Modelos de clasificación de archivos de audio en géneros musicales, a partir de técnicas de machine learning con SVM y de deep learning con CNN

M. A. Echeverri-Carmona* and M. Toro-Escobar†
Instituto de Física, Universidad de Antioquia (UdeA), calle 70 No. 52-21, Medellín, Colombia
(Dated: April 11, 2022)

Technological revolutions over the centuries have led to voice recognition, identification of visual patterns, and primarily classification of audio signals, becoming essential tools provided by artificial intelligence to poveer multiple metaverse users of musical predictions according to their interests. Recognizing the usefulness of this strategy in the commercial sector, it became necessary to deepen the algorithmic processes to achieve the categorization of audio signals in musical genres, which is why in this project the metric between two models of audio classification was compared, through the processing of its spectral signals with a CNN, and the analysis of its statistical parameters with an SVM. Both models were trained to make predictions of the test data sets, for which Mel MFCC diagrams were generated, and data from the CSV file were analyzed, respectively, to finely obtain as a result a metric of 56% for the SVM method and 51% for CNN, and achieving a greater prediction for genres like classic and hip hop, and a lower prediction for rock.

Keywords: musical genres, support virtual machine, convolutional neural network

I. INTRODUCCIÓN

El interés por simular y replicar los procesos naturales del cuerpo, el cerebro, y esencialmente la mente humana, es lo que ha dado origen al concepto de inteligencia artificial (IA), que enfocada en las ciencias de la computación, ha permitido a través de la identificación y clasificación de datos, la predicción de diversos fenómenos, convirtiéndose en una herramienta fundamental para decodificar información útil tanto para el sector comercial como académico.

Siendo así, es imprescindible estudiar las interacciones entre las máquinas y el lenguaje humano para diseñar sistemas inteligentes, y como un tipo de lenguaje es el sonoro, por utilidad en el presente proyecto, aquel formado por datos de audio que se presentan a través de la música, nos convoca el siguiente interrogante: *¿Con qué herramientas computacionales enfrentarnos al problema de la clasificación y etiquetado de géneros musicales?*, puesto que la clasificación sonora es útil por ejemplo, al momento de indexar colecciones de música según sus características de audio y proporcionar recomendaciones basadas en preferencias, para recuperar una canción similar a una que se tenga como referencia, e incluso para generar música sintética, y demás dinámicas del mundo digital y comercial.

En este sentido, nos planteamos el objetivo de **comparar las métricas entre dos modelos de clasificación en géneros musicales de archivos de audio, a través del procesamiento de sus señales espectrales haciendo uso de una red neuronal convolucional (CNN) y, del análisis de sus estadísticas con una máquina de soporte vectorial (SVM).**

A. Support virtual machine (SVM)

Podemos decir que el machine learning se basa en el uso de algoritmos para recopilar, ordenar, buscar patrones en datos, y aprender de ellos sin necesidad de una reprogramación. Con lo que queda claro entonces que, se tienden a utilizar datos que son clasificados y analizados, buscando patrones, mediante algoritmos que "aprenden" por cuenta propia, los cuales en últimas, pueden realizar predicciones y ayudar a tomar decisiones basándose en una serie de modelos [1].

Ahora, para efectos prácticos del proyecto, es primordial comprender que con support virtual machine, se trabaja directamente con las señales de audio, pues basta con proporcionar parámetros característicos del conjunto de datos, como lo son algunas de sus estadísticas, para a través de su análisis, lograr dichas predicciones previamente mencionadas.

B. Convolutional neural network (CNN)

Por otro lado, el deep Learning va un paso más allá del machine learning y utiliza algoritmos avanzados que simulan el comportamiento de las redes neuronales del cerebro humano. En este sentido, este tipo de algortimos son capaces de soportar gran cantidad de datos, conocidos como Big Data y, funcionar como una mente propia con superposición de capas no lineales, para procesar toda aquella información, re-

* alejandra.echeverri3@udea.edu.co
† mariana.toroe@udea.edu.co

alizar predicciones y adicionalmente, tomar decisiones[1].

Así, la red neuronal convolucional cobra relevancia en el proyecto, ya que a diferencia del método de support virtual machine, no se trabaja directamente con la señal de audio, sino que requiere de un proceso de digitalización de la señal, a través del cual se obtiene un espectrograma o variación del espectro de la señal a través del tiempo, como la representación compacta equivalente a la señal de audio original y, es con esta información visual, con lo que se logra la toma de decisiones de la máquina para la clasificación de los datos iniciales.

Este proceso se explica gráficamente en el siguiente esquema:

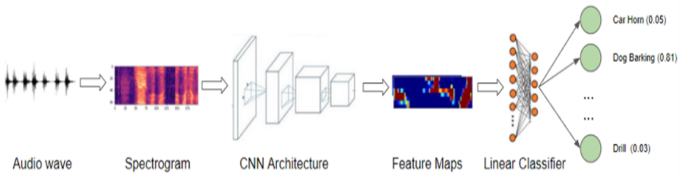

FIG. 1. Modelo de deep learning para el procesamiento de audios[2].

1. *Espectrogramas de Mel*

En lugar de un espectrograma simple que representa la frecuencia con dependencia temporal, los espectrogramas tipo Mel son de mayor utilidad en los modelos de aprendizaje profundo, pues realiza dos modificaciones significativas con las que se optimizan para obtener mejor rendimiento en el procesamiento de audios[3]:

- Se implementa la escala Mel en lugar de la frecuencia en el eje y, lo que permite obtener una representación en escala de tonos, midiendo la percepción humana de esta elevación del sonido.
- Utiliza la escala de decibelios en lugar de la amplitud para indicar los colores.

II. PROCEDIMIENTO EXPERIMENTAL

Se accedió al dataset GTZAN de libre circulación a través de la página de Kaggle[4], de la que se pudo descargar dicho conjunto de datos de manera gratuita y autorizada por los creadores, y teniendo en cuenta el carácter netamente computacional del proyecto, se utilizó el lenguaje de programación Python 3.10, haciendo uso de paquetes de librería como pandas, scikit-learn, librosa, tensorflow, tqdm, entre otras.

A continuación, se desarrollaron dos algoritmos diferentes: El primero de ellos consistió en un algoritmo de deep learning en el que se entrenó una red neuronal convolucional CNN ingresando diagramas tipo MFCC de las señales; Y en el segundo de ellos se desarrolló y entrenó un algoritmo de machine learning conocido como SVM, trabajando con los datos del archivo CSV features_30_sec.csv adjunto en el dataset. Ambos programas se desarrollaron en un Jupyter Notebook y fueron compilados en el editor de código Visual Studio Code.

Para desarrollar el primer algoritmo, inicialmente se procedió a leer e importar las señales de audio con extensión .wav que se encuentran en la carpeta genres-original adjunta en el dataset, para lo que fue necesario hacer uso de librerías como tqdm, os y librosa, con el fin de transformar cada una de esas señales de audio en un array, para posteriormente guardarlo en un dataset. Paso siguiente, se dividió el dataset en un conjunto de prueba, que contenía el 20% de los datos y en un conjunto de entrenamiento, que contenía el otro 80% de los mismos, haciendo uso del módulo train_test_split que provee la librería scikit-learn. En este punto es importante mencionar que no solo se obtuvieron los arreglos de las señales, sino también los respectivos labels que las caracterizan.

Posteriormente, se generaron los diagramas tipo MFCC de las señales con los siguientes argumentos sr = 22050, n_fft = 1024, n_mels = 128 y hop_length = 512, haciendo uso de la librería tqdm y del módulo librosa.feature.melspectrogram. Luego, se definió la red neuronal convolucional utilizando varios módulos de la librería Keras, bajo la siguiente arquitectura:

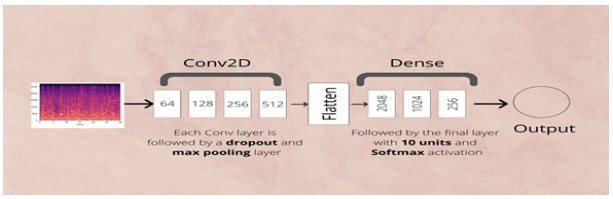

FIG. 2. Arquitectura CNN [5].

Finalmente se entrenó el modelo con 10 neuronas ingresando los diagramas MFCC del dataset de entrenamiento.

Para el segundo algoritmo se desarrolló una SVM, pero a diferencia del modelo anterior, no se utilizaron las señales de audio, sino que se trabajó con el archivo CSV features_30_sec.csv adjunto en el dataset. Se realizó el respectivo pre-procesamiento y limpieza de los datos. Al igual que en el modelo anterior, se dividió el dataset en un conjunto de prueba con un 20% de los datos y en un conjunto de entrenamiento con el otro 80% restante. Paso siguiente, se definió y entrenó la SVM con el conjunto de entrenamiento estandarizado, haciendo uso del módulo SVC con los hiperparámetros C=100, kernel='rbf' y gamma=0.01.

Finalmente, se evaluaron y compararon las métricas de ambos modelos, tanto el de deep learning como el de machine learning.

III. RESULTADOS Y ANÁLISIS

La cantidad de datos con los que se trabajó en el dataset de prueba para ambos modelos son:

TABLA I. Cantidad de señales de audio por género, que constituyen el dataset de prueba de los códigos.

Género	Número de señales de audio
Blues	22
Classical	23
Country	18
Disco	18
Hip hop	22
Jazz	19
Metal	15
Pop	26
Reggae	17
Rock	20

Adentrándonos en el método CNN, uno de los espectrogramas de Mel resultantes se expone en la figura a continuación:

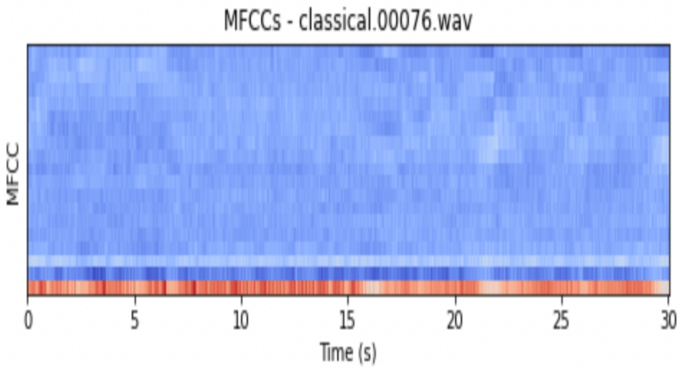

FIG. 3. Espectograma Mel de la señal 76 del género clásico.

Luego, a evaluar los diferentes modelos se consigue el siguiente resultado:

TABLA II. Accuracy de los modelos aplicados.

Modelo	Accuracy	Error
Support Vector Machine	0.56	0.44
Convolutional Neural Network	0.51	0.49

A partir de la *TABLA II.* se determina que el modelo con mayor precisión y mejores resultados es el modelo de machine learning, la SVM, con una accuracy del 56%, pues con respecto a este, el modelo de deep learning, la CNN, arroja una precisión máxima de el 51%. Tal resultado podría deberse a que al enfrentarnos a un problema de clasificación múltiple, y al trabajar con una red convolucional 2D, los datos de entrada, es decir los espectrogramas de Mel para el modelo CNN, deben estar considerablemente mejor procesados para que la red neuronal aprenda por sí misma a distinguir los patrones en dichos diagramas que posibilitan la adecuada clasificación entre géneros; así pues, se atribuye este resultado a la arquitectura de la CNN, pues a diferencia de el modelo SVM que trabaja directamente con las estadísticas de las señales, se sugiere hacer una mejor extrapolación de la señal para la CNN. En relación a lo anterior, la red neuronal convolucional requiere construirse más compleja, con una arquitectura basada en más bloques convolucionales, de forma tal que se puedan ingresar varias imágenes que caractericen la misma señal de audio, como por ejemplo, adicionar a los diagramas de Mel, los diagramas Chroma Feature que proveen información sobre las características armónicas y melódicas según los perfiles de tono de las señales, para que al tener más datos de visuales de estas, las predicciones mejoren considerablemente.

Centrándonos entonces en el modelo CNN, se analiza la dependencia de los dataset con la cantidad de neuronas:

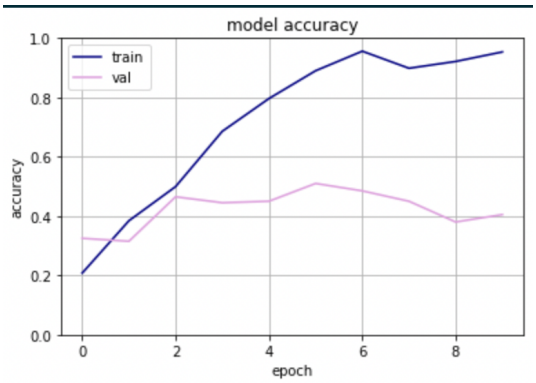

FIG. 4. Accuracy del modelo CNN.

La *FIG. 4.* muestra cómo evoluciona la precisión tanto de los datos de entrenamiento como de los datos de prueba, a medida que se incrementa el numero de neuronas de la red para el modelo de deep learning. En ella se puede observar que la evolución del accuracy para los datos de entrenamiento es muy buena, pues presenta un comportamiento creciente más o menos regular; pero en contraste a esto, la curva correspondiente a los datos de prueba presenta un comportamiento irregular, tanto así, que la mejor precisión del modelo solo se evidencia cuando la red convolucional alcanza las 6 neuronas.

Luego, procediendo con la evaluación de este modelo de clasificación, se presenta para la CNN, una herramienta de estudio importante conocida como la matriz de confusión:

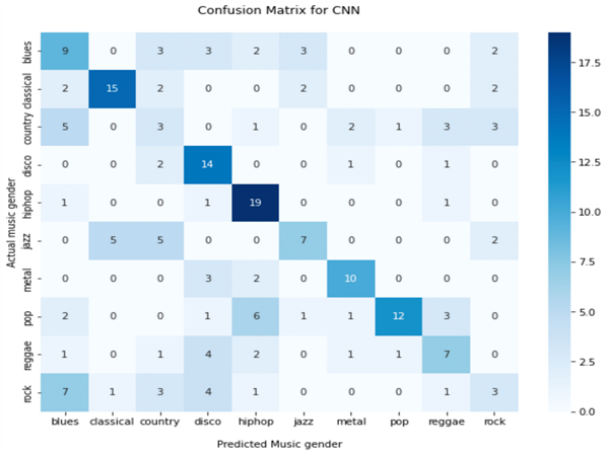

FIG. 5. Matriz de confusión para la CNN.

De dicha matriz representada en la *FIG. 5.* se puede observar que el género con mejor predicción de este algoritmo es el hip hop, pues cerca del 85% del pronóstico de dicho género es acertado; y a diferencia de esto, el género con la peor predicción resulta ser el rock, puesto que solo el 15% de las predicciones son apropiadas, viéndose que la red neuronal clasificó 7 de 20 canciones de rock como blues, lo cual sugiere que el tipo de rock que compone la base de datos es bastante melódico, por lo que los espectrogramas de Mel de este género y del blues tienen una una gran similitud entre tonos, y por consiguiente, la CNN no es capaz de hacer el correcto discernimiento entre ambos géneros; de hecho, también se puede atribuir la predicción poco acertada del country, correspondiente a un 17%, a que en la práctica, este género musical resulta ser una combinación entre géneros como el blues, pop y rock, por lo que la CNN presenta limitaciones para clasificarle. Y en contraposición a esto, se puede apreciar que la CNN aprendió muy bien a clasificar géneros musicales con caracaterísticas muy propias de sí mismos, como por ejemplo el metal y el clásico.

Luego, en cuanto al modelo SVM se tiene respectivamente que:

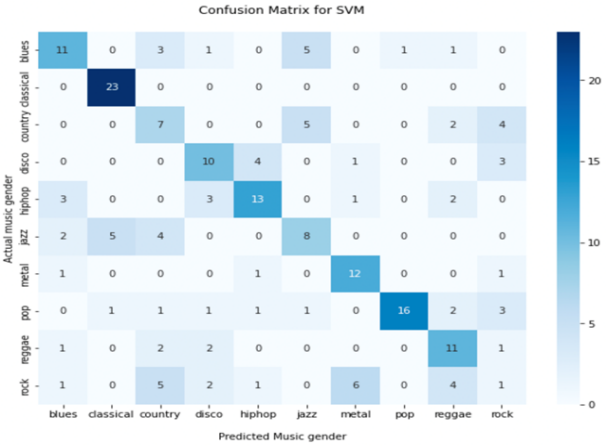

FIG. 6. Matriz de confusión para la SVM.

Al igual que en el modelo anterior, en la *FIG. 6.* se evidencia que la predicción menos efectiva le corresponde al rock, pues únicamente el 10% de las predicciones para este género son acertadas, lo que reafirma lo expuesto con anterioridad, y es que el tipo de rock que se tiene en el dataset es bastante melódico, y por consiguiente ninguno de los algoritmos puede clasificarlo de manera acertada. Y en contraste a esto, se sigue la misma línea predictiva del modelo CNN, pues con la SVM se obtiene una predicción del 100% para el género de música clásica, el cual constituyó ser uno de los mejores predichos también en el método primeramente analizado.

En retrospectiva, el comportamiento de ambos modelos es muy parecido, en cuanto oscilan con una métrica entre 0.5 a 0.6, y además, predicen de manera acertada aproximadamente en la misma proporción, los distintos géneros musicales.

IV. CONCLUSIONES

Con la metodología implementada se concluye que, con respecto al modelo CNN, el modelo SVM posee una mayor métrica, correspondiente a un 56% de categorización acertada de las señales de audio en géneros musicales. Estos resultados ponen en evidencia la ardua tarea de clasificación del sonido, pues incluso con un modelo tan complejo como la CNN y con datos aumentados, someramente se logra cruzar el umbral del 50% de precisión.

También, los géneros mejor predichos son los que tienden identificarse como géneros puros, con simetría y equilibrio entre las notas y tonos, como lo es el clásico. Y en contraposición a esto, cuando convergen características de diversos géneros para dar origen por ejemplo, al country, se requiere de una mayor complejidad del modelo predictivo.

En este sentido, se evidencia que el entrenamiento de los algoritmos, es decir el prepocesamiento de los datos que serán ingresados para entrenar el modelo, constituye la parte más esencial y elemental en todo algoritmo de clasificación, y particularmente en este problema, la limpieza y filtrado de información en las señales al principal rango humano audible es una labor clave para evitar confusiones en las predicciones.

Finalmente, queda claro que si bien los modelos de machine learning están actualmente progresando, aún necesitan intervención del hombre para realizar los ajustes necesarios que permitan buenas predicciones, mientras que los modelos de deep learning se encuentran capacitados para determinar por sí mismos sus malas predicciones, a través de su red neuronal. En este sentido, se puede concluir que este tipo de algoritmos funcionan adecuadamente siempre que la complejidad de los mis-

mos sea acorde al problema y el preprocesamiento de los datos de entrenamiento sea el adecuado.

V. REFERENCIAS

[1] Kadir, A. A., Xu, X., Hämmerle, E. (2011). Virtual machine tools and virtual machining—a technological review. Robotics and computer-integrated manufacturing, 27(3), 494-508.

[2] Doshi, K. (2021, febrero). Audio Deep Learning Made Simple (Part 1): State-of-the-Art Techniques. Towards Data Science. Recuperado de: https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504

[3] Doshi, K. (2021, febrero). Audio Deep Learning Made Simple (Part 2): Why Mel Spectrograms perform better. Towards Data Science. Recuperado de: https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505

[4] Olteanu, A. (2020). GTZAN Dataset - Music Genre Classification. Kagle Audio Files. Recuperado de: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

[5] Moosa, A. (2021). Music Classification Using Deep Learning — Python. Analytics Vidhya. Recuperado de: https://medium.com/analytics-vidhya/music-classification-using-deep-learning-python-b22614adb7a2