UNIVERSIDAD DE GUANAJUATO

CAMPUS IRAPUATO – SALAMANCA

DIVISIÓN DE INGENIERÍAS

"Identificación de Instrumentos Musicales en Audios utilizando Análisis de Señales e Inteligencia Artificial"

TESIS

PARA OBTENER EL GRADO DE:

MAESTRO EN ADMINISTRACIÓN DE TECNOLOGÍAS

PRESENTA:

Alan Salomón Vázquez Robledo

DIRECTORA DE TESIS:
Dra. Rocío Alfonsina Lizárraga Morales

YURIRIA, GTO Agosto, 2025

Agradecimientos

Personales

La música tan sublime y profunda que me ha brindado una hermosa manera de entender y ver la vida, desde un paisaje de la mente y la percepción espiritual y mental humana unidad al cosmos y al universo mismo para darnos un regalo tan profundo de la creación divina en cada una de sus notas musicales.

Quiero expresar mi más profundo y cariñoso agradecimiento a mis padres, a toda mi familia, a mis hermanas, a mi tía Noelia, su presencia y aliento fueron fundamentales en cada paso del camino, además de compañeros de posgrado y a mi compañera Montserrat, por su constante apoyo y por compartir conmigo sus conocimientos a lo largo de este posgrado. A los docentes que me han brindado su conocimiento. A las personas que me apoyaron con su cariño y amor, en seguir mis metas y sueños en mis proyectos relacionados con la música. En particular a la Dra. Roció Alfonsina Lizárraga Morales por sus conocimientos y paciencia.

Institucionales

Al Departamento de Estudios Multidisciplinarios de la Universidad de Guanajuato y al SECIHTI por su gran apoyo, con el número de beca, CVU: 1192441.

Resumen

En la sociedad actual, la música juega un papel crucial en el desarrollo de la identidad cultural de cada persona, ayuda a expresar las emociones de una manera artística, narrando historias a través de melodías y crea conexiones entre individuos y comunidades mediante diversos géneros o sentimientos representados en esta. En la actualidad, el desarrollo digital de la música nos presenta un campo emergente con nueva información, características distintivas y con ello nuevas problemáticas a resolver.

Los Sistemas de Recuperación de Información Musical (MIR, por sus siglas en inglés, Music Information Retrieval) son un campo de investigación emergente basado en sistemas de software diseñados para extraer, analizar y recuperar información de archivos de audio musical. Estos sistemas combinan varias técnicas del procesamiento de señales, Aprendizaje de Máquina (ML por sus siglas en inglés, Machine Learning) e Inteligencia Artificial (IA por sus siglas en inglés, Artificial Intelligence), para trabajar con información musical. Dichos sistemas han surgido como solución innovadora para diferentes problemáticas sobre la organización de características e información de la música digital. Estos sistemas, emplean técnicas de procesamiento de señales y Aprendizaje de Máquina, permiten analizar automáticamente la estructura de las señales de audio y extraer información relevante, como la identificación de intérpretes o artistas, la identificación de canciones, clasificación de géneros musicales, detección de tempo y ritmo, análisis de emociones musicales y la identificación de instrumentos musicales.

En este trabajo, se propone un modelo de identificación de instrumentos musicales, mediante el uso de las técnicas de Aprendizaje de Máquina, como el Perceptrón Multicapa (MLP por sus siglas en inglés, Multi-layer Perceptron), las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés, Support Vector Machines) y los Vecinos más Cercanos (KNN por sus siglas en inglés, K Nearest Neighbours) y a través de la extracción de características como los Coeficientes Cepstrales de Frecuencia Mel (MFCC por sus siglas en inglés, Mel Frequency Cepstral Coefficients).

Palabras clave: Identificación de instrumentos musicales, Coeficientes Cepstrales de Frecuencia Mel, Aprendizaje de Máquina, Clasificación de audio.

Índice de contenido

Índice de Figuras.

Capítulo 1. Introducción.

La música mantiene una gran importancia en las personas, ya que forma parte de su identidad cultural. La identidad cultural de cada persona comprende los aspectos de creencias, los valores, las emociones y su expresión artística. La evolución de la música es una historia, que refleja como las personas han preservado, la manera de comunicar la música a lo largo del tiempo. Esta evolución parte desde la tradición oral hasta los formatos digitales. En la actualidad, el desarrollo tecnológico ha proporcionado una evolución constante, de cómo la música es creada y presentada al público en general. Esta evolución ha permitido una mayor accesibilidad a diferentes contenidos de archivos musicales.

Uno de los principales retos tecnológicos de la música digital, se encuentra relacionado con la organización y análisis de la cantidad de volúmenes de datos. Esta problemática se ha convertido en una tarea con un grado alto de dificultad. Se encuentra relacionada con la labor intensiva y poco eficiente de los profesionales del área musical, al tener que identificar manualmente las diferentes características de los archivos de audio. Esta tarea en especial es costosa, consume mucho tiempo y tiende a ser muy propensa a errores. Por esta razón es necesario el desarrollo de una herramienta para el análisis automático de datos.

Ante este contexto surgen los Sistemas MIR, que se define como el área encargada del análisis de la estructura y extracción de la información necesaria de una señal de audio, como lo plantea Alexandre M. Lucena (2020). El análisis de la recuperación de información musical es sumamente importante, ya que nos permite comprender y organizar la información musical de un solo archivo de audio. Algunas de las tareas principales de los sistemas MIR se centran en funciones como identificación de artistas, clasificación de géneros, clasificación de estados de ánimo, notación musical y la clasificación de instrumentos musicales.

Dentro de la literatura se identifican tres diferentes metodologías para abordar el tema de la clasificación de instrumentos musicales. La primera aproximación es el uso de técnicas de Procesamiento Digital de Señales (DSP, Digital Signal Processing). La segunda metodología se basa en el uso de estrategias de Aprendizaje de Máquina (Machine Learning). Por último, y recientemente, han emergido metodologías basadas en el Aprendizaje Profundo (Deep Learning).

Dentro de las técnicas basadas en Procesamiento Digital de Señales, la herramienta más popular para describir la energía que varía en el tiempo en diferentes bandas de frecuencias, es la Transformada de Fourier de Tiempo Corto (STFT, por sus siglas en inglés). Meinard Müller *et al.* (2011) describen el uso de la Transformada de Fourier de Tiempo Corto (STFT), visualizada como un espectrograma. Esto permite detectar cambios en la composición espectral de la señal a lo largo del tiempo, lo que puede ser útil para el análisis de señales de audio, como en el procesamiento de voz, la música y otras señales acústicas. Seema Ghisingh *et al.* (2016) mencionan que la característica más antigua utilizada en el procesamiento de señales de voz, son los Coeficientes Cepstrales de Frecuencia Mel (MFCC), tiene uno de los mejores resultados para el reconocimiento o identificación de señales, mostrando la

característica más importante para clasificar guitarra, violín y batería son los MFCC, ya que proporciona resultados más exactos. Además, de forma específica, la característica que proporcionan mejores resultados para la batería es Zero Crossing Rate (ZCR, por sus siglas en inglés). Se define, en su forma elemental, como el número de veces que una onda de señal cruza el cero de amplitud. Para dicho trabajo, el primer paso es aplicar la Transformada de Fourier (FT por sus siglas en inglés), que es una herramienta fundamental en el procesamiento de señales y se utiliza para analizar las características de frecuencia de las mismas. Guido (2016). Tepepa *et al.* (2018), emplea la técnica de Centroide Espectral (SC, por sus siglas en inglés, Spectral Centroid). Esta técnica se define como un parámetro utilizado en el análisis de señales de audio para caracterizar la distribución de la energía espectral a lo largo de las frecuencias presentes en una señal. Además, ayuda a representar la energía que se tiene en cada una de las ventanas, obteniendo vectores característicos de las pistas de audio. Dhvani Shah *et al.* (2022) mencionan el uso de la Transformada de Fourier, la cual permite descomponer una señal en sus componentes de frecuencia, lo que es útil para comprender la estructura armónica de una pieza musical, identificar patrones rítmicos o incluso para la compresión de audio, en su trabajo lograron una precisión de clasificación general de 92.38% y 93.19%, los autores emplearon dos clases musicales, específicamente el piano y el violín, para obtener los resultados de clasificación mencionados. Lo que indica el potencial en este dominio de clasificación y reconocimiento inherentemente temporal. Una de las ventajas en el Procesamiento Digital de las Señales, es el análisis específico donde los métodos pueden ser claros y entendibles, lo que permite reconocer cómo se extraen los atributos de las señales. Algunas de sus desventajas pueden ser el estilo de función operado manualmente.

Ahora, considerando las técnicas de Aprendizaje de Máquina, podemos encontrar la Máquina de Soporte Vectorial, que es un modelo de aprendizaje supervisado que se utiliza principalmente para tareas de clasificación y regresión. Por su parte S. Prabavathy (2020) nos propone la clasificación automática de instrumentos musicales como son el trombón, tuba, trompeta y piano utilizando SVM y KNN, como parte de sus resultados se muestra una precisión con SVM del 99.37 %. Sanraga Kingkor (2021) presenta un modelo de Red Neuronal Artificial (RNA), la cual consiste en un Perceptrón Multicapa de 13 neuronas en la capa de entrada y 20 neuronas en la capa de salida con una función de activación Softmax. Este sistema fue entrenado para realizar clasificación en 20 clases diferentes de instrumentos musicales de la Orquesta Filarmónica de Londres en conjunto con los MFCCs. En su trabajo se logró una precisión del 97% en el conjunto de datos completo que contiene las 20 clases de diferentes instrumentos musicales. Asimismo, se puede incluir la técnica de Spiking Neural Network (SNN, por sus siglas en inglés), que se centra en Redes Neuronales que funcionan eficazmente con datos temporales. El uso de SNN puede modelar más estrechamente la forma en que los humanos distinguen los instrumentos musicales al identificar diferencias en las características temporales de los instrumentos a medida que se reciben neuronalmente. La principal ventaja de utilizar Machine Learning en la Identificación de Instrumentos Musicales en Señales de Audio, es la capacidad para identificar patrones complicados que pueden ser difíciles de detectar utilizando otras técnicas.

Recientemente, una de las metodologías con más auge son las basadas en Aprendizaje Profundo (Deep Learning). Dentro de las opciones, una técnica relevante es la aplicación de Redes Neuronales Convolucionales (CNN o ConvNets). Estas se refieren a un tipo especializado de Red Neuronal Artificial diseñada específicamente para procesar datos como imágenes o señales de audio, los espectrogramas se envían a las Redes Neuronales Convolucionales para aprender patrones de cómo se visualizan los diferentes instrumentos musicales. Maciej Blaszke (2022) presenta la construcción de un algoritmo, para la automatización e identificación de instrumentos presentes en un extracto de audio utilizando conjuntos de Redes Neuronales Convolucionales individuales por instrumento, los cuales son bajo, batería, guitarra y piano. En dicho trabajo se logró una precisión aproximadamente del 100%, pero cuando se observan los resultados de reconocimiento de instrumentos musicales en conjunto, los valores métricos son más bajos. Una arquitectura similar es VGGNet, también conocida como Visual Geometry Group Network, es una arquitectura de Red Neuronal Convolucional. Chinmay Relkar (2019) nos presenta una Red Neuronal Convolucional, ConvNet de 4 capas inspirada en AlexNet, el cual lleva el nombre de VGGNet, con esta arquitectura se lograron mejores resultados en la tarea de reconocimiento de instrumentos en música polifónica. Chinmay Relkar (2019) también menciona la técnica de Red Neuronal Convolucional basada en regiones (RCNN, por sus siglas en inglés), esta técnica fue una de las primeras arquitecturas en abordar el problema de detección de objetos en imágenes utilizando Redes Neuronales Convolucionales. Una técnica variante es la Red Neuronal Recurrente Convolucional (CRNN, por sus siglas en inglés), esta técnica combina las ventajas de las Redes Neuronales Convolucionales para la extracción de características especiales y las Redes Neuronales Recurrentes (RNN) para modelar secuencias temporales. El uso del Aprendizaje Profundo en la Identificación de Instrumentos Musicales, tiene la ventaja de tener alta utilidad de modo que se desempeña a un nivel más alto cuando se trata de precisar la Identificación de Instrumentos Musicales, esto sucede cuando el modelo está entrenado en un conjunto adecuado de datos y situaciones. Con respecto a las desventajas, se requieren grandes cantidades de datos etiquetados en el entrenamiento efectivo de modelos de Aprendizaje Profundo para la Identificación de Instrumentos Musicales, lo que consume mucho tiempo en recopilar y etiquetar.

En este trabajo de tesis se presenta un modelo de clasificación de instrumentos musicales a través de la extracción de características de señales de audio y utilizando técnicas de Aprendizaje de Máquina. A continuación, en el Capítulo 2. Se muestran el Marco Teórico de los fundamentos matemáticos necesarios, para el Procesamiento Digital de Señales. Las diferentes fases de la metodología se describen en el Capítulo 3, en el se presenta el conjunto de datos a usar, el preprocesamiento de los archivos de audio, las técnicas del procesamiento digital de señales, la extracción de características, los modelos de clasificación utilizados y las métricas de evaluación. En el Capítulo 4. Se describen los experimentos y resultados, la herramienta desarrollada, los resultados para el experimento con 20 clases y con 31 clases. Finalmente, en el Capítulo 5. Se presentan conclusiones sobre la eficacia del enfoque propuesto y proporcionar recomendaciones para futuras investigaciones y mejoras en el rendimiento del modelo.

Capítulo 2. Marco Teórico.

Antes de comenzar a describir la metodología para este proyecto de tesis, es importante mencionar algunos conceptos teóricos, los cuales se presentan a continuación: Señales analógicas y digitales, Procesamiento Digital de Señales, Transformada de Fourier, Transformada Discreta de Fourier, Transformada Rápida de Fourier, Transformada de Fourier de Tiempo Corto y Cálculo de Coeficientes Cepstrales de Frecuencia Mel.

2.1. Señales.

Una señal se define como cualquier magnitud física que varía con el tiempo, el espacio o cualquier otra variable o variables independientes. Matemáticamente, describimos una señal como una función de una o más variables independientes. Las señales tienen una clasificación en cuanto a las características de la variable independiente, tiempo y valores de la misma. Para fines de este, trabajaremos con señales continuas y discretas, su definición correspondiente se presenta a continuación:

Señal continua (ver Figura 1): También llamadas señales analógicas están definidas para cada instante de tiempo y toman sus valores en el intervalo continuo (a, b), donde a puede ser $-\infty$ y b puede ser ∞. Matemáticamente, estas señales pueden describirse mediante funciones de una variable continua.

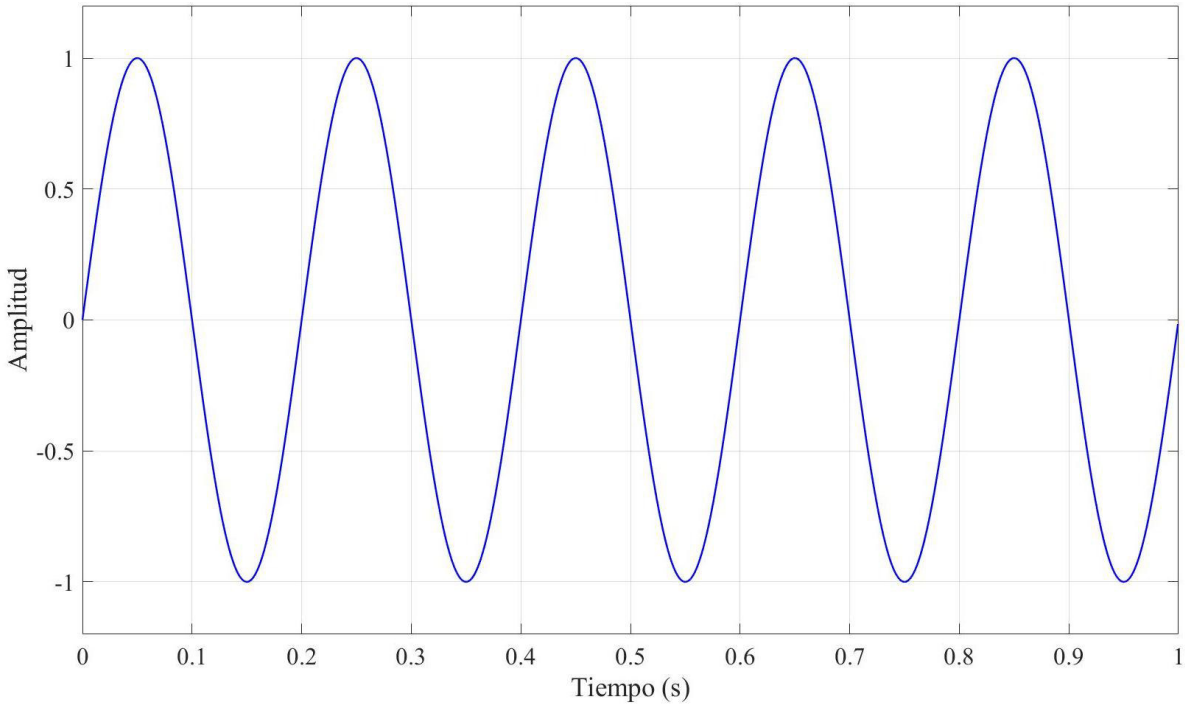


Figura 1. Señal continua.

Señal discreta (ver Figura 2): Sólo están definidas en determinados instantes de tiempo, en la práctica, normalmente están igualmente espaciados para facilitar los cálculos. Alternativamente, si la señal toma valores dentro de un conjunto finito de posibles valores, se dice que la señal es discreta. Estos valores son equidistantes y, por tanto, pueden expresarse como un múltiplo entero de la distancia entre dos valores sucesivos. Una señal discreta en el tiempo que tiene un conjunto de valores discretos es una señal digital. Para que una señal pueda ser procesada digitalmente, debe ser discreta en el tiempo y sus valores tienen que ser discretos (es decir, tiene que ser una señal digital). (Proakis, J. G., & Manolakis, D. G. 2007).

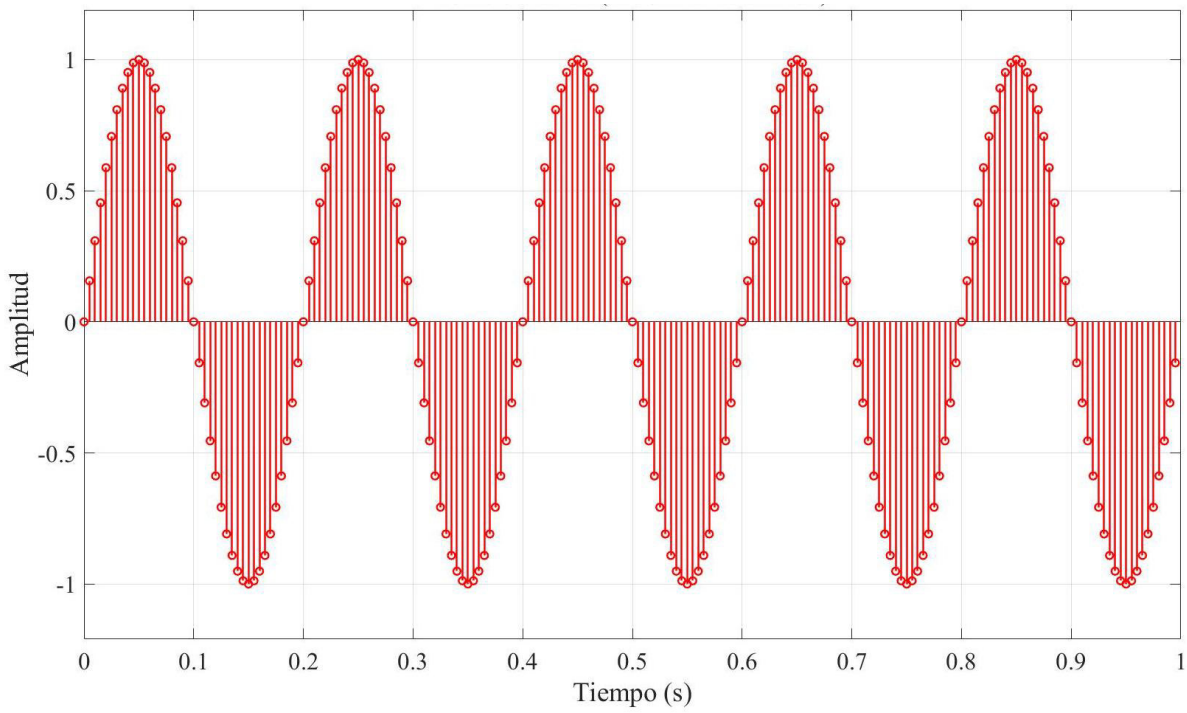


Figura 2. Señal discreta.

2.2. Procesamiento Digital de Señales.

El procesamiento digital de señales es importante en la ingeniería moderna ya que permite analizar, transformar y administrar información proveniente del mundo real como son: sonidos, imágenes y señales biomédicas, convirtiendo estas señales analógicas en digitales para su tratamiento. A diferencia de los sistemas analógicos, los sistemas digitales presentan una mayor facilidad de almacenamiento y la posibilidad de aplicar algoritmos complejos para mejorar o interpretar las señales. Algunas áreas donde ha tomado gran relevancia son las telecomunicaciones, medicina, inteligencia artificial y música. Una de las herramientas más importantes del procesamiento digital de señales es la Transformada de Fourier.

El concepto matemático que conocemos como Transformada de Fourier fue introducido por Joseph B. Fourier en 1811 (Gray & Goodman, 1995), en conexión con un tratado sobre la propagación del calor, mediante un argumento de paso al límite a partir de las series que

también llevan su nombre. No olvidemos el significativo papel que la Transformada de Fourier juega en el terreno de las aplicaciones, fundamentalmente en teoría de la señal, teoría cuántica de campos, tomografía y tratamiento y digitalización de imágenes. A continuación, se presentan las ecuaciones tanto para una señal continua o analógica, como para una señal discreta o digital:

Señal continua, se presenta a continuación en la Ecuación 1.

$$X(f) = \int_{-\infty}^{\infty} x\,(t) e^{-j2\pi ft}\, dt, \qquad (1)$$

Donde $X(f)$ es el resultado de la transformada: es la representación en frecuencia de la señal continua. $x(t)$ es la señal en el dominio del tiempo continuo. t representa el tiempo continuo. $e^{-j2\pi ft}$es el exponencial complejo (una sinusoide compleja), que sirve para "medir" la cantidad de frecuencia f presente en la señal.

Por otro lado, la Transformada Discreta de Fourier (TDF, por sus siglas en inglés), surgida en 1960 (Cooley, J. W. y Tukey, J. W. 1965), nos permite obtener el espectro de una forma de onda. Este espectro está representado por componentes sinusoidales cuyas frecuencias son armónicos de una frecuencia fundamental de análisis. Surge como una adaptación de la Transformada de Fourier para funciones definidas sobre un conjunto finito de puntos. En lugar de trabajar con funciones continuas, la DFT se usa para analizar secuencias de valores, lo que la hace especialmente útil en el análisis de señales digitales, se presenta a continuación en la Ecuación 2.

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j\frac{2\pi}{N}Kn}, \qquad k = 0,1,2,\dots,N-1 \quad (2)$$

Donde $e^{-j\frac{2\pi}{N}Kn}$ es el término exponencial complejo, que representa las componentes sinusoidales de diferentes frecuencias. j es la unidad imaginaria, equivalente a $\sqrt{-1}$. $x(n)$ es la señal de entrada en el dominio del tiempo, con n variando de 0 a $N-1$, donde N es la longitud de la señal. $X(k)$ es el valor de la señal transformada en el dominio de la frecuencia para el índice k, donde k varía de $0\ a\ N-1$. N es la cantidad total de puntos en la secuencia.

La Transformada Rápida de Fourier (FFT, por sus siglas en inglés), es un algoritmo para el cálculo de la Transformada Discreta de Fourier que reduce el tiempo de ejecución del programa en gran medida. Desde 1965, cuando James W. Cooley y John W. Tukey publicaron dicho algoritmo, su uso se ha expandido rápidamente y las computadoras personales han impulsado una explosión de aplicaciones adicionales de la FFT. Algunos ejemplos de la aplicación de la FFT son diseño de circuitos, espectroscopia, cristalografía, procesamiento de señales e imágenes.

En la Figura 3 se muestra una representación visual de la Transformada de Fourier de la señal del acorde “La mayor”, en el cual podemos visualizar tres notas principales que conforman dicho acorde como son: A4 (La4) con una frecuencia de 440 Hz. C#4 (Do sostenido 4) con

una frecuencia de 277.18 Hz y E4 (Mi4) con una frecuencia de 329.63 Hz. En esta misma Figura, en la parte central se visualiza el acorde superpuesto que corresponde a la suma de las señales de cada nota que componen dicho acorde. En la parte inferior se visualiza el resultado de aplicar la Transformada de Fourier en la señal compuesta. En esta representación se observan 3 picos definidos que corresponden exactamente a las frecuencias que conforman el acorde. Esto demuestra como la Transformada de Fourier permite identificar los componentes frecuenciales de una señal compleja como lo es el acorde musical de "La".

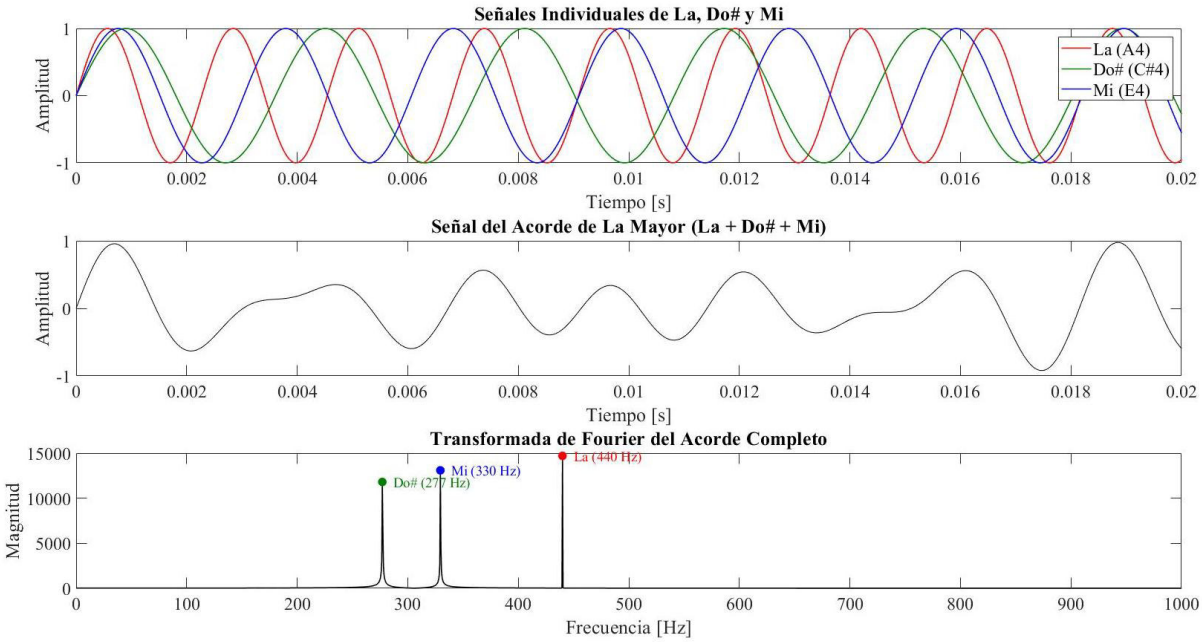


Figura 3. Transformada rápida de Fourier.

La Transformada de Fourier de Tiempo Corto (STFT, por sus siglas en inglés), fue introducida en la década de 1940-1980 es utilizada para el procesamiento de señales, examina las frecuencias presentes en una señal a medida que estas cambian a lo largo del tiempo. Está relacionada con la Transformada de Fourier estándar que permite analizar la señal no estacionaria, es decir, señal cuyas características frecuenciales cambian con el tiempo.

La STFT se puede utilizar para convertir señales cuyo contenido de frecuencia cambia con el tiempo al dominio tiempo-frecuencia, la ecuación de la Transformada de Fourier de Corto Tiempo, se presenta a continuación en la Ecuación 3.

$$STFT\{x(t)\} \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x\,(t)\omega(t-\tau)e^{-j\omega t}\,dt, \qquad (3)$$

Donde $STFT\{x(t)\}$ es la señal de entrada en el dominio del tiempo. Puede ser cualquier señal, como una señal de audio, una señal de vídeo, etc. $X(\tau,\omega)$ es la representación de la

señal $x(t)$ en el dominio del tiempo-frecuencia. En el contexto de la STFT, τ representa la posición en el tiempo y ω representa la frecuencia. Entonces, $X(\tau,\omega)$ representa la amplitud o la fase de la componente de frecuencia ω de la señal $x(t)$ en el tiempo τ. $\omega(t-\tau)$ es una función de ventana. En el contexto de la STFT, se utiliza una función de ventana para realizar un análisis en un intervalo de tiempo finito alrededor del tiempo τ. $e^{-j\omega t}$ es la función exponencial compleja que representa la rotación en el plano complejo. Esta función es parte de la Transformada de Fourier y está presente aquí para llevar a cabo la transformación de la señal $x(t)$ al dominio de la frecuencia.

La presente ecuación representa el valor de la STFT para una ventana centrada en el tiempo m y la frecuencia k. Es decir, que tantas componentes de la frecuencia k, están presentes en la sección de la señal que comienzan en el instante m, se muestra a continuación en la Ecuación 4.

$$x(m,k) = \sum_{n=0}^{N-1} x(n)\,(n+m)\,w(n)\,e^{-j\frac{2\pi}{N}kn}\,, \qquad (4)$$

Donde $x(m,k)$ es el resultado de la STFT, valor complejo que indica la presencia de la frecuencia k en la ventana de tiempo m. $x(n+m)$ es la señal original desplazada por m que indica en qué parte de la señal se está analizando (ventana localizada). $w(n)$ representa la ventana (por ejemplo, Hamming) que da forma a la señal dentro del segmento que se está analizando. $e^{-j\frac{2\pi}{N}kn}$ es la parte de la Transformada Discreta de Fourier que permite convertir la señal en el dominio de frecuencia. N es la longitud de la ventana (cuántas muestras se toman para cada análisis local). n indica el índice de muestra dentro de la ventana, va de 0 a $N-1$.

En la Figura 4 podemos visualizar la representación del acorde mayor de La, con sus respectivas notas musicales que la componen en el dominio de la frecuencia y el tiempo, en el cual se puede apreciar las frecuencias dominantes en pequeños segmentos de tiempo, como estas van cambiando en relación con el dominio del tiempo. En la parte inferior se muestran componentes frecuenciales que conforman el acorde, cuyas frecuencias y armónicos aparecen en diferentes momentos, las líneas horizontales de colores cálidos (amarillo y naranja) indican la presencia de ciertas frecuencias que corresponden a las notas del acorde en un tiempo aproximado de 0 a 2 segundos, en este se observan segmentos de ventanas de aproximadamente 200 milisegundos.

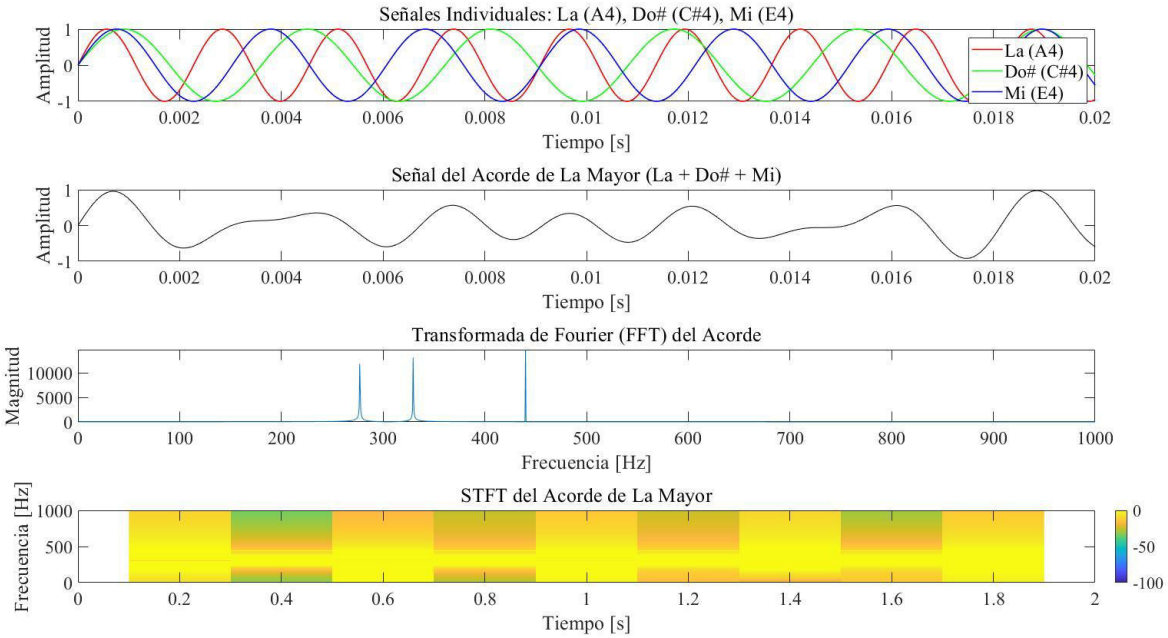


Figura 4. Transformada de Fourier de Tiempo Corto.

2.3. Cálculo de Coeficientes Cepstrales de Frecuencia Mel.

Los Coeficientes Cepstrales de Frecuencia Mel describen la forma espectral de una señal de audio y son ampliamente utilizados en tareas de reconocimiento de voz y clasificación de sonidos. Su cálculo implica diferentes etapas: En primer lugar, las bandas de frecuencia se posicionan logarítmicamente. Esto se denomina escala Mel. Posteriormente se utiliza un método que tiene la capacidad de compactación de energía denominado Transformada de Coseno Discreta (DCT), la cual solo considera los números reales de forma predeterminada, finalmente, se seleccionan los primeros 13 coeficientes resultantes, que son los que generalmente contienen la información más relevante para el análisis. Majeed *et al*. (2015).

En la Figura 5 se puede observar el proceso de cálculo de los Coeficientes Cepstrales de Frecuencia Mel (MFCCs), donde las bandas de frecuencia se distribuyen en la escala Mel de forma logarítmica y se aplica la Transformada de Coseno Discreta (DCT) para obtener los coeficientes, considerando únicamente los primeros 13 componentes como se menciona en el texto.

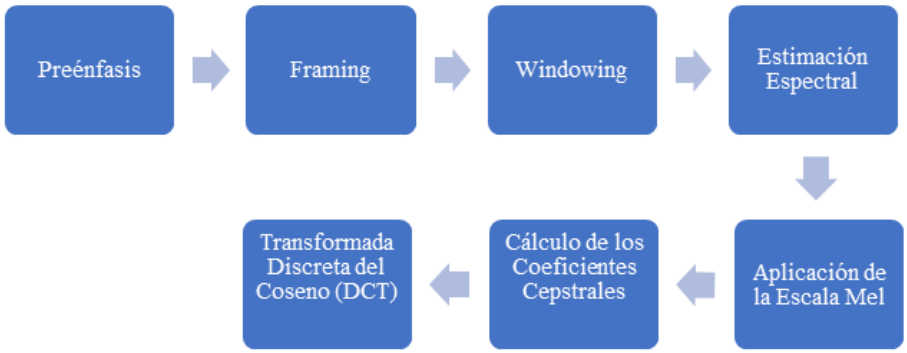


Figura 5. Pasos para calcular los MFCCs.

Para poder obtener los MFCCs, es necesario seguir los siguientes pasos:

Preénfasis: Se aplica un filtro de preénfasis para aumentar la energía de las frecuencias altas y reducir el DC offset que se define como el desplazamiento de corriente continua de una grabación; cuanto más profunda sea la frecuencia de una señal interferente, más se acercará a 0 Hz, lo que significa corriente continua. DC Offset se representa como un desplazamiento de la forma de onda hacia arriba o abajo. La ecuación es: donde α es típicamente 0.95, se presenta a continuación en la Ecuación 5.

$$H(z) = 1 - az^{-1} \quad 0.9 < a < 1, \qquad (5)$$

Donde $H(z)$ es el filtro en el dominio de z (Frecuencia), 1 se refiere a pasar la muestra actual tal cual (ganancia = 1), az^{-1} se refiere a que la salida es igual a la señal actual menos una fracción de la señal anterior. a se refiere a la fuerza del filtro (cuánto énfasis se pone en las frecuencias altas entre 0.9 y 1).

Framing: La señal de cada archivo de audio que tomamos se divide en bloques cortos llamados frames (segmentos o cuadros). La longitud típica del segmento es de 20-30 ms (milisegundos) y el desplazamiento es de 10 ms (milisegundos).

En la Figura 6 se observa el primer frame obtenido del proceso de segmentación de la señal de audio, donde se divide en bloques cortos con una longitud típica de 20-30 ms y un desplazamiento de 10 ms.

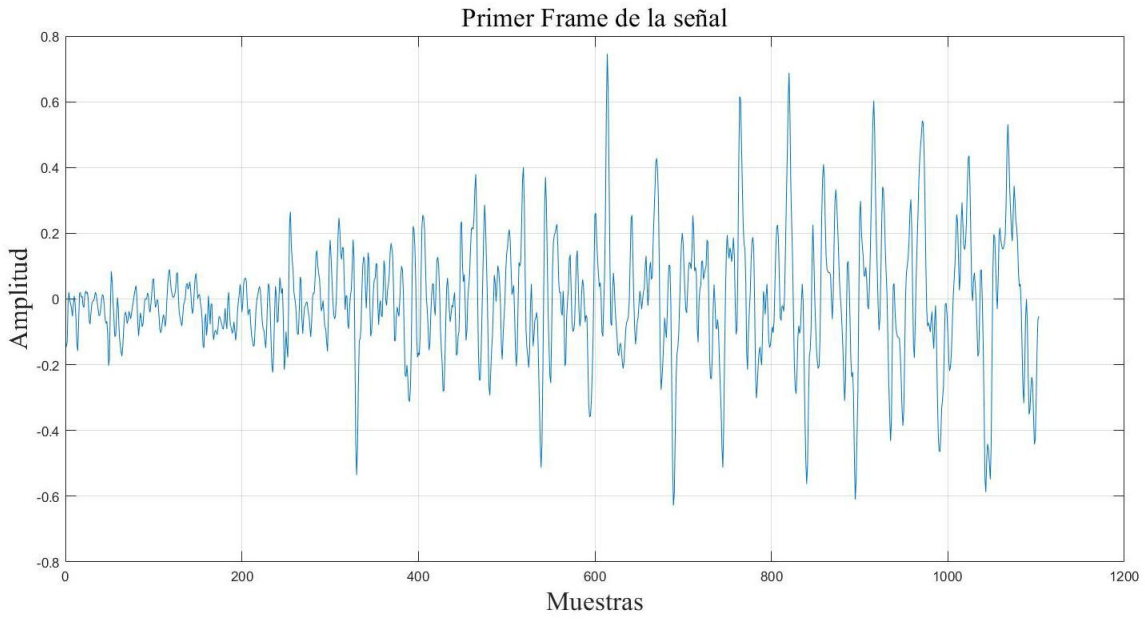


Figura 6. Primer Frame de la señal.

Windowing: Visualización de información en una ventana o recuadro, donde cada segmento se multiplica por una ventana que utiliza la función matemática para suavizar los bordes de este mismo llamada Hamming, para reducir discontinuidades. Donde N es la longitud del segmento, se presenta a continuación en la Ecuación 6.

$$h(n) = x(n)w(n)$$

$$w\,(n) = 0.54 - 0.46\,cos\left(\frac{2\pi}{N-1}\right), \qquad (6)$$

Donde $w\,(n)$ es el valor de la ventana en la muestra. La muestra n oscila entre 0 y $N-1$, 0.54 se refiere a la componente constante de la ventana donde $w\,(n)$ es el valor de la ventana en la muestra n ,donde n oscila entre 0 y $N-1$. $-0.46\,cos\left(\frac{2\pi}{N-1}\right)$ se refiere al componente oscilatorio que da a la ventana una forma suave.

La Figura 7 ilustra el proceso de windowing mediante la función de Hamming aplicada al primer frame, mostrando la transformación de la señal original (en amplitud y muestras) para minimizar las discontinuidades en los bordes del segmento.

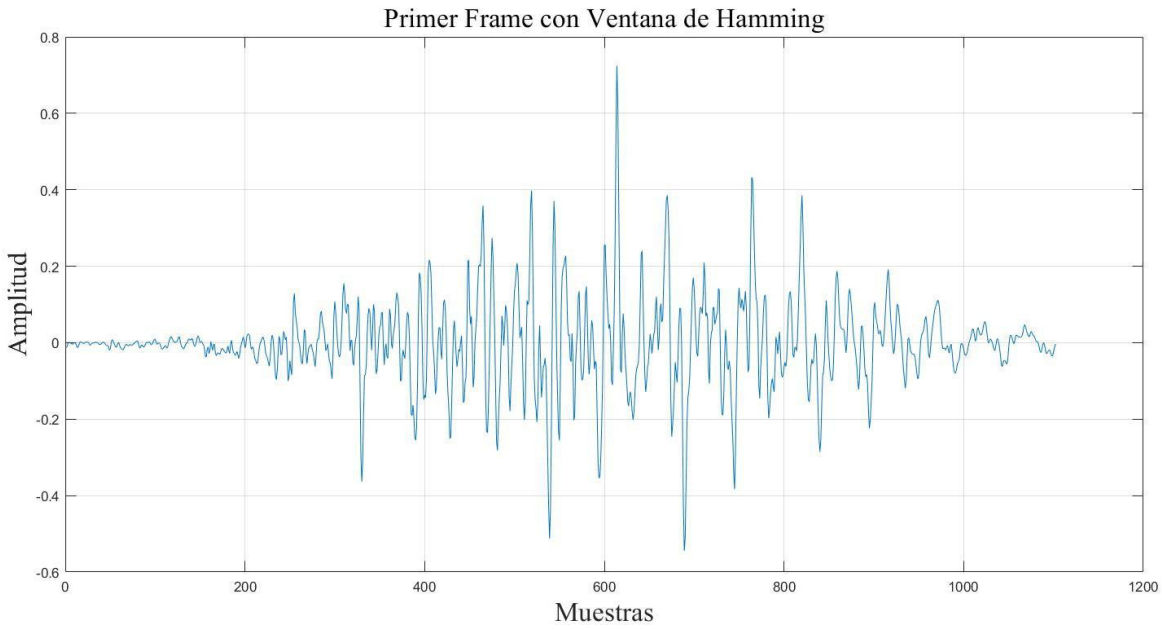


Figura 7. Proceso de Hamming en la señal.

Estimación espectral: Para este paso se aplica la Transformada Discreta de Fourier (DFT por sus siglas en inglés Discrete Fourier Transform) a cada frame (cuadro) para obtener los coeficientes espectrales, ver la Ecuación 3.

En la Figura 8 se observa, la estimación espectral generada por la Transformada Discreta de Fourier representa la descomposición frecuencial del frame procesado, obteniéndose los coeficientes complejos que caracterizan su contenido espectral.

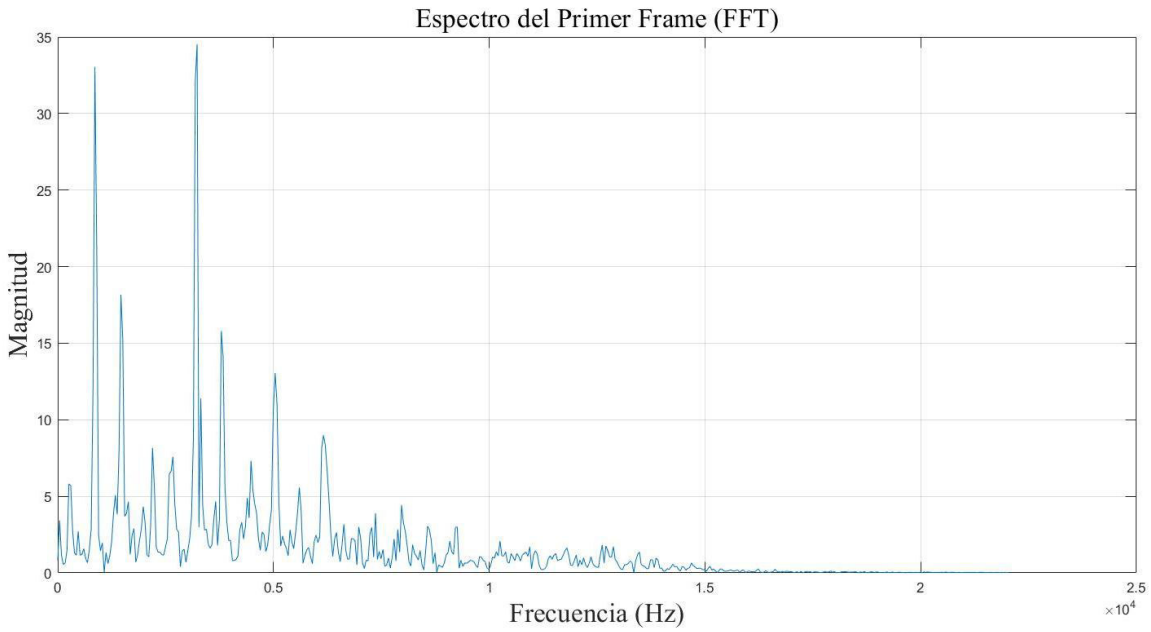


Figura 8. Estimación espectral.

Aplicación de la escala Mel: Se transforma la escala de frecuencia de cada audio a la escala Mel, la cual se centra en como los humanos perciben las frecuencias del sonido, que es más cercana a la percepción humana. Esto se hace utilizando un banco de filtros Mel, donde cada filtro se aplica a la magnitud del espectro, se presenta a continuación en la Ecuación 7.

$$f_M = 2525 \times log\left(1 + \frac{f}{700}\right), \qquad (7)$$

Donde f_M es la frecuencia Mel para la frecuencia lineal f.

La Figura 9 ilustra la aplicación de la escala Mel al espectro de frecuencia, mostrando cómo se transforman las frecuencias lineales a una escala que emula la percepción auditiva humana. En la figura puede observarse el banco de filtros Mel aplicado sobre la magnitud del espectro, donde cada filtro se distribuye logarítmicamente para capturar las características más relevantes de la señal.

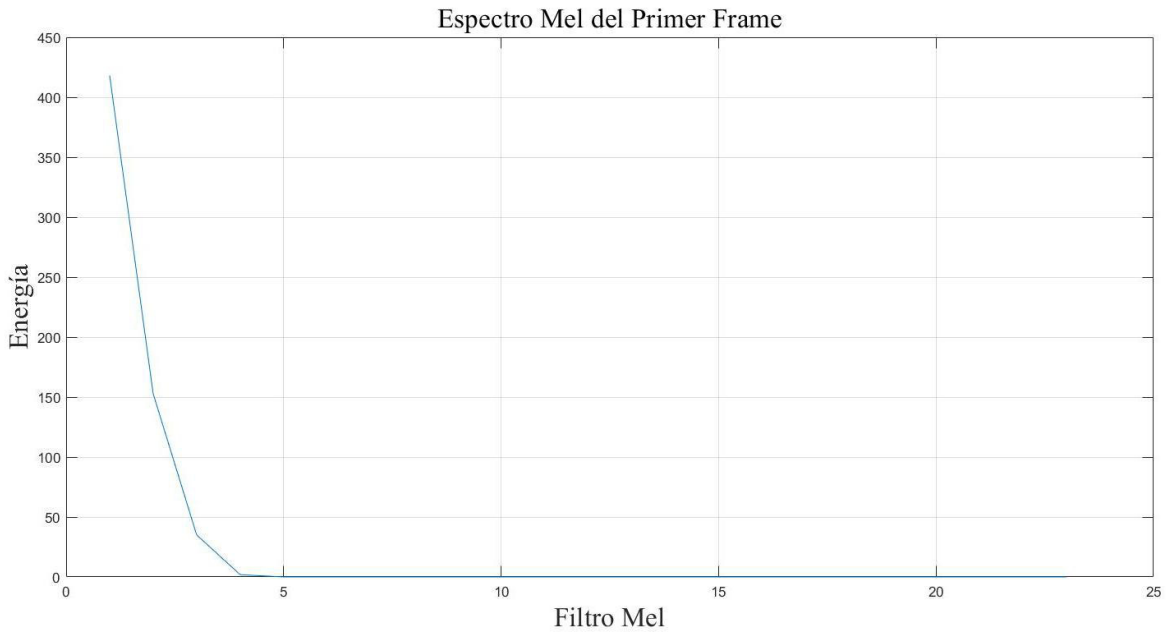


Figura 9. Aplicación de la escala Mel.

Cálculo de los Coeficientes Cepstrales: Se toma el logaritmo de la energía de la magnitud de la respuesta del filtro Mel, se presenta a continuación en la Ecuación 8.

$$E_j^x = \sum_{n-1}^{n} |x(k)|^2 * \psi_i(k), \qquad (8)$$

Donde $|x(k)|$ es el espectro de amplitud, k es el índice de frecuencia, ψ_i es el i^{th} filtro Mel pasa banda, $1 \leq i \leq M$, y M es el número de filtros Mel pasa bandas triangulares. E_j^x es la energía del banco de filtros. Por último, se aplica la Transformada Discreta del Coseno (DCT) a los logaritmos de la energía para obtener los coeficientes ceptrales de cada audio.

Como se puede apreciar en la Figura 10, se presentan los coeficientes MFCCs calculados, mostrando los valores resultantes tras aplicar el logaritmo a la energía del espectro procesado por los filtros Mel. Estos coeficientes, representados gráficamente, capturan las características espectrales más relevantes de la señal en el dominio cepstral, donde se destacan los primeros componentes que contienen la mayor información para el reconocimiento de instrumentos musicales.

Transformada Discreta del Coseno (DCT por sus siglas en inglés, Discrete Cosine Transform): Finalmente, se aplica la Transformada Discreta del Coseno (DCT por sus siglas en inglés, Discrete Cosine Transform) a los logaritmos de la energía para obtener los Coeficientes Cepstrales de cada audio, se presenta a continuación en la Ecuación 9.

$$C_t^x = \sum_{t=1}^{M} \log(E_t^x) \cos\left[l \cdot \frac{(2\pi - 1)\pi}{2M}\right], \qquad (9)$$

Donde C_t^x describe el cálculo de una parte del coeficiente cepstral del proceso MFCCs. $\sum_{t=1}^{M}$ es una suma que recorre t de 1 a M, donde M es el número total de bandas de frecuencia (por ejemplo, bandas mel). $\log(E_t^x)$ es el logaritmo de la energía espectral en la banda t. $\left[l \cdot \frac{(2\pi-1)\pi}{2M}\right]$ se refiere al término ponderado por coseno.

Capítulo 3. Metodología.

La metodología que se implementó para este proyecto es cuantitativa experimental con clasificación supervisada, ya que se hará uso de datos numéricos y métricas para evaluar los resultados obtenidos. En la Figura 10, se muestran las partes de dicha metodología en donde podemos observar las siguientes partes importantes:

- Conjunto de datos.
- Preprocesamiento archivos de audio.
- Extracción de características MFCCs.
- Clasificadores MLP, SVM, KNN.
- Métricas de evaluación.
- Metodologías de entrenamiento y evaluación.

Iniciamos con el conjunto de datos que usamos en el entrenamiento. Este conjunto es sometido a una etapa de preprocesamiento para después extraerle características de MFCCs. Posteriormente, esas características son utilizadas para entrenar 3 diferentes clasificadores (Perceptrón Multicapa, Máquina de Soporte Vectorial y K-Vecinos Más Cercanos). Una vez realizado el entrenamiento, procedemos a la evaluación de los modelos construidos con las métricas de evaluación tales como: Exactitud, Precisión, Recall y F1. Finalmente, se muestran los resultados obtenidos al realizar los experimentos para 20 y 31 clases respectivamente.

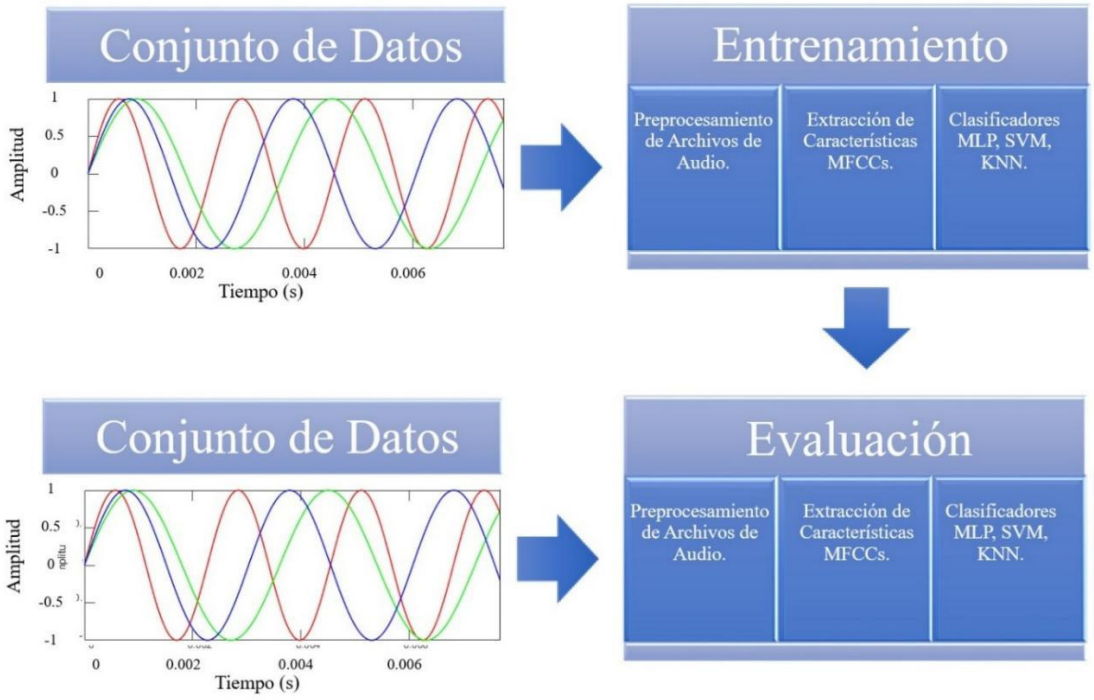


Figura 10. Metodología propuesta para el proyecto.

En las siguientes secciones se describirá de una forma más completa cada una de las siguientes partes.

3.1. Conjunto de datos.

El conjunto de datos que se utilizó, es conocido como AAM: Artificial Audio Multitracks Dataset es introducida por Ostermann *et al*. (2023). Este conjunto de datos contiene pistas de audio de música artificial con anotaciones. Se basa en muestras de instrumentos reales y se genera mediante composición algorítmica respetando la teoría musical. Proporciona mezclas completas de canciones, así como pistas de un solo instrumento. También están disponibles archivos MIDI de cada archivo de audio y archivos de anotación que incluyen: Onsets, Tonos, Instrumentos, Claves, Tempos, Segmentos, Instrumento, Melodía, Compases, y Acordes. Los conjuntos de datos con anotaciones son necesarios para evaluar, comparar y optimizar algoritmos para diversas tareas supervisadas de clasificación musical, regresión y detección.

Las especificaciones técnicas del conjunto de datos son las siguientes: consta de pistas musicales completas, archivos MIDI originales, audios individuales por instrumento, tonalidades, tempos y archivos en formato FLAC con una frecuencia de muestreo de 44.1 kHz.

Para los experimentos desarrollados en este trabajo de tesis, se utilizaron 9300 pistas de audio, correspondientes a 31 clases de diferentes instrumentos musicales. Utilizamos 300 muestras de cada clase: Bajo Eléctrico, Balalaika, Batería, Clarinete, Contrabajo con Arco, Contrabajo Pizzicato, Erhu, Flauta, Flauta de Pan, Flugelhorn, Fujara, Guitarra Acústica, Guitarra Eléctrica Crunch, Guitarra Eléctrica Limpia, Guitarra Eléctrica Solista, Jinghu, Morin Khuur, Órgano Bajo, Piano, Piano Brillante, Piano Eléctrico, Saxofón Alto, Saxofón Tenor, Shakuhachi, Sitar, Trombón, Trompeta, Ukelele, Viola, Violín, Violonchelo.

En la Figura 11 se presentan señales de audio representativas para cada clase de instrumento analizado. La figura permite comparar visualmente las formas de onda generadas por los diferentes instrumentos musicales, destacando patrones en amplitud y variación temporal.

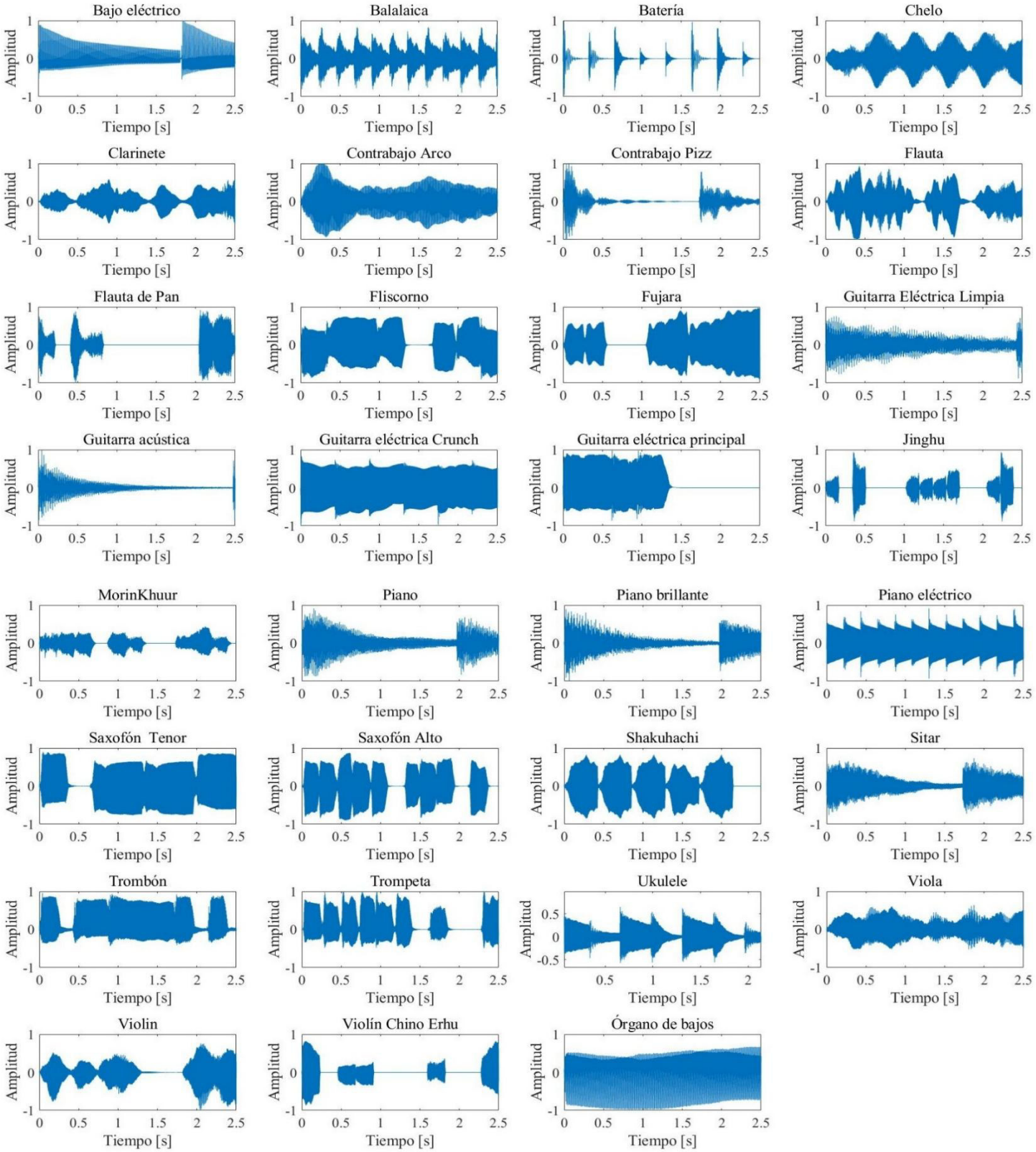


Figura 11. Muestras del conjunto de señales de audio de cada instrumento.

3.2. Preprocesamiento de Archivos de audio.

La eliminación de silencios es importante en el procesamiento de señales de audio, para mejorar calidad y precisión en el análisis. Los silencios no aportan información relevante sobre las características del instrumento musical, su presencia afecta negativamente el rendimiento de los modelos de clasificación. Al no tener los silencios se reduce la cantidad de datos innecesarios y se optimizan los recursos computacionales, representando solo los datos esenciales para obtener mejores resultados. En la Figura 12, se observa una muestra del archivo de audio de una Viola, como podemos darnos cuenta, la señal presenta silencios de lado izquierdo en la señal de color azul. A la derecha de la figura, se presenta en rojo la señal sin los silencios del inicio y del final.

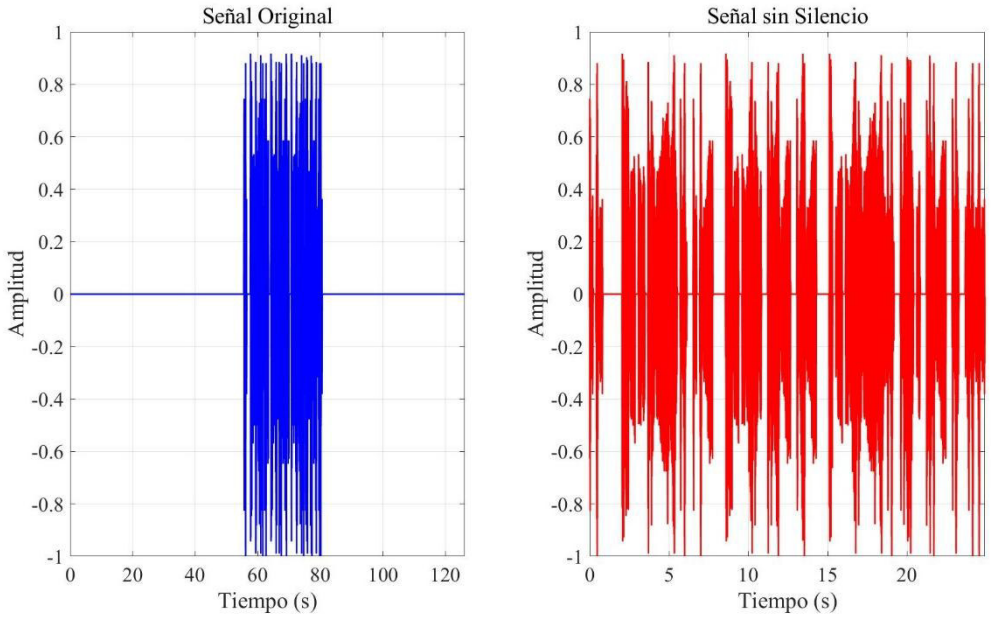


Figura 12. Eliminación de silencios de la señal.

El procedimiento a seguir para la eliminación de silencios inicia con el cálculo de los valores absolutos de la señal analizada. La Ecuación 10 presenta la forma de como calcular los valores absolutos de la señal.

$$A[n] = |x[n]|, \qquad (10)$$

Donde $A[n]$ representa la amplitud absoluta de la muestra n. $x[n]$ es el valor de la señal en la muestra n.|. |representa el valor absoluto. Por ejemplo, si tenemos $x[n] = |1, -2, -6|$, entonces $A[n] = |1,2,6|$.

Posteriormente, el valor mínimo a considerar corresponde al 0.01, es decir al 1% de la amplitud máxima posible, esto se calculó de manera empírica, lo cual es lo suficientemente bajo para ignorar ruidos de fondo. Al contrario, si usamos un valor menor al 0.01, podemos incluir ruido de fondo que no es parte útil del audio. Después, necesitamos validar que no

todo sea silencio en la señal, por lo que analizamos las señales que tengan una duración mínima de al menos 0.01 segundos.

3.3. Extracción de Características.

Después de haber realizado el proceso de eliminación de silencios, la siguiente parte consta de extraer 13 de los MFCCs de cada archivo de audio. Los cuales son necesarios para el análisis del espectro de audio, ya que modelan la forma en que los humanos perciben el sonido, proporcionando una representación compacta y robusta de la señal de cada audio que analizaremos.

Los MFCCs capturan las características espectrales más relevantes de cada audio y son muy útiles para la clasificación de señales de audio, ya que representan tanto la envolvente del espectro como sus cambios a lo largo del tiempo. En la Figura 13 podemos visualizar los 13 MFCCs calculados de un fragmento de un archivo de audio: 002_Viola.flac.

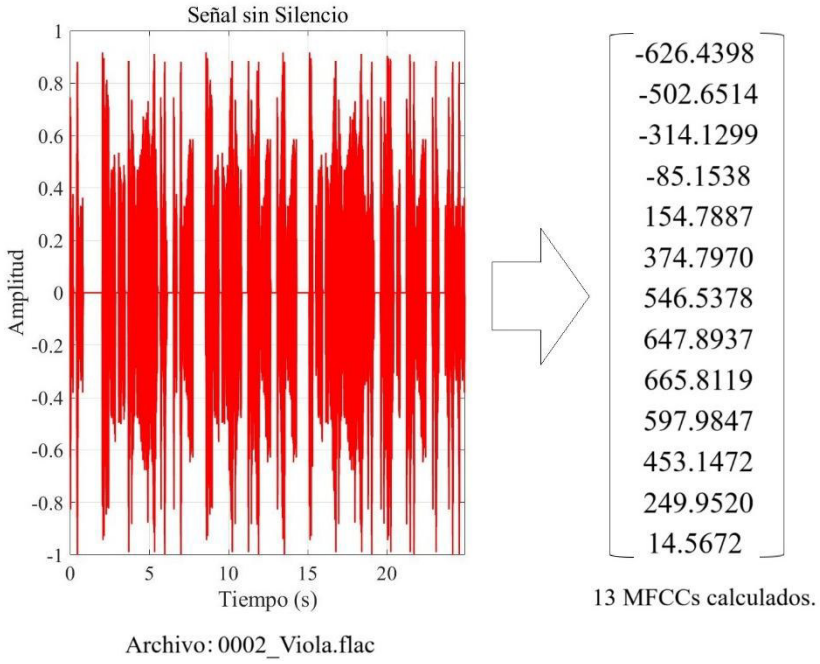


Figura 13. 13 MFCCs, calculados.

3.4. Clasificadores.

El Aprendizaje Automático se encuentra en la intersección de la informática, la ingeniería y la estadística, es una herramienta que se puede aplicar a muchos problemas. Cualquier campo que necesite interpretar y actuar sobre datos puede beneficiarse de las técnicas de Aprendizaje Automático. Existen dos aproximaciones importantes en el Aprendizaje Automático, que son basadas en el Aprendizaje Supervisado y el No Supervisado (Harrington, 2012). El Aprendizaje Supervisado es aquel en el que le estamos dando ejemplos al algoritmo sobre

qué y cómo predecirlo. El Aprendizaje No Supervisado es el que encuentra patrones en los datos sin necesidad de ejemplos.

Para este trabajo se implementó un modelo de clasificación basado en Aprendizaje Supervisado mediante el uso de la herramienta MATLAB (abreviatura de "Matrix Laboratory"). MATLAB es un entorno de computación numérica y un lenguaje de programación de alto nivel, es ampliamente utilizado en ingeniería, matemáticas, ciencias de la computación y otras disciplinas técnicas y científicas. A continuación, se presentan las tres técnicas de Aprendizaje de Máquina que se implementaron en este trabajo: Perceptrón Multicapa, Máquinas de Soporte Vectorial y Vecinos más Cercanos, las cuales presentan diferentes arquitecturas y características, cada una se describe de manera más clara a continuación.

3.4.1. Perceptrón Multicapa (MLP).

El MLP es una forma de red neuronal que consiste en múltiples capas de nodos (llamadas neuronas), donde cada nodo aplica una función de activación. Este modelo es capaz de aprender patrones complejos a partir de datos, Géron, A. (2022). El desarrollo y la popularización del MLP como una arquitectura de red neuronal eficaz y ampliamente utilizada se produjo a partir de la década de 1980 y 1990. El MLP es un discriminador de tipo lineal, debido a que separa dos regiones en el espacio que pertenecen a dos clases diferentes de datos, utilizando como filtro una condición lineal, según Lezama Gutierrez *et al*. (2019). En la Figura 14 se muestra el modelo del MLP, el cual consta de una capa de entrada, una o más capas ocultas y una capa de salida, es fundamental para clasificación, regresión y reconocimiento de patrones.

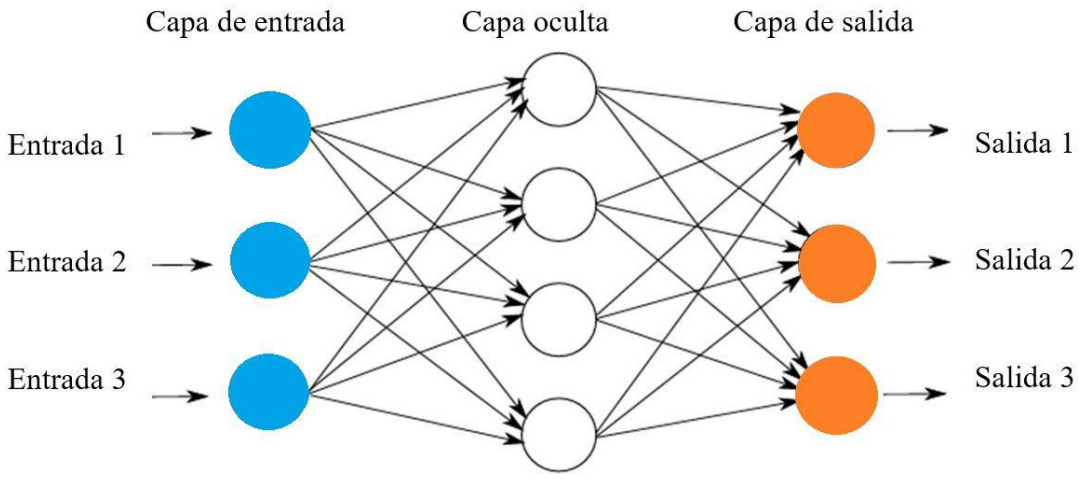


Figura 14. Ejemplo del modelo de un perceptrón multicapa.

El funcionamiento básico de una red neuronal es un proceso mediante el cual la información fluye desde la capa de entrada hacia la salida, atravesando las distintas capas ocultas de la red. Cada neurona en una capa recibe un conjunto de entradas ya sea directamente desde los datos iniciales o desde las salidas activadas de la capa anterior y realiza una operación matemática: multiplica cada entrada por su peso correspondiente, suma todos los resultados y añade un valor de sesgo. Se presenta a continuación en la Ecuación 11.

$$z = (\omega_1 \cdot x_1 + \omega_2 \cdot x_2 + \cdots + \omega_n \cdot x_n) + b, \qquad (11)$$

Donde z representa el valor de entrada a la función de activación, el valor activado generado, se propaga a la siguiente capa. x_1 representa cada característica de entrada. Como por ejemplo valores de MFCCs de una señal de audio. ω_1 representa el peso asociado a cada entrada, durante el proceso de entrenamiento del modelo, estos pesos se ajustan automáticamente para mejorar la precisión del modelo. $\omega_1 \cdot x_1$ representa la multiplicación de cada entrada por su peso correspondiente. Luego, todos esos productos se suman, formando una suma ponderada de las entradas. b es una constante que se suma al final de la operación para que el modelo tenga más flexibilidad al hacer predicciones.

El sesgo permite desplazar la salida hacia arriba o hacia abajo, lo que es útil cuando los datos no están centrados en cero o cuando se necesita ajustar la salida de manera más precisa. Cada una de estas neuronas está activada mediante la función ReLU (Rectified Linear Unit), que a continuación se presenta en la Ecuación 12.

$$f(z) = max(0, z), \qquad (12)$$

Donde z es la entrada a una neurona, es decir, el valor que se obtiene al multiplicar cada entrada por su peso correspondiente, suma todos esos productos y agregar el sesgo. $max(0, z)$ es la función toma el valor máximo entre 0 y z.

Propagación a través de las capas ocultas: Este proceso se repite capa por capa, propagando las activaciones hacia adelante. En cada capa oculta, las neuronas operan de manera similar: reciben las activaciones de la capa anterior, las ponderan con sus respectivos pesos, suman sesgos, aplican la función de activación y generan nuevas salidas.

Capa de salida: En esta etapa se recogen las salidas de la última capa oculta para producir el resultado final del modelo. En tareas de clasificación, se utiliza comúnmente una función de activación que convierte las salidas en probabilidades asociadas a cada clase posible. Esta capa utiliza la función Softmax, la cual convierte el vector de salidas en un conjunto de probabilidades normalizadas que suman 1, lo que resulta especialmente útil para problemas de clasificación multiclase. Se presenta a continuación en la Ecuación 13:

$$\sigma(z)i = \frac{e^{zi}}{\sum_{j=1}^{n} e^{zj}}, \qquad (13)$$

Donde $\sigma(z)i$ es la salida neta de la neurona i en la capa de salida. e^{zi} es la función exponencial aplicada a la salida zi. $\sum_{j=1}^{n} e^{zj}$ representa la suma de todas las salidas de las

n neuronas de la capa de salida. $\sigma(z)i$ representa la salida, donde será un número entre 0 y 1. Todas las salidas juntas suman 1.

La arquitectura del MLP empleada en este trabajo de tesis ha sido diseñada para procesar características extraídas de archivos de audio, particularmente los primeros 13 MFCCs, los cuales constituyen la primera capa de entrada conformada por 13 neuronas. A partir de esta capa, la información se propaga a través de tres capas ocultas con la siguiente configuración: la primera capa oculta consta de 256 neuronas, la segunda capa oculta cuenta con 128 neuronas, y la tercera capa oculta vuelve a tener 256 neuronas. Tras las capas ocultas, la información llega a la capa de salida, que está compuesta por 31 neuronas, una por cada clase que el modelo debe predecir.

En cuanto al proceso de entrenamiento se utilizó el algoritmo backpropagation. El modelo se entrena durante 2000 épocas, lo cual permite una exposición extensa a los datos de entrenamiento y favorece la convergencia del modelo. Para ello se emplean lotes de 32 muestras, lo que significa que los pesos se actualizan después de procesar cada lote. Con esto se permite un ajuste continuo y eficiente de los parámetros a lo largo del proceso de aprendizaje.

3.4.2. Máquina de Soporte Vectorial (SVM).

El clasificador de la SVM, es un modelo de Aprendizaje Supervisado que se utiliza principalmente para tareas de clasificación y regresión, su objetivo es encontrar el hiperplano que mejor separe las diferentes clases de datos en un espacio de características multidimensional. Fueron desarrolladas por Vladimir Vapnik. (1995), y su equipo a finales de la década de 1990.

Las SVM se han convertido en una herramienta fundamental en el campo del Aprendizaje Automático y han sido aplicadas con éxito en una amplia gama de problemas de clasificación y regresión, Géron, A. (2022). Aunque los clasificadores SVM lineales son eficientes y funcionan sorprendentemente bien en muchos casos, muchos conjuntos de datos ni siquiera se acercan a ser linealmente separables, se presenta a continuación en la Ecuación 14. Martínez y Aguilar (2013).

$$f\,(x) = \text{sign}\left(\sum_{i=1}^{n} a_i y_i\ K\left(x, x_i\right) + b\right), \qquad (14)$$

Donde $f\,(x)$ es la función de decisión que clasifica el vector de entrada x. a_i son los coeficientes de los vectores de soporte. y_i son las etiquetas de clase de los vectores de soporte. $K\left(x, x_i\right)$ es la función Kernel que mide la similitud entre el vector de entrada x y los vectores de soporte x_i. b es el sesgo o término de sesgo.

Para este trabajo de tesis, se utilizó el algoritmo de SVM como una estrategia de clasificación supervisada para reconocer distintos instrumentos musicales a partir de características extraídas de archivos de audio. El proceso comienza con la selección de las características

relevantes, específicamente los 13 MFCCs de cada archivo de audio, capturando patrones que diferencian a cada instrumento.

Cada conjunto de características se acompaña de su correspondiente etiqueta de clase, representando uno de los 31 instrumentos musicales considerados. Una vez preparados los datos, el algoritmo de SVM tiene el objetivo de encontrar el hiperplano óptimo que separe de forma eficaz las clases en el espacio de características. En términos simples, este hiperplano actúa como una frontera de decisión que divide los datos de diferentes clases maximizando el margen, es decir, la distancia entre el hiperplano y las muestras más cercanas de cada clase, conocidas como vectores de soporte. Cuanto mayor sea este margen, mejor será la capacidad del modelo para generalizar a datos no vistos.

La Figura 15 ilustra un modelo de clasificación binaria usando SVM, donde podemos observar un hiperplano separador que divide los datos en dos clases, los triángulos y cuadrados. Este hiperplano está ubicado de forma que maximiza la distancia (margen) entre sí y los puntos más cercanos de cada clase, conocidos como vectores de soporte.

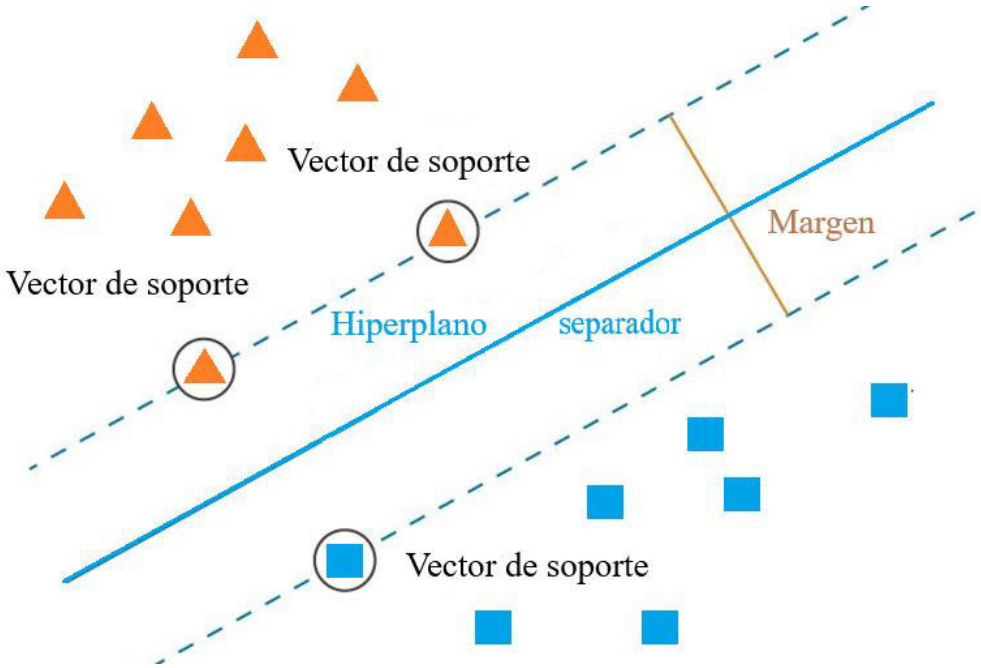


Figura 15. Ejemplo del modelo de Máquinas de Soporte Vectorial.

3.4.3. K-Vecinos Más Cercanos (KNN).

El clasificador de KNN, según los autores Cover, Thomas M. y Hart, Peter E. (1967), lo definen como un punto de datos basado en la mayoría de sus vecinos más cercanos en el espacio de características, ha sido reconocido y utilizado en diferentes formas a lo largo de la historia del Aprendizaje Automático y la estadística. El término "K-Nearest Neighbors" y la formalización del algoritmo tal como lo conocemos hoy en día surgieron de la comunidad de investigación en Aprendizaje Automático en las últimas décadas del siglo XX. Sin embargo, no hay un único individuo al que se le pueda atribuir el descubrimiento de KNN en su forma actual, ya que se basa en principios fundamentales de la estadística y el reconocimiento de patrones que han sido explorados por muchos investigadores a lo largo del tiempo.

El funcionamiento de este algoritmo consta de observar los k datos más similares de nuestro conjunto de datos conocidos; de ahí proviene el k. (k es un número entero y suele ser menor que 20). Por último, tomamos un voto mayoritario de los k datos más similares, y la mayoría es la nueva clase que asignamos a los datos que se pidió clasificar. Harrington, P. (2012).

En este proyecto también se implementó el algoritmo de KNN como técnica de clasificación supervisada para el reconocimiento automático de instrumentos musicales. Consiste en la extracción de características significativas a partir de los archivos de audio, específicamente los 13 primeros MFCCs, que ofrecen una representación robusta de los sonidos. Cada conjunto de características se asocia a una etiqueta correspondiente a una de las 31 clases de instrumentos musicales.

El funcionamiento del modelo es intuitivo pero efectivo. Cuando se presenta un nuevo ejemplo sin etiqueta, el algoritmo compara este nuevo dato, con todos los elementos del conjunto de entrenamiento utilizando la distancia euclidiana como métrica de similitud, Lubis *et al*. (2020). Una vez calculadas las distancias, el algoritmo selecciona los k vecinos más cercanos, en este trabajo k=5. Es decir, observa a qué clases pertenecen esos cinco vecinos más próximos y asigna al nuevo ejemplo a la clase más común entre ellos. Este principio de votación permite que el modelo realice predicciones precisas incluso en situaciones donde la frontera entre clases no es lineal.

La Figura 16 ilustra un modelo de clasificación del algoritmo KNN, el espacio de características se divide en regiones asociadas a cada clase, en este caso los triángulos y cuadrados representan dos clases diferentes. Como se muestra en esta figura, el nuevo dato de prueba (circulo gris) se clasifica de acuerdo con la proximidad a la mayoría de los k datos más cercanos al punto gris.

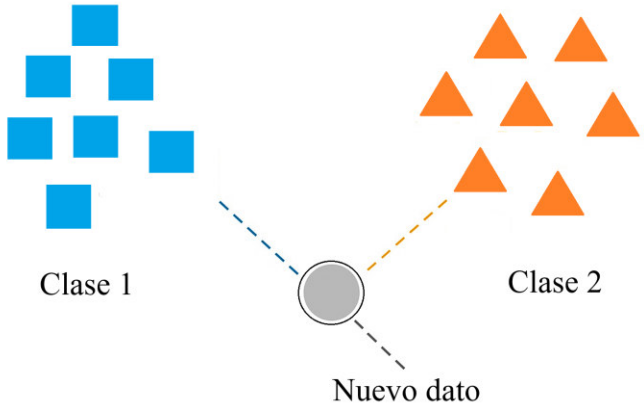


Figura 16. Ejemplo del modelo de Vecinos más Cercanos.

3.5 Métricas de evaluación.

El modelo será evaluado utilizando métricas estándar como la Exactitud, Precisión, Recall y F1-score, dichas métricas y sus ecuaciones se presentan a continuación.

Una métrica para evaluar el rendimiento de un modelo de clasificación es la Exactitud de clasificación. Se calcula con el número total de instancias que un modelo ha predicho correctamente dividido por el número total de instancias predichas, se presenta a continuación en la Ecuación 16.

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}, \qquad (16)$$

La Precisión es la métrica para estimar el rendimiento de un modelo de clasificación. Describe la fracción de instancias positivas reales en comparación con todas las instancias predichas que son positivas, tanto verdaderas como falsas, se presenta a continuación en la Ecuación 17.

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos\ +\ Falsos\ positivos}, \qquad (17)$$

Recall es una métrica empleada en la evaluación de modelos de clasificación. Indica la proporción de instancias positivas que son correctamente identificadas por el modelo en relación con todas las instancias que realmente son positivas, se presenta a continuación en la Ecuación 18.

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ positivos\ +\ Falsos\ negativos}, \qquad (18)$$

La métrica F1 funciona para evaluar el rendimiento de un modelo de clasificación, que combina precisión y Recall en una única medida, se presenta a continuación en la Ecuación 19.

$$F1 = 2 \times \frac{Precision\ x\ Recall}{Precision\ +\ Recall}, \qquad (19)$$

3.5.1 Metodologías de entrenamiento y evaluación.

Para evaluar el rendimiento del modelo, se utilizaron dos metodologías empleadas en el campo de Aprendizaje de Máquina. La primera metodología es Validación por Separación o División Simple (Hold-Out Validation), consiste en dividir el conjunto de datos en dos partes, estos datos se seleccionan aleatoriamente, el modelo se evalúa con los datos de prueba, asegurando que las predicciones sean precisas en nuevos datos. Dicha metodología se utilizó para la evaluación preliminar de los modelos con diferentes hiperparametros. Se llevó a cabo dividiendo el total de datos (9300 archivos de audio, distribuidos equitativamente en 31 clases) en dos subconjuntos: el 80% de los datos (7440 archivos) que se emplearon para el entrenamiento del modelo, y el 20% restante (1860 archivos) se reservaron para la etapa de prueba.

Posteriormente, se implementó la metodología de Validación Cruzada (Cross Validation), consiste en mostrar un resultado promedio de un número determinado de pruebas. Para este trabajo se aplicó una Validación Cruzada con k-folds (k = 5). Este proceso consistió en dividir los 9300 archivos en 5 subconjuntos o folds de igual tamaño para entrenar y validar el modelo 5 veces, rotando en cada iteración el subconjunto usado como validación. El resultado de cada métrica es el promedio de cada experimento realizado con dicha metodología. En el siguiente capítulo, se presentarán y analizarán los resultados obtenidos.

Capítulo 4. Experimentos y Resultados.

En este capítulo se muestran los experimentos y resultados obtenidos en el modelo de identificación de instrumentos musicales, implementando las técnicas de Aprendizaje de Máquina, que analizamos anteriormente.

4.1. Herramienta Desarrollada.

Para el desarrollo de este modelo utilizamos las siguientes características computacionales, el procesador (CPU), que se utilizó consiste en un Intel® Core™ i5-12450H y una memoria (RAM) 32 GB DDR4. El modelo cuenta con una interfaz gráfica donde dicha herramienta permite la fácil lectura, preprocesamiento, análisis y clasificación de los audios. Asimismo, nos muestra las métricas de evaluación. En la Figura 17, se puede apreciar que esta herramienta consta de tres secciones importantes e incorpora las técnicas de Aprendizaje de Máquina: MLP, SVM, KNN. Además, podremos generar diferentes comparaciones con el uso de dichas técnicas. Igualmente incorporar Matrices de Confusión para los diferentes experimentos realizados en este trabajo de tesis.

Figura 17. Aplicación del sistema de clasificación de instrumentos musicales.

4.2. Resultados.

A continuación, se muestran los resultados obtenidos al realizar dos experimentos. En el primer experimento, se prueba el sistema siguiendo una metodología donde se presentan 20 clases diferentes utilizando las métricas de evaluación. Adicionalmente, presentamos una comparación con el trabajo de Mahanta *et al.* (2021). En el segundo experimento, se presenta una extensión utilizando 31 clases.

4.2.1. Experimento con 20 clases.

Para este experimento nos comparamos con un trabajo previo de Mahanta *et al.* (2021). Las pruebas se llevaron a cabo con el concentrado de 300 audios para cada una de las primeras 20 clases de instrumentos musicales, las cuales son: Guitarra acústica, Saxofón Alto, Balalaica, Piano Brillante, Chelo, Clarinete, Contrabajo Arco, Contrabajo Pizz, Batería, Bajo Eléctrico, Guitarra Eléctrica Limpia, Guitarra eléctrica Crunch, 'Guitarra eléctrica principal, Piano Eléctrico, Violín Chino Erhu, Fliscorno, Flauta, Fujara, Jinghu, MorinKhuur. Los cuales fueron procesados para extraer sus 13 MFCCs principales para realizar la posterior clasificación por medio de las técnicas de Aprendizaje de Máquina, como MLP, SVM, KNN.

A continuación, se presentan las Matrices de Confusión de cada uno de los modelos de clasificación. En la Figura 18 se muestra la Matriz de Confusión para el MLP de 20 clases (la imagen con detalles se puede encontrar en la sección de anexos). Podemos visualizar como el porcentaje entre las clases con mayor índice de confusión es la clase Guitarra Eléctrica Limpia, la cual se puede confundir con Piano Eléctrico y Piano Brillante, debido a sus frecuencias medias similares en dichos instrumentos.

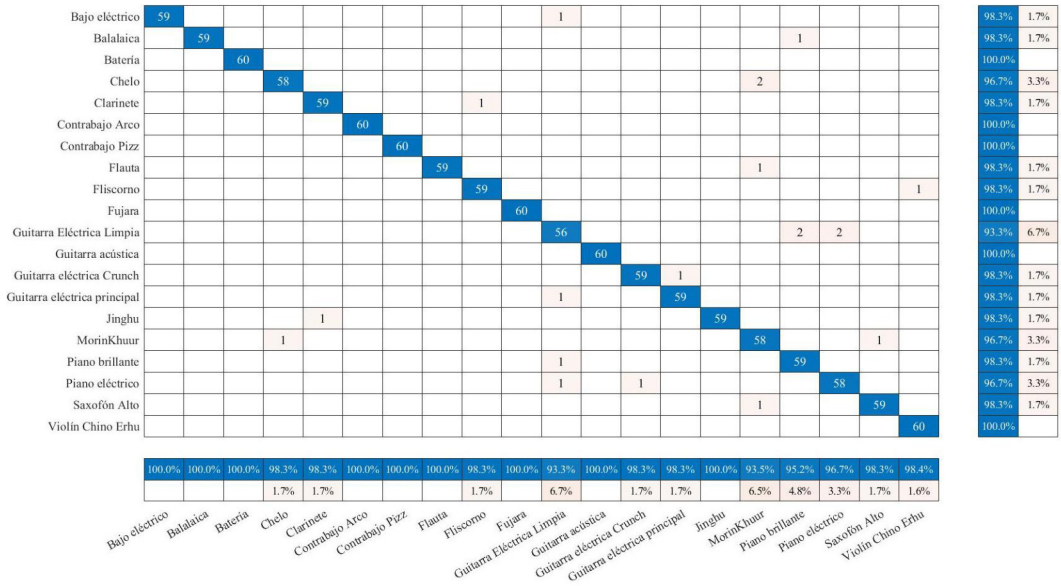


Figura 18. Matriz de Confusión para el MLP de 20 clases.

En la Figura 19 se observa la Matriz de Confusión para SVM de 20 clases (la imagen con detalles se puede encontrar en la sección de anexos). En esta matriz podemos visualizar que las clases con mayor índice de confusión es la clase Guitarra Eléctrica Limpia, la cual se puede confundir con Piano Eléctrico con un total de 7 casos y la clase Piano Eléctrico se confunde con Guitarra Eléctrica Principal en 5 casos. De la misma manera en el caso anterior estas confusiones pueden ser dadas por las frecuencias semejantes entres dichas clases.

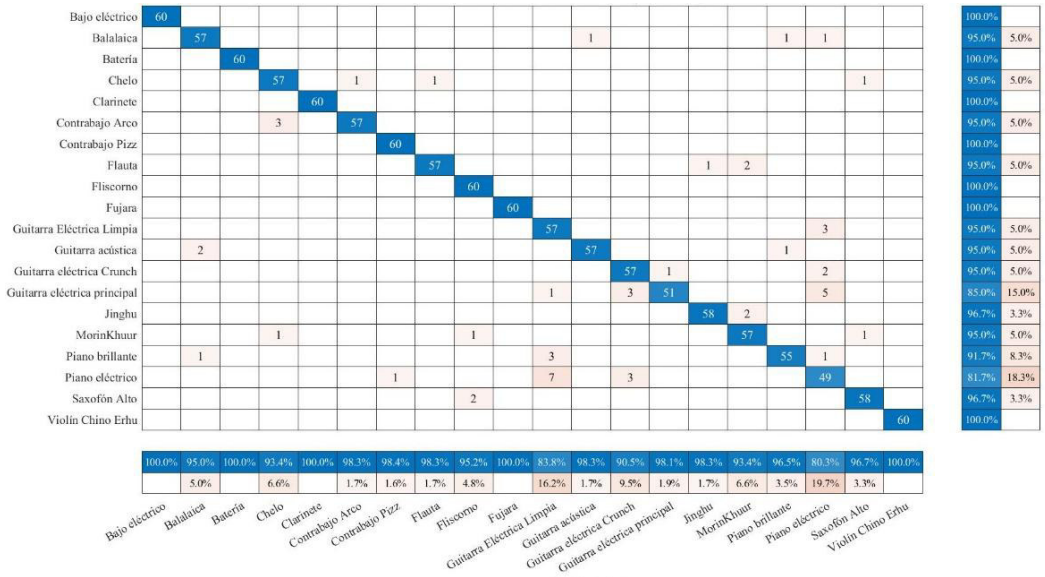


Figura 19. Matriz de Confusión para SVM de 20 clases.

La Figura 20 muestra la Matriz de Confusión para KNN de 20 clases (la imagen con detalles se puede encontrar en la sección de anexos). En la Matriz de Confusión siguiente se muestra como el porcentaje entre las clases con mayor índice de confusión es la clase Guitarra Eléctrica Limpia, la cual se puede confundir con Piano Eléctrico y la clase Piano Brillante.

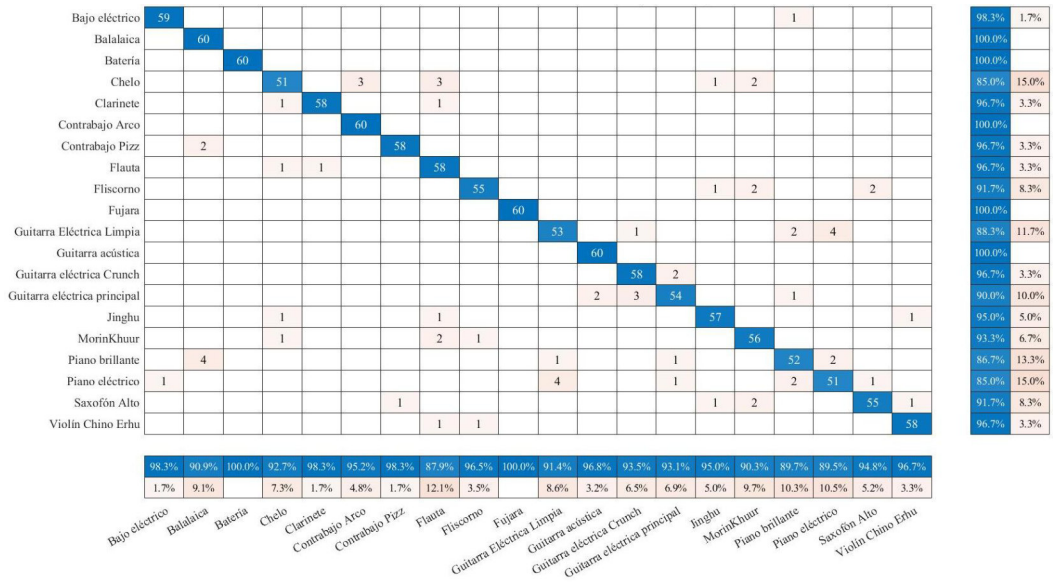


Figura 20. Matriz de confusión para KNN de 20 clases.

En la Tabla 1 podemos ver los resultados obtenidos al clasificar 20 clases de instrumentos musicales. Los resultados de las métricas de evaluación son las siguientes: la Exactitud con un 97.5% se logró con la técnica del MLP , las SVM mostraron un 95.2% y los KNN muestran 93.4%. La Precisión nos muestra que, de las predicciones positivas hechas por el modelo el 97.5% fueron correctas con el MLP, las SVM mostraron un 95.2% y los KNN muestran 93.4%. El Recall identificó correctamente el 97.5% de todos los datos que en realidad pertenecían a las clases correctas, con el MLP, las SVM mostraron un 95.5% y los KNN muestran 93.6%. El F1-score es del 97.4%, con el MLP, las SVM mostraron un 95.3% y los KNN muestran 93.4%, esta métrica es una combinación de la Precisión y el Recall. También podemos observar como las métricas de Precisión, Recall, F1 Score y Exactitud, mostraron un mejor desempeño utilizando el MLP con un 0.5%, en comparación con el trabajo de Mahanta *et al.* (2021).

Tabla 1. *Métricas de evaluación de resultados de las técnicas de Aprendizaje de Máquina para 20 clases.*

Técnicas de Aprendizaje de Máquina (ML).	Precisión	Recall	F1-score	Exactitud
Mahanta *et al.* [2] (20 clases)	97.0%	97.0%	97.0%	97.0%
MLP (20 clases)	**97.5%**	**97.5%**	**97.4%**	**97.5%**
SVM (20 clases)	95.2%	95.5%	95.3%	95.2%
KNN (20 clases)	93.4%	93.6%	93.4%	93.4%

4.2.2. Experimento con 31 clases.

Para el experimento de 31 clases de instrumentos musicales se llevó a cabo con el concentrado de 300 audios para cada una de las 31 clases las cuales son: Guitarra Acústica, Saxofón Alto, Balalaika, Piano Brillante, Violonchelo, Clarinete, Contrabajo con Arco, Contrabajo Pizzicato, Batería, Bajo Eléctrico, Guitarra Eléctrica Limpia, Guitarra Eléctrica Crunch, Guitarra Eléctrica Solista, Piano Eléctrico, Erhu, Flugelhorn, Flauta, Fujara, Jinghu, Morin Khuur, Órgano Bajo, Flauta de Pan, Piano, Shakuhachi, Sitar, Saxofón Tenor, Trombón, Trompeta, Ukelele, Viola, Violín. Los cuales fueron procesados para extraer sus 13 MFCCs principales para realizar la posterior clasificación por medio de las técnicas de Aprendizaje de Máquina, como el MLP, las SVM y los KNN.

A continuación, se presentan las Matrices de Confusión de cada uno de los modelos de clasificación. Se aprecia en la Figura 21 la Matriz de Confusión para el MLP de 31 clases (la imagen con detalles se puede encontrar en la sección de anexos). En la cual podemos

visualizar como el porcentaje entre las clases con mayor índice de confusión es la clase Guitarra Eléctrica Limpia, la cual se puede confundir con Piano Eléctrico, debido a sus frecuencias medias similares en dichos instrumentos.

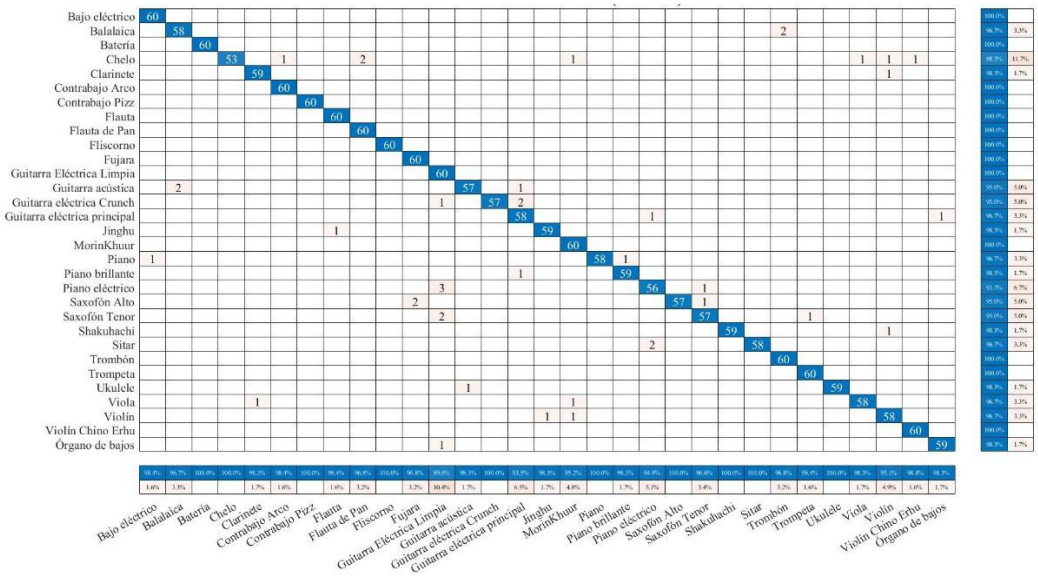


Figura 21. Matriz de Confusión para MLP de 31 clases.

En la Figura 22 se observa la Matriz de Confusión para el Clasificador para SVM de 31 clases (la imagen con detalles se puede encontrar en la sección de anexos). En la cual podemos visualizar como el porcentaje entre las clases con mayor índice de confusión, es la clase Guitarra Eléctrica Limpia, la cual se puede confundir con Sitar, de la misma manera debido a sus frecuencias medias similares.

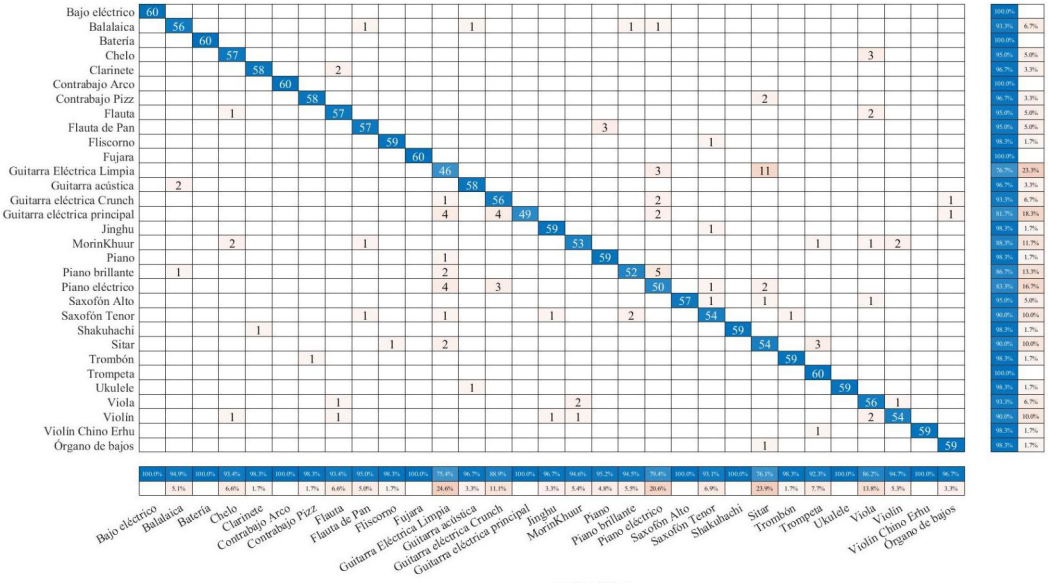


Figura 22. Matriz de Confusión para SVM de 31 clases.

La Figura 23 contiene la Matriz de Confusión para el Clasificador de KNN de 31 clases (la imagen con detalles se puede encontrar en la sección de anexos). En la cual podemos visualizar de la misma manera la confusión entre Piano Eléctrico y Guitarra Eléctrica Limpia.

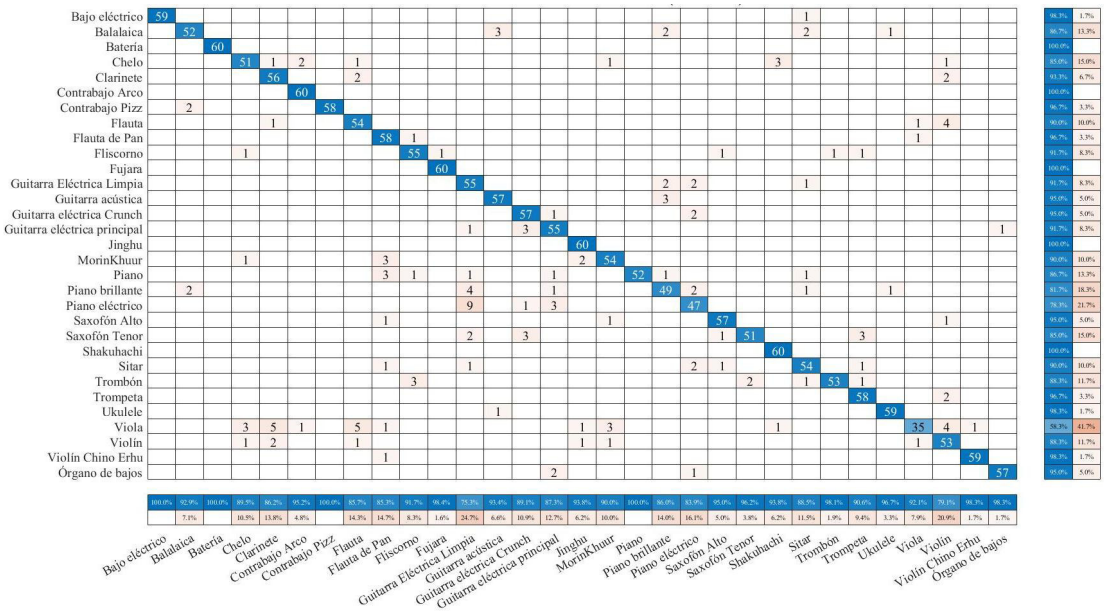


Figura 23. Matriz de confusión para KNN de 31 clases.

Podemos visualizar dichos resultados en la siguiente Tabla 2. Los resultados de las métricas de evaluación son las siguientes: la Exactitud con un 96.4% se logró con la técnica del MLP, las SVM mostraron un 94.0% y los KNN muestran 91.7%. La Precisión nos muestra que, de las predicciones positivas hechas por el modelo el 96.4% fueron correctas con el MLP, las SVM mostraron un 94.0% y los KNN muestran 91.7%. El Recall identificó correctamente el 96.5% de todos los datos que en realidad pertenecían a las clases correctas, con el MLP, las SVM mostraron un 94.3% y los KNN muestran 91.9%. El F1-score es del 96.4%, con el MLP, las SVM mostraron un 94.1% y los KNN muestran 91.6%.

***Tabla 2**. Métricas de evaluación de resultados de las técnicas de Aprendizaje de Máquina (ML) para 31 clases.*

Técnicas de Aprendizaje de Máquina (ML).	Precisión	Recall	F1-score	Exactitud
MLP (31 classes)	**96.4%**	**96.5%**	**96.4%**	**96.4%**
SVM (31 classes)	94.0%	94.3%	94.1%	94.0%
KNN (31 classes)	91.7%	91.9%	91.6%	91.7%

Capítulo 5. Conclusiones y Recomendaciones.

En este capítulo se resume de manera concreta los resultados obtenidos en el modelo de Identificación de Instrumentos Musicales, utilizando las diferentes técnicas de Aprendizaje de Máquina.

Primeramente, se llevó a cabo la clasificación de 20 y 31 clases de instrumentos musicales, comparándonos con un trabajo de Mahanta *et al.* (2021), mediante la extracción de características de 13 de los MFCCs de los 300 archivos digitales de cada clase sumando un total de 9300 audios utilizados, del Conjunto de Datos de Pistas Múltiples de Audio Artificial introducida por Ostermann *et al*. (2023). Además, se compararon las técnicas de Aprendizaje de Máquina, como las SVM, los KNN y el MLP, para la clasificación de las 31 clases.

Para el experimento de 20 clases, se muestra una mejora significativa con el uso de MLP con una precisión del 97.5% en comparación de 97% del trabajo de Mahanta *et al.* (2021), y en el Recall obtenemos un 97.5% en comparación de 97%, en F1-Score 97.4% en comparación de 97% y en Exactitud de 97.5% en comparación de 97%.

En el experimento de 31 clases, se muestran los mejores resultados nuevamente incorporando el MLP con una Precisión 96.4%, Recall 96.5%, F1-score 96.4% y Exactitud del 96.4%. De 9300 archivos de audio de instrumentos musicales se acertaron 8963 archivos de audio.

Algunos aspectos a mejorar son la clasificación de instrumentos musicales que se pueden confundir debido a sus frecuencias tan parecidas tal es el caso de las clases: Guitarra Eléctrica Limpia, Guitarra Eléctrica Crunch, Guitarra Eléctrica Solista, Piano Brillante y Piano Eléctrico ya que podría llevar a un resultado mejor. Además, se puede experimentar con más datos, diferentes instrumentos o ajustar la arquitectura para comprobar si es posible mejorar estos resultados.

Otro aspecto técnico a mejorar de los archivos de audio, es la eliminación total de silencios, no solamente los iniciales y finales. Por otro lado, las características que pudieran incorporarse para obtener mejores resultados son el ZCR (Zero Crossing Rate), RMS (Root Mean Square Energy) y Spectral Centroid (Centro de Gravedad Espectral).

Derivado de este trabajo se realizaron los siguientes artículos:

Vázquez Robledo, A. S., López Ramírez, M., & Lizárraga Morales, R. A. (2024). Clasificación de instrumentos musicales en audios utilizando coeficientes cepstrales y redes neuronales artificiales. Jóvenes en la Ciencia, 33. https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4701

Vázquez, A., Lizárraga, R., & López, M. (2025). Identification of musical instruments in audios using signal analysis and artificial intelligence. Proceedings of the 7th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI 2025) 115–118. Innsbruck, Austria. ISBN 978-84-09-71189-5

Capítulo 6. Referencias Bibliográficas.

Blaszke, M., & Kostek, B. (2022). Musical Instrument Identification Using Deep Learning Approach. Sensors (Basel, Switzerland), 22(8), 3033. https://doi.org/10.3390/s22083033

Cooley, J. W. & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90), 297, https://doi:10.1090/S0025-5718-1965-0178586-1

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dash, S. K., Solanki, S. S., & Chakraborty, S. (2023). A comprehensive review on audio-based musical instrument recognition: Human-machine interaction towards Industry 4.0. Journal of Scientific and Industrial Research, 82(1), 26–37. https://doi.org/10.56042/jsir.v82i1.70251

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media. United States of America. ISBN 8441548285.

Ghisingh, S. V. M. & V. K. Mittal (2016, December 16–18). Classifying musical instruments using speech signal processing methods. In 2016 IEEE Annual India Conference (INDICON). IEEE. https://doi.org/10.1109/INDICON.2016.7839034

Gray, R. M., & Goodman, J. W. (1995). Fourier transforms: An introduction for engineers. Information Systems Laboratory, Department of Electrical Engineering, Stanford University. https://doi.org/10.1007/978-1-4615-2359-8

Guido, R. C. (2016). ZCR-aided neurocomputing: A study with applications. Knowledge-Based Systems, 105, 248–269. https://doi.org/10.1016/j.knosys.2016.05.011

Harrington, P. (2012). Machine learning in action. Manning Publications. United States of America. ISBN 9781617290183.

Kostrzewa, B. K. (2022). Designing a training set for musical instruments identification. In Lecture Notes in Computer Science (LNCS), 13350, 599–610. Springer. https://doi.org/10.1007/978-3-031-08751-6_43

Lezama Gutiérrez, A., Suárez Carreño, F., & Rosales, L. (2019). Computación inteligente y estados emocionales. Suárez Carreño, Franyelit María. Quito, Ecuador. ISBN 978-9942-35-975-9.

Lubis, A., Lubis, M., & Al-Khowarizmi. (2020). Optimization of distance formula in K-Nearest Neighbor method. International Journal of Electrical and Computer Engineering (IJECE), 9(1), 326–338. https://doi.org/10.11591/eei.v9i1.1464

Lucena, A., Pires, C., Nose-Filho, K., & Suyama, R. (2020, October 26). Musical instruments recognition using machine learning. In Brazilian Technology Symposium. Brasil.

Mahanta, S. K., Khilji, A. F. U. R., & Pakray, P. (2021). Deep neural network for musical instrument recognition using MFCCs. Computación y Sistemas, 25(2), 351–360. https://doi.org/10.13053/cys-25-2-3946

Majeed, S. A., Husain, H., Abdul Samad, S., & Idbeaa, T. F. (2015). Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: A comparison study. Journal of Theoretical and Applied Information Technology, 79(1), 2005–2015.

Martínez, G., & Aguilar, G. (2013, julio-diciembre). Reconocimiento de voz basado en MFCC, SBC y espectrogramas. Ingenius, 12–20. https://doi.org/10.13053/ingenius.10.1390-650X

Müller, M., Ellis, D. P. W., Klapuri, A., Richard, G., & Sein, J.-P. (2011, November). Signal processing for music analysis. IEEE Journal of Selected Topics in Signal Processing, 5(6), 1088–1110. https://doi.org/10.1109/JSTSP.2011.2112333

Ostermann, F., Vatolkin, I., & Ebeling, M. (2023). AAM: A dataset of artificial audio multitracks for diverse music information retrieval tasks. Journal of Audio Speech and Music. https://doi.org/10.1186/s13636-023-00278-7

Prabavathy, S. V. R. (2020, May). Musical instruments classification using pre-trained model. International Research Journal of Engineering and Technology (IRJET), 7(5), 585–589.

Proakis, J. G., & Manolakis, D. G. (2007). Tratamiento digital de señales (4ª ed.). Madrid, España: Prentice Hall. ISBN 978-84-8322-347-5.

Relkar, V. T. C. (2019, September). Musical instrument identification using Machine Learning. International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET), 2(9), 1826–1829.

Russell, S. J., & Norvig, P. (2004). Inteligencia artificial: Un enfoque moderno (2ª ed.). Madrid, España: Pearson Educación. ISBN 978-84-205-4003-0.

Shah, D., Narayanan, A., & Espinosa-Ramos, J. I. (2022). Utilizing the neuronal behavior of spiking neurons to recognize music signals based on time coding features. IEEE Access, 10, 37317–37329. https://doi.org/10.1109/ACCESS.2022.3164440

Shinde, P. P., & Shah, S. (2018, August). A review of machine learning and deep learning applications. En Proceedings of the Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 1–6. Pune, India. https://doi.org/10.1109/ICCUBEA.2018.8697857

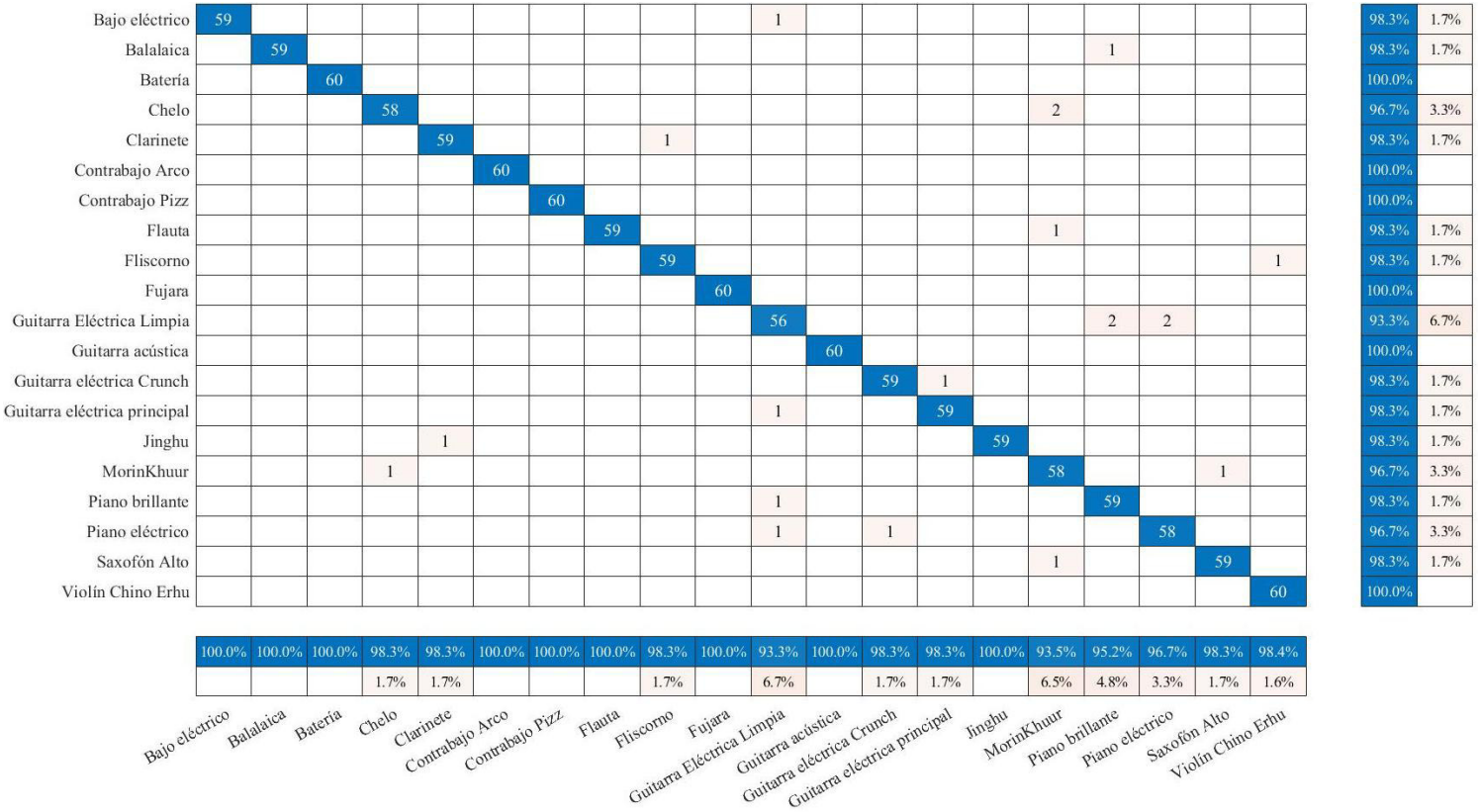
Tepepa, A., Pérez, H. M., & Nakano, M. (2018). Algoritmos de aprendizaje supervisado para la clasificación de géneros musicales caracterizados mediante modelos estadísticos. Research in Computing Science, 147(5), 119–128. Instituto Politécnico Nacional. ISSN 1870-4069. https://doi.org/10.13053/rcs-147-5-9

Vieira, R., Araújo, J., Batista, E., & Schiavoni, F. (2021). Automatic classification of instruments from supervised methods of machine learning. En Anais do XVIII Simpósio Brasileiro de Computação Musical (1–7). Porto Alegre, Brasil: Sociedade Brasileira de Computação (SBC). https://doi.org/10.5753/sbcm.2021.19418

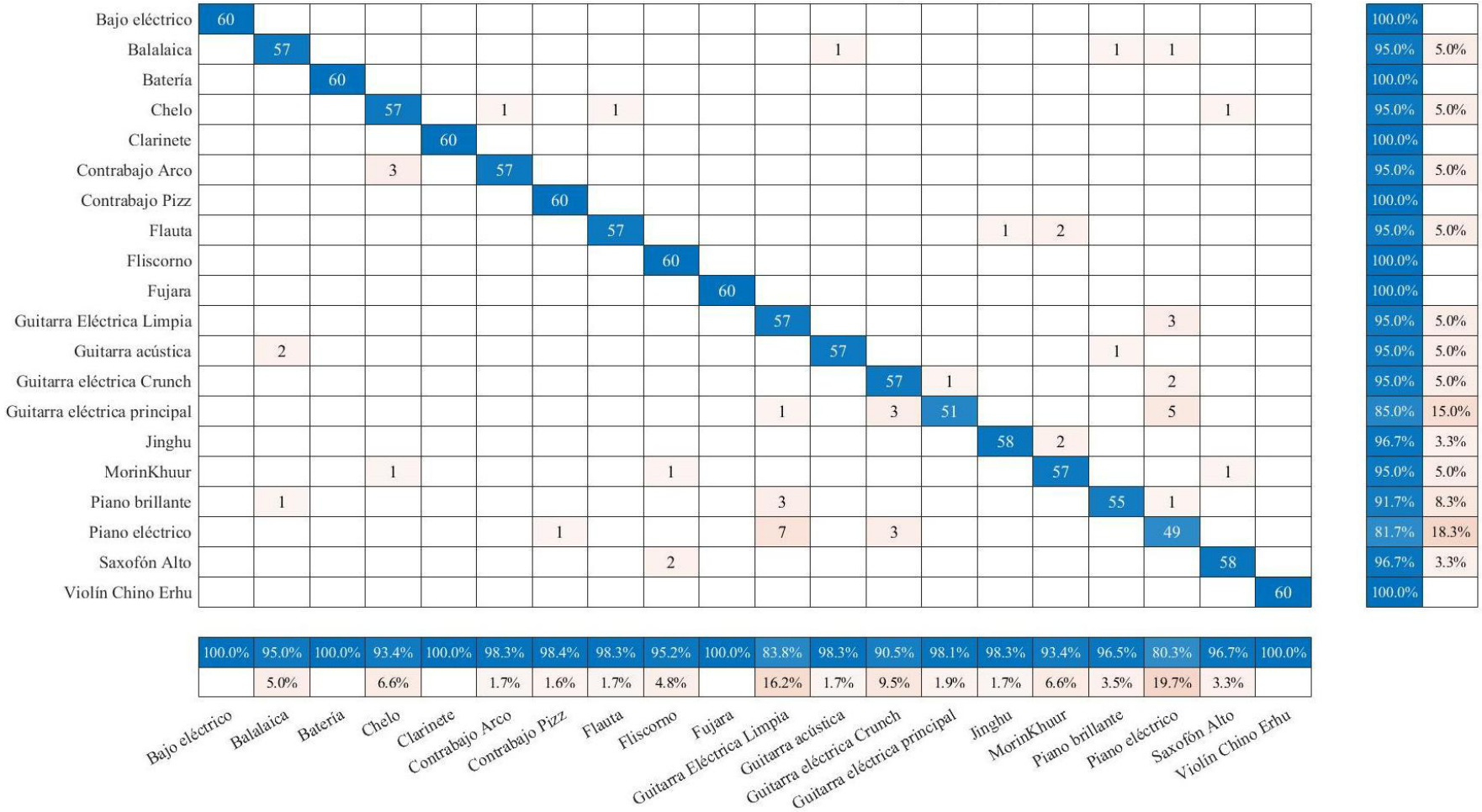
Anexos.

Imágenes detalladas de las Matrices de Confusión.

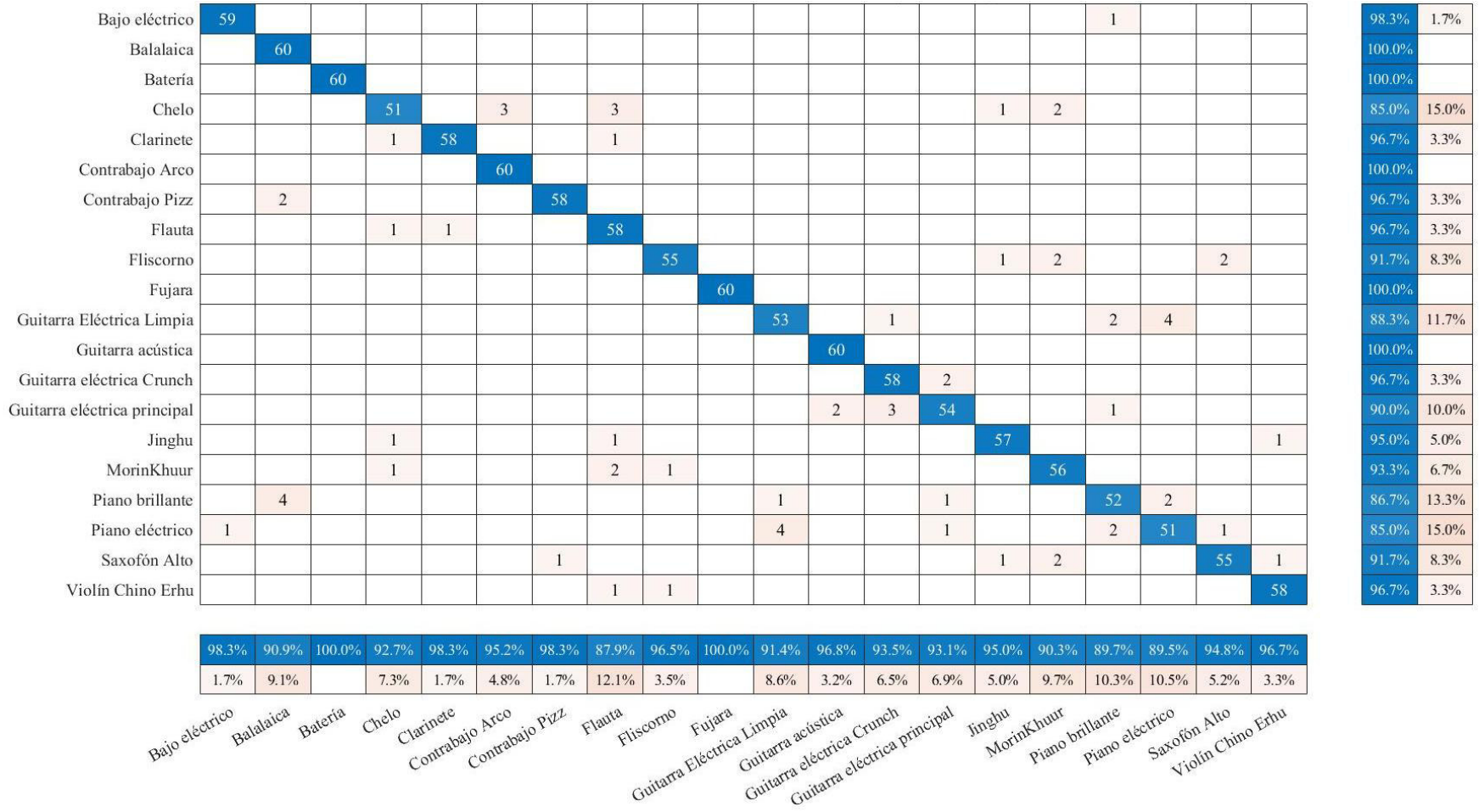
Matriz de Confusión para MLP de 20 clases.



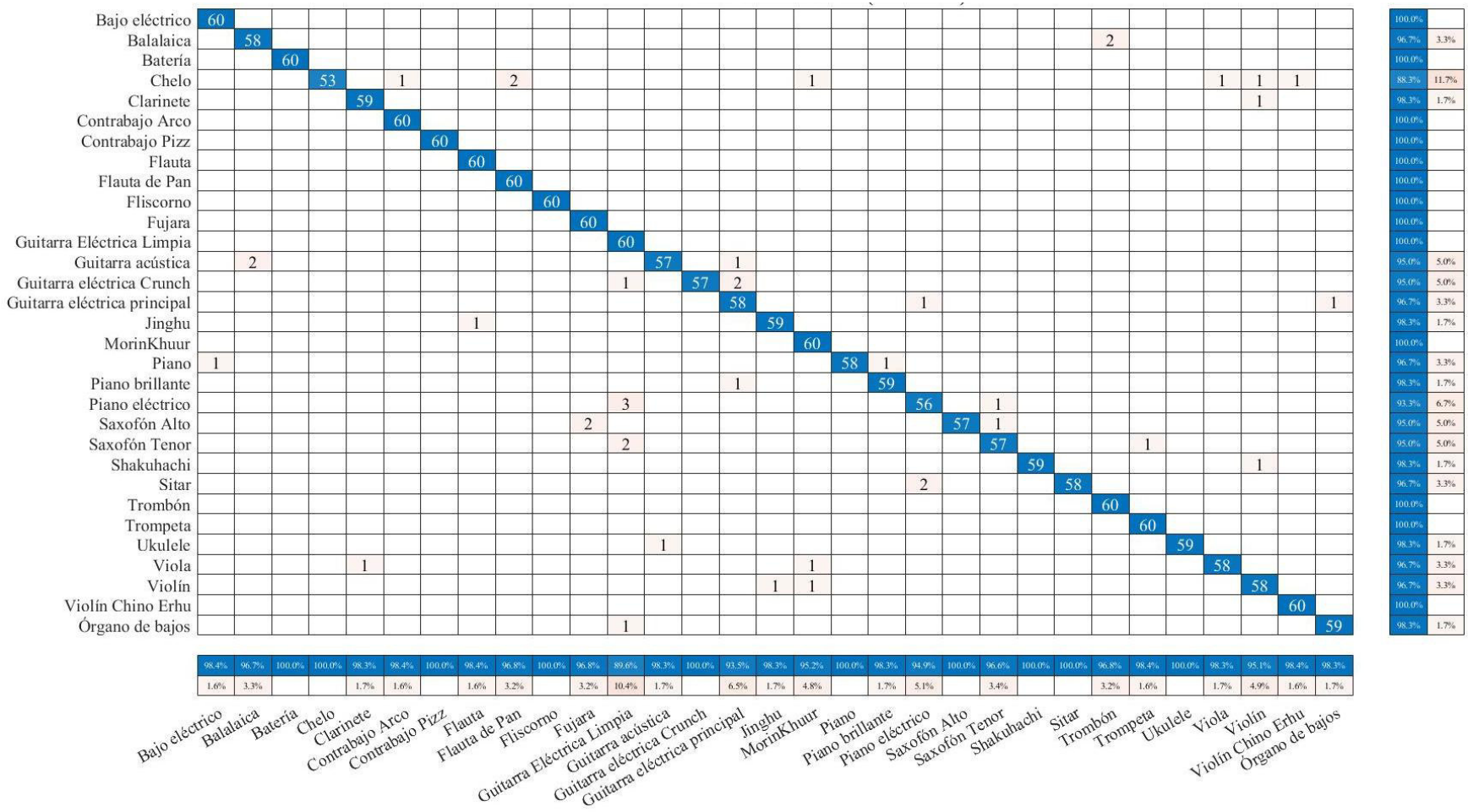
Matriz de Confusión para SVM de 20 clases.



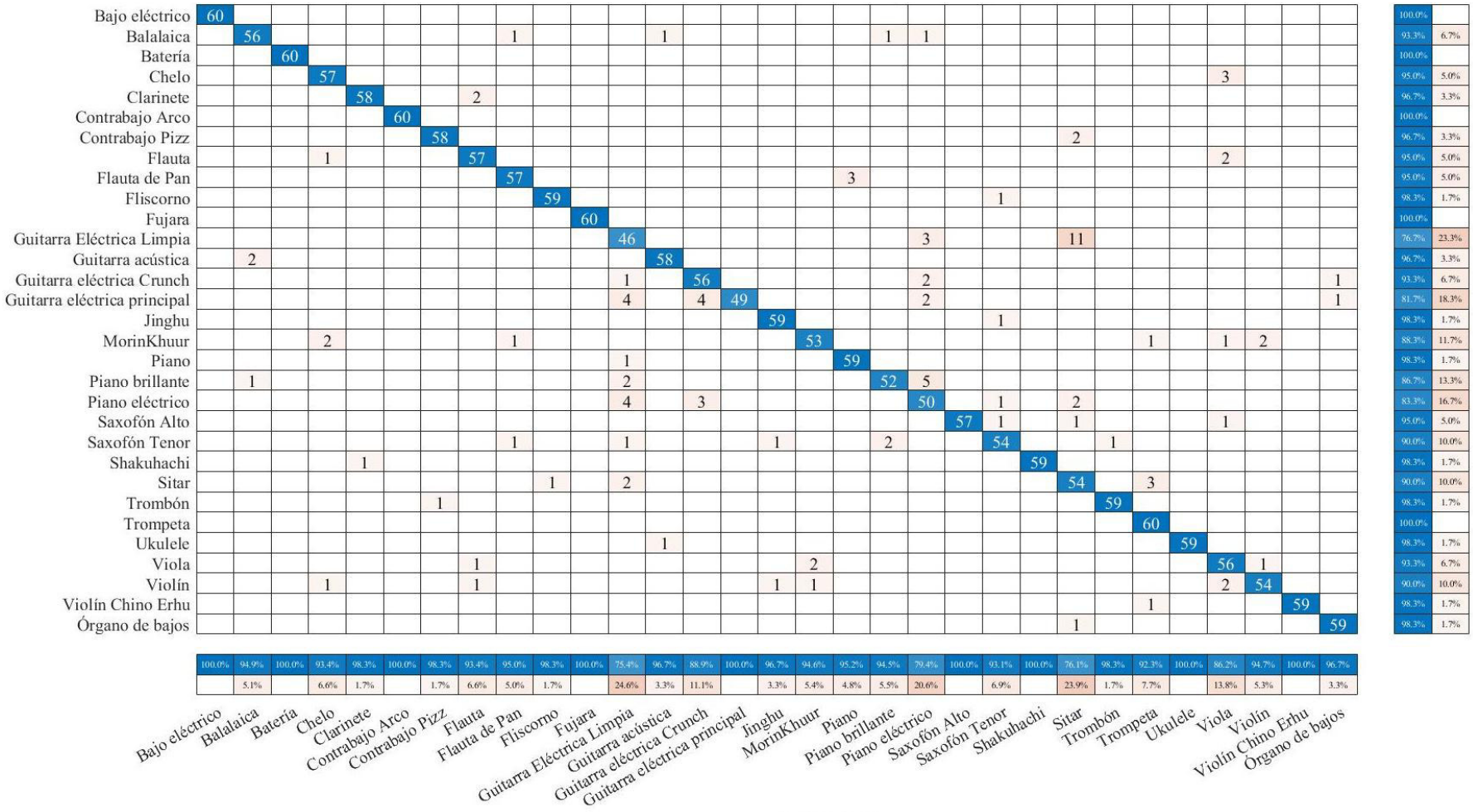
Matriz de Confusión para KNN de 20 clases.



Matriz de Confusión para MLP de 31 clases.



Matriz de Confusión para SVM de 31 clases.



Matriz de Confusión para KNN de 31 clases.