ISSN 1870-4069

Clasificación de la señal de audio cardiaco mediante la transformada de Fourier de tiempo corto y aprendizaje profundo

Leonel Orozco-Reyes[1], Miguel-Ángel Alonso Arévalo[1], Eloísa García-Canseco[2], Roilhi F. Ibarra-Hernández[3]

[1] Centro de Investigación Científica y de Educación Superior de Ensenada, Departamento de Electrónica y Telecomunicaciones, México

[2] Universidad Autónoma de Baja California, Facultad de Ciencias, México

{orozcorl,aalonso}@cicese.edu.mx, roilhi.ibarra@universidaddeensenada.edu.mx eloisa.garcia@uabc.edu.mx

Resumen. La auscultación es una herramienta de diagnóstico no invasiva, de bajo costo y de sencilla implementación, que actualmente provee información importante en el diagnóstico de patologías cardíacas. Con ayuda de la auscultación se obtiene el sonido cardíaco o fonocardiograma, que es el elemento principal del análisis de este trabajo. Los errores de diagnóstico debido a la falta de médicos experimentados y las limitaciones del sistema auditivo humano han llevado al avance en el área de procesamiento digital de señales y el desarrollo de técnicas para el análisis de sonidos cardíacos asistidos por computadora. El presente trabajo tiene como objetivo clasificar señales de fonocardiogramas usando redes neuronales convolucionales. Haciedo uso de la transformada de Fourier de tiempo corto. Para generar matrices de características que mejor representen la señal de audio cardíaco. Las matrices obtenidas serán clasificadas usando redes neuronales convolucionales, en este trabajo se usarán dos arquitecturas de redes neuronales que han demostrado tener un desempeño notable en tareas de clasificación. Estas redes son AlexNet y VGG16. Los resultados obtenidos reflejan un mejor desempeño para la red convolucional AlexNet con una exactitud de 82.2 % sobre otras redes evaluadas.

Palabras clave: Transformada de Fourier de tiempo corto, sonidos cardiacos, redes neuoronales convolucionales.

Cardiac Audio Signal Classification Using the Fourier Transform of Short Time and Deep Learning

Abstract. Auscultation is a non-invasive, low-cost and easy-to-implement diagnostic tool that currently provides important information in the diagnosis of

cardiac pathologies. With the help of auscultation, the heart sound or phonocardiogram is obtained, which is the main element of the analysis of this work. Diagnostic errors due to a lack of experienced clinicians and the limitations of the human auditory system have led to advancement in the area of digital signal prosessing and the development of techniques for computer-aided analysis of heart sounds. The present work aims to classify phonocardiogram signals using convolutional neural networks. Making use of the short-time Fourier transform. To generate feature matrices that best represent the cardiac audio signal. The obtained matrices will be classified using convolutional neural networks, in this work two neural network architectures will be used that have shown to have a remarkable performance in classification tasks. These networks are AlexNet and VGG16. The results obtained reflect a better performance for the AlexNet convolutional network with an accuracy of 82.2% over other evaluated networks.

Keywords: Short-time Fourier transform, heart sounds, networks convolutional neurons.

1. Introducción

De acuerdo a la Organización para la Cooperación y Desarrollo Económicos (OCDE), en México los gastos económicos provocados por la epidemia de enfermedades cardiovasculares y de diabetes representan una amenaza para el futuro del país y para la viabilidad del sistema público de salud [20]. Además, según las estadísticas reportadas por el Instituto Nacional de Estadística y Geografía (INEGI), en México las enfermedades cardiovasculares son la primer causa de mortalidad en el país, las cuales provocaron más de 300 mil decesos en 2020 [12].

A nivel mundial la situación es similar ya que las enfermedades cardiovasculares siguen siendo la principal causa de morbilidad y mortalidad en todo el mundo, con un estimado de más 20 de millones de personas que fallecieron por enfermedades relacionadas con males cardiovasculares en 2020, lo que representa aproximadamente un tercio de todas las muertes a nivel mundial [21].

La auscultación es un método de diagnóstico primario de patologías cardiacas por medio del análisis de sonidos cardiacos. Esta técnica tiene su origen en el uso del tacto y la aplicación del oído en el tórax de los pacientes y no fue hasta la invención del estetoscopio por René Laënnec en 1819 cuando la auscultación como práctica clínica se estableció y diseminó [11]. Existen, además de la auscultación, otros métodos más sofisticados para la detección de patologías cardíacas, tales como la ecocardiografía y la resonancia magnética.

Sin embargo, la auscultación cardíaca persiste debido a ser una técnica de diagnóstico primario, es sencillla, no invasiva y de muy bajo costo. Mediante la auscultación se proporciona una perspectiva de apoyo al médico para conocer de manera inmediata el estado de un paciente y a decidir si es necesario realizar otra prueba más sofisticada. Cardiólogos experimentados pueden incluso distinguir con gran precisión varios tipos de patologías cardiacas y estimar su severidad utilizando como única herramienta un estetoscopio [17].

Sin embargo, el dominio del estetoscopio para lograr un oído clínico bien entrenado requiere de una larga curva de aprendizaje, provocando que el uso de esta herramienta pueda incluso considerarse como un arte en vías de desaparición. Desde finales del siglo XX, se comenzaron a popularizar los estetoscopios electrónicos (también llamados estetoscopios digitales).

Estos aparatos facilitan considerablemente la adquisición digital de los sonidos producidos por el corazón. En la literatura especializada, a la forma de onda de los sonidos producidos por el corazón se le conoce como fonocardiograma (FCG) o también audio cardiaco. Aunque tanto el FCG como el electrocardiograma (ECG) describen la actividad cardiaca, ambos trabajan en dominios diferentes [1].

La señal de ECG es producida por la actividad eléctrica del corazón, mientras que la señal de FCG es producida por la actividad mecánica del corazón. Además, debido a la diferente naturaleza del origen de estas señales, la existencia de un problema en una señal FCG no implica necesariamente la existencia del mismo u otros problemas en el ECG correspondiente. Por estas razones, la mayoría de los algoritmos basados en la señal de ECG no se pueden aplicar directamente al FCG [25].

1.1. Antecedentes

Generalmente, una señal de FCG normal consta de dos sonidos cardíacos fundamentales llamados primer sonido cardíaco (S1) y segundo sonido cardíaco (S2), que se generan debido al cierre de las válvulas auriculoventriculares y semilunares, respectivamente [1]. El FCG es una señal no estacionaria, es decir una señal cuyas propiedades y estadísticas cambian en función del tiempo.

El intervalo desde el punto inicial de S1 hasta el punto de partida de S2 se denomina intervalo de sístole, y el intervalo desde el punto inicial de S2 hasta el punto inicial de S1 se denomina intervalo de diástole. En la Fig. 1-a) se presenta una señal de FCG normal y la Fig. 1-b) ilustra una señal de FCG patológica correspondiente a un daño en la válvula mitral, es visible en la forma de onda el murmullo asociado a este evento.

El intervalo de diástole suele ser más largo que el intervalo de sístole. Los sonidos cardiacos S1 y S2 suelen tener una duración entre 40 y 150 ms y su contenido espectral se encuentra principalmente en el rango de 20 a 150 Hz [2]. En el caso de las patologías (chasquidos, fricciones y murmullos), su duración varía considerablemente dentro del ciclo cardiaco y su contenido espectral se encuentra en el rango de 25 a 700 Hz [5]. Durante la última década han sido propuestos un gran número de trabajos que buscan llevar a cabo la clasificación automática del FCG.

Una revisión exhaustiva de la literatura sobre el análisis y la clasificación de los sonidos cardiacos está fuera del alcance de este artículo, pero esta puede ser consultada en [4, 8]. Una tendencia predominante en el análisis de la señal de FCG es la realizar primeramente una etapa de segmentación. Es decir, localizar S1 y S2, antes de proceder a la clasificación de los sonidos [8].

Con respecto a las técnicas de clasificación, las primeras propuestas se basaron en análisis estadísticos y métodos clásicos de aprendizaje de máquinas tales como máquinas de soporte vectorial (SVM), k-vecino más cercano (k-NN), bosques aleatorios, modelos ocultos de Markov o perceptrón multicapa (MLP) [4, 8].

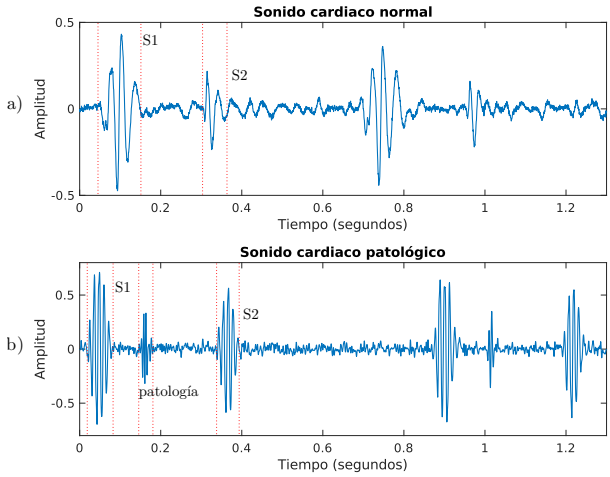

Fig. 1. Dos ejemplos de sonidos cardiacos, en a) se muestra una señal de FCG correspondiente a una persona sin patología, mientras que b) presenta una señal de FGC donde se puede apreciar una patología entre los sonidos S1 y S2.

Estos métodos suelen combinarse con el procesamiento digital de señales para la extracción de características basadas en el dominio del tiempo, en el dominio de frecuencia o en el dominio tiempo-frecuencia [6, 8, 10, 28].

Entre las técnicas de extracción de características las más comunes que han sido previamente utilizadas se encuentran: la transformada rápida de Fourier (FFT), coeficientes cepstrales en escala de frecuencia Mel (MFCC), la Transformada de Fourier de Tiempo Corto (TFTC), la transformada ondeleta discreta (DWT), la transformada ondeleta continua (CWT), la transformada ondeleta a Q-constante (TQWT), la transformada chirplet (CT), la transformada S, Transformada de Hilbert-Huang (HHT), entre otras [8]. En años recientes han aparecido las arquitecturas de aprendizaje profundo (DL), particularmente las de tipo red neuronal convolucional (CNN) y las redes neuronales recursivas (RNN) [4, 8].

Artículos recientes demuestran las ventajas de utilizar técnicas de análisis tiempo-frecuencia y métodos de DL para lograr la detección eficaz de patologías cardíacas únicamente utilizando la señal de FCG [6, 8, 10, 28]. En el presente trabajo se propone utilizar el análisis tiempo–frecuencia del FCG combinado con métodos de clasificación supervisada basados en DL.

Específicamente, se propone utilizar la Transformada de Fourier de Tiempo Corto (TFTC) para convertir la señal de FCG en una señal bidimensional la cual es clasificada por una CNN. Las CNNs que proponemos utilizar han sido específicamente diseñadas para clasificación de imágenes y tienen un alto desempeño: AlexNet, VGG16 [14, 26].

1.2. Bases de datos utilizadas

Para el desarrollo de este trabajo se propone utilizar dos bases de datos de sonidos cardiacos, ambas están disponibles al público. La primera base es la utilizada en “The PhysioNet/Computing in Cardiology (CinC) Challenge 2016” [7].

Tabla 1. Composición base de datos de sonidos cardiacos Physionet CinC [7].

Base de datos	**Pacientes**	**Grabaciones**	**Sonidos patológicos (%)**	**Sonidos sanos (%)**	**Sin clasificación (%)**
A	121	409	67.5	28.4	4.2
B	106	490	14.9	60.2	24.9
C	31	31	64.5	22.6	12.9
D	38	55	47.3	47.3	5.5
E	356	2054	7.1	86.7	6.2
F	112	114	27.2	68.4	4.4
Total	764	3153	18.1	73.0	8.8

Tabla 2. Composición de la base de datos de Yaseen et. al. [30].

Tipo	**Clase**	**Número de grabaciones por clase**
Normal	N	200
Patológico	AS	200
	MR	200
	MS	200
	MVP	200
Total		1000

Esta colección de sonidos proviene de siete grupos independientes de investigación. La base de datos contiene 3,153 grabaciones y un total de 233,512 sonidos cardiacos recolectados de pacientes sanos y pacientes con diferentes patologías cardiacas, como enfermedades de las válvulas del corazón y enfermedades de la arteria coronaria [7], entre otras.

La composición de la base de datos se presenta en la Tabla 1. La segunda base de datos utilizada es la propuesta por [30]. También está compuesta de sonidos patológicos y sonidos sanos. Los sonidos patológicos a su vez están divididos en cuatro categorías que corresponden a las siguientes enfermedades: estenosis aórtica (AS), estenosis mitral (MS), regurgitación mitral (MR) y prolapso de la válvula mitral (MVP). La composición de la base de datos se presenta en la Tabla 2.

Los sonidos de estas bases de datos han sido digitalizados usando diferentes frecuencias de muestreo y 16 bits de resolución. Para este trabajo, se remuestrearon la señales, se uniformizó la frecuencia de muestreo a $Fs = 2000$ Hz con 16 bits de resolución y después se usó un filtro pasa-banda Butterworth con frecuencias de corte de 25 Hz y 900 Hz. Los sonidos de la base de datos de PhysioNet que han sido marcados como no aptos para clasificación no fueron utilizados en este trabajo.

2. Metodología

En la Fig. 2 se describe a grandes rasgos el funcionamiento del método de clasificación propuesto. Utilizando únicamente la forma de onda de los sonidos cardiacos, primeramente se calcula el espectrograma por medio de la transformada de Fourier.

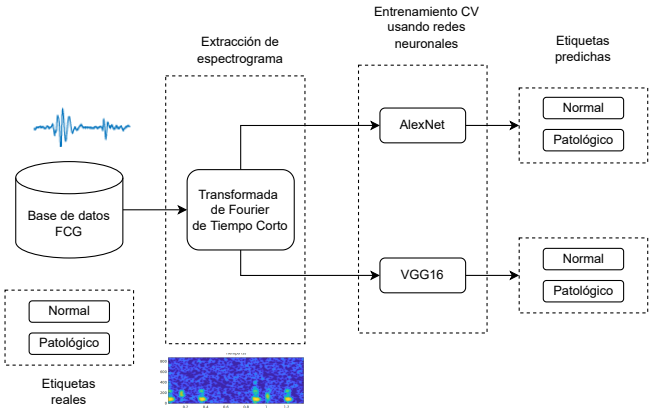

Fig. 2. Proceso de la clasificación de la señal de fonocardiograma usando la TFTC y redes neuronales convolucionales.

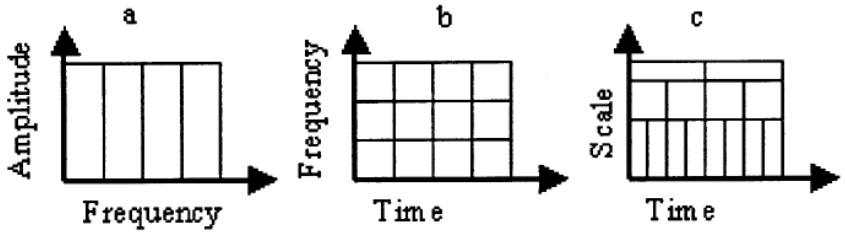

Fig. 3. Las representaciones usadas son (a) transformada de Fourier, (b) trasformada de Fourier de tiempo corto, y (c) transformada ondeleta. Imagen tomada de [19].

Después, el espectrograma es analizado por medio de una red neuronal convolucional que estima si la señal corresponde a un sonido cardiaco normal o a un sonido patológico.

2.1. Transformada de Fourier de tiempo corto

En procesamiento digital de señales es de gran utilidad considerar la dualidad tiempo–frecuencia que tienen las señales, ya que al representarlas en un dominio o en el otro se obtiene información distinta y complementaria a la vez. La herramienta más común para conocer el contenido frecuencial de una señal continua $x(t)$ en el dominio del tiempo es la Transformada de Fourier (TF), la cual puede ser definida como:

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t}\, dt, \tag{1}$$

donde $j = \sqrt{-1}$ y $X(\omega)$ puede considerarse como una medida de cuánto oscila $x(t)$ a la frecuencia angular ω [18].

Este operador matemático, sin embargo, tiene una gran limitante: el tiempo y la frecuencia son excluyentes entre sí. Es decir, al representar una señal en uno de los dominios se pierde la información sobre el otro [3]. De esta manera, al analizar mediante la TF señales de tipo no estacionarias o señales cuyas características en frecuencia varían en función del tiempo, como sucede con el FCG, no sería posible identificar los instantes de tiempo donde se presenten ciertos componentes determinados de la señal; como son S1, S2 y los sonidos patológicos.

Tradicionalmente la forma de solventar esta limitación ha sido mediante una representación en ambos dominios al mismo tiempo. Específicamente, en la transformación de una señal unidimensional en el dominio del tiempo a una bidimensional donde se incorporen tiempo y frecuencia [3]. A este tipo de transformación se les denomina representación tiempo–frecuencia (RTF) [3].

Existen una gran variedad de RTFs, entre los métodos lineales uno de los más utilizados por su sencillez y fácil implementación es la llamada Transformada de Fourier de Tiempo Corto (TFTC o STFT por sus siglas en inglés). En comparación con otras herramientas como la transformada ondeleta es que se puede tener una mayor facilidad en la interpretación con la TFTC, esto puede ser visto en la Figura 3.

La mayoría de los resultados obtenidos usando ondeletas pueden ser obtenidos de la misma manera usando la TFTC [13]. Cuando se usa la transformada ondeleta es común tener dificultades al extraer los componentes fundamentales o cualquier otro componente armónico de la señal. Aunque la TFTC tiene una resolución fija en todas las frecuencias, una vez el tamaño de la ventana está dado, permite una interpretación más fácil en términos de las armónicas [13]. La TFTC se calcula a partir de la traslación (τ) y modulación (ω) de una ventana de análisis $w(t)$ [18]:

$$X(t,\omega) = \int_{-\infty}^{+\infty} x(t)w(t-\tau)e^{-j\omega t}\,dt, \tag{2}$$

donde $w(t) \in \mathbb{R}$, es simétrica y además $\|w_{t,\omega}\| = 1$. Al calcular la TFTC se obtiene un resultado complejo, i.e., $X(t,\omega) \in \mathbb{C}$. En nuestro caso nos interesa particularmente la distribución de la energía en función de la frecuencia y del tiempo, a este parámetro se le conoce como el espectrograma y está definido como $S(\tau,\omega) = |X(\tau,\omega)|^2$ [3, 18]. En la práctica, la señal de FCG, $x(n)$, que utilizamos es de tiempo discreto y puede ser vista como la versión muestreada de $x(t)$ cada nT instantes de tiempo, donde $n \in \mathbb{Z}$ y T es el periodo de muestreo. La versión discreta de la TFTC [27] puede escribirse como:

$$X(m,\omega) = \sum_{n=-\infty}^{+\infty} x(n)w(n-mR)e^{-j\omega n}, \tag{3}$$

donde $w(n)$ es la versión discreta de la ventana de análisis mencionada anteriormente y tiene una longitud de L_w muestras, m es el índice de los intervalos de tiempo correspondientes a cada segmento de la ventana de análisis, R es el número de muestras de traslape entre el instante m y el instante $m+1$. La Fig. 4 ilustra la manera en que se calcula la TFTC por medio de la Transformada Rápida de Fourier (FFT), suponiendo que se calcula una transformada de longitud par de N puntos. El espectrograma se obtiene calculando la expresión $S(m,\omega) = |X(m,\omega)|^2$.

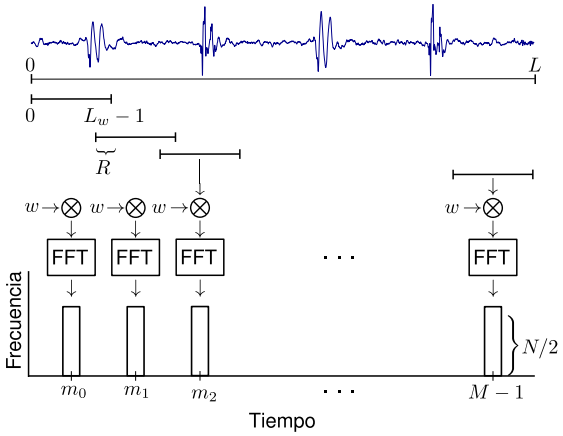

Fig. 4. Representación gráfica de cómo se calcula la TFTC discreta a partir de la FFT. El resultado del cálculo es una matriz compleja de dimensiones $\frac{N}{2} \times M$.

Los parámetros de implementación de la TFTC en este trabajo son los siguientes: se procesaron bloques de FCG de $L = 2800$ muestras, usando una ventana de análisis $w(n)$ de tipo Hamming de longitud $L_w = 100$ muestras, un traslape de $R = 88$ muestras y la longitud de la FFT de $N = 512$ muestras. Finalmente, de la matriz de TFTC se calcula el valor absoluto y se eleva al cuadrado para obtener el espectrograma $S(\tau, \omega)$, el cual tiene dimensiones 257×224.

Dado que las arquitecturas de CNN que se propone utilizar en este trabajo requieren una imagen de entrada de dimensiones distintas a la de $S(\tau, \omega)$, el espectrograma fue recortado para ajustarlo al tamaño 224×224. En este caso se eliminan las filas 225 a la 257, es decir, el análisis posterior únicamente considera frecuencias de hasta ≈871 Hertz. La Fig. 5-a) y b) ilustran el tipo de imagen utilizada para detectar la presencia o ausencia de patologías cardiacas.

La ventaja que presenta la metodología descrita es que no se requiere conocer de antemano la segmentación de la señal cardiaca. Es decir, a diferencia de otras propuestas que existen en la literatura [8], este método no requiere conocimiento de la posición exacta de los sonidos S1, S2 o los silencios de sístole y diástole, para llevar a cabo la detección.

3. Clasificación de las imágenes de sonidos cardiacos mediante aprendizaje profundo

El aprendizaje de máquinas clásico busca que las computadoras puedan actuar con poca intervención humana. Por otro lado, el aprendizaje profundo se trata de que las computadoras aprendan a actuar usando estructuras inspiradas en el cerebro humano y han demostrado que requieren todavía menos intervención humana. Gracias al poder computacional de los dispositivos de cómputo modernos es posible implementar las redes neuronales convolucionales (CNN) para tareas de clasificación relacionadas con la salud.

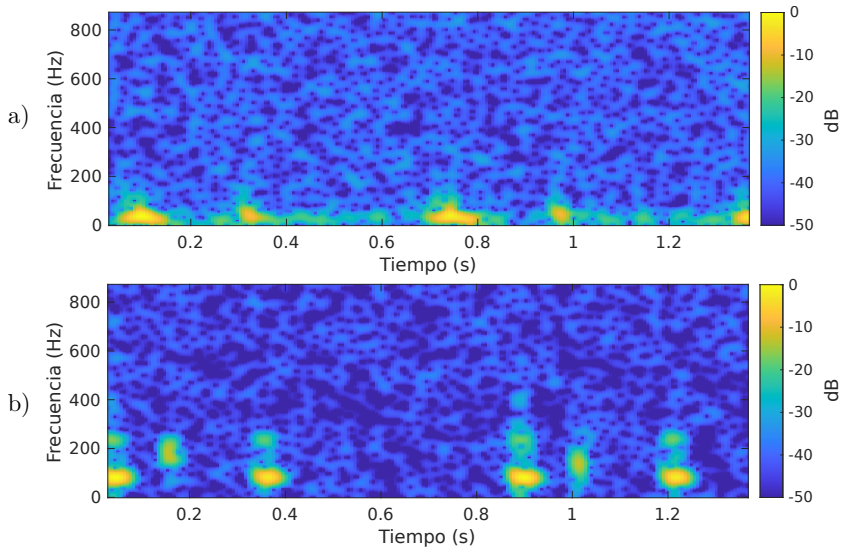

Fig. 5. Ejemplo de espectrograma obtenido de los sonidos cardíacos presentados en la Fig. 1, a) para un sonido cardiaco normal y b) para un sonido cardiaco patológico.

En el caso de este trabajo, la finalidad es desarrollar un clasificador de sonidos cardíacos que tenga un alto nivel de confiabilidad en la detección de patologías.

3.1. Aprendizaje profundo

El aprendizaje profundo es un método de representación-aprendizaje con múltiples niveles de representaciones, obtenidas con la descomposición de módulos simples, pero no lineales que transforman la representación de un nivel bajo a un nivel alto y más abstracto. En esencia son redes neuronales compuestas de tres capas o más usadas para tareas de clasificación, las representaciones de las capas de alto nivel amplifican aspectos de la entrada que son importantes para la discriminación y supresión de variables irrelevantes [15].

Para comprender el funcionamiento de las redes neuronales, estas pueden verse como una versión simplificada del cerebro humano. Las neuronas están organizadas en capas, cada neurona recolecta información de la capa anterior, realiza un cálculo simple y comunica el resultado a la siguiente capa. En las redes más eficientes se pueden tener docenas de capas, por lo que el modelo puede ser llamado de aprendizaje profundo [22]. Una introducción detallada a los conceptos y aplicación del aprendizaje profundo y las CNNs está disponible en [9].

El elemento más importante de las CNN son las capas convolucionales. Se puede entender que una convolución es la aplicación de un filtro, también llamado kernel, a una entrada que da como resultado una activación. La aplicación sistemática del mismo filtro en la entrada resulta en un mapa de activaciones o también llamado mapa de características, que indican la localización y magnitud de una característica detectada en la entrada de los datos.

Las CNN tienen la habilidad de aprender una gran cantidad de filtros que en conjunto pueden llegar a resolver problemas complejos [24]. Esto deriva en su utilidad en tareas de clasificación de imágenes con una capacidad que incluso puede mejorar la pericia de los seres humanos.

Por lo que haciendo uso de las representaciones obtenidas a partir de los sonidos cardíacos, se detectará la presencia de anomalías cardíacas. Las CNNs que fueron seleccionadas para el presente trabajo son AlexNet y VGG16 [14, 26].

Estas redes han sido específicamente diseñadas para clasificación de imágenes y presentan un alto desempeño. El modelo de AlexNet está compuesto por una arquitectura de ocho capas, las primeras cinco son capas convolucionales y las últimas tres son capas densas, fue la primer red en cambiar la función de activación de sigmoide por una de tipo ReLu.

Originalmente fue entrenada con 1.2 millones de imágenes, para evitar el sobre-entrenamiento se implementaron técnicas de Data Augmentation y capas de Dropout. Esta red participó y ganó el reto de ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC), rompiendo así el paradigma de que las características usadas para los filtros tenían que ser hechas a mano.

El modelo de CNN llamado VGG16 también es una arquitectura diseñadas para clasificación de imágenes, la cual ganó la competencia ILSVRC en 2014. Esta red está compuesta por trece capas convolucionales y tres capas densas. La abreviación VGG significa Visual Geometry Group, en referencia al grupo de investigadores que la propuso. A su vez, el número 16 indica el número de capas que la compone [29, 26].

Los modelos de CNN seleccionados fueron diseñados para clasificar cientos de diferentes tipos de imágenes y utilizan matrices de entrada de tres canales para tomar en cuenta los colores rojo, verde y azul. Como se explicó en la sección anterior, las imágenes creadas a partir de los espectrogramas de los sonidos cardiacos tienen solamente un canal.

Por esta razón fue necesario modificar ligeramente las redes en lo que respecta a las capas de entrada y la de salida. Específicamente, las modificaciones que se realizaron fueron la adaptación de la capa de entrada para recibir imágenes de tamaño $224 \times 224 \times 1$ y en la capa de salida se modificó a dos posibles resultados, la predicción de que se trate de una imagen que corresponde a un sonido patológico o a un sonido normal.

Para el entrenamiento de la red se utilizaron un total de 10,200 imágenes, estas fueron balanceadas con una relación 1:1 entre imágenes provenientes de sonidos normales y patológicos. Los hiperparámetros usados fueron: 150 épocas para el entrenamiento, así como el optimizador Stochastic Gradient Descent (SGD) con una tasa de aprendizaje de $0{,}008$, la función de pérdida categorical_crossentropy y un tamaño de lote de 64 muestras.

Estas actividades se realizaron bajo el ambiente virtual en la plataforma de Google Colaboratory en su versión de pago. Las características del sistema utilizado para realizar las pruebas son las siguientes: procesador Intel(R) Xeon(R) @ 2.3GHz, memoria RAM de 26 GB y una tarjeta gráfica Tesla P100 de 16 GB. Los resultados del entrenamiento se presentan en la Fig. 6.

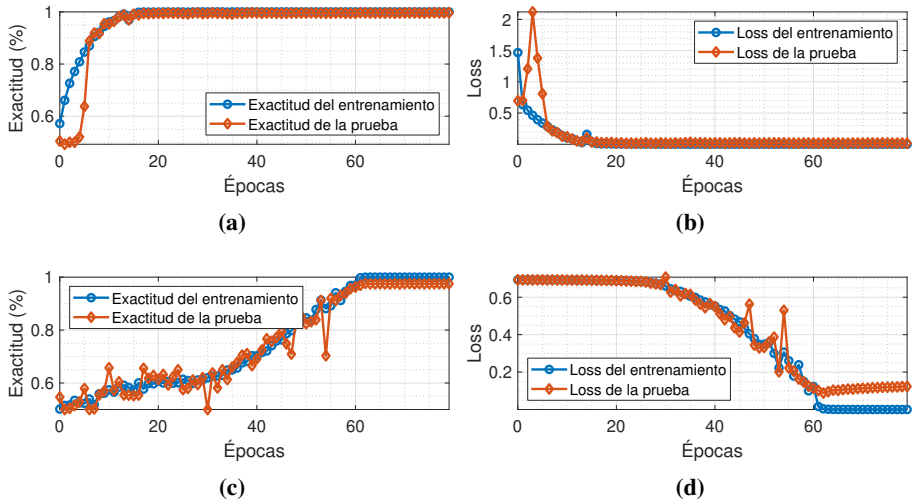

Fig. 6. Exactitud de AlexNet (a) y VGG16 (c); y Loss de AlexNet (b) y VGG16 (d) durante el entrenamiento. Sólo se muestran las primeras 80 épocas para mejorar la visualización de los resultados.

4. Resultados

En esta sección evaluamos el rendimiento de la metodología propuesta con respecto a su capacidad para distinguir con precisión entre sonidos cardiacos normales y patológicos. Se hará uso de la validación cruzada, que es un método estadístico para la evaluación de del rendimiento de un modelo de aprendizaje automático. Este consite en dividir los datos en dos partes, la primera se usará para entrenar al modelo y la segunda se usará para la evaluación del modelo.

Entre los diferentes tipos de validación cruzada, se eligió la validación cruzada de K iteraciones, el valor de K fue establecido a 10, por lo que se entiende que cada iteración era entrenada usando el 90 % de los datos y se usaban 10 % de los datos para la validación, este proceso de entrenamiento/evaluación fue repetido 10 veces. Entre las ventajas de este tipo de validación cruzada está una estimación precisa del rendimiento del modelo, así como evitar el sobre-entrenamiento [23].

La obtención de la exactitud, precisión, sensibilidad, y especifidad serán usados para la evaluación del rendimiento de los modelos usados. La exactitud (Accuracy) hace referencia a lo cerca que está el resultado de una medición del valor verdadero. La sensitividad (Sensitivity) mide la proporción de sonidos patológicos que fueron correctamente identificadas como positivos.

La especificidad (Specificity) mide la proporción de sonidos normales que fueron correctamente identificados como negativos. La precisión (Precision) hace referencia a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Estos valores están detallados en la Tabla 3. Se puede observar que los mejores resultados fueron obtenidos por el modelo de AlexNet con una exactitud de 84.2 %.

Tabla 3. Métricas de la matriz de confusión de las imágenes espectrograma.

	Exactitud	**Precisión**	**Sensibilidad**	**Especificidad**
AlexNet	0.828	0.842	0.807	0.848
VGG16	0.746	0.755	0.729	0.763

Tabla 4. Matrices de confusión de los modelos AlexNet y VGG16 en la clasificación de las imágenes espectrograma.

	Verdadero negativo	**Falso positivo**	**Falso negativo**	**Verdadero positivo**
AlexNet	865	155	197	823
VGG16	778	242	276	744

Para poder comprender de mejor manera los resultados obtenidos, se hará uso de la matriz de confusión. Esta matriz permite analizar el desempeño de un algoritmo de clasificación, decribiendo cómo se distribuyen los valores reales así como las predicciones. Una predicción correcta positiva es un verdadero positivo, una predicción incorrecta positiva es un falso positivo, una predicción correcta negativa es un verdadero negativo y una predicción incorrecta negativa es un falso negativo.

Estos valores están detallados en la Tabla 4. El modelo de AlexNet cuenta con una mayor cantidad de verdaderos negativos en comparación del modelo VGG16 con un 11.18 % más de casos verdaderos negativos. Los falsos positivos del modelo VGG16 son mayores en un 56 % que AlexNet. Los falsos negativos del modelo VGG16 son mayores también en un 40 % que Alexnet.

Finalmente, los casos verdaderos positivos de AlexNet son mayores en un 10 % que el modelo VGG16. Esto nos da un mejor entendimiento de los resultados obtenidos de la Tabla 3, en la que el modelo AlexNet tiene mejor rendimiento al clasificar las patologías cardiacas, al mismo tiempo que tiene menos errores que el modelo VGG16.

Los resultados obtenidos por el algoritmo propuesto se consideran satisfactorios ya que concuerdan con los reportados por otros algoritmos en la página oficial del challenge de PhysioNet realizado en 2016 [7]. Sin embargo, actualmente hay otros algoritmos del estado del arte que reportan valores de exactitud mayores [4, 8].

Es necesario explorar otros métodos de analisis tiempo-frecuencia tales como la Transformada Ondeleta [16, 25] o métodos cuadráticos como la distribución de Wigner-Ville [3, 18, 28]. Con respecto a las CNNs evaluadas, AlexNet exhibe un mejor desempeño que VGG16.

5. Conclusiones y trabajo futuro

En este trabajo se ha propuesto la comparación de dos modelos de redes neuronales convolucionales para la detección de patologías cardíacas usando el espectrograma obtenido a partir de la transformada de Fourier de tiempo corto aplicado a la señal de fonocardiograma. Los resultados obtenidos por el modelo de AlexNet son de una exactitud de 82.8 %, y el modelo VGG16 obtuvo una exactitud de 74.6 %, mostrando que el modelo de AlexNet es más apropiado para llevar a cabo esta clasificación.

La utilidad de este trabajo es resaltar la importancia de una correcta elección de la red neuronal convolucional. Otra ventaja de AlexNet es su complejidad computacional durante la etapa de entrenamiento, ya que es considerablemente más rápida.

En el trabajo futuro sería interesante explorar el uso de otras representaciones tiempo-frecuencia, así como otros modelos de redes neuronales convolucionales adecuados a la clasificación de imágenes médicas.

Los resultados obtenidos por el algoritmo propuesto se consideran buenos ya que concuerdan con los reportados por otros algoritmos sometidos al PhysioNet Challenge CinC 2016 [7]. Sin embargo, actualmente existen algoritmos que reportan mejor desempeño [4, 8].

Referencias

1. Abbas, A. K., Bassam, R.: Phonocardiography signal processing. Synthesis Lectures on Biomedical Engineering, vol. 4, no. 1, pp. 1–194 (2009) doi: 10.1007/978-3-031-01637-0
2. Arnott, P. J., Pfeiffer, G. W., Tavel, M. E.: Spectral analysis of heart sounds: relationships between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives. Journal of biomedical engineering, vol. 6, no. 2, pp. 121–128 (1984) doi: 10.1016/0141-5425(84)90054-2
3. Boashash, B.: Time-frequency signal analysis and processing: a comprehensive reference. Academic press (2015)
4. Chen, W., Sun, Q., Chen, X., Xie, G., Wu, H., Xu, C.: Deep learning methods for heart sounds classification: A systematic review. Entropy, vol. 23, no. 6, pp. 667 (2021) doi: 10.3390/e23060667
5. Choi, S., Jiang, Z.: Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique. Computers in biology and medicine, vol. 40, no. 1, pp. 8–20 (2010) doi: 10.1016/j.compbiomed.2009.10.003
6. Chowdhury, T. H., Poudel, K. N., Hu, Y.: Time-frequency analysis, denoising, compression, segmentation, and classification of pcg signals. IEEE Access, vol. 8, pp. 160882–160890 (2020) doi: 10.1109/ACCESS.2020.3020806
7. Clifford, G. D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., Mark, R. G.: Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. Computing in Cardiology, vol. 43, pp. 609–612 (2016) doi: 10.22489/cinc.2016.179-154
8. Dwivedi, A. K., Imtiaz, S. A., Rodriguez-Villegas, E.: Algorithms for automatic analysis and classification of heart sounds–a systematic review. IEEE Access, vol. 7, pp. 8316–8345 (2018) doi: 10.1109/ACCESS.2018.2889437
9. Géron, A.: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media Inc. (2019)
10. Ghosh, S. K., Ponnalagu, R. N., Tripathy, R. K., Acharya, U. R.: Automated detection of heart valve diseases using chirplet transform and multiclass composite classifier with PCG signals. Computers in biology and medicine, vol. 118, pp. 103632 (2020) doi: 10.1016/j.compbiomed.2020.103632

11. Hanna, I. R., Silverman, M. E.: A history of cardiac auscultation and some of its contributors. The American journal of cardiology, vol. 90, no. 3, pp. 259–267 (2002) doi: 10.1016/S0002-9149(02)02465-7
12. INEGI: Características de las defunciones registradas en México durante 2020 (2021)
13. Jurado, F., Saenz, J. R.: Comparison between discrete STFT and wavelets for the analysis of power quality events. Electric Power Systems Research, vol. 62, no. 3, pp. 183–190 (2002) doi: 10.1016/S0378-7796(02)00035-4
14. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM, vol. 60, no. 6, pp. 84–90 (2017) doi: 10.1145/3065386
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature, vol. 521, no. 7553, pp. 436–444 (2015) doi: 10.1038/nature14539
16. Lu, J., Jiang, Q., Li, L.: Analysis of adaptive synchrosqueezing transform with a time-varying parameter. Advances in Computational Mathematics, vol. 46, no. 5 (2020) doi: 10.1007/s10444-020-09814-x
17. Mahnke, C. B.: Automated heartsound analysis/computer-aided auscultation: a cardiologist's perspective and suggestions for future development. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3115–3118. IEEE (2009) doi: 10.1109/IEMBS.2009.5332551
18. Mallat, S.: A wavelet tour of signal processing: The sparse way (2009)
19. Messer, S. R., Agzarian, J., Abbott, D.: Optimal wavelet denoising for phonocardiograms. Microelectronics Journal, vol. 32, no. 12, pp. 931–941 (2001) doi: 10.1016/S0026-2692(01)00095-7
20. OECD: Obesity update (2017)
21. Organization, W. H.: World health statistics 2022: Monitoring health for the SDGs, sustainable development goals. Tech. rep., World Health Organization (2022)
22. Peyré, G.: Mathematics of neural networks. École Normale Supérieure PSL (2020)
23. Refaeilzadeh, P., Tang, L., Liu, H., Angeles, L., Scientist, C. D.: Cross-validation. Springer New York (2016)
24. Rosebrock, A.: Deep learning for computer vision with python - Starter. Pyimageserach (2017)
25. Safara, F., Doraisamy, S., Azman, A., Jantan, A., Ramaiah, A. R. A.: Multi-level basis selection of wavelet packet decomposition tree for heart sound classification. Computers in biology and medicine, vol. 43, no. 10, pp. 1407–1414 (2013) doi: 10.1016/j.compbiomed.2013.06.016
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, vol. 6, pp. 1–14 (2014) doi: 10.48550/arXiv.1409.1556
27. Smith, J. O.: Spectral audio signal processing. W3K (2011)
28. Vyas, S., Patil, M. D., Birajdar, G. K.: Classification of heart sound signals using time-frequency image texture features. Computational Intelligence and Healthcare Informatics, pp. 81–101 (2021) doi: 10.1002/9781119818717.ch5
29. Wei, J.: AlexNet: The architecture that challenged CNNs. Towards Data Science, (2019)

30. Yaseen, Son, G. Y., Kwon, S.: Classification of heart sound signal using multiple features. Applied Sciences (Switzerland), vol. 8, no. 12, pp. 2344 (2018) doi: 10.3390/app8122344