ISSN 2395-8618

Reconocimiento de instrumentos musicales a partir de señales de audio mediante aprendizaje profundo

Erick Michel Ramírez Rodríguez, Hind Taud, Magdalena Saldana Pérez, José Luis Oropeza Rodríguez

***Resumen*—Actualmente el aprendizaje profundo se ha involucrado en diferentes áreas de la ciencia, resultando sumamente útil en tareas relacionadas con la visión por computadora. Sin embargo, hay otras áreas de la tecnología que se han visto beneficiadas, entre ellas podemos mencionar el procesamiento digital de señales. El procesamiento de señales de audio generadas por instrumentos musicales es de gran relevancia para el mercado de la música, pero lo es también para la preservación del acervo musical. La propuesta tiene como objetivo identificar las ventajas de la implementación del aprendizaje profundo en el análisis de señales de audio, mediante la clasificación de instrumentos musicales. Se implementaron dos redes neuronales que son entrenadas mediante señales de audios e imágenes de características de audio. Las señales de audio empleadas forman parte de un repositorio de dominio público de la orquesta filarmónica de Londres. Los resultados obtenidos indican que la red profunda propuesta arroja resultados que difieren un 0.01% con respecto a los obtenidos con una red convolucional. Quedando demostrado que la efectividad de la propuesta presentada está cercana a la existente en trabajos similares presentes en el estado del arte. En donde se emplea la extracción de características como parte de los resultados que se obtienen, a diferencia de solamente tomar la señal de entrada como en este trabajo se plantea.**

***Palabras clave*—Procesamiento de señales, audio, señales de audio, aprendizaje profundo, redes neuronales.**

Recognition of Musical Instruments from Audio Signals Using Deep Learning

***Abstract*—Currently, deep learning has been involved in different areas of science, being useful in tasks related to computer vision. However, there are other areas of technology that have benefited, among them we can mention digital signal processing. The processing of audio signals generated by musical instruments is of great relevance for the music market, but also for the preservation of the musical heritage. The proposal aims to identify the advantages of implementing deep learning in the analysis of audio signals, through the classification of musical instruments. Two neural networks were implemented that are trained using audio signals and audio feature images. The audio signals used are part of a public domain repository of the London Philharmonic Orchestra. The results obtained indicate that the proposed deep network neural result that differ by 0.01% with respect to those obtained with a convolutional network. It has been demonstrated that the effectiveness of the proposal presented is close to that existing in similar works present in the state of the art. In which feature extraction is used as part of the results obtained, as opposed to only taking the input signal as in this work.**

***Index Terms*—Signal processing, audio, audio signals, deep learning, neural networks.**

I. Introducción

Poder identificar sonidos específicos en señales de audio es de utilidad en diferentes ámbitos, por ejemplo: identificar señales de audio generadas en salones de clase, sonidos que mejoran la experiencia educativa; entre otros. En el caso de la música, resulta útil separar sonidos de instrumentos que fueron grabados en conjunto, para mejorar las grabaciones, o bien sustituir los sonidos por nuevos. En el caso de conversaciones grabadas, diferenciar voces de forma individual, permite identificar las características vocales de los participantes.

Los trabajos existentes del aprendizaje máquina o el aprendizaje profundo para el reconocimiento, tienen la necesidad de tomar como datos de entrada de la red las características de las señales extraídas con la intervención de un experto.

Se han construido sistemas que extraen información relevante de los sonidos de los instrumentos musicales con el fin reconocer sus fuentes (ver cap. 2.1). Los métodos utilizados en la implementación de sistemas de reconocimiento de instrumentos musicales provienen de diferentes enfoques, particularmente el procesamiento de señales y reconocimiento de patrones. Por un lado, estos enfoques tradicionales necesitan la definición y extracción de las características con las que depende la exactitud del modelo de clasificación; por otro lado, el mundo de los medios digitales y de audio en específico, está creciendo a un ritmo muy vertiginoso. Por este crecimiento existe la necesidad de obtener de manera automática las características de audio.

En este trabajo se investiga la respuesta de un clasificador utilizando señales sin la necesidad de extraer las características propias de la señal. Con este fin, se propone un clasificador utilizando los datos originales de la señal; tarea que ha sido utilizada en los trabajos existentes del aprendizaje máquina o el aprendizaje profundo.

Por lo anterior, se propone una metodología de red neuronal multicapa Deep Neural Network (DNN), donde los datos de entrada son presentados en forma de vector de una dimensión.

Manuscript received on 18/05/2023, accepted for publication on 02/09/2023.

E.M. Ramírez Rodríguez, M. Saldana Perez, J.L. Oropeza Rodríguez are with the Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), CDMX, México ({rramirezr2023, amagdasaldana, joropeza}@cic.ipn.mx).

Hind Taud is with the Instituto Politécnico Nacional (IPN), Centro de Innovación y Desarrollo Tecnológico en Cómputo (CIDETEC), CDMX, México (htaud@ipn.mx).

ISSN 2395-8618

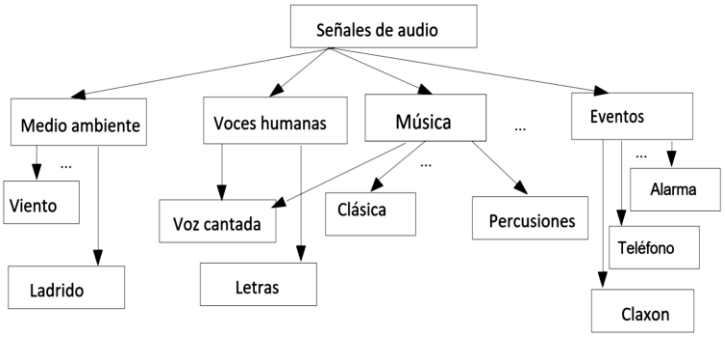


Fig. 1. Categorías de señales de audio

TABLA I
TRABAJOS PREVIOS PARA EL RECONOCIMIENTO DE INSTRUMENTOS

Autores	Esquemas de características	Algoritmos de clasificación
A. Wieczorkowska and A. Czyzewski [2]	Wavelets y transformada de Fourier (FT)	Árboles de decisión
S.Essid, G.Richard, and B. David [3]	Algoritmos genéticos de 150 características	Modelos de mezcla Gaussianas (GMM) y Máquinas de Soporte Vectorial (SVM)
M. E. Özbek, C. Delpha, and P. Duhamel [4]	Coeficientes Cepstrales de las Frecuencias Mel (MFCC)	Máquinas SVM
J. D. Deng, C. Simmermacher, and S. Cranefield [5]	Técnicas estándar de reducción de dimensiones PCA e ISOMAP	Algoritmos Bayes, KNN, MLP, RBF y SVM.
C. Joder, S. Essid, and G. Richard [6]	Segmentación y características temporales	Modelo de Markov HMM y SVM, GMM
N. Senan, R. Ibrahim, N. M. Nawi, I. T. R. Yanto, and T. Herawan [7]	MFCC	SVM, K-Nerest Neighbor (KNN) y Naïve Bayes.
D. G. Bhalke, C. B. R. Rao, and D. S. Bormane [8]	MFCC y FT	Redes neuronales

Con el fin de evaluar su rendimiento de la red mencionada en el párrafo anterior, y debido a que en el estado del arte las bases de datos utilizadas no se encuentran disponibles. Se eligió hacer una comparación con una red neuronal convolucional CNN (Convolutional Neural Network, por sus siglas en inglés), que tiene como entrada la imagen del espectrograma de la señal de audio modificado por un banco de filtros que responden al comportamiento de la frecuencia de Mel. [1].

El trabajo se estructura de la siguiente forma, en la primera sección se presenta la introducción, en la segunda sección se describen trabajos relacionados a la investigación, la tercera sección describe los métodos empleados, en la cuarta sección se describe la metodología propuesta, en la quinta sección se muestran los resultados obtenidos; y finalmente se presentan las conclusiones de la propuesta.

II. TRABAJOS RELACIONADOS

Las señales de audio pueden ser divididas en varias categorías como se muestra en Fig. 1. Cabe señalar que los límites entre categorías no siempre se definen de forma inequívoca. Por ejemplo, que la clase "música" y la clase "voces humanas" se superponen ya que la clase "voz cantada" pertenece a estas dos categorías. Por lo tanto, la definición de categorías disjuntas puede ser complicada en contextos de aplicación particulares.

El análisis de señales de audio en general tiene muchas aplicaciones prácticas, entre ellas se encuentra el contenido musical que incluye tareas difíciles como la codificación estructurada, anotaciones musicales; entre otros. El reconocimiento automático de instrumentos musicales es una subtarea crucial para resolver estas tareas a partir de señales musicales.

Esta recopilación de trabajos previos se presenta separándola en métodos tradicionales para el reconocimiento de instrumentos y aplicaciones del aprendizaje profundo para tareas relacionadas con audio.

A. Reconocimiento de instrumentos

Se han propuesto y adoptado varios esquemas de características en la literatura del análisis de sonido de instrumentos. Además de los esquemas de características adoptados, se han empleado diferentes modelos computacionales o algoritmos de clasificación con el fin de detectar y clasificar instrumentos a partir de señal de audio.

A continuación, se presenta una recopilación de los trabajos que obtienen la clasificación de instrumentos, sus esquemas de características empleados y algoritmos de clasificación.

Los métodos de clasificación de instrumentos musicales que combinan la reconstrucción optimizada del espacio característico con redes neuronales utilizan un análisis de componentes principales y un coeficiente de correlación para la optimización. Tal como se describe en [8].

En los trabajos del estado descritos en la Tabla 1, tal como se observa, se requiere la extracción de características para realizar el reconocimiento de instrumentos musicales, por lo que realizan dicha labor de forma diferente a la presente propuesta.

ISSN 2395-8618

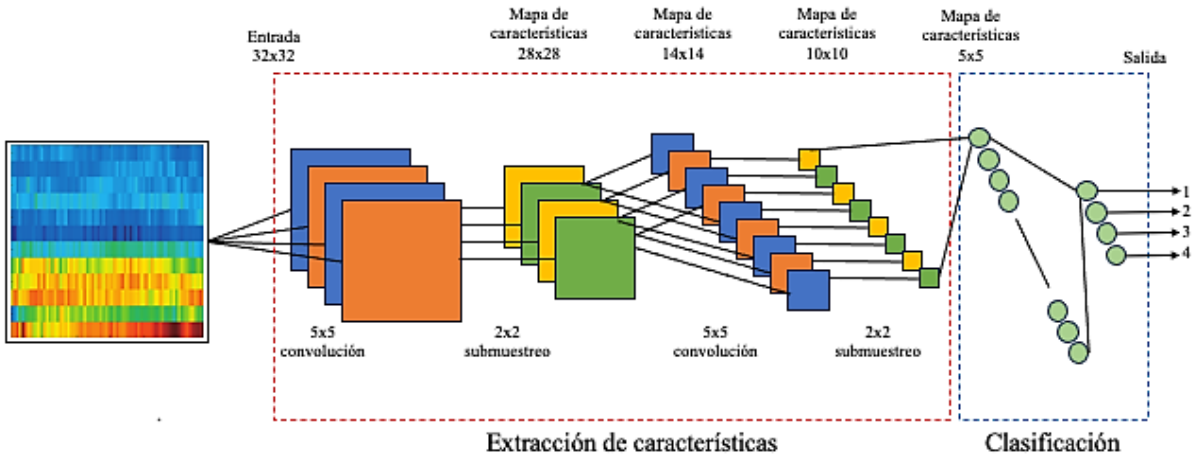


Fig. 1. Esquema general del aprendizaje profundo

TABLA II
USO DE APRENDIZAJE PROFUNDO PARA AUDIO

Autores	Campo de investigación
Z. Ling, S. Kang, H. Zen, A. W. Senior, M. Schuster, X. Qian, H. M. Meng, and L. Deng [9]	Generación de voz
M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani [10]	Extracción de características para CNN
C. Kereliuk, B. L. Sturm, and J. Larsen [11]	Análisis de contenido musical para CNN
D.YuandJ.Li [12]	Reconocimiento de voz para RNN y CNN
P. Bell, P. Swietojanski, and S. Renals [13]	Reconocimiento automático de voz mediante entropía cruzada
Y. Kim, M. J. Kim, J. Goo, and H. Kim [14]	Reconocimiento automático del habla junto LSTM
S. Oramas, F. Barbieri, O. Nieto, and X. Serra [15]	Etiquetado de género musical
X. Li, Y. Guan, Y. Wu, and Z. Zhang [16]	Estimación de pasos de piano
Y. Shen, J. Cao, J. Wang, and Z. Yang [17]	Clasificación acústica urbana basado en CNN y DBN

B. Aprendizaje profundo

Áreas de la computación como el procesamiento de señales mediante algoritmos computacionales complejos, se beneficia poco a poco con los avances de otras áreas de la computación, adaptando los nuevos algoritmos y metodologías para que sean capaces de trabajar con señales.

En los últimos años, el aprendizaje profundo no solo ha crecido en los campos de investigación de la visión por computadora y el reconocimiento de voz, sino también en campos que analizan las señales acústicas o señales de audio. Campos tales como: la detección de eventos acústicos, el procesamiento del lenguaje y el habla por computadora, la generación de voz, análisis de contenido musical mediante, género musical, recuperación de información musical, detección de sonido ambiental y separación de fuente, mejora de audio; entre otros.

A continuación, se describen algunos trabajos que han tenido relevancia en los distintos campos mencionados, con el objetivo de ampliar la visión de las aplicaciones y mejoras que tiene el uso del aprendizaje profundo en distintos campos relacionados con las señales de audio.

Como se puede observar en la Tabla 2, la aplicación de las redes neuronales profundas comienza a ser relevante para fines de investigación en trabajos relacionados con el audio. Siendo pertinente el hecho de ampliar su aplicación en tareas de reconocimiento de instrumentos, tal como sucede en el caso de algoritmos de inteligencia artificial que se mencionaron en la Tabla 1.

Dado lo anterior, resulta adecuado realizar el estudio tanto del uso de las redes DNN para el reconocimiento de instrumentos, así como la utilización directa de la señal sin tener que extraer parámetros característicos con el objetivo de evaluar el comportamiento de dicha propuesta, tal como se desarrolla en el presente trabajo.

III. MATERIALES Y MÉTODOS

A continuación, se describen conceptos tales como el aprendizaje profundo para realizar la tarea de clasificación; redes neuronales convolucionales y el espectrograma modificado por un banco de filtros que responden al comportamiento de la frecuencia de Mel para la comparación de resultados. Así también se describen tanto la base de datos utilizada como las herramientas para el procesamiento de datos.

El Aprendizaje Profundo [18, 19] es una rama del aprendizaje automático relacionado con las redes neuronales artificiales. Durante los últimos años, esta rama ha tenido un

ISSN 2395-8618

TABLA III
NÚMERO DE MUESTRAS Y DISTRIBUCIÓN

Instrumentos	Número de muestras			
	Totales	Entrenamiento	Prueba	Validación
Flauta	877	529	174	174
Violín	1502	904	299	299
Guitarra	106	66	20	20
Saxofón	733	441	146	146
Total	3218	1940	639	639

Entrenamiento	Prueba	Validación
~60%	~20%	~20%

gran impacto en varios campos de la ciencia como la clasificación de imágenes, el reconocimiento de objetos, comprensión del lenguaje natural, análisis de sentimientos, la respuesta a preguntas y la traducción de idiomas entre otros.

Los avances en hardware, específicamente las Unidades de Procesamiento Gráfico o GPU (Graphics Processing Unit), han permitido abordar problemas con redes profundas en un tiempo razonable y la disponibilidad de una gran cantidad de datos etiquetados o no etiquetados lleva a un aprendizaje eficiente.

La Fig. 2 muestra un esquema general de la arquitectura de una red DNN5 que se usa como referencia al empleado en el presente trabajo. El esquema de aprendizaje profundo se puede representar mediante dos partes entre los datos de entrada y los datos clasificados de salida.

Compuesto por las capas ocultas o una estructura de red profunda, la primera parte realiza la extracción de características y la segunda involucra un clasificador como Regresión Logística o Softmax que predice la etiqueta de clase en una forma de distribución de probabilidad. Dependiendo del problema que las redes neuronales puedan abordar, la primera parte se puede lograr mediante un enfoque de aprendizaje automático supervisado como las redes neuronales convolucionales (CNN), uno no supervisado como Auto-Encoders (AE) o ambos como AE convolucionales.

Las redes neuronales convolucionales (CNN), son redes neuronales multicapa o profundas utilizadas en el procesamiento y clasificación de imágenes. Específicamente, las CNN tienen una gran utilidad para encontrar patrones en las imágenes para reconocer ciertos objetos, rostros, escenas, etc. Son redes capaces de aprender directamente de los datos, ya que toman en cuenta la información espacial de los objetos. Las CNN contienen tres capas: convolucional, de agrupación (pooling) y completamente conectada (FC); que logran extraer las características y modificando la dimensionalidad, hasta obtener una salida de cada una de las clases.

Un banco de filtros que responden al comportamiento de la frecuencia de Mel es un análisis de señal ampliamente usado en tareas de reconocimiento de voz y que está basado en la psicoacústica.

La base de datos Philharmonia Orchestra (PHO). Es una colección de grabaciones de varios instrumentos musicales de la orquesta filarmónica de Londres. A diferencia de las otras fuentes de datos, el conjunto de datos PHO proporciona grabaciones de notas individuales en lugar de escalas musicales, lo que permite omitir el paso de división de archivos. La colección contiene 20 instrumentos musicales diferentes. Para cada instrumento, las muestras abarcan todo su conjunto de tonos tocados en cada octava con diferentes niveles de fuerza (piano, forte) y duración.

Los archivos están publicados en formato MP3 a una frecuencia de muestreo de 44,1 kHz con una tasa de bits que varía entre 64 y 96 KiloBytes por segundo.

IV. METODOLOGÍA PROPUESTA

Un concepto central en los estudios de señales musicales es la calidad del sonido. Un sonido musical emitido por un instrumento tiene cuatro atributos que lo caracterizan: tono, volumen o intensidad, duración y timbre. Estos cuatro atributos hacen posible que un oyente distinga instrumentos musicales entre sí. El tono, el volumen y la duración se comprenden mejor que el timbre y tienen claras contrapartes físicas.

Para los sonidos musicales, el tono está bien definido y es casi igual a la frecuencia fundamental. La contraparte física de la sonoridad es la intensidad, que es proporcional al cuadrado de la amplitud de la presión acústica. La tercera dimensión, la duración percibida, se corresponde estrechamente con la duración física con tonos no muy cortos. El timbre es el menos comprendido entre los cuatro atributos. Tradicionalmente, el timbre se define como la calidad de un sonido por la cual un oyente puede decir que dos sonidos del mismo volumen y tono son diferentes [8, 9].

Los seres humanos tienen una habilidad aprendida para identificar un instrumento musical con solo escuchar el sonido. A pesar de la complejidad de esta tarea, se puede identificar cierto número de instrumentos y ciertas configuraciones de mezcla. La tarea de reconocimiento automático de instrumentos tiene como objetivo intentar construir algoritmos que identifiquen la información a partir de señales acústicas digitales.

Con el fin de limitar el alcance de este trabajo, se seleccionaron los siguientes 4 instrumentos para entrenar el modelo: flauta, saxofón, violín y guitarra, (ver Tabla 3), de la base de datos de audio del dominio público de Philharmonia Orchestra (PHO).

Se eligió una red neuronal convolucional basada en el modelo VGGNet, debido a que puede tomar como entrada una imagen de tamaño reducido, donde se le aplica una succión de filtros de

ISSN 2395-8618

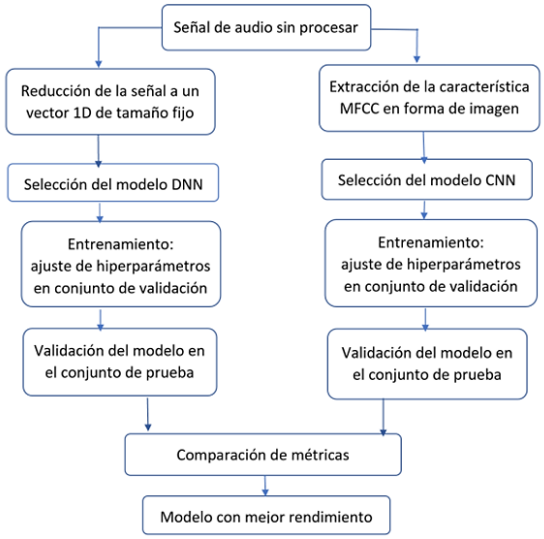


Fig. 2. Diagrama de flujo que presenta la secuencia de la metodología propuesta para el proyecto

Num_labels 4
Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	10496
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 256)	65792
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 256)	65792
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 256)	65792
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 256)	65792
dropout_4 (Dropout)	(None, 256)	0
dense_5 (Dense)	(None, 4)	1028

Total params: 274,692
Trainable params: 274,692
Non-trainable params: 0

Fig. 3. Modelo DNN implementado con cinco capas en las que se aplica el dropout, se cuenta con 274,692 parámetros

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 38, 172, 32)	320
leaky_re_lu (LeakyReLU)	(None, 38, 172, 32)	0
batch_normalization (BatchNo	(None, 38, 172, 32)	128
spatial_dropout2d (SpatialDr	(None, 38, 172, 32)	0
conv2d_1 (Conv2D)	(None, 36, 170, 32)	9248
leaky_re_lu_1 (LeakyReLU)	(None, 36, 170, 32)	0
batch_normalization_1 (Batch	(None, 36, 170, 32)	128
max_pooling2d (MaxPooling2D)	(None, 18, 85, 32)	0
spatial_dropout2d_1 (Spatial	(None, 18, 85, 32)	0
conv2d_2 (Conv2D)	(None, 16, 83, 64)	18496
leaky_re_lu_2 (LeakyReLU)	(None, 16, 83, 64)	0
batch_normalization_2 (Batch	(None, 16, 83, 64)	256
spatial_dropout2d_2 (Spatial	(None, 16, 83, 64)	0
conv2d_3 (Conv2D)	(None, 14, 81, 64)	36928
leaky_re_lu_3 (LeakyReLU)	(None, 14, 81, 64)	0
batch_normalization_3 (Batch	(None, 14, 81, 64)	256
global_average_pooling2d (Gl	(None, 64)	0
dense (Dense)	(None, 4)	260

Total params: 66,020
Trainable params: 65,636
Non-trainable params: 384

Fig. 4. Modelo de CNN implementado con 66,020 datos empleando VGGNet

convolución de tamaño 3x3, permitiendo así lograr una red profunda.

El modelo de la red multicapa consiste en varias capas ocultas, y para observar efectos de profundidad, se hará la comparación de redes DNN de 5 capas y de 8 capas similares.

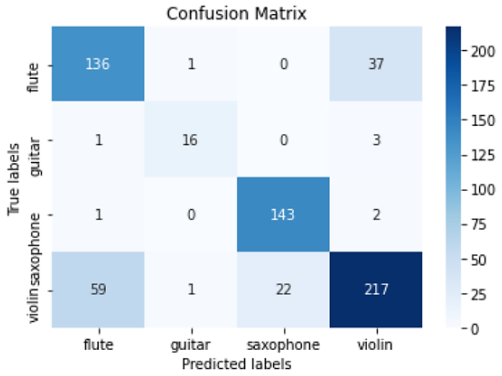
En la Fig. 3 se presenta el diagrama de bloques de cada una de las etapas por las que las señales de audio son procesadas hasta su reconocimiento, del lado izquierdo se muestra para el uso de la red DNN y del lado derecho para el de la CNN. En el caso de las redes multicapa en el bloque de selección del modelo de DNN se hace uso de dos arquitecturas con diferentes profundidades denotado como DNN5 y DNN8.

Para ambas redes multicapa DNN5 y DNN8 se utiliza la técnica de abandono (dropout), la cual permite reducir el posible sobreajuste de la red, se basa en omitir el funcionamiento de algunas neuronas del modelo aleatoriamente durante el proceso de entrenamiento. El funcionamiento del dropout permite que la red sea más pequeña, ajustando su tamaño, para que se evite el sobre entrenamiento. En la Fig. 4 se observan las características del modelo implementado.

ISSN 2395-8618

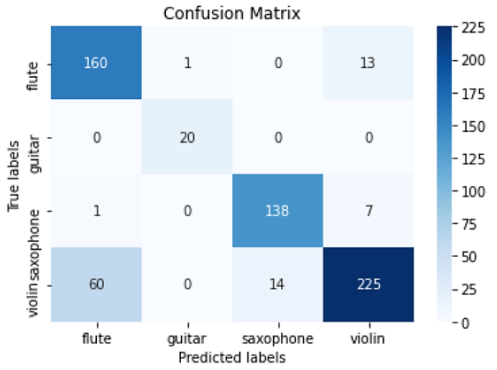
TABLA IV
ECUACIONES DE MÉTRICAS DE EVALUACIÓN

Métricas de evaluación	
Precision	$\frac{TP_i}{TP_i + FP_i}$
Recall	$\frac{TP_i}{TP_i + FN_i}$
F1	$\frac{2 \times Precision \times Recall}{Precision + Recall}$



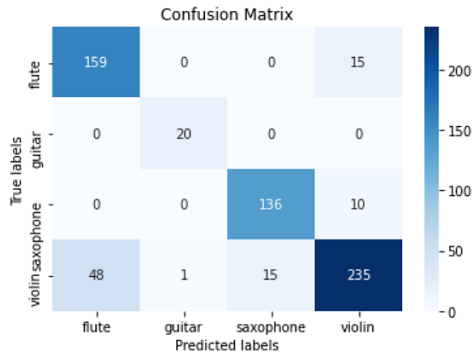
(c) Matriz de confusión

Fig. 5. Red DNN5



(c) Matriz de confusión

Fig. 6. Red DNN8



(c) Matriz de confusión

Fig. 7. Red CNN

De manera similar se implementó una red CNN para ello, se utilizó una versión de VGGNet que es una arquitectura de red neuronal convolucional, tal como se mencionó previamente. El modelo tiene varias capas formadas por convolución con filtro de tamaño 3, función de activación, la normalización de los pesos en cada Batch, Dropout, y Pooling.

El propósito de usar VGGNet es aumentar la profundidad de la red para mejorar la precisión de la clasificación. En la Fig. 5 se muestra el modelo de CNN planteado.

Se seleccionaron los siguientes hiper-parámetros para realizar el entrenamiento de los modelos DNN5, DNN8 y CNN; función de pérdida Categorical Crossentropy, Optimizador Adam, Dropout 0.5. Como se mencionó anteriormente se hace uso de la técnica de Dropout para evitar el sobre entrenamiento.

V. RESULTADOS

Las métricas de evaluación para este trabajo son las más comunes usadas en la clasificación de reconocimiento de patrones, tales como la tasa de exactitud que es la proporción entre el número de predicciones correctas y el total de predicciones.

Para las métricas de evaluación, se aplican las métricas de precisión de tasa de exactitud, precisión y recall. En la Tabla 4, se presenta la metodología propuesta para el proyecto.

Aunque la tasa de exactitud es una métrica importante, no siempre es adecuada para medir el desempeño de un modelo en caso de que haya un desequilibrio en el número de muestras en las clases lo que implica la necesidad de tener otras métricas.

Se destaca la medida F1 que se calcula a partir de la media armónica entre las medidas recall y precisión. La exactitud se calcula para el modelo en su conjunto, pero recall y precisión se calculan para las clases individualmente.

En este caso:

- TP es el número de veces que una clase es verdadera y su predicción es verdadera.
- TN es el número de veces que una clase es falsa y su predicción es falsa.
- FP es el número de veces que una clase es falsa y su predicción es verdadera.
- FN es el número de veces que una clase es verdadera y su predicción es falsa.
- i es índice de la clase
- N es el número total de las clases.

El modelo de red multicapa presentado en la metodología está formado de 5 capas ocultas (DNN5), una capa de la entrada, y una de la salida. El tamaño del Batch empleado es de 50 y la de iteraciones es 200 (ver Fig. 6). Con el aumento de número de capas ocultas a 8 (DNN8), con el mismo tamaño de Batch de 50 y de iteraciones 200 (ver Fig. 7).

Para el entrenamiento de modelo de Convolución (CNN) con un tamaño de Batch de 30 y de iteraciones 200 (ver Fig. 8).
La comparación de las tasas de exactitud, de MacroF1 y MicroF1 (ver Tabla 5), nos indica que la red DNN8, red propuesta en este trabajo para clasificar señales sin procesar, arroja resultados casi similares a la de convolución. De otro lado, si comparamos el salto de las tasas entre DNN5 y DNN8, implicará que una DNN12 o una DNN con más capas superará la de CNN.

ISSN 2395-8618

TABLA V
COMPARACIÓN DE RESULTADOS

	Precisión			Recall		
Instrumentos	**DNN5**	**DNN8**	**CNN**	**DNN5**	**DNN8**	**CNN**
Flauta	0.6903	0.7239	0.7681	0.7816	0.9195	0.9137
Guitarra	0.8888	0.9523	0.9523	0.8	1	1
Saxofón	0.8666	0.9078	0.9006	0.9794	0.9452	0.9315
Violín	0.8378	0.9183	0.9038	0.7257	0.7525	0.7859

Métricas	**DNN5**	**DNN8**	**CNN**
Exactitud	0.8012	0.8497	0.8607
Macro F1	0.8181	0.8847	0.8917
Micro F1	0.8181	0.8497	0.8607

TABLA VI
EXACTITUD EN EL CONJUNTO DE PRUEBA

Modelos	**Tasa de exactitud**
DNN5	0.75
DNN8	0.81
CNN	0.82

Lo más importante en el aprendizaje automático es la etapa de generalización. Se prueba el modelo ya entrenado en datos nunca vistos antes, que son los datos de validación.

Los resultados de la exactitud se describen en la Tabla 6. La comparación de las tasas de exactitud, refuerza los resultados anteriores relacionados a que la red DNN8, red propuesta en este trabajo para clasificar señales sin procesar, arroja resultados casi similares a la de convolución.

VI. CONCLUSIONES

En este trabajo, se muestran los resultados de los clasificadores DNN5, DNN8 (redes multicapa) y CNN. Para el caso de las redes multicapa los datos de entrada son los datos originales de la señal, es decir sin la necesidad de extraer las características de la señal.

Los resultados muestran el gran potencial que tiene el aprendizaje profundo para la clasificación de instrumentos musicales a partir de los datos originales de la señal de audio, específicamente las redes neuronales multicapa.

Los resultados demuestran que para todas las métricas utilizadas (Precision, Recall, F1), al incrementar el número de número de capas internas de una red tipo DNN, los resultados se aproximan adecuadamente a los obtenidos con CNN.

Con la elección, obtención y arreglo del formato de la base de datos, el desarrollo de un clasificador de instrumentos musicales a partir de los datos originales de la señal, su implementación y su comparación con la red neuronal convolucional CNN fue posible demostrar el alcance inicial que se planteó en el presente trabajo de investigación.

Se proponen algunas mejoras al presente trabajo de investigación:

- Emplear mayor número datos con el fin de mejorar los resultados obtenidos, ya que está demostrado que el aprendizaje profundo funciona mejor entre mayor sea el número de ejemplos.
- Emplear mayor número de capas para mejorar el desempeño, ya que está demostrado que el aprendizaje profundo funciona mejor entre mayor sea el número de capas de la red.
- Escalar a otros trabajos como la clasificación de instrumentos musicales con señal de audio compuesto por varias notas de varios instrumentos.
- Escalar el uso de señal original, sin transformarlo a características con los métodos de procesamiento de señal, a otras aplicaciones de clasificación de señales acústicas aprovechando mejor la capacidad de los modelos de aprendizaje profundo en la extracción de características.

REFERENCIAS

[1] G. Todor, F. Nikos, and G. Kokkinakis, *Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task*, University of Patras, Greece, (2005).

[2] A. Wieczorkowska and A. Czyzewski, “Rough set based automatic classification of musical instrument sounds,” *Electronic Notes in Theoretical Computer Science*, vol. 82, no. 4, pp. 298–309, 2003.

[3] S. Essid, G. Richard, and B. David, “Musical instrument recognition by pairwise classification strategies,” in *IEEE Trans. Speech Audio Process*., vol. 14, no. 4, pp. 1401–1412, (2006).

[4] M.E. Özbek, C. Delpha, and P. Duhamel, “Musical note and instrument classification with likelihood-frequency-time analysis and support vector machines,” in 15*th European Signal Processing Conference, EUSIPCO´07*, Poland, pp. 941–945, 2007.

[5] J.D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analy- sis for mu sical instrument classification," *IEEE Trans. Syst. Man Cybern. Part B*, vol. 38, no. 2, pp. 429–438, 2008.

ISSN 2395-8618

[6] C. Joder, S. Essid, and G. Richard, “Temporal integration for audio classification with application to musical instrument classification,” *IEEE Trans. Speech Audio Process*., vol. 17, no. 1, pp. 174–186, 2009.
[7] N. Senan, R. Ibrahim, N.M. Nawi, I.T.R. Yanto, and T. Herawan, “Rough and soft set approaches for attributes selection of traditional malay musical instrument sounds classification,” *Int. J. Softw. Sci. Comput. Intell.*, vol. 4, no. 2, pp. 14–40, 2012. 9
[8] D. G. Bhalke, C. B. R. Rao, and D. S. Bormane, “Automatic musical instru- ment classification using fractional Fourier transform based- MFCC features and counter propagation neural network,” *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 425–446, 2016.
[9] Z. Ling, S. Kang, H. Zen, A. W. Senior, M. Schuster, X. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends”, *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, 2015.
[10] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro temporal locality in deep learning based acoustic event detection," in *EURA- SIP J. Audio Speech Music. Process.*, vol. 2015, p. 26, 2015.
[11] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Trans. Multim*., vol. 17, no. 11, pp. 2059–2071, 2015.
[12] D. Yu and J.Li, "Recent progresses in deep learning based acoustic models," *IEEE CAA J. Autom. Sinica*, vol. 4, no. 3, pp. 396–409, 2017.
[13] P. Bell, P. Swietojanski, and S. Renals, "Multitask learning of context- dependent targets in deep neural network acoustic models," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 2, pp. 238–247, 2017.
[14] Y. Kim, M. J. Kim, J. Goo, and H. Kim, "Learning self-informed feature contribution for deep learning-based acoustic modeling," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 11, pp. 2204–2214, 2018.
[15] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," Trans. Int. Soc. Music. Inf. Retr., vol. 1, no. 1, pp. 4–21, 2018.
[16] X. Li, Y. Guan, Y. Wu, and Z. Zhang, "Piano multipitch estimation using sparse coding embedded deep learning," *EURASIP J. Audio Speech Music. Process.*, vol. 2018, p. 11, 2018.
[17] Y. Shen, J. Cao, J. Wang, and Z. Yang, "Urban acoustic classification based on deep feature transfer learning," *J. Frankl. Inst.*, vol. 357, no. 1, pp. 667–686, 2020.
[18] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning,” *Neural Comput. Appl.*, vol. 18, no. 7, pp. 436–444, 2015.
[19] Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press book, 2016.