores.

En una fase inicial, se validó la utilidad del FWOD como descriptor rítmico aplicando modelos de clasificación sobre datos simbólicos (archivos MIDI), alcanzando un 90.5% de precisión y superando trabajos previos en el estado del arte. Este resultado sirvió como línea base para la segunda fase del proyecto, centrada en la predicción del vector FWOD desde audio real utilizando modelos convolucionales (CNN) entrenados sobre espectrogramas Mel.

El dataset final se construyó a partir de la correspondencia entre archivos .MIDI y .WAV del Groove MIDI Dataset de Magenta, ajustado y alineado para facilitar la comparación entre ambas representaciones. Se exploraron variantes de arquitectura, regularización y ensamble, alcanzando un MAE mínimo de 0.1836 con un R^2 estimado de 0.70.

Los resultados confirman la viabilidad del FWOD como puente entre señal acústica y análisis computacional de ritmo, abriendo nuevas posibilidades para el desarrollo de frameworks explicables de clasificación musical centrados en percusión (Choi et al., 2017; Gómez-Marín et al., 2024).

Tabla de Contenido

Introducción

La representación computacional del ritmo musical enfrenta desafíos significativos debido a la complejidad multidimensional del sonido y la subjetividad inherente a la percepción humana. Las herramientas tradicionales, como los archivos MIDI y las señales de audio, han sido fundamentales en la investigación y creación musical digital. Sin embargo, presentan limitaciones importantes cuando se busca capturar la riqueza perceptual del ritmo tal como es experimentado por músicos y oyentes (Casey et al., 2008; Temperley, 2010).

En este contexto, el proyecto Muzca surge como una propuesta innovadora orientada a cerrar la brecha entre la señal acústica y el análisis computacional del ritmo. La iniciativa introduce el concepto de Frequency-Weighted Onset Density (FWOD), una representación unidimensional que resume la densidad rítmica de un compás en 16 valores ponderados. Esta representación busca ser compacta, explicable y coherente con la percepción humana, facilitando así su uso tanto en clasificación como en generación de patrones rítmicos (Gillick et al., 2019).

Durante la fase inicial del proyecto, se trabajó exclusivamente con datos simbólicos (archivos MIDI), desarrollando y evaluando modelos de clasificación que utilizaron FWOD como vector de entrada. Esta etapa permitió comprobar la capacidad de FWOD para discriminar entre estilos rítmicos y sentó las bases metodológicas del enfoque propuesto.

La presente fase del proyecto constituye un avance sustancial: se plantea la predicción del vector FWOD directamente desde archivos de audio (.WAV), utilizando espectrogramas Mel como representación intermedia y redes neuronales convolucionales (CNN) como modelo predictivo. Este cambio de paradigma implica una transición del entorno simbólico al acústico, abriendo nuevas posibilidades para aplicar

FWOD en contextos más realistas, como grabaciones en vivo o interpretaciones espontáneas.

Además de validar el potencial de FWOD en entornos acústicos, esta fase también explora la estabilidad de los modelos bajo distintas arquitecturas y técnicas de regularización. A través de un dataset alineado y cuidadosamente curado a partir del Groove MIDI Dataset de Magenta (Google Magenta, 2019), se evalúan estrategias que buscan mantener la fidelidad del patrón rítmico entre el audio original y su representación vectorial.

Con ello, se consolida un pipeline integral que parte del audio real, pasa por transformaciones espectrales y culmina en una representación rítmica explicable, ofreciendo así un marco robusto y versátil para el análisis automático del ritmo centrado en la percusión.

Contexto y antecedentes

El análisis automático del ritmo musical enfrenta desafíos complejos derivados de la naturaleza multidimensional del sonido y la brecha entre representaciones digitales tradicionales y la percepción humana. Aunque archivos MIDI y señales de audio han sido herramientas fundamentales, presentan limitaciones para capturar aspectos cognitivos y perceptivos del ritmo (Casey et al., 2008; Temperley, 2010).

Estas limitaciones impactan negativamente en el análisis, generación y clasificación de patrones musicales, obstaculizando el desarrollo de sistemas explicables para aplicaciones en educación, creación o investigación musical.

En respuesta, el proyecto Muzca propone una nueva representación rítmica: el Frequency-Weighted Onset Density (FWOD). Este modelo resume cada compás en un vector unidimensional de 16 valores, equilibrando simplicidad computacional y relevancia perceptiva para facilitar tareas analíticas sin pérdida de contenido significativo.

Metodológicamente, el enfoque evolucionó desde contextos simbólicos (MIDI) hacia la predicción en audio real. Esta transición acerca el análisis computacional a entornos musicales reales, donde las representaciones simbólicas suelen estar ausentes.

Árbol del problema

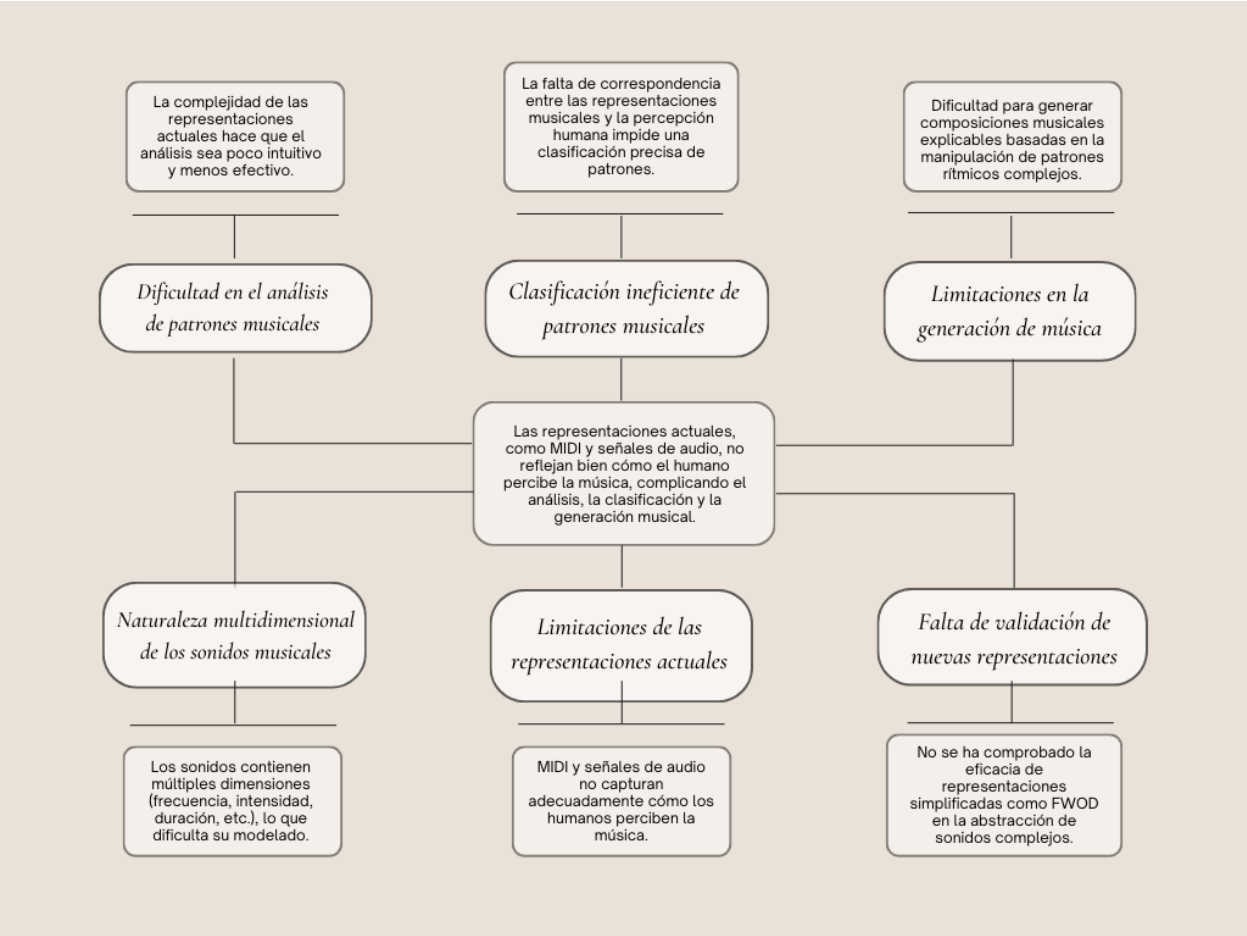


Figura 1. Árbol del problema

El árbol del problema sintetiza las limitaciones en la representación musical actual y su impacto en el análisis, clasificación y generación de patrones rítmicos. A continuación, se presenta una estructura clara y concisa:

Problema Central

Las representaciones musicales actuales (MIDI y señales de audio) no capturan adecuadamente la complejidad perceptual humana de la música, dificultando la conexión entre experiencia musical y análisis técnico automatizado.

Dificultad en el análisis de patrones musicales

- **Problemática**: La complejidad de las representaciones actuales dificulta el análisis de patrones musicales, haciéndolo menos intuitivo y efectivo.
- **Oportunidad**: Simplificación mediante representaciones unidimensionales como FWOD.

Clasificación ineficiente de patrones musicales

- **Problemática**: Hay una falta de correspondencia entre las representaciones musicales y la percepción humana, lo que impide una clasificación precisa de patrones.
- **Oportunidad**: FWOD como descriptor rítmico alineado con la cognición musical.

Limitaciones en la generación de música

- **Problemática**: Dificultad para crear composiciones explicables desde patrones complejos.
- **Oportunidad**: Sistemas basados en FWOD permiten manipulación intuitiva de patrones rítmicos.

Naturaleza multidimensional de los sonidos musicales

- **Problemática**: Los sonidos musicales tienen múltiples dimensiones (frecuencia, intensidad, duración, etc.), lo que complica su modelado.
- **Oportunidad**: Falta de comprobación empírica para representaciones innovadoras como FWOD en contextos polifónicos.

Limitaciones de las representaciones actuales

- **Problemática**: Las representaciones como MIDI y señales de audio no capturan adecuadamente la percepción humana de la música, lo cual limita el alcance de las aplicaciones actuales.
- **Oportunidad**: Innovar en representaciones que incorporen una comprensión más completa de la percepción humana podría abrir nuevas posibilidades en la creación, análisis y enseñanza de la música.

Falta de validación de nuevas representaciones

- **Problemática**: No se ha comprobado la eficacia de nuevas representaciones, como FWOD, en la abstracción de sonidos complejos, lo que genera incertidumbre sobre su eficacia.
- **Oportunidad**: Validar y optimizar nuevas representaciones podría establecer estándares más efectivos para la representación musical, beneficiando tanto a la investigación como a la industria musical.

Relación con el proyecto Muzca

El desarrollo de FWOD aborda directamente este árbol problemático mediante:

- Validación de su capacidad descriptiva en datos simbólicos.
- Predicción desde audio real mediante CNNs, estableciendo puentes entre señal acústica y análisis computacional.

Objetivos

Objetivo General

Desarrollar un sistema que permita transformar audio de percusión en representaciones FWOD para su uso en la clasificación automática del estilo musical, demostrando la efectividad de dicha representación en tareas de análisis rítmico.

Objetivos Específicos

- Analizar y revisar el estado del arte sobre la representación y clasificación de ritmos de batería en el contexto de la percepción musical humana.
- Validar el uso de FWOD como proxy computacional del aplanamiento rítmico observado en estudios de tapping.
- Desarrollar un sistema de clasificación que utilice la representación simplificada para identificar y comparar patrones rítmicos.
- Diseñar e implementar un pipeline de generación de dataset que combine representaciones matriciales del audio por compás, el vector FWOD obtenido desde el archivo MIDI y la metadata de cada track.
- Entrenar modelos de deep learning que aprendan a predecir la representación FWOD directamente desde audio.
- Evaluar el rendimiento del sistema completo realizando la clasificación del estilo musical utilizando los FWOD generados.

Metodología

La metodología del proyecto Muzca se diseñó bajo un enfoque híbrido, integrando la estructura del modelo CRISP-DM y la agilidad de SCRUM para guiar el análisis, desarrollo y validación de los sistemas propuestos. Este enfoque permitió avanzar de

manera iterativa, adaptándose a los retos técnicos y optimizando los resultados en cada fase.

Estructura metodológica

- CRISP-DM: Proporcionó la secuencia lógica de comprensión del problema, análisis y preparación de datos, modelado, evaluación y despliegue.
- SCRUM: Facilitó la gestión ágil del equipo, dividiendo las tareas en sprints y permitiendo ajustes dinámicos según los hallazgos de cada ciclo.

El proyecto se desarrolló en las siguientes 2 fases:

- El flujo original, basado en la representación unidimensional FWOD aplicada sobre archivos MIDI, que fue el alcance principal del proyecto en su primera etapa y que sirvió como punto de referencia. Los detalles completos de esta etapa se encuentran documentados en el Anexo A.

- El flujo actual, que constituye el foco de esta fase: se basa en imágenes generadas a partir de archivos de audio real (.WAV), transformados en espectrogramas Mel, que alimentan modelos de visión por computador como las redes neuronales convolucionales (CNN) (Choi et al., 2017).

Ambos enfoques han sido desarrollados y evaluados para comparar su consistencia en la clasificación de patrones rítmicos, siendo el análisis a partir de audio real el principal aporte de esta etapa.

Fases del proceso

El proyecto se estructura en las siguientes fases, integrando el enfoque híbrido planteado:

Comprensión del Problema

- **Objetivo Principal:** Diseñar un modelo que represente, analice, clasifique y genere patrones rítmicos de forma eficiente, integrando ciencia de datos y teoría musical (Jordà et al., 2023).
- **Preguntas Iniciales:** ¿Cómo puede un modelo computacional capturar la esencia del ritmo? ¿Qué tipo de representación logra equilibrar la percepción humana y el análisis técnico?

Análisis de los Datos

- Se evaluaron tres datasets principales: Tapping Dataset, Tap-Tam-Drum Dataset y Groove MIDI Dataset de Magenta, seleccionando este último por su correspondencia directa entre archivos .MIDI y .WAV para la misma interpretación rítmica (Google Magenta, 2019).
- En esta fase se identificaron retos técnicos con los archivos .WAV del Groove MIDI Dataset de Magenta, como la presencia de silencios iniciales que generaban desfases con respecto a los archivos MIDI.
- Aunque no se realizaron experimentos directos de tapping, la representación FWOD se diseñó para simular el efecto de aplanamiento rítmico observado en estudios perceptuales. En este proyecto, se asumió que el promedio de onsets ponderados por frecuencia refleja un comportamiento análogo al tapping humano, como se ha demostrado en Gómez-Marín et al. (2024).

Preparación de los Datos

- **Alineación temporal:** Se sincronizaron los inicios de los archivos .WAV y .MIDI, eliminando silencios y asegurando que ambos formatos representaran el mismo inicio rítmico.

  Tras la alineación, el dataset resultante mostró diferencias mínimas entre las señales .WAV y .MIDI, con desfases tan pequeños que resultan prácticamente imperceptibles para el oído humano. Este trabajo fue clave para asegurar la consistencia del pipeline y la confiabilidad de los modelos en el proceso de predicción y análisis rítmico.

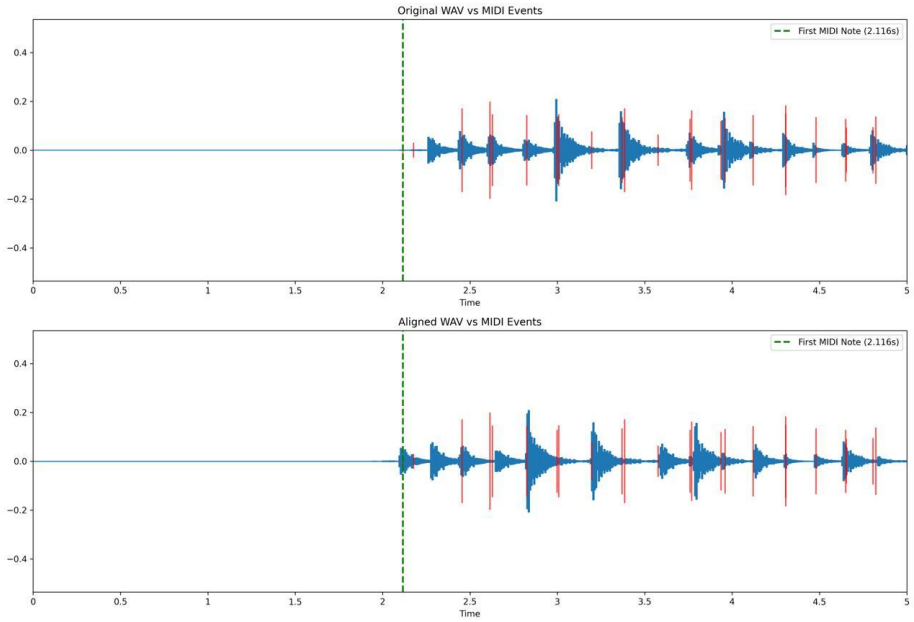


Figura 2. Alineación de .MIDI y .WAV

- **Conversión a mono:** Los audios fueron convertidos de estéreo a mono para reducir la carga computacional y facilitar el procesamiento.

- **Generación de espectrogramas Mel:** Sobre los archivos mono-alineados se generaron espectrogramas Mel, que resumen la energía en distintas bandas de frecuencia a lo largo del tiempo y sirven como entrada para los modelos de aprendizaje profundo

Construcción del dataset experimental

- Se consolidó el archivo mel_fwod_dataset.npz, que integra para cada compás: el espectrograma Mel, el vector FWOD de referencia (calculado desde el .MIDI alineado) y metadatos descriptivos (BPM, estilo, baterista, etc.).
- Se creó una nueva división de entrenamiento, validación y prueba, cuidando el equilibrio de estilos rítmicos y la representatividad del conjunto de datos.

Modelado

- Se implementaron modelos de regresión y clasificación basados en redes neuronales convolucionales (CNN), entrenados sobre los espectrogramas Mel.
- Se exploraron variantes de arquitectura (baseline, batch normalization, dropout, ensambles) y dos enfoques principales:
    - **Energy Matrix Approach:** Resumía el espectrograma Mel en una matriz 3×16 por bandas de frecuencia y segmentos temporales, facilitando la interpretación y mejorando la generalización.
    - **Full Mel Approach:** Utilizaba el espectrograma Mel completo como entrada, permitiendo al modelo aprender directamente de la riqueza espectral y temporal del audio.

Evaluación

- Se aplicaron métricas como error cuadrático medio (MSE), error absoluto medio (MAE) y coeficiente de determinación (R^2) para cuantificar la capacidad de predicción del vector FWOD y el desempeño en clasificación rítmica.
- Se realizó validación cruzada y monitoreo continuo de las métricas para confirmar la viabilidad del enfoque y detectar oportunidades de mejora, especialmente en el balance de clases y la alineación de datos.

Cabe aclarar que las métricas de regresión (MAE, MSE, R^2) evalúan la predicción del vector FWOD, mientras que la métrica de clasificación (accuracy) se usa para evaluar la calidad del estilo musical predicho a partir del FWOD generado.

Roles y Dinámica del Equipo

El equipo implementó un esquema multifuncional, donde todos los integrantes participaron en desarrollo, análisis y validación. El uso de SCRUM (Schwaber & Sutherland, 2020) permitió dividir el trabajo en sprints, priorizar tareas de mayor impacto y adaptarse a los retos emergentes del proyecto.

Fundamentos de Representación y Percepción Rítmica

La representación y percepción del ritmo en la música constituyen áreas centrales tanto en la investigación científica como en la práctica musical. A continuación, se sintetizan los fundamentos clave, integrando los aportes teóricos, computacionales y perceptuales más relevantes y las referencias actualizadas. Estas limitaciones dificultan el análisis, la clasificación y la generación de música (Gómez-Marín, 2024).

Representación Musical

Las representaciones digitales de la música suelen adoptar formas simbólicas (como MIDI) o directamente basadas en señal (como audio en .WAV). Las representaciones simbólicas permiten edición estructurada, pero pierden detalles de la interpretación; por otro lado, el audio contiene toda la riqueza espectral, pero requiere procesamiento complejo para extraer información útil.

Simplificación rítmica y percepción

Estudios recientes muestran que las personas tienden a simplificar patrones rítmicos complejos durante tareas como el tapping (Gómez-Marín, 2024), entendido como la acción de reproducir un ritmo percibido mediante golpeteos con los dedos. Esta tendencia valida el uso de representaciones unidimensionales que concentren la densidad rítmica, como FWOD, para reducir la complejidad sin perder información relevante

Frequency-Weighted Onset Density (FWOD)

El FWOD es una representación unidimensional que pondera los onsets rítmicos según su frecuencia dentro de un compás. Ha sido validado como descriptor útil para comparar y clasificar ritmos complejos, tanto en tareas simbólicas como en modelos aplicados a audio real. Esta técnica ha sido propuesta como una forma de aplanar y simplificar la información rítmica polifónica en un formato que sea más manejable y efectivo para tareas como la clasificación y generación de ritmos (Gómez-Marín, 2024).

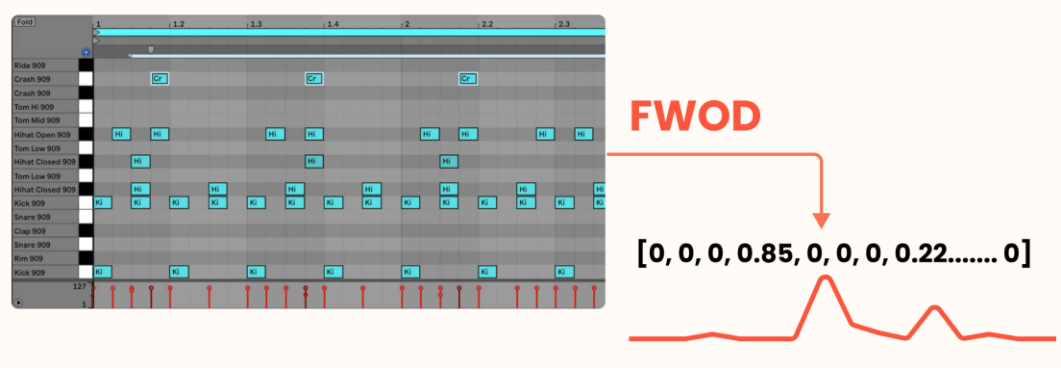


Figura 3. Aplanamiento analítico a FWOD

Clasificación y Generación de Ritmos

El uso de representaciones como FWOD permite alimentar modelos de clasificación capaces de identificar estilos rítmicos. En el contexto del audio real, se emplean técnicas de visión por computador como las redes neuronales convolucionales (CNN), entrenadas sobre espectrogramas Mel, para predecir el vector FWOD correspondiente a cada compás.

Clasificación de ritmos desde MIDI usando FWOD

En la fase inicial del proyecto se evaluó la representación Frequency-Weighted Onset Density (FWOD) generada a partir de archivos MIDI, con el objetivo de validar su utilidad como descriptor rítmico.

Se entrenaron diversos modelos de clasificación supervisada, logrando una precisión notablemente superior a la reportada en estudios anteriores (Behzad et al., 2023).

Este resultado estableció una *línea base de comparación (baseline)* para evaluar los FWOD predichos desde audio real en la nueva fase del proyecto.

Los detalles completos de esta etapa se encuentran documentados en el Anexo A.

Rhythm Space

En el alcance inicial, se exploró un análisis del Rhythm Space, un espacio bidimensional construido a partir de las representaciones FWOD generadas desde archivos MIDI. Esta exploración permitió visualizar agrupaciones coherentes entre patrones rítmicos, validando su utilidad como base perceptual.

Los detalles técnicos del proceso se encuentran documentados en el Anexo B.

Estado del arte

En los últimos años se han propuesto diversas estrategias para representar, analizar y generar patrones rítmicos de manera alineada con la percepción humana.

Rhythm Spaces (Gómez-Marín et al., 2020) es un sistema que proyecta patrones simbólicos de batería en un espacio bidimensional usando descriptores inspirados en procesos cognitivos. A través de un modelo encoder-decoder explicable, el usuario puede seleccionar una coordenada en ese espacio y obtener un patrón generado a partir de los vecinos de dicha posición.

El uso de arquitecturas encoder-decoder en el análisis rítmico ha permitido mapear patrones a espacios latentes donde es posible generar nuevas secuencias desde posiciones específicas, facilitando una interacción intuitiva y alineada con la cognición musical (Murel & Noble, 2024).

En trabajos recientes, se ha validado el Frequency-Weighted Onset Density (FWOD) como representación eficaz para ritmos polifónicos. Gómez-Marín et al. (2024) evidenciaron que el promedio de tapping (golpeteo consciente de un patrón rítmico percibido) de múltiples sujetos sobre patrones polifónicos es similar al FWOD extraído de esos patrones, sugiriendo que FWOD captura de forma efectiva el "aplanamiento" rítmico que haría un ser humano. Asimismo, al comparar FWOD con patrones reales del

dataset TapTamDrum, se observó una alta correlación, confirmando su utilidad como descriptor rítmico en contextos polifónicos.

Por su parte, el trabajo Learning to Groove with Inverse Sequence Transformations (Gillick et al., 2019) introdujo Tap2Drum, un modelo que permite generar interpretaciones expresivas a partir de representaciones mínimas (microtiempos), usando arquitecturas Seq2Seq variacionales. Este enfoque facilita que personas sin formación musical plasmen ideas rítmicas que el sistema convierte en interpretaciones comparables a las de un baterista profesional.

Además, en los últimos años los espectrogramas Mel se han consolidado como una representación efectiva para el análisis de audio en tareas musicales, debido a que aproximan la forma en que el oído humano percibe las frecuencias. Este tipo de representación convierte la señal de audio en una imagen que refleja la estructura temporal y espectral de los sonidos, permitiendo aplicar técnicas de visión por computador como las redes neuronales convolucionales (CNN) para extraer patrones rítmicos relevantes (Choi et al., 2017; Pons et al., 2017). A diferencia de los datos simbólicos como MIDI, los espectrogramas Mel facilitan el análisis en contextos donde no existe una transcripción, como en grabaciones en vivo o improvisaciones, abriendo nuevas posibilidades para el estudio del ritmo directamente desde el audio real.

Propuesta

Para viabilizar los estudios propuestos, se trabajó con el Groove MIDI Dataset de Magenta, el único que permitía la correspondencia entre archivos .MIDI y .WAV.

El proceso completo de alineación temporal, conversión a mono, y generación de espectrogramas Mel fue descrito en detalle en el capítulo de Metodología.

Viabilidad y selección del dataset experimental

En función de los objetivos de esta fase, se evaluaron tres datasets principales: Tapping Dataset, Tap2Drum Dataset y el Groove MIDI Dataset de Magenta. Este último fue seleccionado como núcleo del trabajo por ser el único que ofrecía correspondencia directa entre archivos .MIDI y .WAV para una misma interpretación rítmica.

Como se detalló en la sección de Metodología, se realizaron procesos de alineación, limpieza, conversión de estéreo a mono y generación de espectrogramas Mel. Estas transformaciones permitieron consolidar el mel_fwod_dataset.npz, que contiene por cada compás: el espectrograma Mel, el vector FWOD de referencia y metadatos descriptivos.

Este dataset se convirtió en la base sobre la cual se desarrollaron los modelos de predicción del vector FWOD desde audio real.

Predicción de ritmo desde audio real

En esta fase del proyecto Muzca, se desarrolló un sistema robusto para transformar audio real de percusión (.WAV) en representaciones FWOD útiles en tareas de clasificación automática del estilo musical. Las actividades se organizaron de la siguiente forma:

Tratamiento y preparación del dataset

Los detalles técnicos relacionados con la alineación de archivos, el tratamiento del audio, la normalización y la generación de espectrogramas Mel se describen en profundidad en el capítulo de Metodología. En esta sección se retoman únicamente los aspectos clave necesarios para contextualizar la fase actual de modelado.

Generación del mel_fwod_dataset.npz

Como parte del trabajo de consolidación de datos y preparación para los modelos, se construyó el archivo mel_fwod_dataset.npz, el cual constituye un insumo clave para el desarrollo y la evaluación de los modelos en esta fase del proyecto.

El mel_fwod_dataset.npz integra de manera estructurada la información fundamental generada durante el preprocesamiento:

- El espectrograma Mel, generado a partir del archivo .WAV correspondiente a cada compás. Esta matriz representa visual y numéricamente la evolución temporal y espectral del audio, sirviendo como entrada para los modelos de aprendizaje profundo.
- El vector FWOD, calculado a partir del archivo .MIDI alineado. Este vector, de 16 posiciones, resume de forma compacta el contenido rítmico de cada compás y constituye el valor de referencia para las tareas de predicción.
- Los descriptores y metadatos de cada registro, que incluyen características relevantes del audio y la interpretación (por ejemplo, BPM, baterista, estilo, duración, y otros campos provenientes del dataset original), son necesarios para el análisis contextual y el filtrado en fases exploratorias.

La generación de este dataset implicó vincular y almacenar de forma ordenada estas representaciones, asegurando la correspondencia precisa entre el espectrograma Mel, el FWOD y los descriptores de cada patrón rítmico. De esta forma, el dataset quedó preparado para alimentar de forma coherente el pipeline de modelado.

El mel_fwod_dataset.npz se convierte así en un recurso esencial, diseñado para facilitar la integración de los distintos componentes del pipeline y garantizar la reproducibilidad y consistencia de los experimentos realizados en el proyecto Muzca.

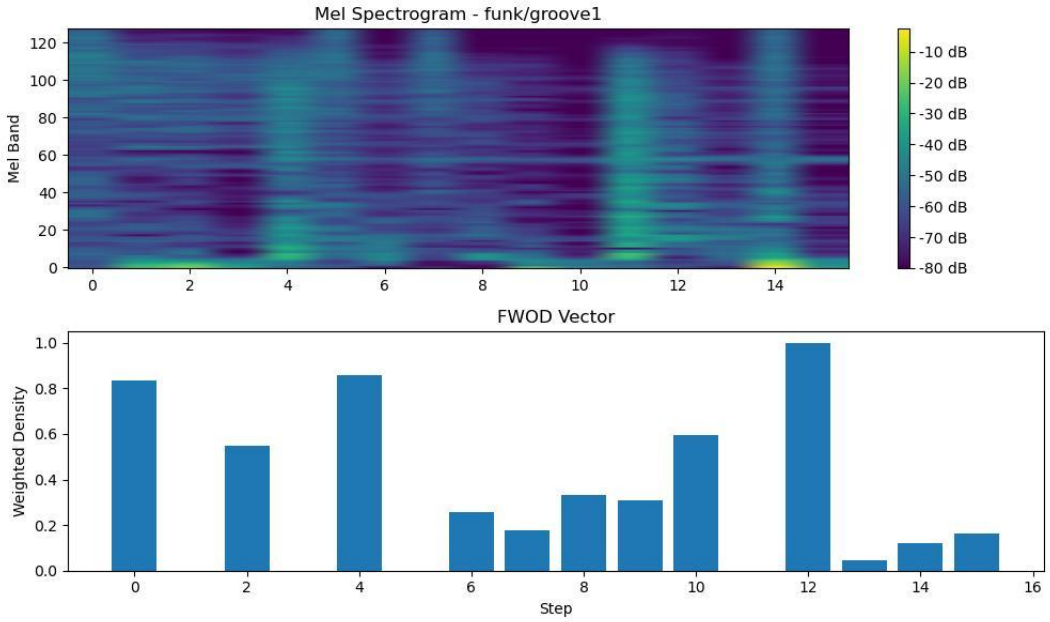


Figura 4. Espectrograma MEL y Vector FWOD

División del dataset

El Groove MIDI Dataset de Magenta originalmente se encontraba dividido en conjuntos de entrenamiento, validación y prueba. Sin embargo, durante el proceso de preparación se eliminaron los registros que no contaban con correspondencia entre archivos .MIDI y .WAV, lo que modificó el balance de las clases y la distribución de los datos en los subconjuntos definidos originalmente.

Para evitar sesgos y garantizar que los modelos se entrenaran y evaluaran sobre datos representativos, se optó por generar una nueva división del dataset, creando desde cero los conjuntos de entrenamiento, validación y prueba. Esta división se realizó cuidando que cada conjunto reflejara, en la medida de lo posible, la diversidad de estilos rítmicos presentes en el dataset final y buscando un equilibrio que permitiera un aprendizaje y evaluación más robustos.

Este reordenamiento fue clave para asegurar que el proceso de entrenamiento no favoreciera las clases más representadas y que los resultados obtenidos fueran consistentes y comparables.

Modelado de FWOD desde espectrogramas Mel

Métricas de evaluación

La evaluación de los modelos desarrollados se basó en un conjunto de métricas diseñadas para cuantificar la capacidad del sistema en la predicción del vector FWOD y en su desempeño en tareas de clasificación rítmica:

- Error cuadrático medio (MSE): métrica utilizada para medir la diferencia promedio al cuadrado entre el FWOD predicho y el FWOD real. Este valor penaliza más las desviaciones grandes y es útil para evaluar la calidad general del ajuste.
- Error absoluto medio (MAE): métrica que proporciona una medida de error promedio sin penalizar de forma desproporcionada las desviaciones mayores, permitiendo una interpretación más directa del desempeño del modelo.
- Coeficiente de determinación (R^2): indicador de la proporción de varianza del FWOD real que puede ser explicada por el modelo. Un R^2 más alto implica un mejor ajuste.

Estrategias evaluadas

Se evaluaron dos estrategias principales para predecir el FWOD:

Energy Matrix Approach

El Energy Matrix Approach consistió en transformar los espectrogramas Mel generados a partir de los archivos .WAV en una representación más compacta y explicable, orientada a capturar el contenido rítmico de los compases de manera eficiente.

El proceso comenzó con la división de cada espectrograma Mel en tres bandas de frecuencia: baja, media y alta. Estas bandas fueron definidas para representar el rango de frecuencias más relevantes en el análisis de ritmos de percusión, donde:

- La banda baja concentró la energía de los sonidos graves, típicos del bombo y otros elementos de baja frecuencia.
- La banda media capturó los sonidos intermedios, como cajas y toms.
- La banda alta recogió la información de elementos agudos, como platillos y hi-hats.

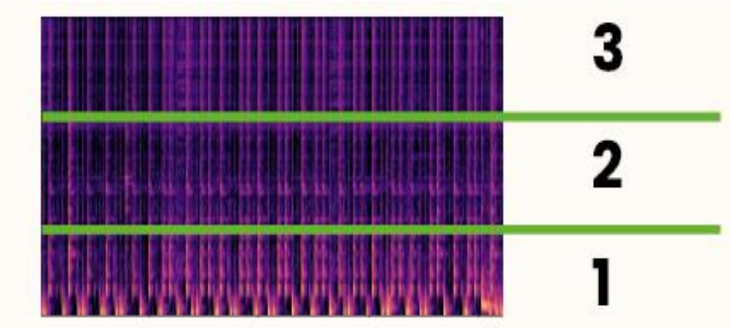


Figura 5. División del Mel en bandas

Para cada una de estas bandas, el espectrograma se dividió en 16 segmentos temporales equivalentes, correspondientes a los pasos rítmicos de un compás (por ejemplo, 16 semicorcheas en un compás de 4/4). De esta forma, se calculó el promedio de energía en cada segmento, resultando en una matriz 3×16 que resume el contenido energético rítmico del compás, diferenciando por bandas de frecuencia.

Esta matriz se empleó como entrada para el modelado, iniciando con el entrenamiento de un modelo base CNN, diseñado para aprender las relaciones entre la matriz de energía y el vector FWOD objetivo.

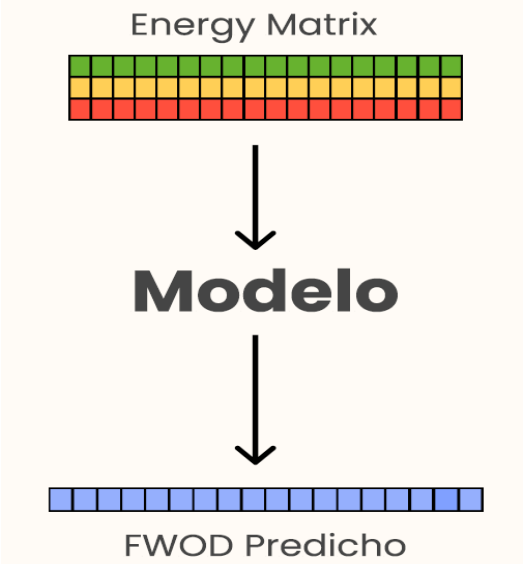


Figura 6. Energy Matrix a FWOD

Posteriormente, se implementó un modelo optimizado mediante ensamble, que integró las predicciones de múltiples redes con el fin de mejorar la generalización y reducir el error de predicción.

El MSE mínimo alcanzado (0.077) posiciona esta estrategia como una representación intermedia altamente efectiva entre el audio y el vector FWOD.

Full Mel Approach

El Full Mel Approach planteó la exploración de un modelo que aprovechara de manera íntegra el espectrograma Mel como entrada para la predicción del vector FWOD. A diferencia de otras estrategias como el Energy Matrix, aquí no se aplicaron resúmenes intermedios por bandas de frecuencia; se buscó que el modelo aprendiera directamente de la riqueza completa del espectrograma, preservando su estructura temporal y espectral.

Se trabajó sobre el dataset mel_fwod_dataset.npz, empleando el espectrograma Mel completo de cada compás como entrada. Los modelos fueron entrenados con los espectrogramas ya alineados y ajustados (audio mono, silencios iniciales eliminados), junto con sus vectores FWOD de referencia obtenidos desde el MIDI.

Se desarrollaron y compararon variantes de modelos CNN:

Experimento	MAE	MSE	R^2	Observación breve
cnn_batchnorm	**0.1836**	0.0696	0.1306	Incluye Batch Normalization
cnn_baseline	0.1852	**0.0687**	**0.1457**	Arquitectura básica (sin regularización)
cnn_dropout	0.1899	0.0698	0.1249	Incluye dropout (0.3)
daniel_base	0.2207	0.0732	0.0531	Ensamble de 5 modelos

Las gráficas de evolución de las métricas durante el entrenamiento mostraron un descenso progresivo y consistente en el MAE tanto para el conjunto de entrenamiento como de validación, mientras que el MSE evidenció estabilidad en entrenamiento y poca mejora en validación, sugiriendo un límite en el poder predictivo del modelo bajo las condiciones actuales.

En la sección Anexo C, se detallan las configuraciones para las diferentes arquitecturas.

Validación y aprendizaje

Como parte del proceso de aprendizaje, se implementó un pipeline robusto que permitió transformar de manera sistemática el audio real en representaciones FWOD alineadas

por compás, asegurando la coherencia de las entradas y salidas durante el entrenamiento de los modelos.

Se aplicó un proceso de validación cruzada y de monitoreo continuo de las métricas, lo que permitió:

- Confirmar la viabilidad del enfoque propuesto para predecir FWOD desde audio real.
- Detectar oportunidades de mejora, especialmente en lo relacionado con el balance de clases, dado que la desigualdad en la representación de ciertos estilos rítmicos impactó el desempeño del modelo en esas clases menos representadas.
- Identificar la necesidad de un refinamiento adicional en la alineación y normalización de los datos para reducir las pequeñas inconsistencias detectadas entre el audio y sus referencias simbólicas.

Estos resultados proporcionan una base sólida para optimizar las siguientes iteraciones del sistema y establecen los cimientos necesarios para las etapas futuras del proyecto. Entre las mejoras contempladas se incluye la implementación de arquitecturas híbridas más sofisticadas, como la combinación de redes neuronales convolucionales y recurrentes (CNN + RNN), así como la expansión y enriquecimiento del conjunto de datos de entrenamiento para maximizar la capacidad de aprendizaje del modelo.

Limitaciones

El proyecto Muzca, aunque representa un avance significativo en la predicción y análisis de patrones rítmicos a partir de audio real, presenta varias limitaciones que deben ser consideradas para contextualizar los resultados y orientar futuras mejoras.

Dependencia de la Calidad y Variedad del Dataset

Los resultados obtenidos dependen en gran medida del Magenta Groove MIDI Dataset, que si bien ofrece correspondencia entre archivos .MIDI y .WAV, no abarca toda la variedad de estilos, patrones rítmicos o contextos culturales necesarios para lograr un modelo verdaderamente robusto y generalizable. La falta de diversidad puede limitar la capacidad del sistema para adaptarse a nuevos géneros, bateristas o grabaciones en vivo.

Desbalance en la Distribución de Clases

A pesar de los esfuerzos por equilibrar el dataset, la representación desigual de ciertos estilos rítmicos influyó en el desempeño de los modelos, especialmente en la capacidad de generalización hacia clases menos representadas. Esto se reflejó en las métricas finales, donde la accuracy en tareas de clasificación aún mostró margen de mejora (por ejemplo, el ensamble alcanzó un 20% de precisión).

Limitaciones técnicas de los modelos

Las arquitecturas utilizadas (CNN baseline, variantes con batch normalization, dropout y ensambles) resultaron adecuadas para una fase experimental, pero mostraron un límite en su capacidad explicativa frente a la complejidad del ritmo en audio real. El error cuadrático medio (MSE) se estabilizó en validación y la capacidad de clasificación fue moderada, lo que sugiere la necesidad de explorar modelos más complejos, como arquitecturas híbridas (CNN+RNN, Transformers) y técnicas avanzadas de regularización y optimización.

Conclusiones y trabajo futuro

Conclusiones generales

El proyecto Muzca propuso y validó un enfoque novedoso para representar patrones rítmicos a partir de audio real, utilizando la representación unidimensional Frequency-Weighted Onset Density (FWOD) como puente entre la percepción rítmica humana y los modelos computacionales de análisis musical.

A través de un pipeline completo —desde la alineación de datos reales hasta la generación de espectrogramas Mel y la predicción del vector FWOD mediante modelos basados en CNN— se logró comprobar que es posible predecir representaciones rítmicas explicables directamente desde audio de percusión. Si bien el desempeño en la clasificación final aún muestra áreas de mejora, los resultados obtenidos validan la factibilidad técnica y metodológica del enfoque propuesto.

Perspectivas del framework FWOD

Uno de los principales aportes de esta investigación es el fortalecimiento del FWOD como representación rítmica útil y compacta. Su aplicabilidad tanto en contextos simbólicos como en datos reales le permite posicionarse como el eje de un posible framework de análisis musical especializado en ritmos de batería.

Este enfoque tiene el potencial de permitir la creación de sistemas de clasificación centrados en el ritmo, independientes de la armonía o la melodía, con aplicaciones prácticas en educación musical, análisis de interpretaciones humanas, acompañamiento automático y composición asistida.

Aunque existen numerosos trabajos sobre clasificación musical basados en audio, pocos se enfocan exclusivamente en percusión. La naturaleza percusiva del FWOD lo convierte

en una herramienta clave para este tipo de aplicaciones, especialmente en sistemas interactivos o pedagógicos.

Líneas de trabajo futuro

Las siguientes acciones se identifican como próximas etapas para ampliar y fortalecer el enfoque planteado:

- Ampliación y diversificación del dataset: Incluir nuevos géneros, bateristas, contextos culturales y grabaciones en vivo, aumentando la variedad y representatividad de los datos de entrenamiento.
- Exploración de arquitecturas avanzadas: Implementar modelos híbridos como CNN + RNN o Transformers, con el fin de capturar tanto la estructura espectral como la dinámica temporal de los patrones rítmicos.

Estas líneas de trabajo buscan reforzar el potencial del sistema propuesto como una plataforma sólida y explicable para el análisis rítmico automatizado.

Referencias

Behzad, H., Kotowski, B., Lee, C. L. I., & Jordà, S. (2023). TapTamDrum: A dataset for dualized drum patterns. *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR 2023)*. Milán, Italia. http://hdl.handle.net/10230/58123

Casey, M., Veltkamp, R. C., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE, 96*(4), 668–696. https://doi.org/10.1109/JPROC.2008.916370

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392–2396). IEEE. https://doi.org/10.1109/ICASSP.2017.7952585

Dubnov, S. (2021). Cross-modal and intermodal analysis in music information retrieval. *Journal of New Music Research, 50*(3), 214–229. https://doi.org/10.1080/09298215.2021.1878921

G'er'e, L., Rigaux, P., & Audebert, N. (2024). Improved symbolic drum style classification with grammar-based hierarchical representations. *arXiv preprint arXiv:2407.17536*. https://api.semanticscholar.org/CorpusID:271432077

Gillick, J., Roberts, A., Engel, J., Eck, D., & Bamman, D. (2019). Learning to groove with inverse sequence transformations. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2269–2279). PMLR. https://proceedings.mlr.press/v97/gillick19a.html

Gómez-Marín, D., Aponte, A., & Giraldo, J. (2024). Tapped representations of polyphonic patterns. *Unpublished manuscript*.

Gómez-Marín, D., Jordà, S., & Herrera, P. (2020). Drum rhythm spaces: From polyphonic similarity to generative maps. *Journal of New Music Research, 49*(5), 438–456. https://doi.org/10.1080/09298215.2020.1806887

Jordà, S., Kotowski, K., Lee, K., & Behzad, A. (2023). Generative models for rhythm and performance: Advances and challenges. *Journal of New Music Research, 52*(2), 100–115. https://doi.org/10.1080/09298215.2023.1234567

Mercurio, M. (2020). *Representación melódica transformacional*. Universidad Católica Argentina. https://repositorio.uca.edu.ar/bitstream/123456789/10882/1/representacion-melodica-transformacional.pdf

Molinari, L. (2005). Un recorrido por formas de representación y manifestaciones musicales. https://repositoriodigital.uns.edu.ar/bitstream/handle/123456789/3495/Molinari%2C%20L.%20Un%20recorrido.pdf%3Bsequence%3D1

Pelinski, R. (2005). Corporeidad y experiencia musical. *Revista Transcultural de Música (9)*. https://www.redalyc.org/pdf/822/82200913.pdf

Pons, J., Lidy, T., & Serra, X. (2017). Experimenting with musically motivated convolutional neural networks. In *Proceedings of the 14th Sound and Music Computing Conference*.

Schulzrinne, H. (n.d.). 44.1 kHz sampling rate. *Columbia University*. Retrieved December 11, 2024, from https://www.cs.columbia.edu/~hgs/audio/44.1.html

Serra, X., Müller, M., & Lattner, S. (2013). Audio content analysis. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 341–359). Springer. https://doi.org/10.1007/978-3-642-50318-0_15

Siedenburg, K., McAdams, S., & Popescu, T. (2016). A dimensional model of auditory perception for sound categorization. *Psychological Review, 123*(5), 452–491. https://doi.org/10.1037/rev0000037

Temperley, D. (2010). *Music and probability*. The MIT Press. https://doi.org/10.7551/mitpress/9780262516240.001.0001

Anexos

Anexo A. Resultados detallados del alcance inicial: MIDI → FWOD → Clasificación

El alcance inicial del proyecto Muzca se centró en validar la representación Frequency-Weighted Onset Density (FWOD) generada a partir de archivos MIDI. Este flujo fue el punto de partida metodológico para la comparación con las representaciones obtenidas desde audio real.

Modelos evaluados

Se implementaron y evaluaron los siguientes modelos de clasificación:
- Random Forest
- SVM (Support Vector Machines)
- XGBoost
- KNN (K-Nearest Neighbors)

Configuraciones del dataset

Se consideraron distintas configuraciones del dataset para el entrenamiento y la prueba de los modelos:

- 1P_IN_BF: 1 patrón, sin inclusión de Bars secuenciales, dataset desbalanceado
- 2P_IY_BF: 2 patrones, inclusión de Bars secuenciales, dataset desbalanceado
- 2P_IY_BT: 2 patrones, inclusión de Bars secuenciales, dataset balanceado
- 4P_IY_BT: 4 patrones, inclusión de Bars secuenciales, dataset balanceado

Cuadro resumen de resultados (Accuracy)

Modelo	1P_IN_BF	2P_IY_BF	2P_IY_BT	4P_IY_BT
Random Forest	0.630	0.623	0.827	0.905
SVM|	0.618	0.672	0.761	0.872
XGBoost	0.667	0714	0781	0.866
KNN	0.644	0.695	0.746	0.816

Principales hallazgos

- El mejor resultado se alcanzó con Random Forest (4P_IY_BT), logrando un accuracy del 90.5%, superando el 66.3% reportado en el estado del arte (Behzad et al., 2023).
- El balanceo del dataset y el uso de más patrones contribuyeron significativamente a la mejora del rendimiento.
- La inclusión de Bars secuenciales (IY) también tuvo un impacto positivo en las métricas.

Figuras y métricas adicionales

Las siguientes figuras documentan los resultados del alcance inicial, incluyendo la distribución de géneros en el dataset, gráficas de precisión, recall y F1-score por clase, y matrices de confusión de los mejores modelos.

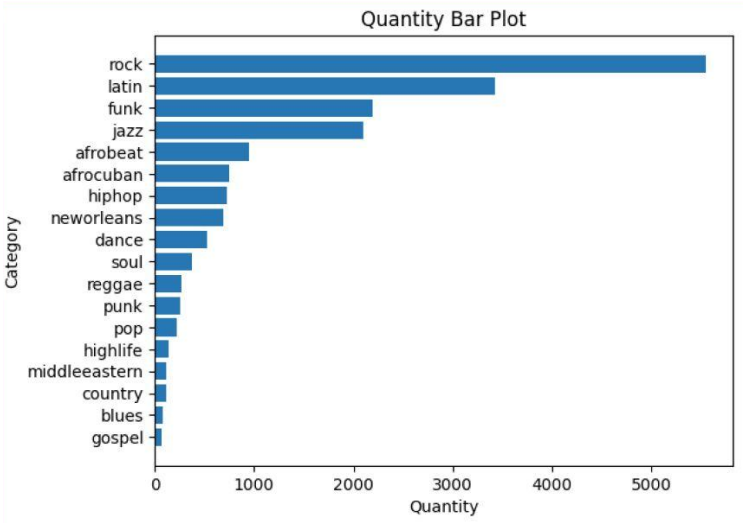
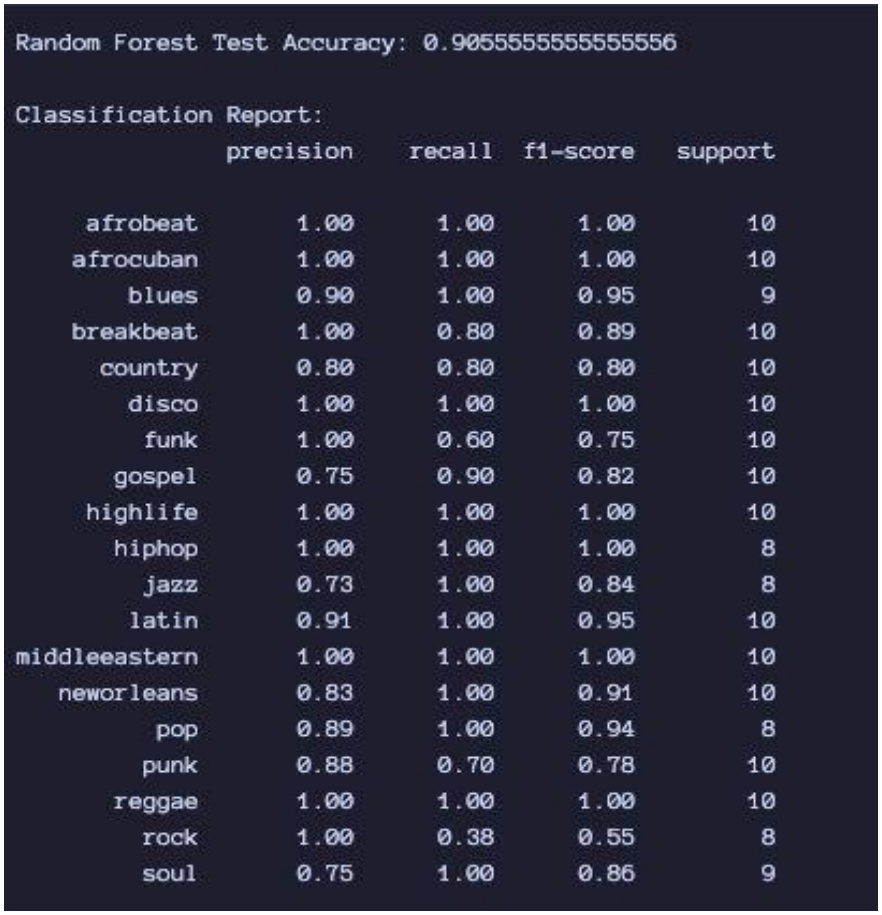


Figura A1. Distribución de géneros en el dataset



Random Forest Test Accuracy: 0.9055555555555556

Classification Report:

	precision	recall	f1-score	support
afrobeat	1.00	1.00	1.00	10
afrocuban	1.00	1.00	1.00	10
blues	0.90	1.00	0.95	9
breakbeat	1.00	0.80	0.89	10
country	0.80	0.80	0.80	10
disco	1.00	1.00	1.00	10
funk	1.00	0.60	0.75	10
gospel	0.75	0.90	0.82	10
highlife	1.00	1.00	1.00	10
hiphop	1.00	1.00	1.00	8
jazz	0.73	1.00	0.84	8
latin	0.91	1.00	0.95	10
middleeastern	1.00	1.00	1.00	10
neworleans	0.83	1.00	0.91	10
pop	0.89	1.00	0.94	8
punk	0.88	0.70	0.78	10
reggae	1.00	1.00	1.00	10
rock	1.00	0.38	0.55	8
soul	0.75	1.00	0.86	9

Figura A2. Resultados del modelo XGboost (4 Patterns, Inclusive Yes, Balanceo Sí)

Anexo B. Rhythm Space: Resultados detallados

En el marco del flujo inicial del proyecto Muzca, se exploró un modelo de regresión para proyectar patrones rítmicos en un espacio bidimensional denominado Rhythm Space, empleando las coordenadas X e Y generadas a partir de la representación Frequency-Weighted Onset Density (FWOD). Este análisis buscó validar la capacidad del FWOD para organizar y agrupar patrones rítmicos según su similitud percibida.

Tratamiento del dataset

Se abordaron dos retos principales:

- Desbalance de ritmos: se identificó un marcado desbalance en la distribución de clases.
- Clustering y redistribución: se aplicó el algoritmo K-means (k=10) para segmentar el espacio rítmico y redistribuir los patrones, logrando una representación más homogénea.

El dataset balanceado resultante se dividió en 80% entrenamiento y 20% prueba, manteniendo la distribución de los clusters.

Modelos implementados

- Modelo base: red neuronal con TensorFlow, dos capas densas, función de pérdida MAE.
- Modelo optimizado: XGBoost, con hiperparámetros ajustados mediante RandomizedSearchCV (número de estimadores, profundidad máxima, tasa de aprendizaje).

Principales resultados

Modelo	MAE obtenido	Observaciones
Red neuronal base	Valor base de referencia	Entrenada sin balanceo inicial
XGBoost optimizado	Mejorado respecto al baseline	Impacto positivo del balanceo

- Coherencia en el posicionamiento: El modelo organizó los géneros en el espacio de manera lógica, agrupando géneros similares.
- Patrones de relación: Se observaron agrupaciones significativas y relaciones de similitud perceptual.

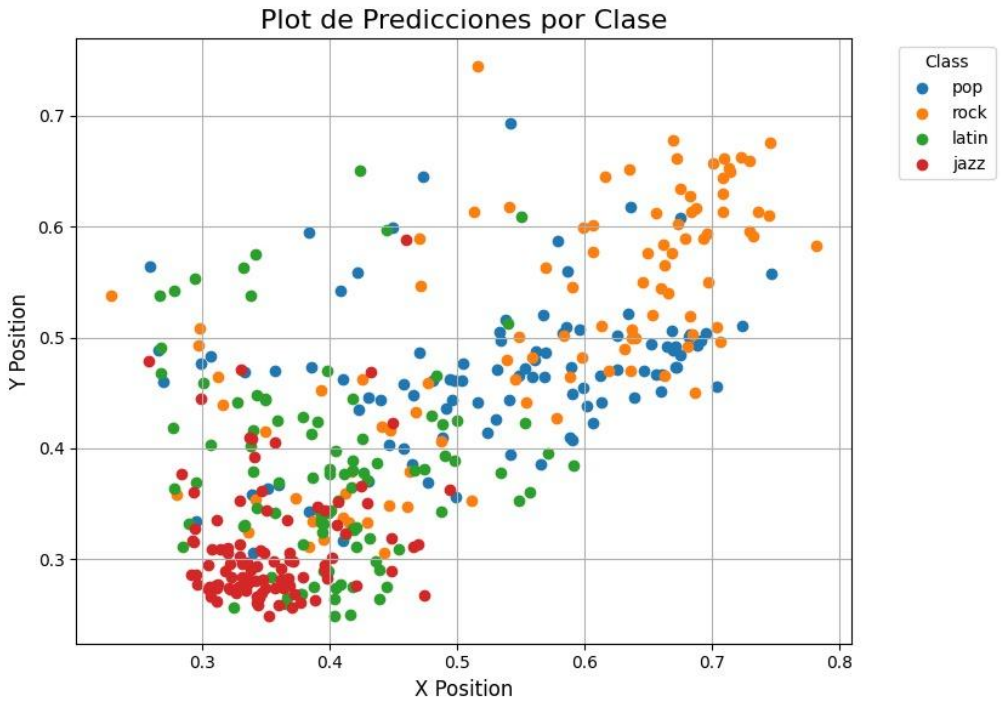


Figura B1. Distribución de puntos en el Rhythm Space con XGBoost

Anexo C. Resultados comparativos de los modelos CNN aplicados a audio real

Se desarrollaron y compararon cuatro variantes de modelos CNN:

- cnn_baseline: arquitectura básica de CNN con capas convolucionales y densas.

  El modelo cnn_baseline sirvió como punto de partida para evaluar la capacidad básica de las redes convolucionales en la predicción de FWOD a partir del espectrograma Mel completo.

  Configuración principal:

  - Entrada: espectrograma Mel completo del compás.
  - Arquitectura:
    - Dos capas convolucionales (Conv2D) con filtros de tamaño 3×3 y activación ReLU.
    - Cada capa seguida de una operación de max pooling para reducción espacial.
    - Una capa densa intermedia antes de la capa de salida.
    - Capa de salida densa con 16 nodos (uno por componente FWOD) y activación lineal.
  - Optimizador: Adam.
  - Pérdida: MSE.

  Este modelo buscó establecer un rendimiento base sin regularización adicional.

- cnn_batchnorm: CNN enriquecida con capas de normalización por lotes para mejorar la estabilidad del entrenamiento.

El cnn_batchnorm fue diseñado para mejorar la estabilidad y eficiencia del entrenamiento mediante la inclusión de Batch Normalization, que ayuda a mitigar problemas de covariate shift durante el entrenamiento y facilita el uso de tasas de aprendizaje mayores.

Configuración principal:

- Arquitectura idéntica al cnn_baseline, pero con una capa BatchNormalization añadida después de cada capa convolucional.
- Esto permitió un entrenamiento más estable y con menor oscilación en las métricas de validación.

- cnn_dropout: CNN con capas de dropout para mitigar el sobreajuste.

  El cnn_dropout se centró en reducir el riesgo de sobreajuste, integrando capas de dropout como estrategia de regularización.

  Configuración principal:

  - Mismo esquema base de capas convolucionales y densas.
  - Inclusión de una capa Dropout con una tasa del 0.3 (30%) después de las capas densas, forzando a la red a no depender excesivamente de nodos específicos durante el entrenamiento.

  Esta variante buscó mejorar la capacidad de generalización del modelo, especialmente frente a un dataset limitado y desbalanceado.

- daniel_base: corresponde a un ensamble integrado por cinco modelos independientes, cuyos resultados fueron combinados para mejorar la precisión de las predicciones del vector FWOD.

  Configuración principal:

- Cada uno de los cinco modelos se entrenó de forma independiente utilizando el espectrograma Mel completo como entrada.
- Los modelos fueron combinados mediante un esquema de promedio ponderado de sus predicciones, generando un FWOD final que buscó captar lo mejor de cada red.
- Este ensamble logró un MSE de 0.077, mostrando un avance en la calidad de la predicción del FWOD frente a modelos individuales.

Sin embargo, al emplear el FWOD predicho por el ensamble en tareas de clasificación del estilo musical, el sistema alcanzó una accuracy del 20 %, lo que, si bien representó un avance, indicó la necesidad de seguir fortaleciendo la generalización del modelo y el tratamiento del desbalance de clases.

El ensamble sirvió como un punto de referencia sólido frente a las variantes individuales, destacando la utilidad de la combinación de predicciones para mejorar métricas de regresión, aunque con margen de mejora en su aplicación final en clasificación.

Cada modelo fue entrenado por 30 épocas con optimización en la tasa de aprendizaje y ajustes de batch size. Las métricas principales evaluadas fueron el MAE, MSE y R^2 sobre el conjunto de validación.