Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias
en Electrónica y Telecomunicaciones
con orientación en Telecomunicaciones

Clasificación de la señal de audio cardíaco mediante análisis tiempo-frecuencia y aprendizaje de máquinas

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestro en Ciencias

Presenta:

Leonel Orozco Reyes

Ensenada, Baja California, México

2022

Tesis defendida por

Leonel Orozco Reyes

y aprobada por el siguiente Comité

Dr. Miguel Ángel Alonso Arévalo

Director de tesis

Dr. Roberto Conte Galván

Dr. Roilhi Frajo Ibarra Hernández

Dr. Israel Marck Martínez Pérez

Dra. María del Carmen Maya Sánchez

Coordinadora del Posgrado en Electrónica y Telecomunicaciones

---

Dr. Pedro Negrete Regagnon

Director de Estudios de Posgrado

Resumen de la tesis que presenta Leonel Orozco Reyes como requisito parcial para la obtención del grado de Maestro en Ciencias en Electrónica y Telecomunicaciones con orientación en Telecomunicaciones.

Clasificación de la señal de audio cardíaco mediante análisis tiempo-frecuencia y aprendizaje de máquinas

Resumen aprobado por:

Dr. Miguel Ángel Alonso Arévalo
Director de tesis

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 principal la aplicación del análisis de técnicas de aprendizaje de máquinas para la clasificación de audio cardíaco. Haciendo uso de tres representaciones tiempo-frecuencia siendo estas el espectrograma, el espectrograma en escala de Mel y la transformada ondeleta *Synchrosqueezing*, 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 tres arquitecturas de redes neuronales que han demostrado tener un desempeño notable en tareas de clasificación. Los resultados obtenidos indican que la correcta variación de la combinación de las tres representaciones tiempo-frecuencia, así como la correcta elección del clasificador influye en el rendimiento y el tiempo necesario para la clasificación.

Palabras clave: Auscultación, patologías cardíacas, transformada ondeleta *Synchrosqueezing*, redes neuronales convolucionales

Abstract of the thesis presented by Leonel Orozco Reyes as a partial requirement to obtain the Master of Science degree in Electronics and Telecommunications with orientation in Telecommunications.

Classification of heart sounds using time-frequency analysis and machine learning

Abstract approved by:

Dr. Miguel Ángel Alonso Arévalo
Thesis Director

Auscultation is a non-invasive, low-cost, easy-to-implement diagnostic tool that provides valuable information for diagnosing cardiac pathologies. With the help of a digital stethoscope, a heart sound or phonocardiogram signal is obtained, which is the main element of study in this dissertation. Misdiagnosis due to a lack of experienced clinicians and the human auditory system's limitations have led to modern computer-aided analysis of heart sounds. The present work has the main objective of applying audio signal processing and machine learning techniques to classify heart sounds as normal or abnormal. The proposed methodology uses three time-frequency representations, the spectrogram, the mel scale spectrogram, and the wavelet synchrosqueezed transform. We generate a feature matrix that best represents the heart sound based on these representations. The obtained matrices are classified using convolutional neural networks; three neural networks architectures are used AlexNet, VGG16, and Ullah. These architectures have shown a remarkable performance in classification tasks. The obtained results indicate that the proper combination of the three time-frequency representations and the appropriate choice of the classifier have a significant repercussion on the performance and the time required for classification.

Keywords: Auscultation, cardiac pathology, wavelet synchrosqueezed transform, convolutional neural network

Dedicatoria

A mis padres y a mis hermanos; a mis profesores y a mis amigos.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada por brindarme la oportunidad de realizar mis estudios de posgrado.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría (CVU: 994947).

A mi padre y madre, Leonel y Lucerito, que siempre me han apoyado y amado de manera incondicional, y me han dado un espacio seguro y cómodo para poder trabajar en esta tesis. Ustedes son la razón de todo lo que he logrado en la vida.

A mis hermanos, Marcos y Diana, con los que he pasado momentos increíbles y siempre me sacan una sonrisa, además de la ocasional plática amena, gracias por su tiempo.

A mi perrita Canelita por acompañarme en varias noches mientras trabajaba y por ser demasiado adorable.

A mi director de tesis, el Dr. Miguel Alonso Arévalo, por su constante y paciente orientación, por las pláticas amenas que tenemos y por darme la confianza de trabajar con él.

A los miembros de mi comité de tesis, el Dr. Roberto Conte Galván, el Dr. Roilhi Frajo Ibarra Hernández y el Dr. Israel Marck Martínez Pérez, por sus comentarios, orientación y apoyo.

A los profesores del CICESE, por todos los conocimientos transmitidos durante estos dos años, así como por su disponibilidad y apoyo.

A mis compañeros y amigos, por apoyarme durante el transcurso de la maestría, por los buenos momentos y las risas.

Tabla de contenido

Página

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Lista de figuras (continuación)

Figura Página

Lista de tablas

Capítulo 1. Introducción

Actualmente en diversos países se tiene un severo problema de enfermedades cardiovasculares y de diabetes que presenta un reto a la salud de los ciudadanos, así como a los sistemas de salud que deben lidiar con estos problemas. De acuerdo con la Organización Mundial de la Salud (OMS), en el 2016, 15.2 millones de defunciones fueron registradas en todo el mundo debido a enfermedades relacionadas con el corazón, así como, accidentes cerebrovasculares. Éstas han sido las principales causas de mortalidad durante los últimos 15 años (Organización Mundial de la Salud, 2018).

En México, las estadísticas no son más favorables. De acuerdo con los datos de la Dirección General de Información de Salud (DGIS) (Dirección General de Información de Salud, 2020), cerca del 20 % del total de defunciones en 2018 se debió a enfermedades del corazón. La ecografía cardíaca y la resonancia magnética han desplazado a la auscultación en las economías más ricas, sin embargo la auscultación cardíaca sigue proporcionando un diagnóstico sólido al médico ambulatorio. La auscultación es una técnica sencilla, no invasiva y de bajo costo que ayuda al médico a conocer de manera inmediata el estado del paciente y a decidir si otras pruebas más especializadas son necesarias.

Cardiólogos experimentados, pueden distinguir con gran precisión varios tipos de patologías cardíacas y estimar su severidad utilizando como única herramienta un estetoscopio. La mala interpretación de un sonido cardíaco al momento de realizar la auscultación no es sorprendente, pues el sistema auditivo humano solo es capaz de detectar una fracción de la energía acústica generada por el corazón (Mahnke, 2009).

En México, el problema es que el número de especialistas con oído clínico entrenado para realizar este tipo de diagnóstico es reducido, lo cual vuelve difícil el acceso a un diagnóstico experto. La literatura científica reciente ha demostrado que el análisis de la señal de audio cardíaco, también conocida como fonocardiograma (FCG), es una técnica con un enorme potencial para diagnosticar enfermedades cardíacas de manera eficaz y de muy bajo costo (Mahnke, 2009).

Como consecuencia se han desarrollado diversas técnicas de detección automática, y se han propuesto pasos para llevar a cabo una buena clasificación, incluyendo, pero no limitados a: la segmentación de sonidos cardíacos, la extracción de las ca-

racterísticas más relevantes y la clasificación de los sonidos cardíacos. Esto es con el fin de poder dar apoyo para el problema de la falta de especialistas con oído clínico (Dwivedi *et al.*, 2019).

La importancia de una buena detección de patologías cardíacas deriva del impacto que tiene para reducir los costes financieros y el uso más efectivo de médicos especialistas, ya que debido a la propia naturaleza de bajo costo y no invasiva de la auscultación, provee la oportunidad de realizar un análisis a un gran número de personas, resultando en una disminución del tiempo que toma un diagnóstico, así como la reducción de los tiempos de traslado de los médicos (Bozkurt *et al.*, 2018).

En recientes trabajos, se ha demostrado la eficacia de la combinación de técnicas de análisis tiempo-frecuencia con modelos aprendizaje automático para lograr la detección de patologías cardíacas (Ghosh *et al.*, 2019; Zhang *et al.*, 2017; Dwivedi *et al.*, 2019), por lo que, en el presente trabajo, se extenderá el análisis de la clasificación haciendo uso de técnicas tiempo-frecuencia con aprendizaje profundo.

Los sonidos cardíacos utilizados en este trabajo de investigación para la clasificación se obtuvieron de la base de datos de *The PhysioNet/Computing in Cardiology Challenge 2016* (Clifford *et al.*, 2016) y de *Classification of heart sound signal using multiple features* (Yaseen *et al.*, 2018).

1.1. Objetivos

1.1.1. Objetivo general

- Desarrollar un método de detección de patologías en la señal de audio cardíaco, mediante técnicas de aprendizaje de máquinas y representaciones tiempo-frecuencia de la señal de fonocardiograma (FCG).

1.1.2. Objetivos específicos

- Revisar los diferentes métodos de clasificación de imágenes y seleccionar las mejores arquitecturas para la clasificación.

- Calcular las representaciones tiempo-frecuencia del fonocardiograma (FCG) a través de la implementación del espectrograma, espectrograma en escala Mel y la transformada ondeleta *Synchrosqueezed* (Wavelet Synchrosqueezed Transform, WSST).
- Clasificar las representaciones tiempo-frecuencia obtenidas usando redes neuronales convolucionales.
- Evaluar y comparar la clasificación realizada por diferentes arquitecturas de redes neuronales convolucionales.

1.2. Justificación

Las enfermedades cardíacas han ido en aumento a lo largo de estos años, volviéndose una de las principales causas de morbilidad en México, esto es debido a diversos factores sociales, económicos y culturales. También se tiene un problema de falta de médicos especialistas. Además de limitaciones fisiológicas en el sistema auditivo humano al reconocimiento de gran parte de la energía de los sonidos cardíacos, que pueden llegar a ser determinantes al momento de llevar a cabo una evaluación médica.

La necesidad de la creación de diversas técnicas para el cuidado de la salud ha cobrado importancia con el avance de la tecnología en el área médica. Diversas soluciones han sido propuestas, entre ellas está la auscultación, que debido a la propia naturaleza no invasiva es muy factible el análisis de sonidos cardíacos para la obtención de características que puedan ser discriminatorias entre un sonido normal y el sonido de una patología cardíaca.

Gracias al poder computacional de los dispositivos de cómputo modernos es posible implementar las redes neuronales para tareas de clasificación relacionadas con la salud, en este caso para desarrollar un clasificador de señales cardíacas que tenga un alto nivel de confiabilidad en la detección de patologías cardíacas.

Se hará uso de técnicas de representación tiempo-frecuencia para calcular la representación espectral del fonocardiograma, obteniendo así, características represen-

tativas que serán clasificadas con diversas arquitecturas de redes neuronales convolucionales.

1.3. Metodología

Para poder cumplir con los objetivos establecidos anteriormente, el presente trabajo ha sido organizado en diferentes etapas. Los siguientes puntos ofrecen una recapitulación de los más relevantes:

- Se revisó una gran cantidad de literatura relacionada a señales de FCG.
- Se realizó una investigación sobre las técnicas de representación tiempo-frecuencia.
- Se realizó una investigación sobre las técnicas de clasificación automática.
- Se obtuvieron las bases de datos que contienen los sonidos cardíacos que serán clasificados.
- Se realizó un preprocesamiento de los sonidos cardíacos.
- Se obtuvieron las características tiempo-frecuencia de los sonidos cardíacos.
- Se crearon las representaciones tiempo-frecuencia de las señales de FCG.
- Se entrenaron y obtuvieron los resultados de los clasificadores.
- Se analizaron y compararon los resultados obtenidos.
- Se evaluaron y validaron los rendimientos obtenido de los diferentes clasificadores.

1.4. Estructura de la tesis

El desarrollo de la tesis inicia en el Capítulo 2, con la explicación fisiológica del funcionamiento del corazón, así como el origen de los sonidos cardíacos. Se explicará la composición del ciclo cardíaco y las características que serán relevantes para realizar la clasificación de los sonidos cardíacos. Se hará un breve repaso a los trabajos que se

han realizado con anterioridad en CICESE, así como un repaso al estado del arte de la clasificación automática de sonidos cardíacos.

En el Capítulo 3, se detallarán las técnicas de análisis tiempo-frecuencia que se utilizarán a lo largo de la tesis, para obtener representaciones de los sonidos cardíacos que serán utilizados para su clasificación. Se empezará explicando las bases de datos con las que se trabajará, detallando la cantidad de sonidos (patológicos y sanos) que las componen.

Se hará una explicación detallada de las tres técnicas de análisis tiempo-frecuencia que se usarán, siendo estas el espectrograma, el espectrograma en escala Mel y la transformada ondeleta *Synchrosqueezing*. Cada técnica de análisis será acompañada de la visualización de las imágenes creadas para facilitar el entendimiento del proceso de representación. Se incluirán las imágenes adicionales que fueron creadas como una combinación de las tres representaciones tiempo-frecuencias previamente explicadas.

En el Capítulo 4, se explicarán de manera detallada la composición y estructura de las redes neuronales convolucionales, así como, los detalles técnicos de los modelos de redes neuronales que se usarán para la clasificación de las imágenes obtenidas en el Capítulo 3. Cada red neuronal usada, mostrará su rendimiento con las diferentes representaciones tiempo-frecuencia obtenidas.

En el Capítulo 5, se expondrán los resultados obtenidos de cada modelo de red neuronal y cada representación tiempo-frecuencia de manera particular, los resultados de la clasificación serán presentados usando un diagrama de confusión para una mejor visualización de los mismos.

En el Capítulo 6, se expondrán las conclusiones obtenidas del trabajo realizado, así como ideas para el desarrollo de trabajos futuros en la clasificación de sonidos cardíacos usando redes neuronales.

La estructura del presente trabajo se puede visualizar en la Figura 1.

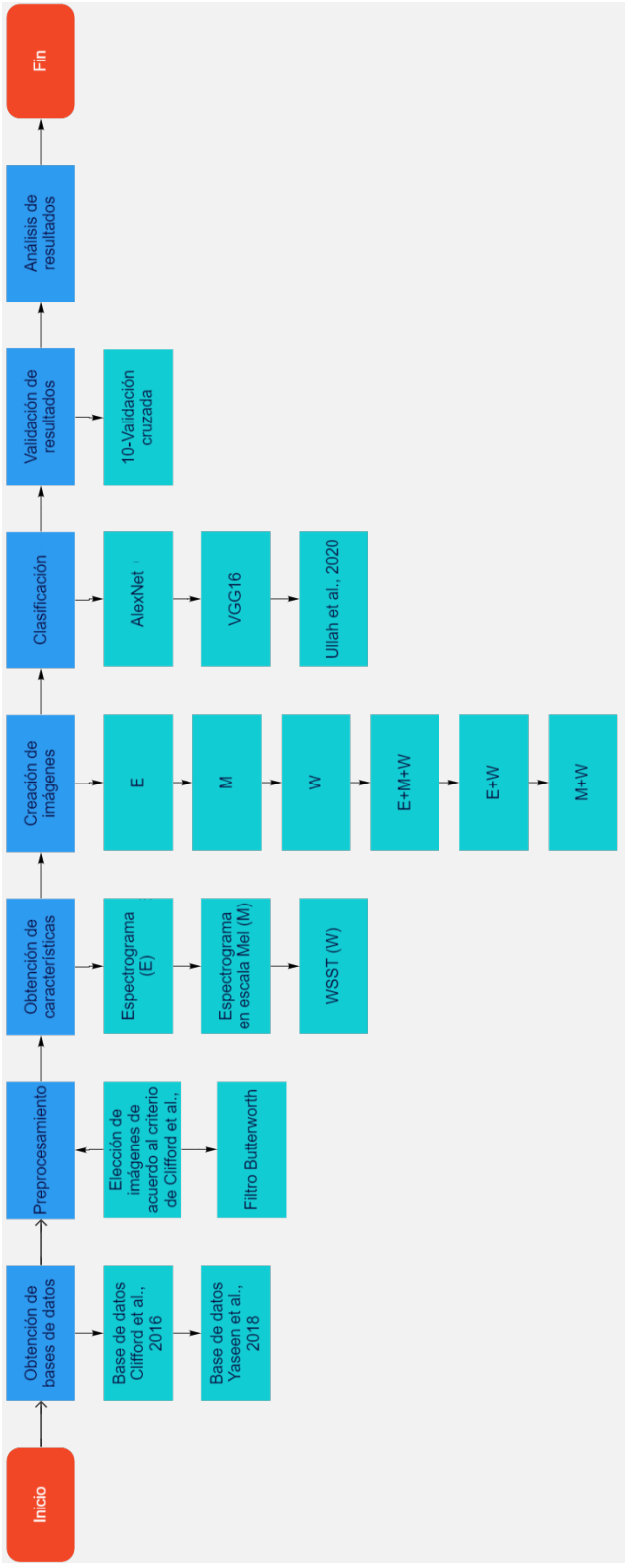


Figura 1. Estructura de la tesis.

Capítulo 2. Antecedentes

En este capítulo se explicarán los antecedentes y la información relacionada a los sonidos cardíacos que serán usados durante el presente trabajo. Se comenzará presentando información relevante a la fisiología cardiovascular, describiendo la anatomía del corazón, así como las características de los sonidos cardíacos que serán de ayuda para la tarea de clasificación de los mismos.

Se explicará el ciclo cardíaco desde un punto de vista mecánico, lo cual facilita el entendimiento del origen de los sonidos cardíacos que se usarán, así como el origen de las características patológicas en los fonocardiogramas.

Se presentarán los trabajos relacionados realizados en CICESE. Así como el Estado-del-arte de la clasificación automática de anomalías cardíacas utilizando el FCG.

2.1. Fisiología cardiovascular

El corazón es uno de los órganos más importantes del cuerpo humano. Es una bomba muscular que tiene dos funciones: la primera, es la de recibir sangre baja en oxígeno de los órganos y tejidos, para ser enviada a los pulmones para su oxigenación. Y, la segunda, es la obtención de la sangre oxigenada de los pulmones para ser enviada hacia el resto del cuerpo.

La anatomía interna del corazón revela la presencia de dos lados con dos cámaras huecas cada uno: una aurícula y un ventrículo que están conectadas. Las cámaras superiores (las aurículas) son bombas débiles y tienen como función el almacenamiento de la sangre para su ventrículo. Las cámaras inferiores (los ventrículos) son bombas fuertes que tiene como función el bombeo de la sangre que es logrado con la contracción de las aurículas seguido de la contracción de los ventrículos.

El papel del lado derecho del corazón conformado por la aurícula y ventrículo derecho, es la recolección de la sangre baja en oxígeno proveniente del cuerpo, así como su envío a los pulmones para su oxigenación. Y, el papel del lado izquierdo del corazón, conformado por la aurícula y ventrículo izquierdo, es el de recoger la sangre oxigenada de los pulmones y enviarla al cuerpo.

Este flujo de sangre unidireccional, es logrado gracias a cuatro válvulas en el corazón que evitan que la sangre regrese cuando las presiones de las cámaras cambian. Estas válvulas son:

- Válvulas auriculoventriculares: son la tricúspide y bicúspide, permiten el flujo de la sangre de las aurículas hacia los ventrículos.
- Válvulas semilunares: siendo la aórtica y pulmonar, permiten el flujo de sangre de los ventrículos hacia las arterias mayores para ser enviadas a todo el cuerpo.

Esta anatomía del corazón, así como el flujo sanguíneo son ilustradas en la Figura 2, donde las flechas azules indican la dirección de la sangre dentro del corazón (Abbas y Bassam, 2009; Weinhaus y Roberts, 2005).

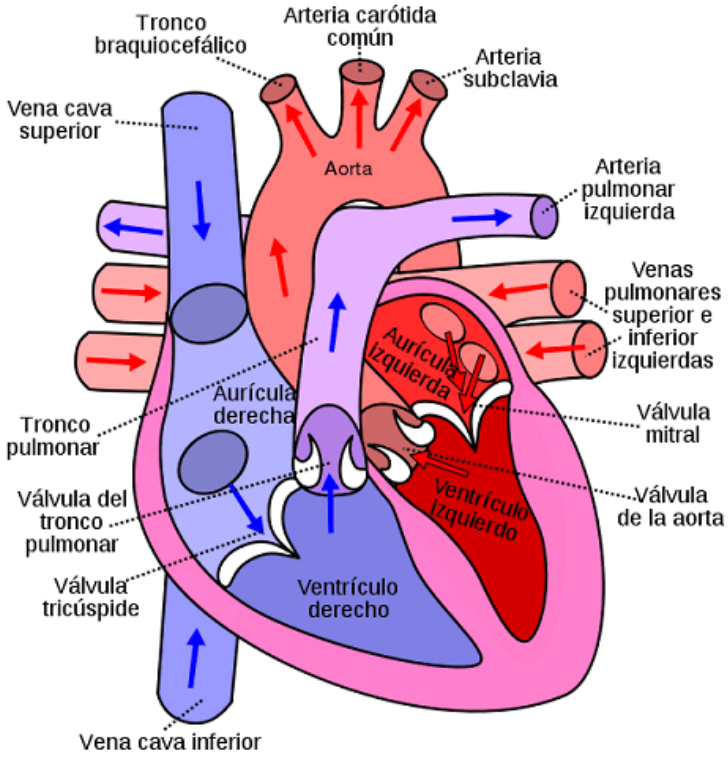


Figura 2. Anatomía del corazón, imagen obtenida de: `https://www.lifeder.com/corazon-partes-funciones/`.

2.2. Ciclo cardíaco

Las acciones mecánicas de las válvulas del corazón producen sonidos cardíacos incluyendo sonidos cardíacos, fundamentales (Fundamental Heart Sounds, FHSs) que

son señales eléctricas complejas y no estacionarias. El primer sonido cardíaco es (S_1) que se define como el sonido resultante de la sístole ventricular, está presente en todas las personas. Tiene un rango de frecuencia entre 10 y 200 Hz. El segundo sonido cardíaco es (S_2) Figura 3 y resulta de el cierre valvular aórtico y el sigmoideo pulmonar, está presente en todas las personas. Tiene un rango de frecuencia entre 20 y 250 Hz (Dwivedi *et al.*, 2019; Reed *et al.*, 2004).

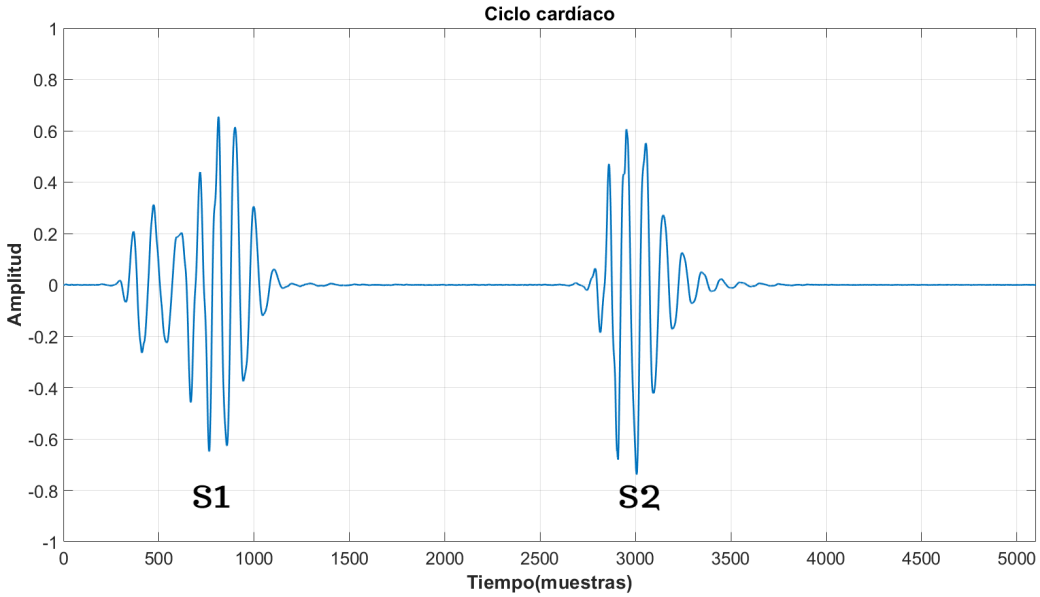


Figura 3. Forma de onda de un ciclo cardíaco sano.

Durante la operación cardíaca normal, un patrón claro es observado de S_1-S_2, con un período sistólico (S_1 a S_2) y un período diastólico (S2 a S1). Sin embargo cuando se presentan anormalidades se hace presente un tercer sonido cardíaco (S_3) que se puede apreciar en la diástole por disfunción ventricular, es habitual en la niñez, es frecuente en adolescentes y es muy raro después de los 40 años, donde puede considerarse como patológico. Es posible considerar un cuarto sonido cardíaco (S_4) el cual es un ruido auricular ocasionado por la tensión de las válvulas auriculoventriculares como del miocardio ventricular, debido a un llenado acelerado, pueden ser galopes, clics, chasquido de apertura y murmullos podrían ocurrir (Dwivedi *et al.*, 2019; Cruz Ortega *et al.*, 2016).

La Tabla 1 resume las características patofisiológicas de los sonidos del corazón (Dwivedi *et al.*, 2019), y la Figura 4 muestra un murmullo presente en un ciclo cardíaco.

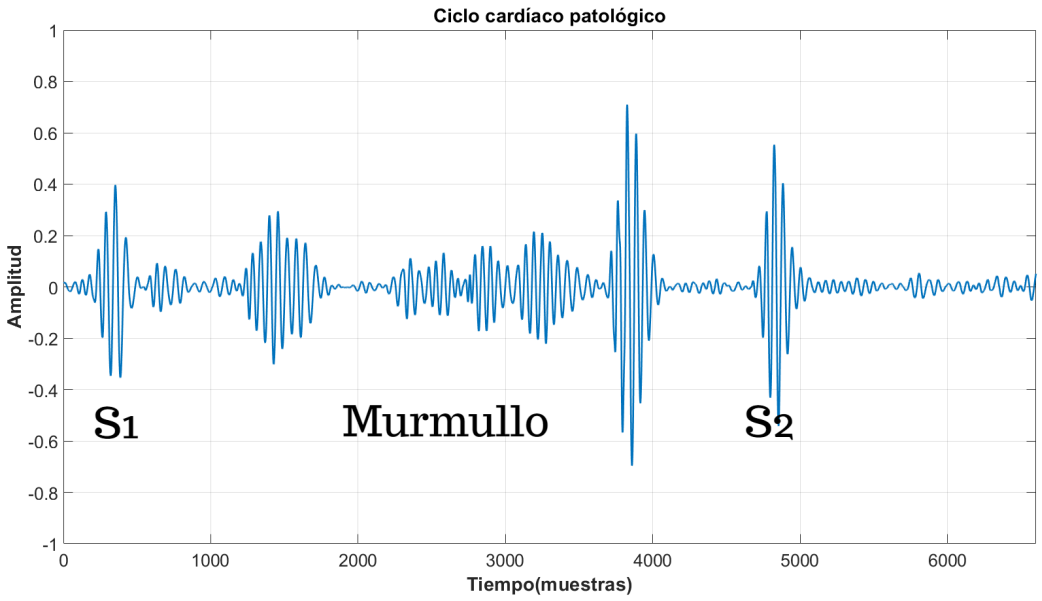


Figura 4. Forma de onda de un ciclo cardíaco patológico.

Tabla 1. Patofisiología de sonidos cardíacos normales y patológicos.

Sonido del corazón	Rango de Frecuencia (Hertz)	Duración (segundos)
S1	10-200	0.12-0.15
S2	20-250	0.08-0.12
S3	25-75	0.004
S4	15-70	0.004

Como se ha mencionado anteriormente la señal de FCG es no estacionaria, esto es ya que los parámetros estadísticos que caracterizan a la señal varían. Por lo que se tienen que analizar ambas partes para la extracción de características, en el dominio del tiempo se tienen las características de ritmo cardíaco, tasa de cruces por cero y duración del ciclo cardíaco. En el dominio de la frecuencia se tienen las características de potencia total de la señal de PCG, ancho de banda y factor-Q (Khan *et al.*, 2020; Homsi *et al.*, 2016).

Para la extracción de características, varios autores han considerado diversas formas de obtenerlas. Como por ejemplo, de acuerdo a Soeta y Bito (2015) se puede usar la *Short-Time Fourier Transform (STFT)*, Obaidat (1993) utiliza la distribución Wigner-Ville (WVD), Addison *et al.* (2009) utiliza la transformada ondeleta y Zhang *et al.* (2017) utiliza ondeletas de Gabor.

2.3. Antecedentes en CICESE

En el CICESE ya se ha explorado anteriormente y con éxito el uso de técnicas de procesamiento de señales y de aprendizaje automático para analizar el FCG.

- Se ha propuesto un esquema de segmentación de audio cardíaco basado en *Matching Pursuit* (Nieblas *et al.*, 2013).
- Se ha desarrollado un método para representar sonidos cardíacos usando un modelo armónico + ruido (*Harmonic plus noise*), basado en Matching Pursuit y técnicas de codificación de predicción lineal (*Linear Prediction Coding*) (Ibarra *et al.*, 2016)
- Se han llevado a cabo trabajos de obtención de métodos para la segmentación de audio cardíaco mediante análisis tiempo-frecuencia (Gutiérrez, 2016).
- Modelos de clasificación automática usando métodos de extracción de características basados en Matching Pursuit con átomos tiempo-frecuencia (Hernández, 2019).
- Así como la clasificación del audio cardíaco usando representación escasa y átomos de Gabor usando diferentes clasificadores automáticos (Gutiérrez Uribe, 2019).

Durante estas investigaciones se encontró que el audio cardíaco puede ser representado de manera precisa mediante el uso de algoritmos de análisis de tiempo-frecuencia. En general, los métodos de representación de tiempo-frecuencia (TFR) han sido utilizados para estudiar FCG dado que se trata de una señal biológica no estacionaria.

2.4. Estado-del-arte de la clasificación automática

La invención del estetoscopio en el año de 1816, por el médico francés Laënnec, fue la primera herramienta para el análisis de los sonidos cardíacos y pulmonares. La prueba de la importancia de la invención del estetoscopio es que hoy en la actualidad se sigue usando como una herramienta para el diagnóstico médico. Sin embargo, aunque el estetoscopio es un invento revolucionario tuvo un periodo de evolución a lo

largo de los siglos XIX y XX, en el que diversos inventores hacían aportaciones para el mejoramiento del estetoscopio (Roguin, 2006).

A pesar de que la auscultación ha sido una herramienta efectiva, se deben considerar diversos factores para demostrar que la auscultación puede ser una herramienta limitada por la capacidad auditiva del médico. Para la mejora de la detección y diagnóstico de patologías cardíacas, se tiene la ayuda de la auscultación asistida por computadora *(Computer Aided Auscultation, CAA)*, la cual ha demostrado tener el potencial para la detección de los sonidos S_3 y S_4, dando como resultado un mejor resultado en el diagnóstico del sonido cardíaco (Mahnke, 2009; Nieblas *et al.*, 2014).

Para la solución a este problema se propone el uso de técnicas de análisis tiempo-frecuencia, ya que estas son especialmente adecuadas para el análisis de señales no estacionarias en conjunto con clasificadores automáticos, gracias al poder de cómputo en la nube así como al poder computacional de los sistemas físicos actuales las redes neuronales han tenido un gran auge en el campo de la clasificación.

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 se puede llamar de aprendizaje profundo (Peyré, 2020).

Las ventajas que tiene el aprendizaje profundo sobre el aprendizaje de máquinas son el uso de una mayor cantidad de datos de muestra y que no es necesaria la extracción manual de las características.

Los métodos de aprendizaje profundo son métodos 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. Para tareas de clasificación, las representaciones de capas de alto nivel amplifican aspectos de la entrada que son importantes para la discriminación y supresión de variables irrelevantes (LeCun *et al.*, 2015).

En el caso del preprocesamiento de la señal FCG se tienen trabajos usando Coefi-

cientes Cepstrales en la frecuencia Mel (*MFCC*) que son administradas a un Modelo Oculto de Markov (*HMM*) para la clasificación de las señales, donde se han encontrado resultados prometedores (Chauhan *et al.*, 2008).

Otro trabajo en el que se usan los *MFCC* y el espectrograma en escala Mel, para la obtención de características tiempo-frecuencia que serán suministradas a un modelo de red neuronal convolucional puede ser visto en (Bozkurt *et al.*, 2018).

Existen artículos que agrupan diferentes tipos de representaciones, así como clasificadores automáticos para poder comparar su rendimiento en tareas de clasificación de sonidos cardíacos (Dwivedi *et al.*, 2019).

En el trabajo de Ghosh *et al.* (2020) se usan transformadas Chirplet y clasificadores multiclase en señales de anomalías en las válvulas del corazón.

Se hace uso de espectrogramas para obtener características tiempo-frecuencia que serán clasificadas con una red neuronal convolucional propia (Ullah *et al.*, 2020).

En el trabajo de Yaseen *et al.* (2018) se hace uso de *MFCC* y de la transformada ondeleta discreta *(DWT)* para la obtención de características, que serán clasificadas usando redes neuronales y máquinas de soporte vectorial *(SVM)*.

En Azmy (2016) se hace uso de la transformada madre y *SVM* para la tarea de clasificación de sonidos cardíacos en sanos y patológicos.

Se tiene la transformada ondeleta *Synchrosqueezing (Wavelet Synchrosqueezed Transform, WSST)* que debido a su alta resolución, tanto en tiempo como en frecuencia, recientemente se ha convertido en una herramienta muy utilizada en el área de procesamiento digital de señales, dado el enorme aumento en la capacidad de cálculo y de memoria de almacenamiento de las computadoras actuales (Ghosh *et al.*, 2019).

Se ha usado el espectrograma Mel junto con la transformada de la constante-Q para la obtención de las representaciones tiempo-frecuencia, que en conjunto con *SVM* han dado buenos resultados para la clasificación de sonidos cardíacos es analizado en Vyas *et al.* (2021).

En la presente revisión del estado-del-arte no se ha encontrado la aplicación de las

tres representaciones tiempo-frecuencia que se usarán para el preprocesamiento del FCG en combinación con las redes neuronales convolucionales para el desarrollo de un clasificador de sonidos cardíacos.

Capítulo 3. Análisis tiempo-frecuencia de la señal de PCG

En este capítulo, se describirán las representaciones tiempo-frecuencia que serán usadas a lo largo de este trabajo. Cada una de las representaciones tendrán dimensiones idénticas (224 x 224) para ser clasificadas usando redes neuronales convolucionales, esto será explicado con más detalle en el capítulo 4.

Se comenzará el capítulo con la presentación de las dos bases de datos disponibles, de las que se obtendrán sonidos cardíacos necesarios para la obtención de las imágenes de las representaciones tiempo-frecuencia, estas bases de datos cuentan con la facilidad de que ya están etiquetados los sonidos cardíacos patológicos y los sanos.

La primera representación que se revisará será el espectrograma, que es una forma de mostrar de manera visual de la magnitud de una señal en el tiempo, así como sus componentes espectrales. La segunda representación analizada es el espectrograma en escala Mel, que usa una diferente unidad en su escala a diferencia de la anterior, se hace uso de la relación perceptual entre las frecuencias percibidas por el oído humano y las frecuencias físicas reales de los sonidos. La última representación tiempo-frecuencia que se usará será una transformada ondeleta, debido a su habilidad para representar los componentes una señal y caracterizarlos en el tiempo o la frecuencia.

Siendo estas tres representaciones las que se usarán, también se ha planteado la idea de la combinación de una o más representaciones al preprocesamiento del FCG con el fin de mejorar el rendimiento de la clasificación, los parámetros para la obtención de cada representación, así como su participación dentro de la clasificación será explicado junto con una figura para mejorar la visualización de la representación.

3.1. Bases de datos

En este trabajo se usarán dos bases de datos para la obtención de los sonidos cardíacos, la primera base es parte de ”The PhysioNet/Computing in Cardiology(CinC) Challenge 2016” que provee una colección de sonidos cardíacos públicos, obtenidos de ocho fuentes por siete grupos de investigadores independientes a nivel mundial, entre ellos participaron instituciones prestigiosas como el *MIT*. La base de datos inclu-

ye 4,430 grabaciones tomadas de 1,072 sujetos, haciendo un total de 233,512 sonidos cardíacos recolectados de pacientes sanos y pacientes con diferentes condiciones médicas, como enfermedades de las válvulas del corazón y enfermedades de la arteria coronaria (Clifford *et al.*, 2016). La composición de la base de datos se explica en la Tabla 2.

Tabla 2. Composición de la base de datos de (Clifford *et al.*, 2016)

Base de datos	pacientes	grabaciones	Proporción de Patológicos	grabaciones (%) Normales	Desconocido
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

La segunda base de datos que se usará está extraída del trabajo de Yaseen *et al.* (2018), está compuesta de dos conjuntos de datos; los sonidos patológicos y los sonidos sanos. Estos conjunto de datos componen cinco categorías:

- Señales normales (N).
- Estenosis aórtica (AS).
- Estenosis mitral (MS).
- Regurgitación mitral (MR).
- Prolapso de la válvula mitral (MVP).

La composición de la base de datos está explicada en la Tabla 3.

Tabla 3. Composición de la base de datos de (Yaseen *et al.*, 2018)

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

Las señales de audio cardíaco fueron preprocesadas de dos maneras:

1. Las señales de la base de datos (Clifford *et al.*, 2016) tienen una etiqueta que indica si una señal de audio cardíaco es adecuada para tareas de clasificación. Por ello, sólo se tomaron en cuenta las señales de audio cardíaco que son aptas para tareas de clasificación. La cantidad de sonidos cardíacos no usados se detallan en la Tabla 4.

2. De las señales restantes, se usó un filtro paso-banda Butterworth con frecuencias de corte en 25 Hz y 900 Hz, siendo la respuesta del filtro representada en la Figura 5.

Tabla 4. Sonidos cardíacos no usados de la base de datos (Clifford *et al.*, 2016)

Base de datos	Sonidos no usados
A	17
B	122
C	4
D	3
E	128
F	5

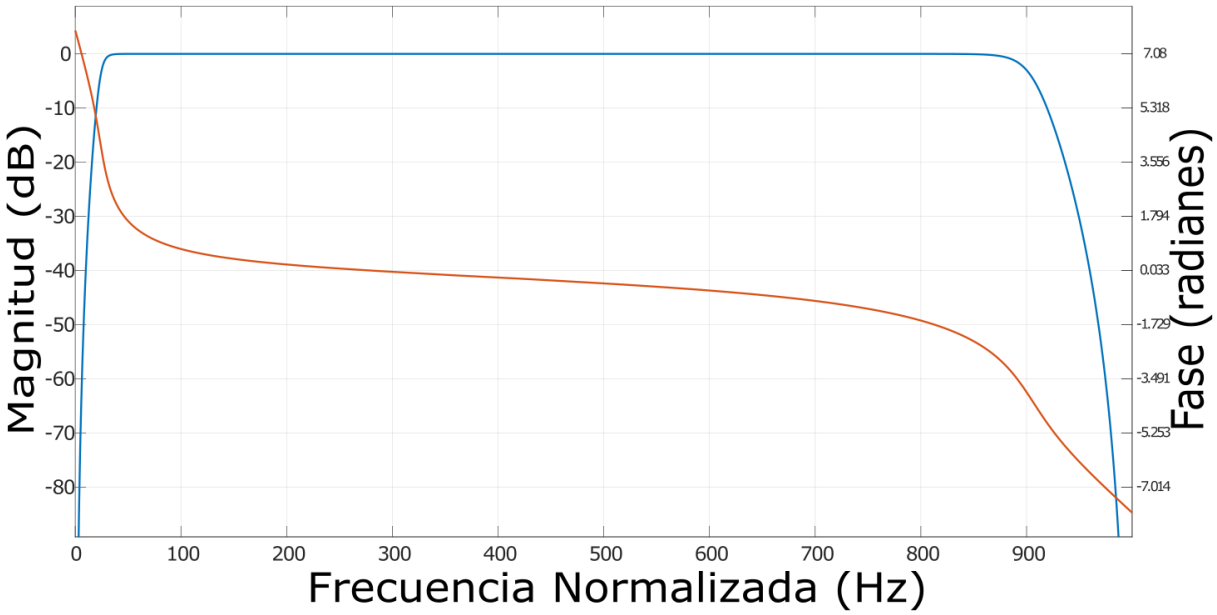


Figura 5. Respuesta en Magnitud (dB) y Respuesta en Fase del filtro.

3.2. Transformada de Fourier de tiempo corto (STFT)

La transformada de Fourier de tiempo corto, o también conocida como STFT (*Short-Time Fourier Transform*), es una herramienta muy utilizada en el procesamiento de se-

ñales en el dominio tiempo-frecuencia. En comparación con otras herramientas como la transformada ondeleta es que se puede tener una mayor facilidad en la interpretación con la STFT. La mayoría de los resultados obtenidos usando ondeletas pueden ser obtenidos de la misma manera usando la STFT (Jurado y Saenz, 2002). 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 STFT 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 (Jurado y Saenz, 2002).

El análisis de Fourier descompone una señal en sus componentes frecuenciales. La relación entre la representación en el tiempo $f(t)$ y la representación en la frecuencia $F(\omega)$ se define como la transformada de Fourier y matemáticamente se expresa en la Ecuación 1 (Kiymik *et al.*, 2005):

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \Longleftrightarrow f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{j\omega t} d\omega. \tag{1}$$

Esta transformada se aplica a señales estacionarias, es decir, señales cuyas propiedades estadísticas no cambian en el tiempo. Cuando la señal es no estacionaria se puede introducir un parámetro de frecuencia local para que la transformada de Fourier local vea a la señal dentro de una ventana en la que se puede aproximar para ser estacionaria. El tamaño de la ventana (N) se puede determinar con la relación entre la frecuencia de muestreo (f_s) y la resolución espectral deseada de la forma $N = \frac{fs}{resolución\ deseada}$. Entonces se aplica la STFT a una función ventaneada $\psi(t)$ en un eje de tiempo τ, resultando en la Ecuación 2:

$$F(\omega, \tau) = \int_{-\infty}^{\infty} f(t) \psi^*(t - \tau) e^{-j\omega t} dt. \tag{2}$$

Cuando la ventana $\psi(t)$ es una función Gaussiana, la STFT es llamada una transformada de Gabor. Las funciones básicas de la transformada son generadas por la modula-

ción y la transformación de la función ventana $\psi(t)$, donde ω y τ son parámetros de modulación y traslación, respectivamente. La limitación de la STFT es la ventana fija de tiempo $\psi(t)$ ya que causa una resolución tiempo-frecuencia fija. Esto es explicado por el principio de incertidumbre de Heisenberg, que quiere decir que sólo podemos obtener una mejor resolución en el dominio del tiempo o de la frecuencia pero no en ambos a la vez (Kiymik *et al.*, 2005).

3.2.1. STFT discreta

La STFT presentada, en la Ecuación 3, es usada para el análisis tiempo-frecuencia de señales no estacionarias, donde el uso de la transformada de Fourier por sí misma no es del todo adecuada. La STFT discreta, descompone la señal variante en el tiempo, en componentes en el dominio tiempo-frecuencia, por lo que nos permite un conocimiento de la evolución en el tiempo de cada componente de la señal (Jurado y Saenz, 2002).

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

donde $x(n)$ es la señal de entrada en el instante n, $w(n)$ es la longitud M de la ventana, m es el índice de los instantes de tiempo de cada ventana, R es el tamaño de los saltos (en muestras) entre cada STFTs (Smith, 2011).

3.3. Espectrograma

Un espectrograma, es una manera de representar de manera visual la magnitud de una señal en el tiempo en varias frecuencias presentes en una forma de onda particular. Se puede apreciar si se tiene una gran concentración de energía o la falta de la misma, así como la visualización de la variación de los niveles energéticos a lo largo del tiempo. Esta visualización suele estar representada por una escala de colores, y resulta muy útil para la facilidad en la interpretación de la representación tiempo-frecuencia de una señal.

Para su obtención se toman muestras de longitud fija de la señal de entrada en función del tiempo, siendo un ejemplo de la base de datos Yaseen *et al.* (2018) la Figura 6

y de la base de datos Clifford *et al.* (2016) la Figura 7, a la que se le aplica una función ventana para su posterior conversión al dominio de la frecuencia usando la transformada rápida de Fourier (FFT). La ventana se deslizará con un factor de longitud de traslape para luego repetir el proceso. Esto se repetirá hasta que toda la señal ha sido analizada, este procedimiento se explicará gráficamente en la Figura 8, obteniendo como resultado las Figuras 12 y 13.

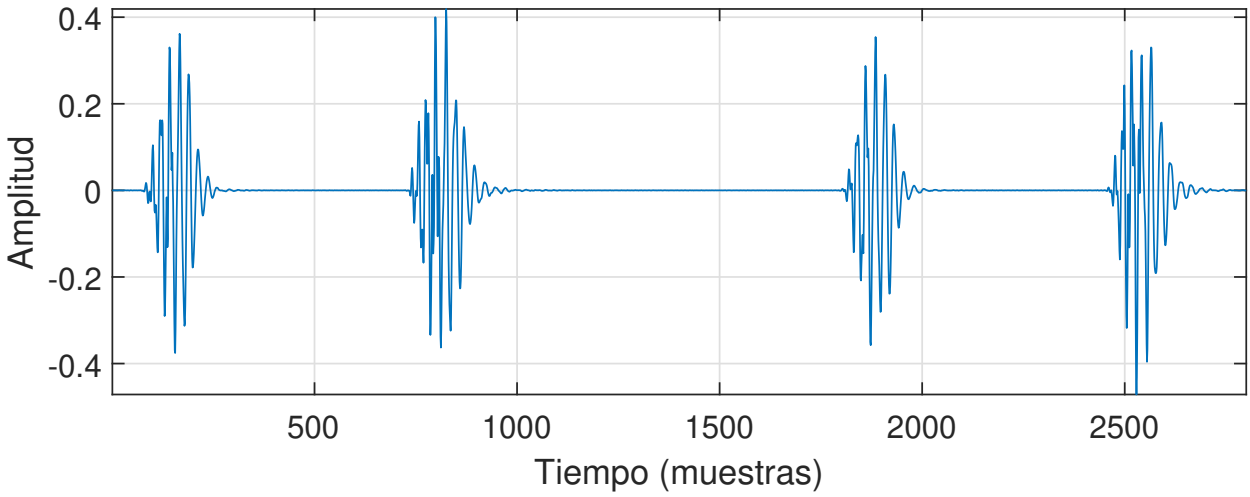


Figura 6. Señal de FCG normal tomada de la base de datos del artículo Yaseen *et al.* (2018), con frecuencia de muestreo de 8000 Hz.

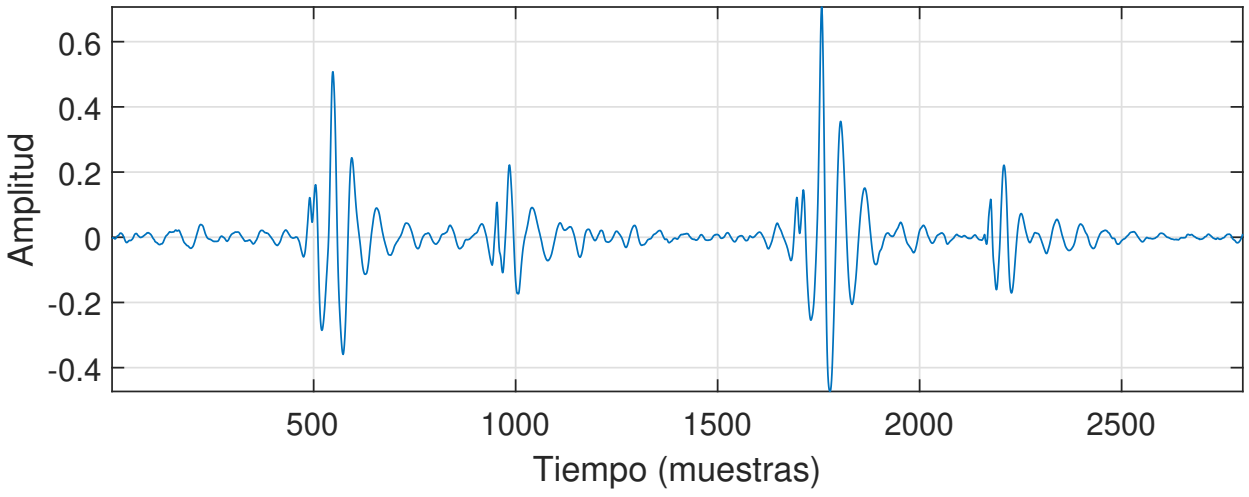


Figura 7. Señal de FCG anormal tomada de la base de datos Physionet Cinc Challenge estudiada en Clifford *et al.* (2016), con frecuencia de muestreo de 2000 Hz.

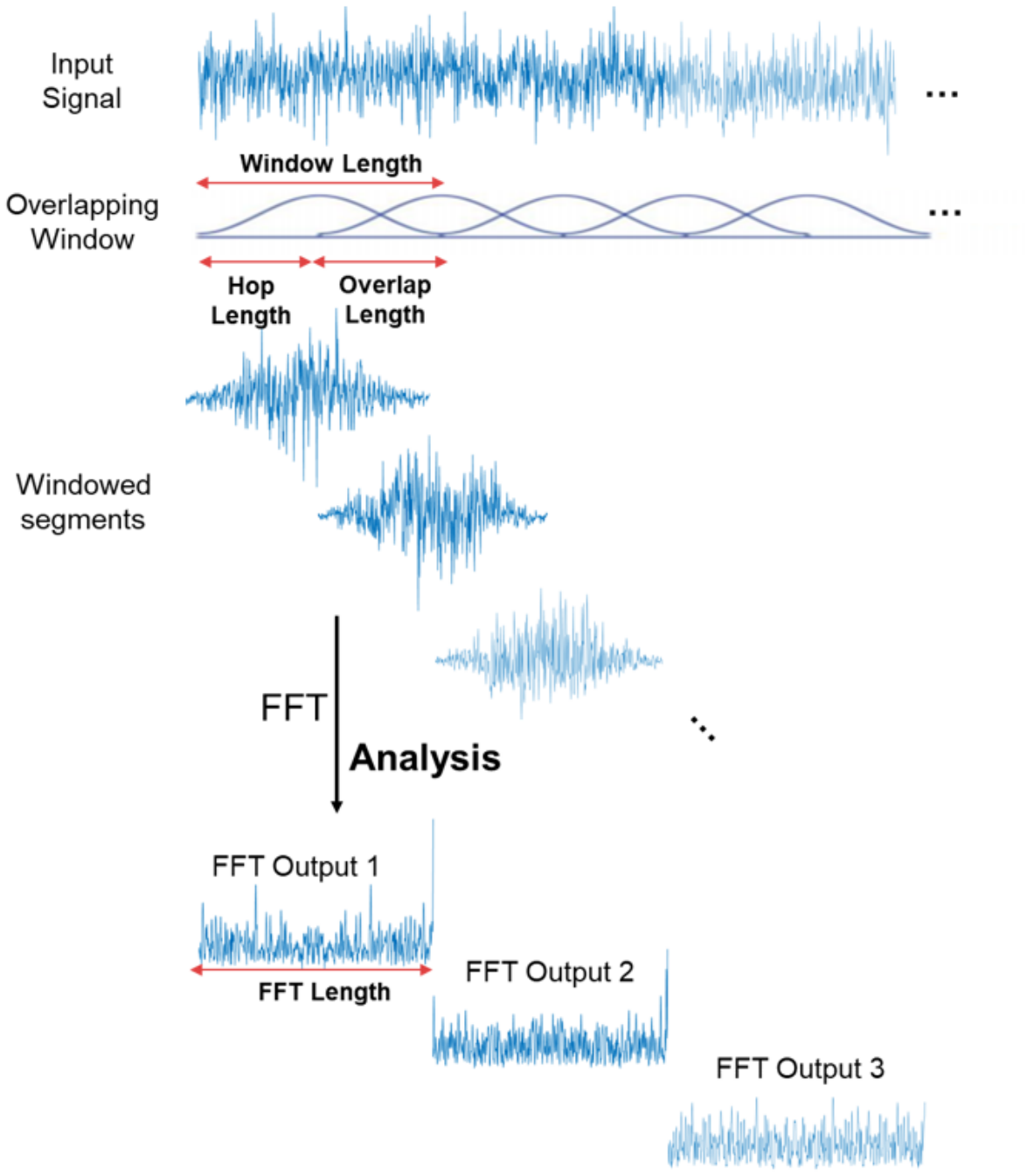


Figura 8. Proceso de obtención de un espectrograma, imagen tomada de: `https://www.mathworks.com/help/dsp/ref/dsp.stft.html`.

3.3.1. Ventanas de análisis

Las ventanas son funciones de ponderación aplicadas a datos para reducir la filtración espectral con intervalos de observación finita. Se puede decir que la ventana

es aplicada a la información como una multiplicación ponderada para reducir el orden de la discontinuidad en el límite de la extensión periódica. Esto se logra al emparejar tantas órdenes posibles de la derivada en el límite como se explica a continuación. La manera más sencilla de lograr esta igualdad es al establecer los valores de las derivadas a cero o muy cerca de cero. Por lo tanto, la información ventaneada será llevada de manera gentil a cero en los límites, por lo que la extensión periódica de la información será continua en muchos órdenes de la derivada (Harris, 1978).

Existen cientos de ventanas, algunas de las más utilizadas se presentan a continuación:

- Hanning Figura 9 y Hamming Figura 10: las dos ventanas presentan un pico amplio, pero con lóbulos laterales bajos. Una diferencia entre ambas ventanas es que la ventana de Hanning llega a cero en ambos extremos, eliminando así, toda discontinuidad. Caso contrario, es la ventana de Hamming, que no llega a cero por lo que se presenta una discontinuidad en la señal.
  La ecuación de la ventana de Hamming es (Harris, 1978):

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \quad \text{para } 0 \leq n \leq M-1. \tag{4}$$

  La ecuación de la ventana de Hanning es:

$$\omega(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{M-1}\right) \quad \text{para } 0 \leq n \leq M-1. \tag{5}$$

  Donde M es un entero que indica el número de muestras de la ventana.

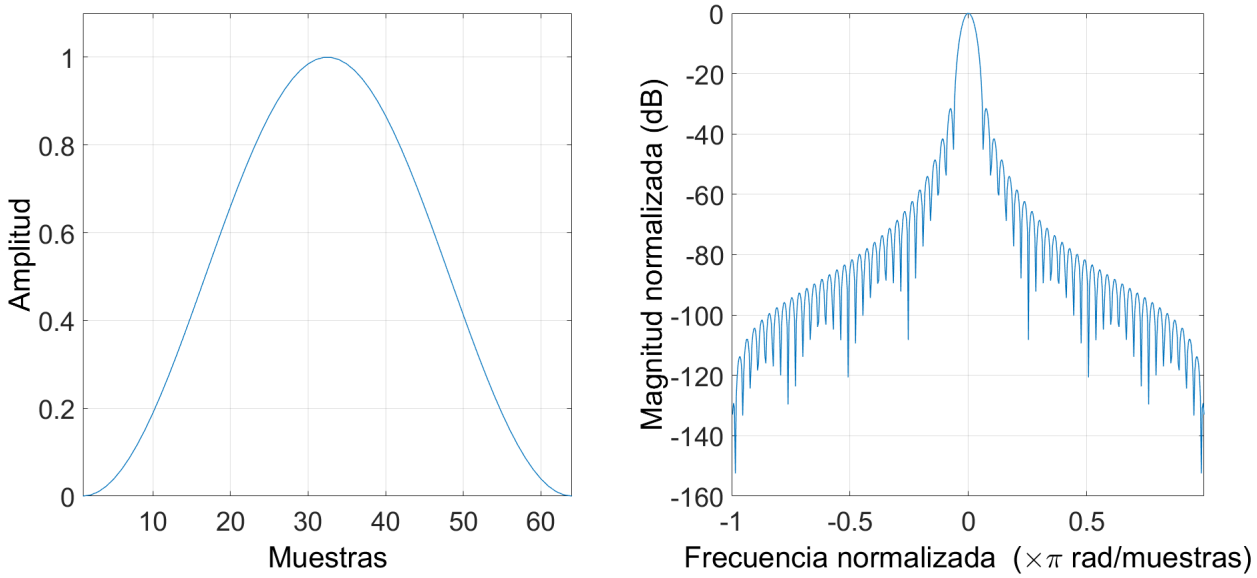


Figura 9. Ventana de Hanning de 64 muestras de longitud y su transformada de Fourier para frecuencia normalizada.

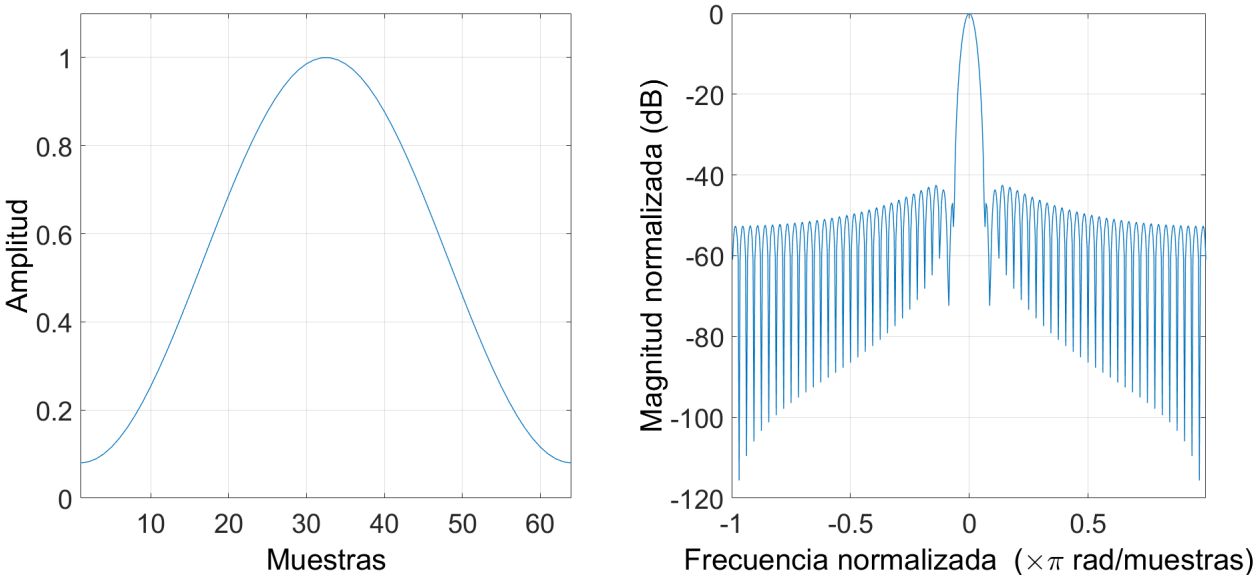


Figura 10. Ventana de Hamming de 64 muestras de longitud y su transformada de Fourier para frecuencia normalizada.

- La Blackman (Figura 11): esta ventana es similar a las ventanas anteriores, con la diferencia que se conforma de la suma de los primeros tres términos de una sumatoria de cosenos, con la intención de minimizar la filtración espectral. La ecuación de la ventana de Blackman es (Harris, 1978):

$$\omega(n) = 0.42 - 0.5\cos\left(\frac{2\pi n}{M-1}\right) + 0.08\cos\left(\frac{4\pi n}{M-1}\right) \quad \text{para } 0 \leq n \leq M-1, \qquad (6)$$

donde M es un entero que indica el número de muestras de la ventana.

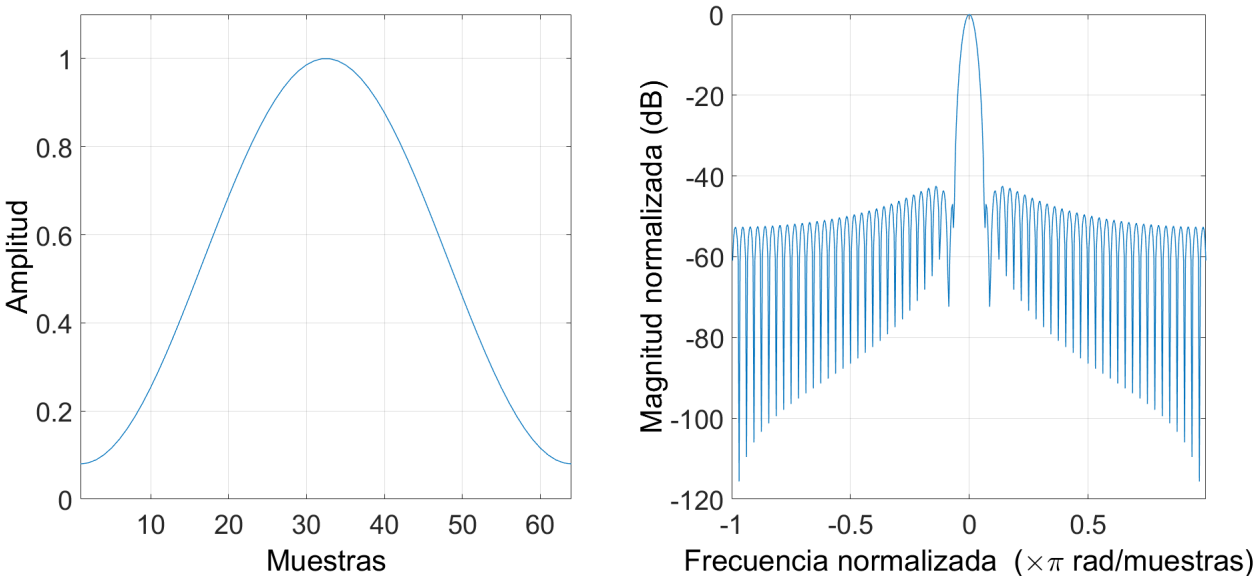


Figura 11. Ventana de Blackman de 64 muestras de longitud y su transformada de Fourier para frecuencia normalizada.

Se hicieron pruebas iniciales para determinar qué ventana sería la más conveniente para nuestro caso, teniendo en mente la facilidad de implementación, las limitaciones de tiempo y equipo, así como para mantener una similitud en los datos, entre las distintas representaciones, se optó por la ventana de Hamming para nuestros análisis.

3.3.2. Imágenes del espectrograma

Teniendo en cuenta lo anteriormente explicado, se procedió a la obtención de las imágenes usando el espectrograma como la herramienta para el análisis tiempo-frecuencia. Los parámetros usados para obtener el espectrograma son:

- Frecuencia de muestreo: 2000 Hz
- Longitud de la ventana de análisis: 2800 muestras
- Longitud de la ventana de Hamming: 100 muestras
- Longitud del traslape: 88 muestras
- Longitud de la FFT: 512 muestras

Para tener un mejor control sobre la obtención de las imágenes se decidió homogeneizar la frecuencia de muestreo de las bases de datos disponibles a 2000Hz. La matriz obtenida fue analizada y se recortó para tener una dimensión de 224x224 muestras. Se calculó el valor absoluto de la matriz, y los valores resultantes fueron divididos entre el valor máximo de la matriz, es decir fue normalizada en magnitud.

Con estas consideraciones se obtuvieron 11,287 (24.17%) imágenes correspondientes a sonidos patológicos y 35,418 (75.83%) imágenes que corresponden a sonidos sanos, distribuidas como se muestra en la Tabla 5:

Tabla 5. Total de imágenes de la representación del espectrograma

Base de datos	Imágenes patológicas	Imágenes Sanas
A (Clifford *et al.*, 2016)	6,318	2,632
B (Clifford *et al.*, 2016)	362	1,474
C (Clifford *et al.*, 2016)	703	194
D (Clifford *et al.*, 2016)	330	205
E (Clifford *et al.*, 2016)	1,909	28,894
F (Clifford *et al.*, 2016)	708	1,815
(Yaseen *et al.*, 2018)	957	204

Un ejemplo de la imagen obtenida usando una señal tomada de la base de datos (Yaseen *et al.*, 2018) se puede apreciar en la Figura 12, otro ejemplo usando un sonido tomado de de la base de datos (Clifford *et al.*, 2016) se puede apreciar en la Figura 13.

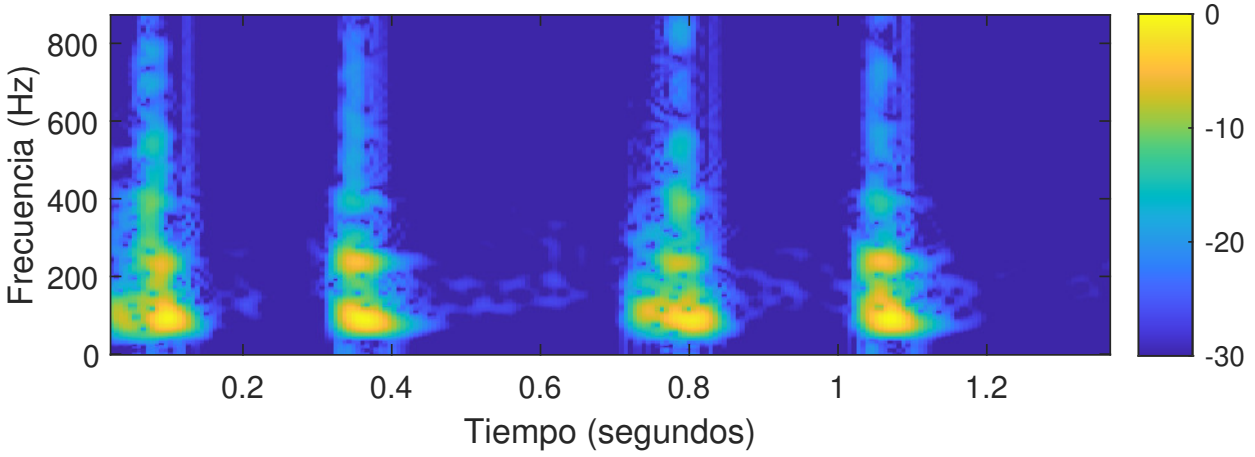


Figura 12. Espectrograma de la señal sana tomada de la base de datos del artículo (Yaseen *et al.*, 2018), con frecuencia de muestreo de 8,000 Hz.

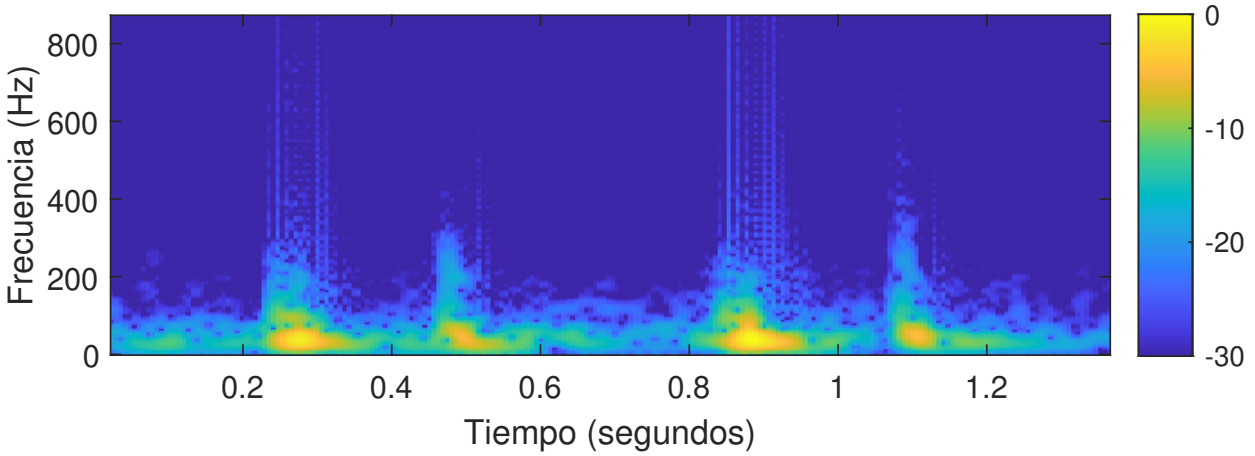


Figura 13. Espectrograma de la señal patológica tomada de la base de datos (Clifford *et al.*, 2016), con frecuencia de muestreo de 2,000 Hz.

3.4. Espectrograma en escala Mel

Existe evidencia de la psicofísica auditiva, que el oído humano percibe el sonido de una forma no lineal en función de la frecuencia, esto puede ser explicado con la Ley de Weber–Fechner, que expresa la relación logarítmica entre un estímulo y su percepción (Wu y Lin, 2000; Allen, 2008). Por lo que se propone usar la escala Mel dentro del análisis del espectrograma de la señal de audio cardíaco.

3.4.1. Escala Mel

La escala de Mel es derivada de los experimentos psicoacústicos que resultaron de la medida perceptual del sonido. Es una escala de audición que relaciona las frecuencias percibidas y las frecuencias físicas reales. Entonces, se puede definir que un Mel es una unidad de tono y de magnitud, para discriminar el aspecto y tono de varias frecuencias (Stevens y Volkmann, 1940; Umesh *et al.*, 2002).

El punto de referencia entre esta escala y la frecuencia, se define equiparando un tono de 1000 Hz, 40 dBs por encima del umbral de audición del oyente, con un tono de 1000 mels. Por encima de 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente (Figura 14).

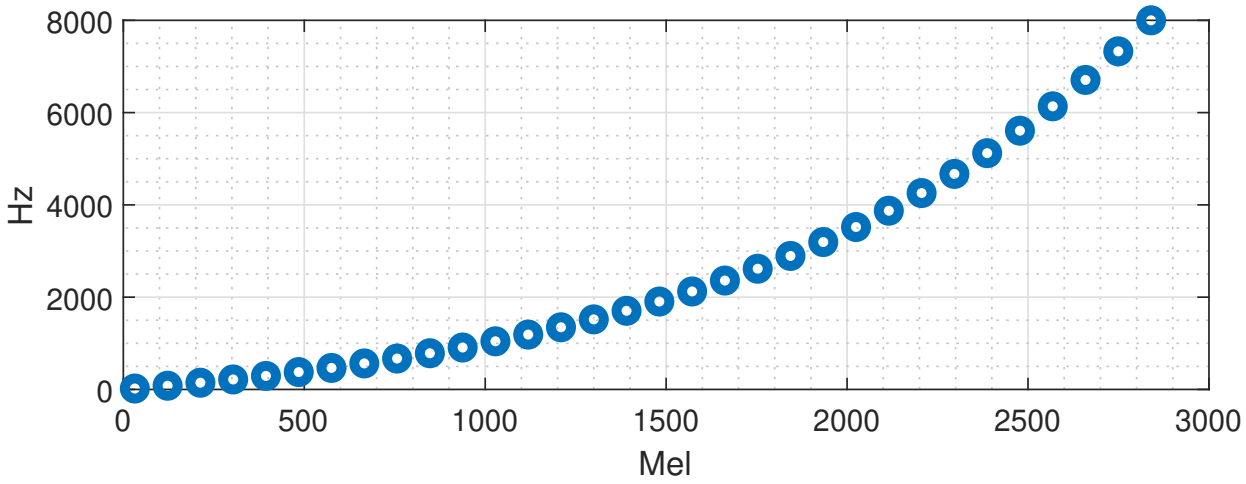


Figura 14. Relación entre Hertz y la escala Mel.

La relación entre Mel $\phi(f)$ y la frecuencia lineal l_f está dada por la ecuación 7 (Kopparapu y Laxminarayana, 2010):

$$\phi(f) = 2595 * \log_{10}\left(1 + \frac{l_f}{700}\right). \tag{7}$$

Un espectrograma en la escala de Mel se obtiene aplicando una transformada no lineal al eje de la frecuencia de la trasformada de Fourier de tiempo corto, inspirada por respuestas cuantificadas del sistema auditivo humano, por lo que simplifica la visualización del contenido frecuencial al ser de dimensiones menores (Shen *et al.*, 2018).

3.4.2. Banco de filtros de Mel

Los parámetros que definen un banco de filtros Mel (M) son el número de filtros Mel F, la frecuencia mínima l_{fmin} y la frecuencia máxima l_{fmax}. Estos filtros calculan el espectro promedio en cada frecuencia central con un ancho de banda creciente. La Ecuación 8 del banco de filtros de Mel está dada por (Kopparapu y Laxminarayana, 2010):

$$\mathrm{M(m,\ k)} = \begin{cases} 0 & \text{para } lf_f(k) < lf_c(m-1) \\ \frac{lf(k)-lf_c(m-1)}{lf_c(m)-lf_c(m-1)} & \text{para } lf_f(m-1) \leq lf_k < lf_c(m) \\ \frac{lf(k)-lf_c(m+1)}{lf_c(m)-lf_c(m+1)} & \text{para } lf_f(m) \leq lf_k < lf_c(m+1) \\ 0 & \text{para } lf_f(k) \geq lf_c(m+1) \end{cases} \tag{8}$$

El banco de filtros de Mel $M(m, k)$, el cual se ilustra en la Figura 15, es una matriz de dimensiones F x N. Donde k corresponde a la frecuencia $f(k) = k * f_s/N$, f_s corresponde a la frecuencia de muestreo en Hertz, N es la longitud de la ventana de tiempo de una señal discreta (Sigurdsson *et al.*, 2006). El número de filtros en el banco será de 24 para poder lograr las dimensiones necesarias para el entrenamiento, esto es tomando en cuenta el estudio de (Abdollahpur *et al.*, 2017).

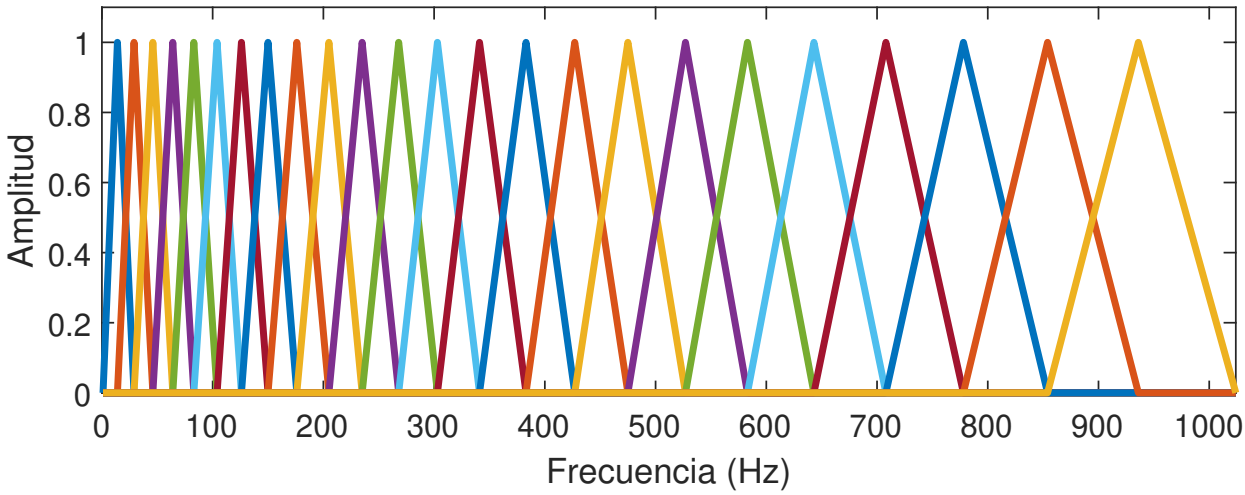


Figura 15. Filtros Mel.

3.4.3. Imágenes del espectrograma en escala de Mel

Considerando lo anteriormente explicado se propone la creación de las imágenes del espectrograma en escala Mel con los siguientes parámetros:

- Frecuencia de muestreo: 2000 Hz
- Longitud de la ventana de análisis: 2800 muestras
- Longitud de la ventana de Hamming: 100 muestras

- Longitud del traslape: 88 muestras
- Tamaño de la FFT: 512 muestras
- número de filtros de Mel: 24

Se consideró para tener una homogeneidad en la frecuencia de muestreo de los sonidos cardíacos entre la base de (Clifford *et al.*, 2016) y (Yaseen *et al.*, 2018) una frecuencia de muestreo de 2000Hz, por lo que los sonidos cardíacos de la base de datos de (Yaseen *et al.*, 2018), fueron remuestreados usando MATLAB para cumplir este criterio.

La matriz obtenida de la aplicación de 24 filtros de Mel fue de 24x224, para evitar tener que rellenar con ceros de 200x224 muestras se optó por utilizar el espectrograma, recortado de tal manera que tuviese el complemento de la dimensión deseada, la matriz resultante de 224x224 fue normalizada.

Se obtuvieron 11,287 imágenes patológicas y 35,418 imágenes sanas, como se muestra en la Tabla 6. Un ejemplo de la imagen obtenida usando un sonido tomado de la base de datos de (Yaseen *et al.*, 2018) se presenta en la Figura 16, y un ejemplo de la imagen obtenida usando un sonido tomado de la base de datos de (Clifford *et al.*, 2016) se presenta en la Figura 18, se debe tener en cuenta que el eje de las ordenadas se expresará en unidades de frecuencia, esto es debido a la combinación de dos representaciones tiempo-frecuencia en una misma imagen.

Tabla 6. Total de imágenes de la representación del espectrograma en escala Mel

Base de datos	Imágenes patológicas	Imágenes sanas
A (Clifford *et al.*, 2016)	6,318	2,632
B (Clifford *et al.*, 2016)	362	1,474
C (Clifford *et al.*, 2016)	703	194
D (Clifford *et al.*, 2016)	330	205
E (Clifford *et al.*, 2016)	1,909	28,894
F (Clifford *et al.*, 2016)	708	1,815
(Yaseen *et al.*, 2018)	957	204

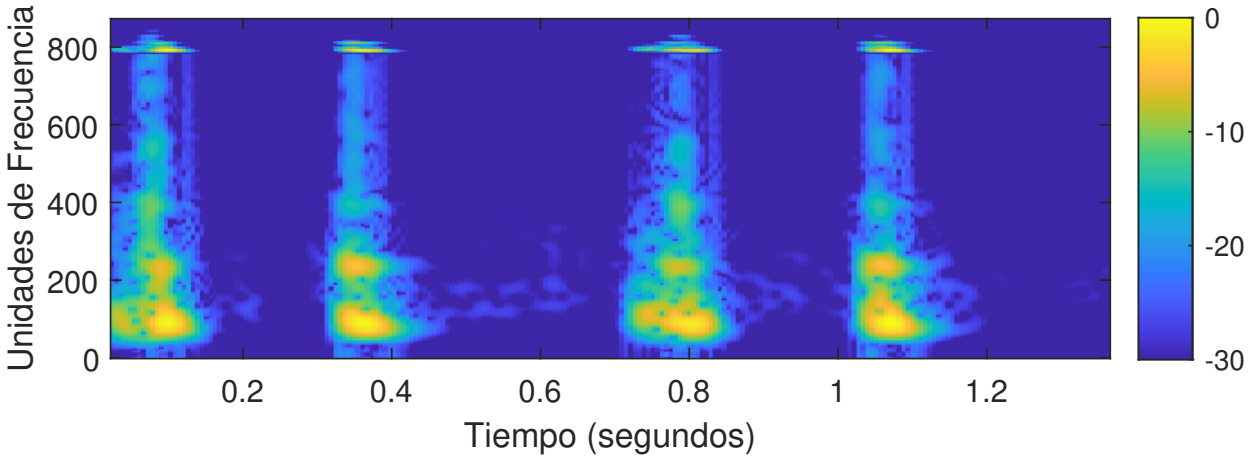


Figura 16. Espectrograma en escala de Mel de la señal sana tomada de la base de datos del artículo (Yaseen *et al.*, 2018), con frecuencia de muestreo de 8000 Hz.

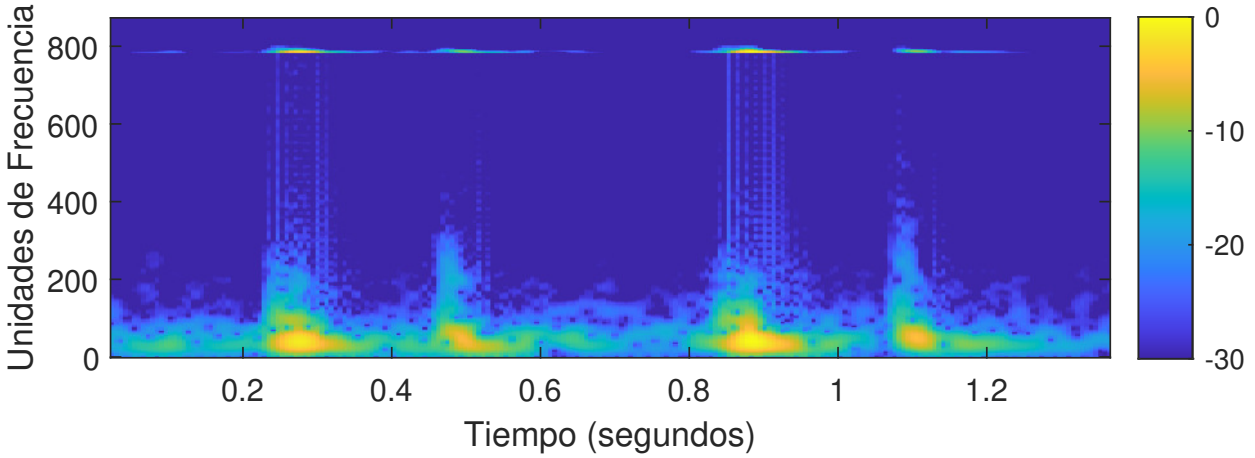


Figura 17. Espectrograma en la escala de Mel de la señal patológica tomada de la base de datos (Clifford *et al.*, 2016), con frecuencia de muestreo de 2000 Hz.

3.5. Transformada ondeleta *Synchrosqueezing* (Wavelet Synchrosqueezing Transform)

3.5.1. Análisis tiempo-frecuencia

Las representaciones tiempo-frecuencia proveen información muy útil en el análisis de señales no estacionarias, es decir que presentan cambios o saltos en su comportamiento. Existen diversos algoritmos para el análisis tiempo-frecuencia, la gran mayoría de ellos pueden ser métodos lineales o cuadráticos.

En los métodos lineales la señal analizada es caracterizada por sus productos inter-

nos con una familia de modelos preasignados, generados a partir de un modelo básico. En esta categoría está la transformada de Fourier de tiempo corto (*windowed Fourier transform*), y la transformada ondeleta (*wavelet tansform*). Estas representaciones tienen la habilidad de manifestar los componentes de una señal y caracterizarlos en el dominio del tiempo o en la frecuencia, a esta caracterización se le llama resolución en el tiempo o en la frecuencia. Debido a que existen señales cuya firma espectral es la misma, se vuelve necesario el uso de técnicas de análisis tiempo-frecuencia (Boashash, 2015).

El principio de incertidumbre de Heisenberg relaciona el límite en la precisión con la que cierto parámetro puede ser conocido. De manera análoga, el principio de incertidumbre de Gabor establece que los componentes espectrales no pueden ser definidos claramente en cualquier instante de tiempo. Lo que quiere decir, que se puede tener una alta localización en tiempo o en la frecuencia, pero no en ambas de manera simultánea (Tary *et al.*, 2018; Daubechies *et al.*, 2010).

En métodos cuadráticos para construir una representación tiempo-frecuencia, se puede evitar la utilización de la familia de modelos con los que la señal es comparada o medida. Por lo que, algunas características pueden ser más nítidas o definidas en un plano tiempo frecuencia.

Por lo que se puede decir que los métodos lineales y cuadráticos tienen problemas en el análisis de señales no estacionarias, ya que los métodos cuadráticos obscurecen la representación tiempo frecuencia con los términos de interferencia, y los métodos lineales son muy rígidos o proveen una imagen borrosa (Daubechies *et al.*, 2010).

En Messer *et al.* (2001) se hace una comparación del análisis de una señal y su pertinente análisis en diferentes dominios.

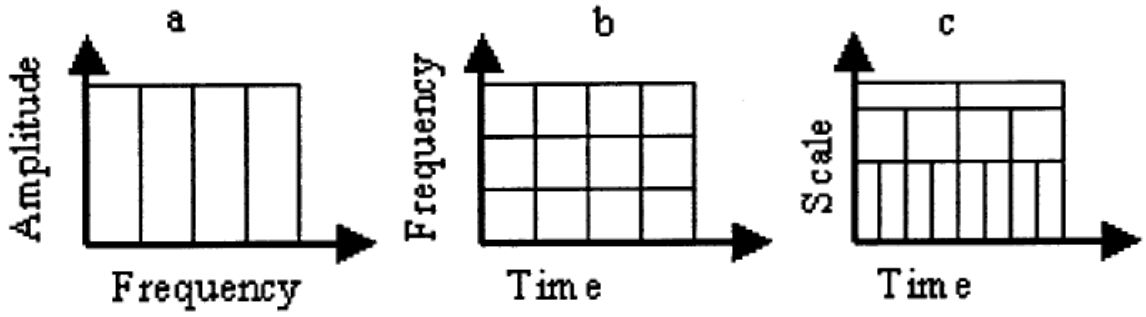


Figura 18. Las representaciones usadas son (a) transformada de Fourier, (b) trasformada de Fourier de tiempo corto, y (c) transformada ondeleta. Imagen tomada de Messer *et al.* (2001).

3.5.2. Ondeletas (Wavelets)

La idea de la transformada ondeleta, es el poder representar una función arbitraria f como una superposición de ondeletas. Las ondeletas son funciones generadas de una función ψ por dilataciones (a) y traslados (b).

$$\psi^{a,b}(t) = |a|^{-1/2}\,\psi\left(\frac{t-b}{a}\right), \tag{9}$$

donde t es una variable de una dimensión. La ondeleta madre ψ tiene que satisfacer la condición de $\int \psi(x)dx = 0$ (Antonini *et al.*, 1992).

3.5.3. Transformada ondeleta continua

Se puede describir como la transformada ondeleta continua a la relación entre una señal $s(t)$ y una familia de ondeletas:

$$W_s(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} s(t)\Psi^*\left(\frac{t-b}{a}\right)dt, \tag{10}$$

donde a representa una dilatación en el tiempo y b representa un traslado en el tiempo. Ψ y Ψ^* es su complejo conjugado, t es tiempo, y $W_s(a,b)$ es una representación en la escala del tiempo de la señal. La ondeleta de referencia es llamada la ondeleta madre (Tary *et al.*, 2018; Shensa, 1992).

La longitud temporal de las ondeletas usadas en la relación cruzada varía depen-

diendo de la componente frecuencial que se va a analizar. Las ondeletas de mayor tamaño son usadas en bajas frecuencias para poder tener una mejor resolución en la frecuencia a cambio de una disminución en la resolución en el tiempo. Las ondeletas de menor tamaño son usadas para mejorar la resolución temporal a cambio de la resolución espectral. El diccionario de las ondeletas se crea al estirar y trasladar la ondeleta de referencia (Tary *et al.*, 2018).

Para mejorar la resolución, la transformada ondeleta *Synchrosqueezing* (WSST) aplica tres pasos (Ghosh *et al.*, 2019; Tary *et al.*, 2018):

1. La aplicación computacional de la transformada ondeleta continua (CWT) para asegurar la resolución tiempo-frecuencia variable.

$$C_X^{\Psi}(n,k) = \frac{1}{k}\sum_{\tau=0}^{N-1} z(n)\Psi^*\left(\frac{\tau-n}{k}\right), \tag{11}$$

   donde $z(n)$ es una señal PCG, evaluada en los puntos $n = 0, 1, 2.....N-1$, $\Psi(n)$ es la ondeleta madre.

2. Cálculo de las frecuencias instantáneas para mejorar la interpretación, k es la escala de la ondeleta.

$$(w_X)^*(n,k) = R\left(\frac{1}{2\pi j}\frac{\Delta C_X^{\Psi}(n,k)}{C_X^{\Psi}(n,k)}\right), \tag{12}$$

   donde $R(\cdot)$ representa la parte real.

3. Reasignación de las frecuencias para minimizar el efecto de fuga espectral usando una nueva ubicación como $(n, \overline{w_X}(n,k)$ en la matriz de la escala de tiempo.

La WSST es representada matemáticamente como:

$$WS_z(n,w) = \frac{1}{K}\sum_{k=1}^{K} C_X^{\Psi}(n,k)\delta(w-(w_X)^*(n,k)). \tag{13}$$

La $WS_z(n,w)$ es la matriz tiempo frecuencia de la señal $z(n)$.

Algunas características importantes a tener en cuenta son que las representaciones con la WSST no son muy sensitivas a la ondeleta de referencia que se haya elegido. Se debe considerar también la cantidad del umbral aplicado a la representación de la CWT para poder definir el nivel base del cálculo de las frecuencias instantáneas. Un umbral muy agresivo puede ayudar a remover sonidos no deseados a costa de la posibilidad de eliminar algunas partes de la señal. Se debe tener en cuenta que tanto la CWT como la WSST son invertibles (Tary *et al.*, 2018).

3.5.4. Imágenes de la transformada WSST

Además de los parámetros anteriormente explicados, para la obtención de la representación tiempo-frecuencia usando la WSST se hizo la consideración de remuestrear ambas bases de datos a 1000Hz, esto es para poder mejorar la visualización de la complejidad numérica de los datos. Esta consideración se mantendrá cuando se hable sobre las representaciones de la WSST. Los parámetros usados para la obtención de la WSST son:

- Longitud de la ventana de análisis: 1400 muestras
- Frecuencia de muestreo: 1000 Hz
- Número de escalas por octava: 32

De la matriz obtenida se cambió su tamaño a 224x224, los valores menores a cero fueron igualados a cero y los valores de la matriz resultante fueron divididos entre su valor máximo, es decir fueron normalizados.

Se obtuvieron 11,287 imágenes patológicas y 35,418 imágenes sanas de las bases de datos disponible, distribuidas conforme a la Tabla 7:

Tabla 7. Total de imágenes de la representación usando la WSST

Base de datos	Imágenes patológicas	Imágenes Sanas
A (Clifford *et al.*, 2016)	6,318	2,632
B (Clifford *et al.*, 2016)	362	1,474
C (Clifford *et al.*, 2016)	703	194
D (Clifford *et al.*, 2016)	330	205
E (Clifford *et al.*, 2016)	1,909	28,894
F (Clifford *et al.*, 2016)	708	1,815
(Yaseen *et al.*, 2018)	957	204

La imagen de la Figura 19 corresponde a la representación tiempo-frecuencia usando la WSST de la base de datos (Yaseen *et al.*, 2018), y la Figura 20 corresponde a la base de datos (Clifford *et al.*, 2016).

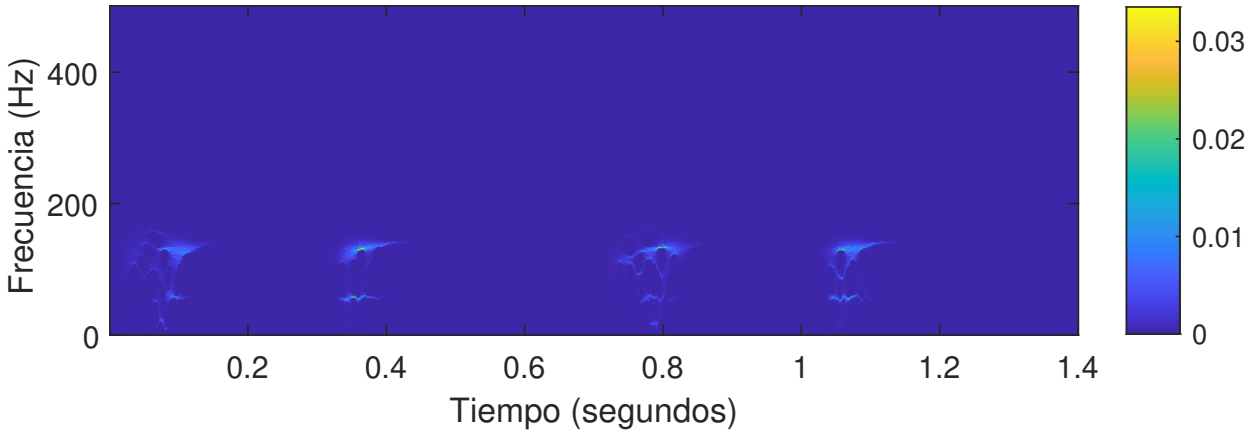


Figura 19. Representación tiempo-frecuencia de la WSST de la señal de FCG normal de la base de datos (Yaseen *et al.*, 2018), con frecuencia de muestreo de 1,000 Hz.

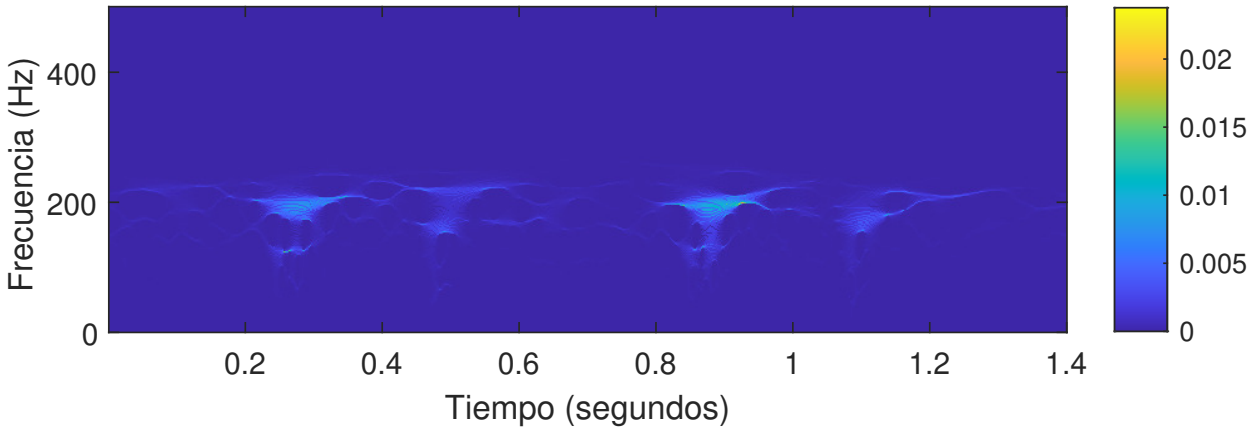


Figura 20. Representación tiempo-frecuencia usando la WSST de la señal de FCG anormal de la base de datos (Clifford *et al.*, 2016), con frecuencia de muestreo de 1,000 Hz.

3.6. Imágenes adicionales

Una vez analizadas y explicadas las tres representaciones tiempo-frecuencia que se usarán se procedió a crear más combinaciones tiempo-frecuencia, bajo la premisa de que combinando representaciones es posible mejorar la clasificación de los sonidos cardíacos.

Debido a la naturaleza de lo que se pretende hacer con la combinación de diferentes tipos de representaciones tiempo-frecuencia, se estará usando unidades de frecuencia para la visualización de las imágenes, esto es por las variaciones en las unidades usadas para medir la frecuencia entre las distintas representaciones.

El total de las imágenes obtenidas fue de 46,705 en cada combinación de las representaciones. De estas 11,287 corresponden a imágenes patológicas y 35,418 a imágenes sanas, distribuidas conforme a la Tabla 8.

Tabla 8. Total de imágenes obtenidas

Base de datos	Imágenes patológicas	Imágenes Sanas
A (Clifford *et al.*, 2016)	6,318	2,632
B (Clifford *et al.*, 2016)	362	1,474
C (Clifford *et al.*, 2016)	703	194
D (Clifford *et al.*, 2016)	330	205
E (Clifford *et al.*, 2016)	1,909	28,894
F (Clifford *et al.*, 2016)	708	1,815
(Yaseen *et al.*, 2018)	957	204

3.6.1. Imágenes WSST + espectrograma + Mel

La primera combinación de imágenes que se obtuvo fue la combinación de las tres representaciones tiempo-frecuencia (WSST + Espectrograma + Espectrograma en escala Mel). Para ello se usaron los siguientes parámetros:

- Frecuencia de muestreo para la WSST: 1000 Hz
- Longitud de la ventana de análisis: 1400 muestras
- Frecuencia de muestreo para los espectrogramas: 2000 Hz
- Longitud de la ventana de análisis para los espectrogramas: 2800 muestras

- Longitud de la ventana de Hamming para los espectrogramas: 100 muestras
- Longitud del traslape para los espectrogramas: 88 muestras
- Tamaño de la FFT para los espectrogramas: 512 muestras
- número de filtros de Mel para el espectrograma en escala Mel: 24

Las matrices obtenidas fueron procesadas de la siguiente manera:

- Espectrograma: La matriz obtenida fue analizada y se recortó para tener una dimensión de 224x224 muestras, posteriormente la matriz fue normalizada.
- Espectrograma en escala Mel: La matriz obtenida de la aplicación de 24 filtros de Mel fue de 24x224, esta matriz fue normalizada.
- WSST: De la matriz obtenida se cambió su tamaño a 224x224, los valores menores a cero fueron igualados a cero y la matriz final fue normalizada.

La relación entre las distintas representaciones tiempo-frecuencia quedaron de la siguiente manera:

- El espectrograma compone el 37.5 % de la imagen (84x224 muestras).
- El espectrograma en escala Mel compone 10.71 % de la imagen (24x224 muestras).
- La WSST compone el 51.79 % de la imagen (116x224 muestras).

Ejemplos de la combinación de estas representaciones se puede observar en las Figuras 21 y 22.

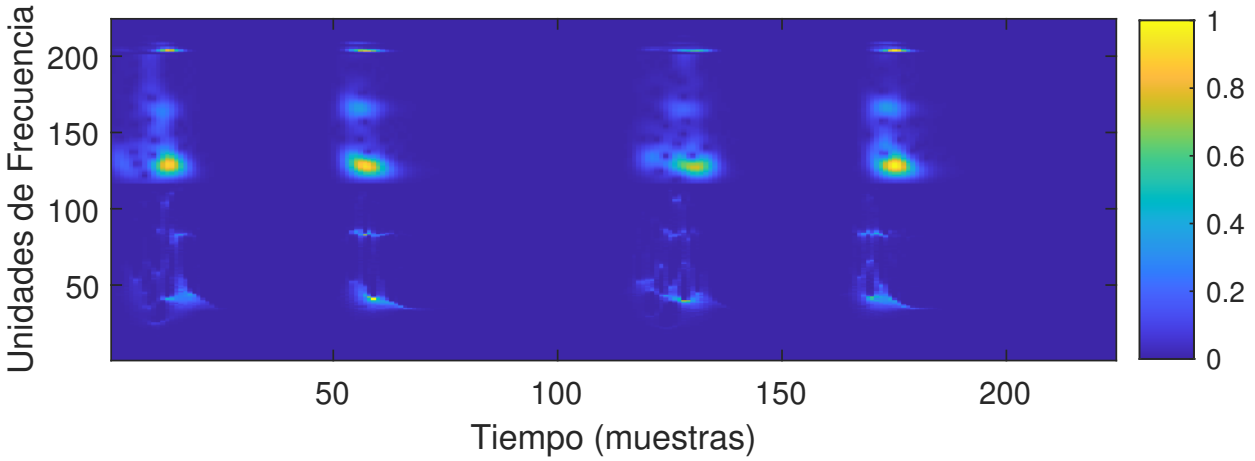


Figura 21. Representación tiempo-frecuencia usando una combinación del espectrograma, el espectrograma en escala Mel y la WSST de la señal de FCG normal tomada de la base de datos del artículo (Yaseen *et al.*, 2018).

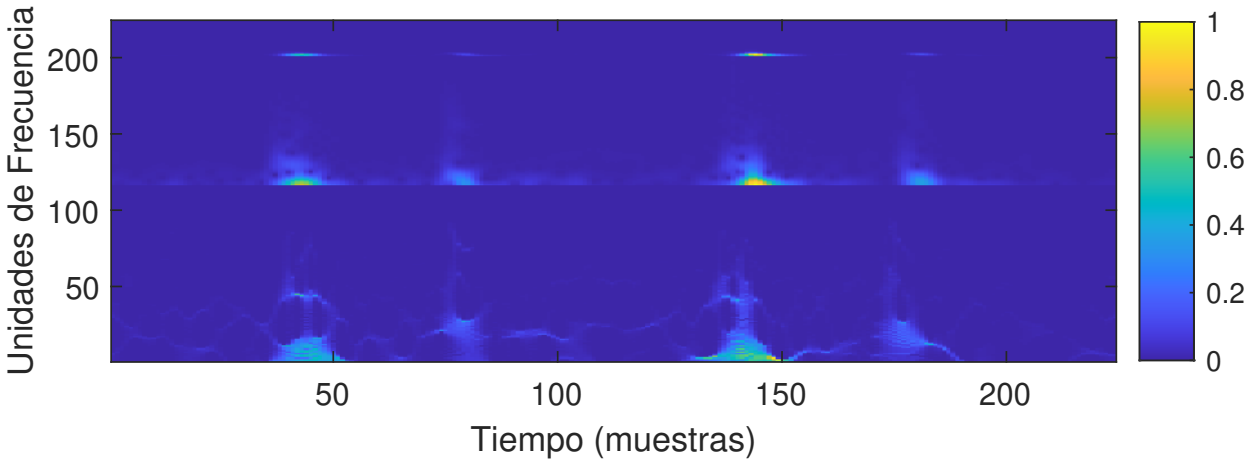


Figura 22. Representación tiempo-frecuencia usando una combinación del espectrograma, el espectrograma en escala Mel y la WSST de la señal de FCG anormal tomada de la base de datos Physionet Cinc Challenge estudiada en (Clifford *et al.*, 2016).

3.6.2. Imágenes WSST + espectrograma

La segunda combinación de imágenes que se obtuvo fue la combinación de las representaciones tiempo-frecuencia (WSST + Espectrograma). Para ello se usaron los siguientes parámetros:

- Frecuencia de muestreo para la WSST: 1000 Hz
- Longitud de la ventana de análisis: 1400 muestras
- Frecuencia de muestreo para los espectrogramas: 2000 Hz

- Longitud de la ventana de análisis para los espectrogramas: 2800 muestras
- Longitud de la ventana de Hamming para los espectrogramas: 100 muestras
- Longitud del traslape para los espectrogramas: 88 muestras
- Tamaño de la FFT para los espectrogramas: 512 muestras

Las matrices obtenidas fueron procesadas de la siguiente manera:

- Espectrograma: La matriz obtenida fue analizada y se recortó para tener una dimensión de 224x224 muestras, posteriormente la matriz fue normalizada.
- WSST: De la matriz obtenida se cambió su tamaño a 224x224, los valores menores a cero fueron igualados a cero y la matriz final fue normalizada.

La relación entre las distintas representaciones tiempo-frecuencia quedaron de la siguiente manera:

- El espectrograma compone el 48.21 % de la imagen (108x224 muestras).
- La WSST compone el 51.79 % de la imagen (116x224 muestras).

Ejemplos de las visualización de la combinación de estas representaciones se puede observar en las Figuras 23 y 24.

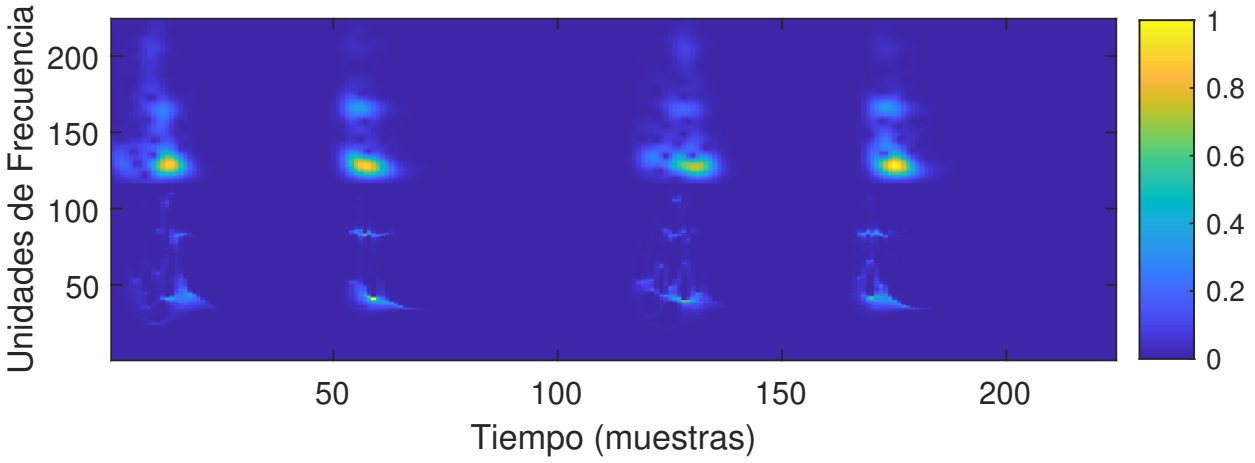


Figura 23. Representación tiempo-frecuencia usando una combinación del espectrograma y la WSST de la señal de FCG normal tomada de la base de datos del artículo (Yaseen *et al.*, 2018).

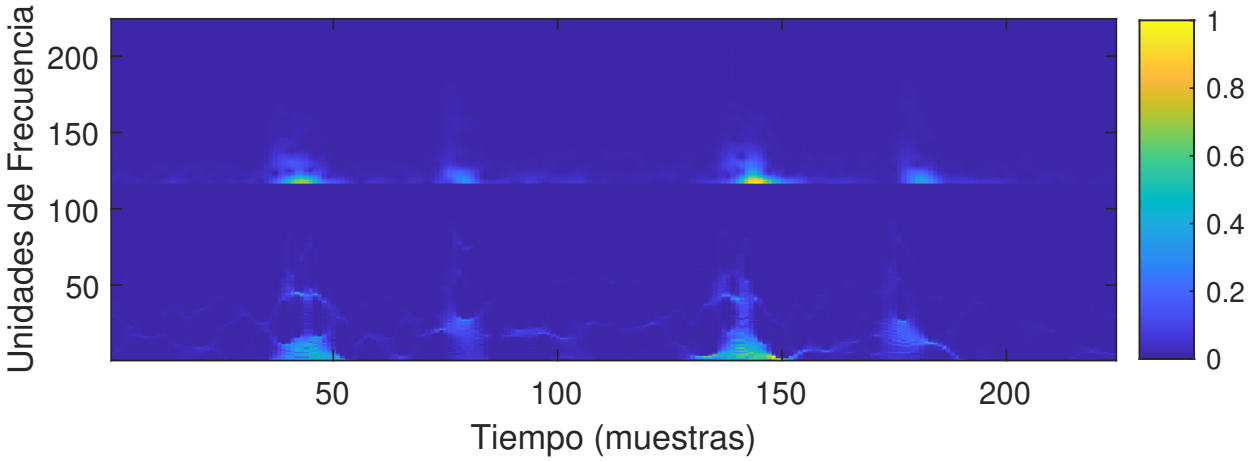


Figura 24. Representación tiempo-frecuencia usando una combinación del espectrograma y la WSST de la señal de FCG anormal tomada de la base de datos Physionet Cinc Challenge estudiada en (Clifford *et al.*, 2016).

3.6.3. Imágenes WSST + Mel

La tercera combinación de imágenes que se obtuvo fue la combinación de las representaciones tiempo-frecuencia (WSST + Espectrograma en escala Mel). Para ello se usaron los siguientes parámetros:

- Frecuencia de muestreo para la WSST: 1000 Hz
- Longitud de la ventana de análisis: 1400 muestras
- Frecuencia de muestreo para los espectrogramas: 2000 Hz
- Longitud de la ventana de análisis para los espectrogramas: 2800 muestras
- Longitud de la ventana de Hamming para los espectrogramas: 100 muestras
- Longitud del traslape para los espectrogramas: 88 muestras
- Tamaño de la FFT para los espectrogramas: 512 muestras
- número de filtros de Mel para el espectrograma en escala Mel: 24

Las matrices obtenidas fueron procesadas de la siguiente manera:

- Espectrograma en escala Mel: La matriz obtenida de la aplicación de 24 filtros de Mel fue de 24x224, esta matriz fue normalizada.

- WSST: De la matriz obtenida se cambió su tamaño a 224x224, los valores menores a cero fueron igualados a cero y la matriz final fue normalizada.

La relación entre las distintas representaciones tiempo-frecuencia quedaron de la siguiente manera:

- El espectrograma en escala Mel compone 10.71 % de la imagen (24x224 muestras).
- La WSST compone el 89.28 % de la imagen (200x224 muestras).

Ejemplos de las visualización de la combinación de estas representaciones se puede observar en las Figuras 25 y 26.

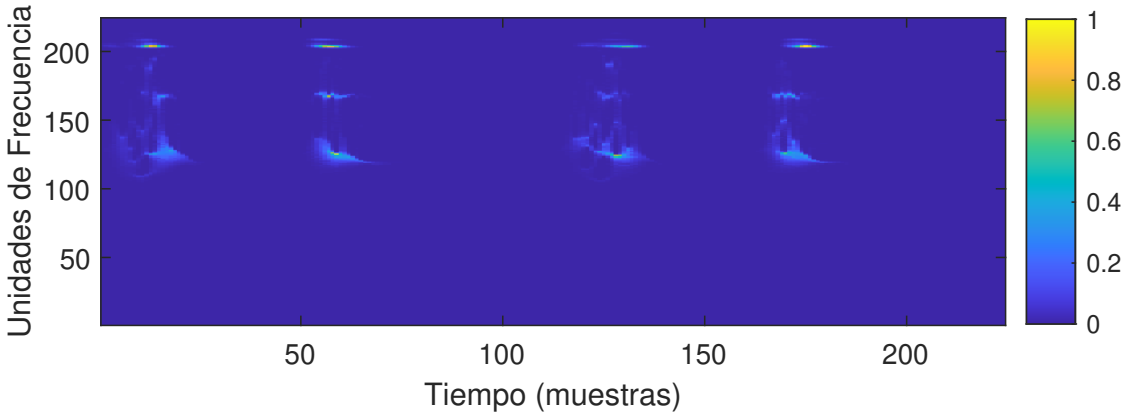


Figura 25. Representación tiempo-frecuencia usando una combinación del espectrograma en escala Mel y la WSST de la señal de FCG normal tomada de la base de datos del artículo (Yaseen *et al.*, 2018).

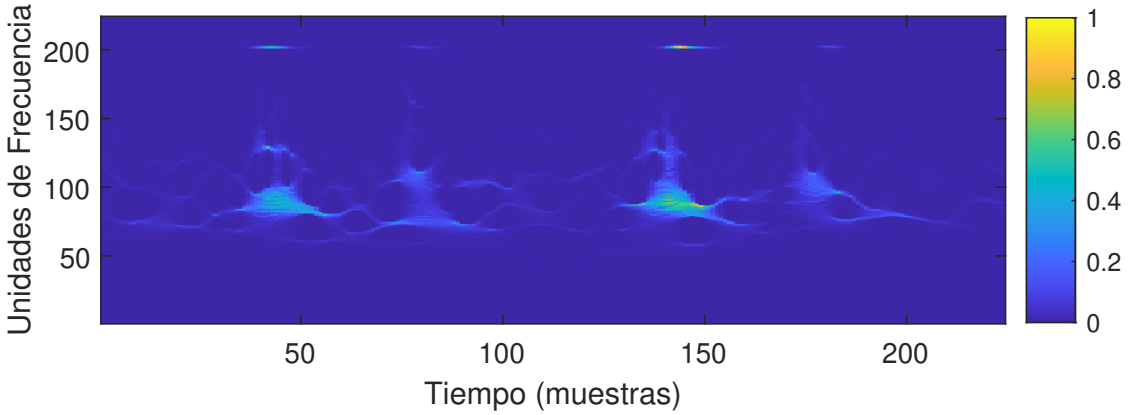


Figura 26. Representación tiempo-frecuencia usando una combinación del espectrograma en escala Mel y la WSST de la señal de FCG anormal tomada de la base de datos Physionet Cinc Challenge estudiada en (Clifford *et al.*, 2016).

Capítulo 4. Clasificación de las representaciones tiempo-frecuencia

En este capítulo se describirán los modelos de redes neuronales convolucionales que se usarán para la clasificación de las imágenes obtenidas en el Capítulo 3.

Se empezará con la definición de conceptos relacionados al aprendizaje profundo, así como describir las características estocásticas que caracterizan este tipo de algoritmos. Se definirá el tipo de aprendizaje elegido para esta tarea de clasificación, así como sus características y alternativas.

Posteriormente se definirá el concepto de redes neuronales convolucionales, así como la estructura que los conforman, describiendo las partes necesarias para entender su funcionamiento y los hiperparámetros necesarios para realizar su correcto entrenamiento.

Finalmente, se describirán los modelos de redes neuronales convolucionales que se usarán, estos modelos fueron elegidos debido a la popularidad y habilidad en tareas de clasificación, así como los resultados para cada modelo obtenidos a partir de las representaciones tiempo-frecuencia obtenidas en el Capítulo 3.

4.1. Deep Learning

El concepto de aprendizaje profundo (*Deep Learning*) puede ser visto como un sinónimo de aprendizaje automático (*Machine Learning*), por lo que es necesario revisar las diferencias entre dichas definiciones. Todo parte de la Inteligencia Artificial (*Artificial Intelligence*), que es una rama de la ciencia que se encarga de la teoría y desarrollo de sistemas computacionales que son capaces de realizar una tarea que normalmente requeriría inteligencia humana. Dentro de este campo está el aprendizaje automático, que es una rama que estudia la idea de que los sistemas pueden aprender de datos para identificar patrones y tomar decisiones con la mínima intervención humana. Y a su vez, dentro de este campo está el aprendizaje profundo que es una combinación de capas de procesamiento no lineal para la extracción y procesamiento de variables para la toma de decisiones por cuenta propia (Rosebrock, 2017).

El aprendizaje de máquinas en la práctica es la ejecución de un algoritmo en un conjunto de datos para obtener un modelo. Este modelo después puede ser evaluado en datos que no se hayan visto previamente durante su entrenamiento o bien, se pueden obtener predicciones sobre datos nuevos. Se pueden definir estos conceptos como:

- Algoritmo: Procedimiento ejecutado en un conjunto de datos para la obtención de un modelo.
- Modelo: Estructura de datos y coeficientes usados para hacer predicciones sobre datos.

Algunos algoritmos de aprendizaje de máquinas son deterministas. Esto quiere decir que si se tiene el mismo conjunto de datos, el modelo aprenderá el mismo modelo cada vez que se ejecute. Ejemplos de este tipo de comportamiento son la regresión lineal y los algoritmos de regresión logística. Otros algoritmos no son deterministas, en su lugar son estocásticos. Esto quiere decir que a su comportamiento se le incluye un elemento de aleatoriedad. Esto no quiere decir que el aprendizaje del algoritmo es aleatorio en su completo, sino que pequeñas decisiones específicas del algoritmo durante el entrenamiento pueden variar de manera aleatoria (Rosebrock, 2017).

Lo anterior se traduce a que cada vez que un algoritmo de aprendizaje de máquina estocástico es ejecutado sobre los mismos datos producirá un modelo diferente, esto produce una variación en las predicciones, en el error, la precisión y el rendimiento del algoritmo. Este elemento de aleatoriedad en algunas decisiones del algoritmo puede mejorar el rendimiento en problemas de alta complejidad (Rosebrock, 2017).

Existen diferentes maneras en que un algoritmo encontrará la solución para un problema, esto depende de la interacción de dicho algoritmo con los datos del problema. En general se puede decir que existen dos tipos de aprendizaje automático que los algoritmos pueden seguir dependiendo de los datos que se le proporcionen en la entrada (Aggarwal, 2018), mostrados por la Figura 27.

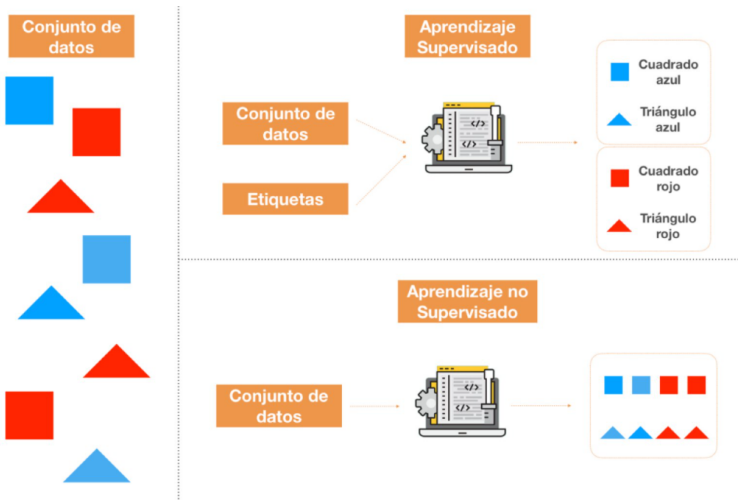


Figura 27. Tipos de aprendizaje automático, imagen obtenida de: `https://aprendeia.com/diferencia-entre-aprendizaje-supervisado-y-no-supervisado/`.

- **El aprendizaje supervisado** significa que tenemos datos en la entrada y etiquetas en la salida. No es factible escribir el código que relacione las entradas con las salidas por lo que usamos algoritmos de aprendizaje automático para que aprenda a predecir las salidas en base a las entradas. Esto es llamado una función de aproximación, en la que se busca o aprende una función que pueda relacionar las entradas con las salidas.
- **El aprendizaje no supervisado** significa que tenemos datos en la entrada pero no etiquetas en la salida. Esto obligará a la red a buscar reglas generales para realizar la tarea de la clasificación.

Debido a que se tiene un problema de clasificación se hará uso del aprendizaje supervisado en este trabajo. Además, las bases de datos que se tienen disponibles ya cuentan con un etiquetado previo sobre el estado (patológico o sano) de los sonidos cardíacos.

4.2. Redes neuronales convolucionales

En la redes neuronales artificiales tradicionales cada neurona en la capa de entrada está conectada a cada neurona de la capa siguiente, a esto se le llama una capa densa (*fully-connected*). Pero en las redes neuronales convolucionales (*CNN*) no se suelen

ocupar capas densas sino hasta las últimas capas de la red.

Las capas convolucionales son el elemento más importante de las CNN. Una convolución es la aplicación de un filtro a una entrada que resulta en una activación. La aplicación sistemática del mismo filtro en la entrada resulta en un mapa de activaciones llamado mapa de características, que indica la localización y magnitud de una característica detectada en la entrada. Lo innovador de las CNN es la habilidad de aprender una gran cantidad de filtros en paralelo que son específicos para el problema que se está resolviendo. Lo que deriva en la detección de características muy específicas para el conjunto de datos en la entrada (Rosebrock, 2017).

De manera más profunda dentro del contexto de las CNN una convolución es una operación lineal que involucra la multiplicación de un conjunto de pesos con la entrada. Esta multiplicación es realizada entre un arreglo de datos de entrada y un arreglo de pesos con dos dimensiones llamado filtro o kernel. El filtro es de menor tamaño que los datos de entrada. La operación de la multiplicación es un producto punto, entre el filtro y una porción de los datos de entrada del tamaño del filtro. La razón por la que se ocupa un filtro de menor tamaño que los datos de entrada es para poder aplicar el filtro varias veces a los datos de entrada en diferentes puntos. Específicamente el filtro es aplicado de manera sistemática a cada parte de los datos de entrada del tamaño del filtro de izquierda a derecha, y de arriba hacia abajo (Rosebrock, 2017).

Esto nos permite la detección de características concretas en los datos de entrada, esta habilidad es llamada (*translation invariance*), que nos dice que estamos más interesados en la presencia de una característica más que en su posición. La salida de la multiplicación del filtro con la entrada una vez es un valor. Conforme se vaya aplicando el filtro se irá construyendo un arreglo de dos dimensiones que representará los valores de la entrada filtrados. Este arreglo bidimensional es llamado un mapa de características (Rosebrock, 2017; Aggarwal, 2018).

4.3. Estructura de una red neuronal convolucional

Las redes neuronales convolucionales son modelos complejos y están compuestas de diferentes parámetros Figura 28, que serán explicados a continuación.

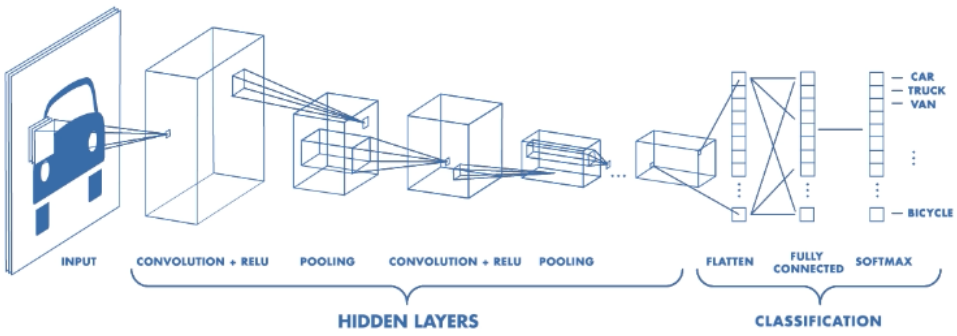


Figura 28. Arquitectura de una red neuronal convolucional,imagen tomada de: `https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/`.

4.3.1. Nodos y capas

Las redes neuronales tienen dos hiperparámetros para controlar la arquitectura o topología de la red, el número de capas y el número de nodos en cada capa escondida. La forma más habitual de poder conseguir el mejor número de capas y nodos es por medio de una experimentación de forma sistemática. Ya que cada problema para resolver tiene un grado único de complejidad.

Un nodo, también llamado neurona o perceptrón, es una unidad computacional que tiene una o más conexiones en la entrada, una función de transferencia que modifica la entrada y una conexión para la salida. Una red de una capa, tiene sólo una capa de nodos, entonces este concepto puede ser extendido a redes de varias capas, dando lugar al concepto de un perceptrón multicapa (*MLP*), que se describe como una cascada de percerptrones de una capa (Aggarwal, 2018).

Podemos resumir los tipos de capas en un MLP como:

- Capa de entrada: se encuentran las variables de entrada, en ocasiones se le llama como capa visible.
- Capas ocultas: capas con nodos que se encuentran entre la capa de entrada y la capa de salida. Pueden haber una o más de estas capas.
- Capa de salida: Una capa con nodos que producen las variables de salida.

Estos términos dan paso a conceptos más técnicos como:

- Tamaño (*Size*): El número de nodos en el modelo.
- Ancho (*Width*): El número de nodos en una capa específica.
- Profundidad (*Depth*): El número de capas de la red neuronal.
- Capacidad (*Capacity*): el tipo de estructura o funciones que pueden ser aprendidas por la configuración de la red.
- Arquitectura (*Architecture*): El arreglo específico de las capas y los nodos en la red.

No hay un paradigma concreto para estimar de una manera óptima o semi óptima la estructura que debe tener una red. De manera habitual se tienen las siguientes soluciones a este problema.

1. Prueba y error. Es el camino más primitivo, pero si se usa una experimentación sistemática se puede llegar a tener un modelo que tenga un rendimiento deseado.
2. Aproximación heurística. Cuyo objetivo principal es encontrar una fórmula que estime el número de nodos en las capas escondidas en función del número de entradas y de salidas. El estimado puede tomar la forma de una topología exacta o de un rango de topologías que deberían ser probadas. Como el algoritmo genético de optimización bayesiana.

3. Búsqueda exhaustiva. Se puede probar todas las combinaciones posibles en el número de capas y nodos, podría ser factible para redes pequeñas. Pero puede ser muy desafiante para redes más grandes con bases de datos muy extensas. Es una solución válida si se tiene el tiempo y los recursos.

4.3.2. Tipos de capas

Dentro de las capas escondidas, existen diferentes tipos de capas que cumplen con funciones específicas y que deben ser entendidas y usadas de manera consciente (Aggarwal, 2018; Rosebrock, 2017), estas son:

- Capa convolucional: La capa convolucional es el elemento principal dentro de una CNN. Esta capa consiste en un número establecido de filtros, donde cada filtro tiene un ancho y una altura, suelen ser de forma cuadrada. Estos filtros son pequeños (hablando en términos de su dimensión espacial).
- Capa de activación: Después de cada capa convolucional se aplica una función de activación no lineal, como ReLU, ELU u otras como una Leaky ReLU. Las capas de activación no son técnicamente capas (debido al hecho de que ningún parámetro o peso es aprendido dentro de la capa de activación), y en algunas ocasiones son omitidas de los diagramas de la arquitectura de la red ya que se asume que después de cada capa convolucional debe serguir una capa de activación.
- Capa densa: Una capa densa es una capa en la que cada una de sus neuronas tienen conexiones con todas las neuronas de la capa anterior, es un estándar en las redes neuronales convencionales. Las capas densas en el caso de las redes convolucionales se encuentran siempre al final de la red. Se pueden aplicar una o más capas densas al modelo.
- Normalización de lotes (*Batch normalization*): El concepto apareció por primera vez en el documento de 2015 escrito por Ioffe y Szgedy *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* , las capas de normalización de lotes son usadas para normalizar las activaciones dado un volumen de entradas antes de pasarlas a la siguiente capa. La normalización

de lotes ha probado ser un método efectivo para reducir el número de épocas que toma entrenar una red neuronal.

- Dropout: El *Dropout* es una forma de regularización que busca prevenir el sobre entrenamiento al incrementar la precisión de la prueba, quizá a expensas de la precisión del entrenamiento. Funciona con cada mini-lote en cada conjunto de entrenamiento, las capas de *Dropout* tienen una probabilidad de desconectar entradas de manera aleatoria de la capa anterior a la capa siguiente.
- Capas de agrupación (*Pooling layers*): Las capas de agrupación cumplen la función de reducir de manera progresiva el tamaño espacial (ancho y altura) de los datos de entrada. Haciendo esto se logra reducir la cantidad de parámetros y requerimientos computacional en la red, además de ayudarnos con el sobre entrenamiento. Dentro de los métodos de pooling más usados están:
  - Average Pooling: Calcula el promedio de cada valor de la porción en el mapa de características.
  - Maximun Pooling: Calcula el valor máximo de cada valor de la porción del mapa de características.

4.3.3. Lotes y épocas (*Batches and epochs*)

Los lotes y las épocas son hiperparámetros de suma importancia dentro de las redes neuronales, sirven para controlar aspectos relacionados al entrenamiento y rendimiento de la red. Un correcto manejo de estos parámetros puede mejorar el tiempo de entrenamiento así como la precisión del modelo. Por lo que es importante dar una definición (Aggarwal, 2018; Rosebrock, 2017).

Comenzaremos definiendo qué es una muestra. Una muestra se puede definir de manera sencilla como una fila de datos. Contiene información que será procesada por la red neuronal, y da como resultado una salida que nos servirá para comparar la capacidad de predicción y calcular el error de la predicción. Un conjunto de datos para entrenamiento está compuesto de muchas filas de datos. Otros nombres que se le

puede dar son: instancia, una observación, un vector de entrada o un vector de características.

El tamaño del lote (*Batch size*) es un hiperparámetro que nos indica el número de muestras con el que la red neuronal trabaja antes de actualizar sus parámetros internos. Se puede ver como un ciclo *for* que itera sobre una o más muestras y realizando predicciones, al final del lote las predicciones son comparadas con las salidas esperadas y se calcula el error de predicción. Usando el valor del error de predicción se actualiza y mejora el modelo (Rosebrock, 2017).

Un conjunto de datos puede ser divido en uno o más lotes. Cuando todas las muestras de entrenamiento se usan para hacer un sólo lote, el algoritmo se llama *batch gradient descent*. Cuando el tamaño del lote es de una sola muestra, el algoritmo es llamado *stochastic gradient descent*. Cuando el tamaño del lote es más de una muestra pero menor al tamaño total del número de muestras el algoritmo se llama *mini-batch gradient descent* (Aggarwal, 2018; Rosebrock, 2017).

Podemos resumir la clasificación del tamaño de los lotes en:

- *Batch Gradient Descent*: El tamaño del lote es igual al tamaño del conjunto de entrenamiento.
- *Stochastic Gradient Descent*: El tamaño del lote es igual a una muestra.
- *Mini-Batch Gradient Descent*: El tamaño del lotes es mayor a una muestra pero menor al número total de muestras.

Por otra parte, el número de épocas es un hiperparámetro que describe el número de veces que el algoritmo aprenderá con un conjunto de datos. Una época nos dice que cada muestra en el conjunto de datos ha tenido oportunidad de mejorar los parámetros internos del modelo. Las épocas pueden estar compuestas de uno o más lotes, y se pueden visualizar como un ciclo *for* en el que se iteran las muestras de tamaño del lote establecido para mejorar el modelo.

Se puede ejemplificar el rol que tienen estos dos hiperparámetros si se imagina que se tienen 300 muestras, se elige un tamaño de lote de 10 y 20 épocas. Esto significa que el conjunto de datos será divido en 30 lotes, cada uno con 10 muestras. Los parámetros del modelo serán actualizados después de cada lote de 10 muestras. Esto significa que cada época tiene 30 lotes o 30 actualizaciones del modelo. Si se tienen 20 épocas el modelo será expuesto al conjunto de datos 20 veces, haciendo un total de 600 lotes durante el entrenamiento.

4.3.4. Funciones de activación

La importancia de las funciones de activación viene de la propia naturaleza de la red neuronal, puesto que cada neurona realiza una operación lineal sencilla, al momento de la obtención de los resultados se tiene otro resultado lineal, para evitar ese problema se introduce una función no lineal en las capas escondidas (Aggarwal, 2018; Rosebrock, 2017), estas funciones pueden ser:

- Función identidad: Esta función también conocida como función lineal, permite que la entrada sea igual a la salida, por lo que se dice que a una red neuronal de varias capas a la que se le aplica un función lineal se vuelve una regresión lineal. Por lo tanto esta función se ocupa cuando se desea emplear una regresión lineal, como en la imagen 29 y determinada por la ecuación 14.

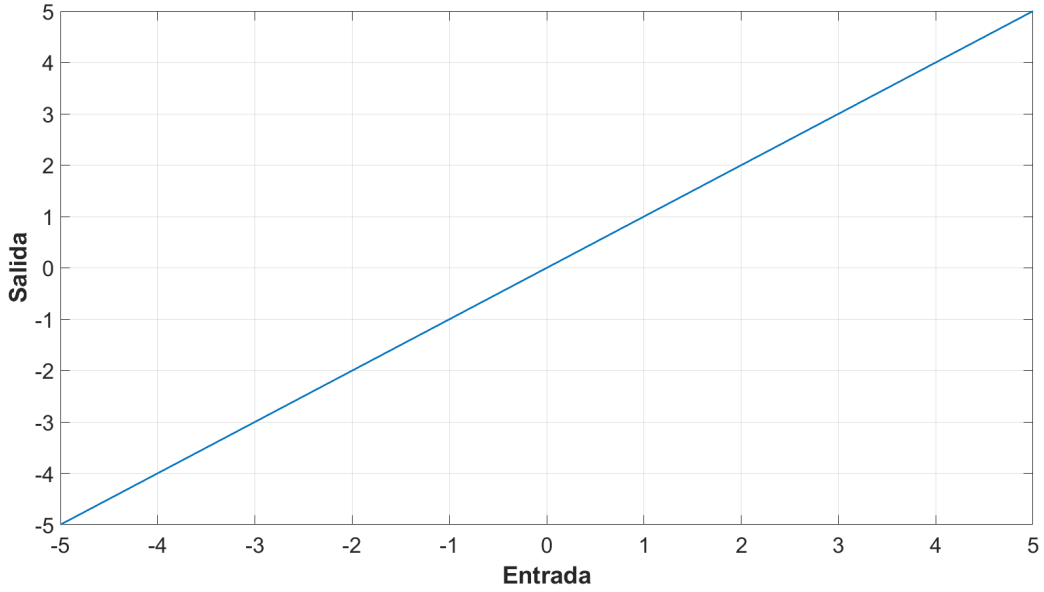


Figura 29. Diagrama de la función identidad.

$$y = x, \tag{14}$$

donde y es la salida de la neurona y x es la entrada de la neurona.

- Función escalón: Esta función también conocida como función umbral, indica que si la entrada es menor que cero la neurona va a ser dará como salida cero, en caso contrario la salida será uno. Esta función es útil para clasificación con salidas categóricas, mostrada en la imagen 30, y determinada por la ecuación 15.

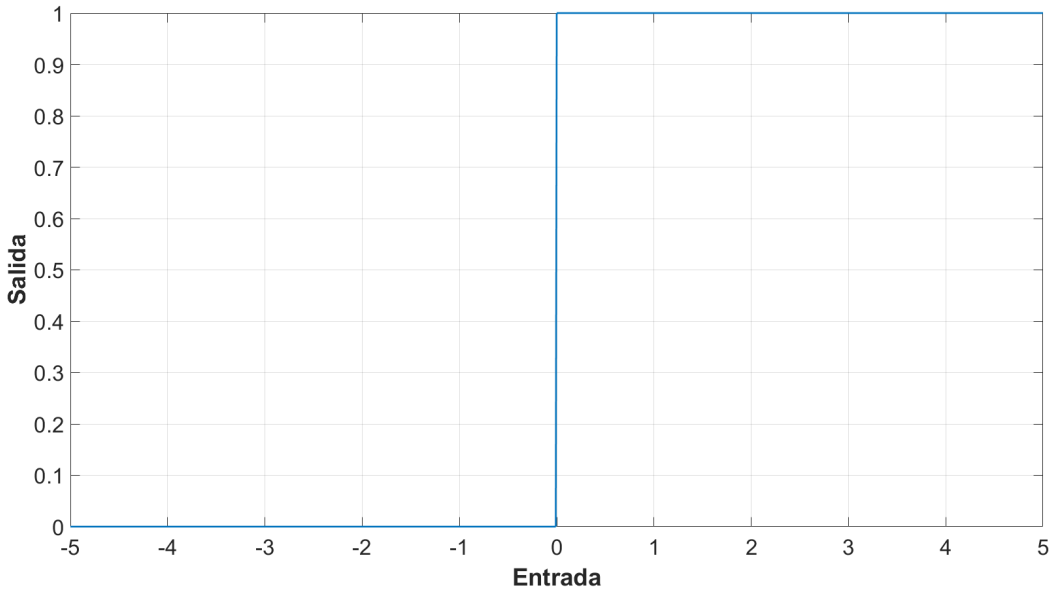


Figura 30. Diagrama de la función escalón.

$$y = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases} \tag{15}$$

donde x es la entrada de la neurona.

- Función sigmoide: Esta función también conocida como logística, está en un rango de valores de salida entre cero y uno, por lo que la salida es interpretada como una probabilidad. Si se evalúa la función con valores de entrada muy negativos, la función será igual a cero, si se evalúa en cero la función dará 0.5 y en valores altos su valor sería aproximadamente 1, como mostrado en la Figura 31, y determinada por la Ecuación 16.

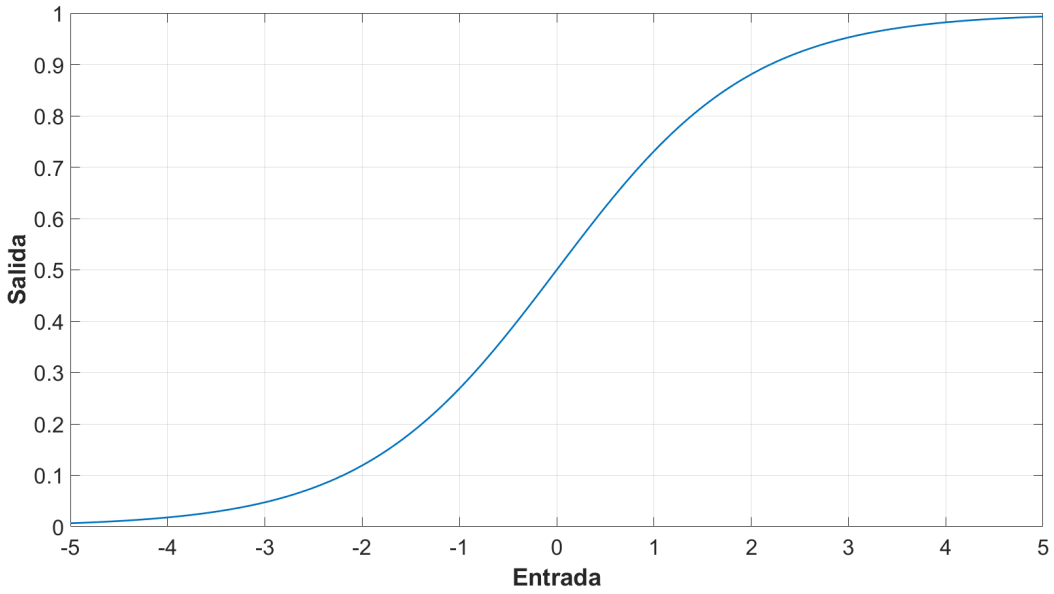


Figura 31. Diagrama de la función sigmoide.

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{16}$$

donde σ es la salida de la neurona y x es la entrada de la neurona.

- Función TanH: La función tangente hiperbólica tiene un rango de valores de salida entre -1 y 1. Se dice que esta función es un escalamiento de la función logística, por lo que a pesar que está centrada tienen un problema similar a la sigmoide debido al problema de desaparición del gradiente, que se da cuando en el entrenamiento se genera un error con el algoritmo de propagación hacia atrás y debido a esto el error se va propagando entre las capas, por lo que en cada iteración toma un valor pequeño y la red no puede obtener un buen aprendizaje, como en la Figura 32, y determinada por la Ecuación 17.

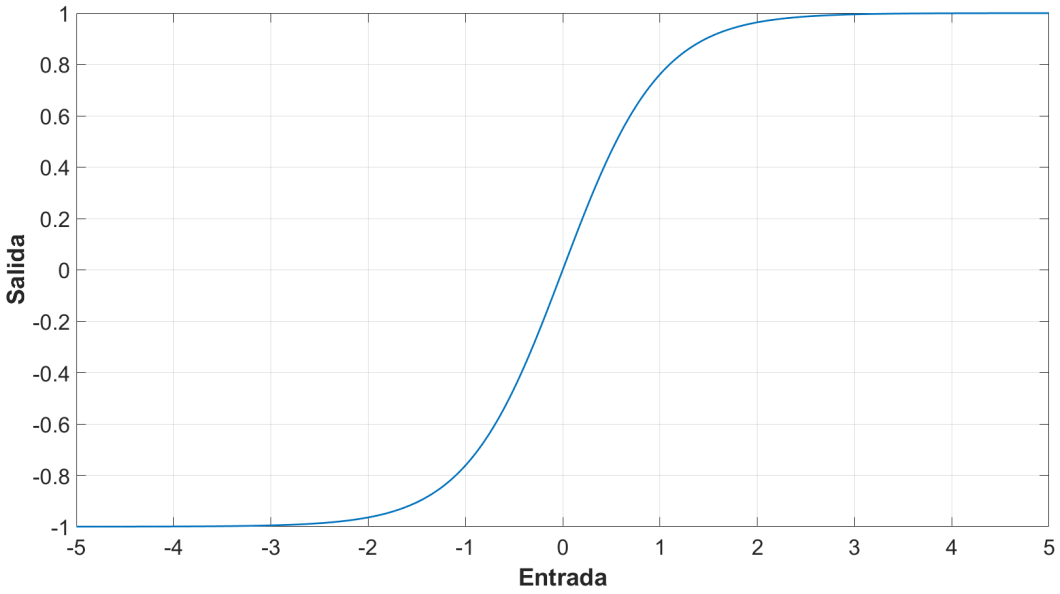


Figura 32. Diagrama de la función TanH.

$$tanh(X) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \tag{17}$$

donde x es la entrada de la neurona.

- Función ReLu: Esta es la función más utilizada debido a su facilidad para facilitar el entrenamiento en las redes neuronales. Si a esta función se le da valores de entrada muy negativos el resultado es cero pero si se le da valores positivos quedan igual, como en la Figura 33 y mostrada por la Ecuación 18.

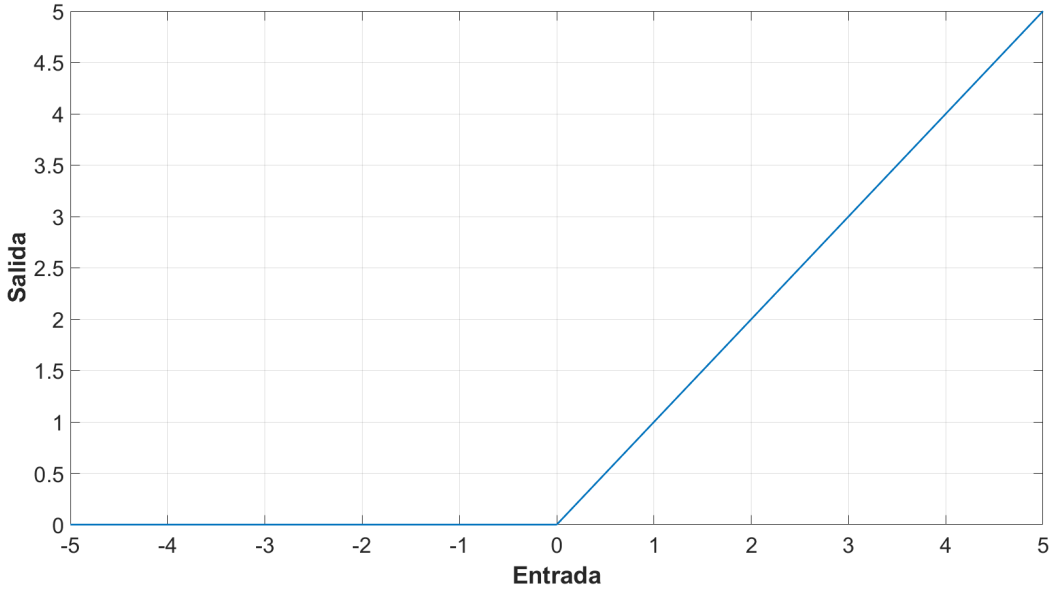


Figura 33. Diagrama de la función ReLU.

$$y = \begin{cases} 0 & \text{si } x < 0, \\ x & \text{si } x \geq 0, \end{cases} \tag{18}$$

Los modelos de las redes neuronales que se usarán, ocupan la función de activación *ReLu*.

4.3.5. *Descenso del gradiente (*Gradient Descent*)*

El descenso del gradiente es un algoritmo de optimización que nos permite encontrar los mínimos de una función, en el caso de las redes neuronales nos permite encontrar el punto mínimo en la relación de los pesos con las pérdidas del sistema (Aggarwal, 2018; Rosebrock, 2017). La Figura 34 ilustra una visualización del proceso del descenso del gradiente. Se observa la presencia de un mínimo local y global.

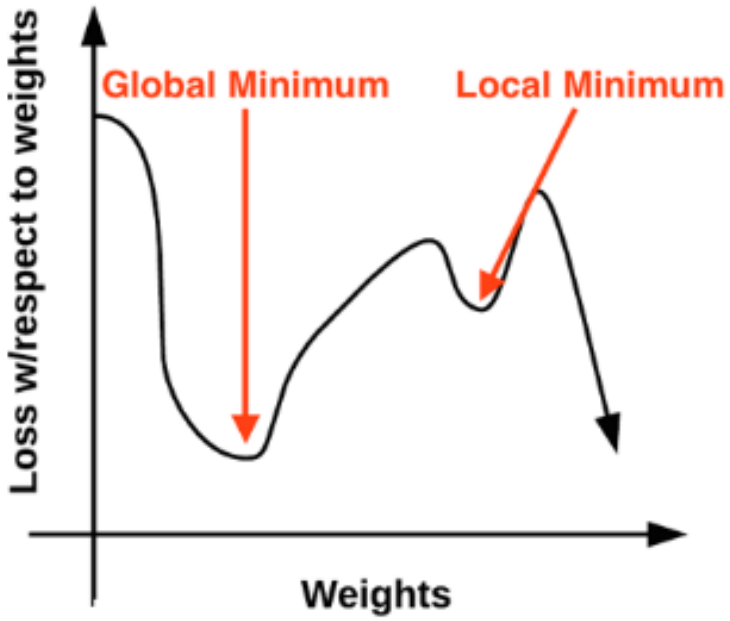


Figura 34. Visualización del descenso del gradiente (Rosebrock, 2017).

El algoritmo del descenso del gradiente realiza el cálculo del gradiente de una función con valores específicos en su entrada. Siendo el gradiente un vector de derivadas parciales de una función con respecto a sus variables de entrada. Este proceso es repetido hasta que se encuentra el mínimo de la función objetivo, un número máximo posible de soluciones o alguna otra condición para detenerse.

El algoritmo del descenso del gradiente puede ser adaptado para minimizar la función de pérdida de un modelo predictivo, a esta adaptación se le conoce como *stochastic gradient descent*. Este algoritmo puede ser utilizado para entrenar y optimizar muchos tipos de modelos, como regresión lineal y regresión logística. El reto al usar este algoritmo es el cálculo del gradiente en nodos dentro de las capas escondidas de la red.

La función de pérdida representa el error del modelo o la función de error, los pesos son las variables de la función y los gradientes de la función de error con respecto a los pesos son referidos como errores del gradiente.

En resumen, el algoritmo de optimización *stochastic gradient descent* puede ser usado durante el entrenamiento de las redes neuronales. Este algoritmo requiere del

cálculo de gradientes para cada variable en el modelo para poder producir nuevas variables.

4.3.6. *Tasa de aprendizaje*

La tasa de aprendizaje es un hiperparámetro que controla el cambio del modelo en respuesta al error estimado cada vez que los pesos son actualizados. La elección de un correcto valor para la tasa de aprendizaje es difícil debido a que un valor muy pequeño resulta en tiempos de entrenamiento más largos, además de que podría quedarse perdido, y valores muy grandes podrían tener valores sub-óptimos debido al rápido e inestable entrenamiento.

4.4. Modelos de redes neuronales

En este trabajo se usaron tres modelos de redes neuronales convolucionales que han probado ser de gran utilidad para tareas de clasificación de imágenes, así como tareas de clasificación de patologías cardíacas.

Para el entrenamiento de los modelos se seleccionaron al azar 10,200 imágenes de cada representación tiempo-frecuencia obtenidas en el Capítulo 3, este número total de imágenes fue seleccionado para poder entrenar de manera eficiente los modelos de redes neuronales convolucionales. Considerando las limitaciones del equipo con el que se entrenó, el equipo usado fue una combinación de equipo local y equipo virtual usando el ambiente de *Google Colaboratory* (Google, 2021).

El equipo local usado tiene las características siguientes:

- Procesador Ryzen 7 4800H.
- Memoria RAM de 16 GB DDR4@3200 MHz.
- Tarjeta gráfica Nvidia GTX 1660 Ti de 6 GB.

El equipo de *Google Colaboratory* varía dependiendo del tipo de suscripción que se tenga a la plataforma, en un inicio se usó la versión gratuita y posteriormente se compró una suscripción mensual después, debido al cambio de equipos que se prestaban en el ambiente gratuito.

Las características del ambiente gratuito son:

- Procesador Intel(R) Xeon(R) @ 2.3GHz.
- Memoria RAM de 14 GB.
- Tarjeta gráfica Tesla K80 de 12 GB.

Las características del ambiente de pago son:

- Procesador Intel(R) Xeon(R) @ 2.3GHz.
- Memoria RAM de 26 GB.
- Tarjeta gráfica Tesla P100 de 16Gb.

4.4.1. AlexNet

La red de AlexNet está compuesta 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 sigmoide por la ReLu. Originalmente fue entrenada con 1.2 millones de imágenes, tomando 6 días para ser entrenada completamente, para evitar el sobreentrenamiento se implementaron técnicas de *Data augmentation* y capas de Dropout. Participando en el reto de *ImageNet Large Scale Visual Recognition Challenge 2012* ganó por un gran margen, rompiendo así el paradigma de que las características tenían que ser hechas a mano (Wei, 2019).

La arquitectura del modelo está descrita en la Figura 37.

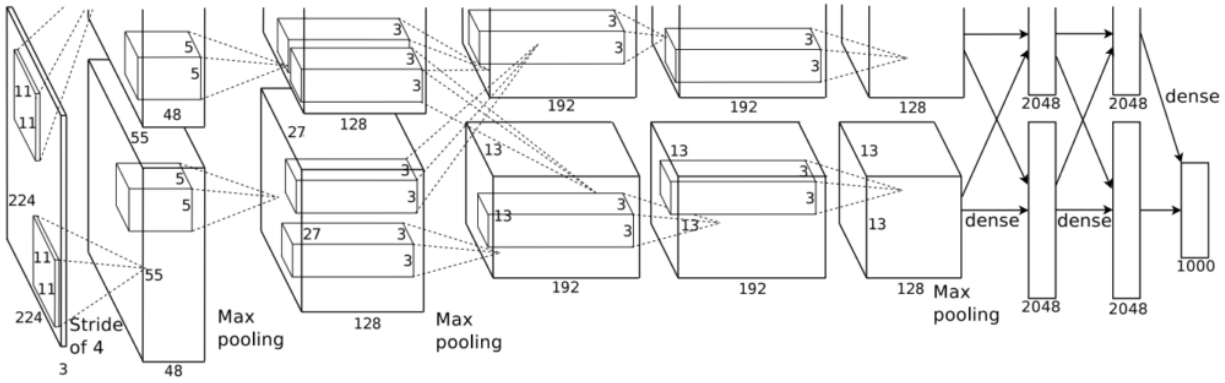


Figura 35. Arquitectura de la red Alexnet, imagen tomada de (Krizhevsky *et al.*, 2012).

La composición de cada capa de la estructura de la red usada está detallada en la Figura 36, teniendo un total de 46,737,474 parámetros.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 54, 54, 96)	11712
batch_normalization (BatchNo	(None, 54, 54, 96)	384
max_pooling2d (MaxPooling2D)	(None, 26, 26, 96)	0
conv2d_1 (Conv2D)	(None, 26, 26, 256)	614656
batch_normalization_1 (Batch	(None, 26, 26, 256)	1024
max_pooling2d_1 (MaxPooling2	(None, 12, 12, 256)	0
conv2d_2 (Conv2D)	(None, 12, 12, 384)	885120
batch_normalization_2 (Batch	(None, 12, 12, 384)	1536
conv2d_3 (Conv2D)	(None, 12, 12, 384)	1327488
batch_normalization_3 (Batch	(None, 12, 12, 384)	1536
conv2d_4 (Conv2D)	(None, 12, 12, 256)	884992
batch_normalization_4 (Batch	(None, 12, 12, 256)	1024
max_pooling2d_2 (MaxPooling2	(None, 5, 5, 256)	0
flatten (Flatten)	(None, 6400)	0
dense (Dense)	(None, 4096)	26218496
dropout (Dropout)	(None, 4096)	0
dense_1 (Dense)	(None, 4096)	16781312
dropout_1 (Dropout)	(None, 4096)	0
dense_2 (Dense)	(None, 2)	8194

Figura 36. Estructura de las capas de la red AlexNet implementada.

Durante la fase del entrenamiento, como se explicó anteriormente, se usaron 10, 200 imágenes, de las cuales 5, 100 imágenes tienen etiqueta "patológica", y 5, 100 imágenes tienen etiqueta "sana". Los hiperparámetros usados para el entrenamiento son:

- Optimizador *Stochastic Gradient descent* (SGD)
- Tasa de aprendizaje de 0.008
- Función de pérdida: *categorical_crossentropy*
- 150 épocas
- Un tamaño de lote de 64 muestras

La determinación de los hiperparámetros usados para el entrenamiento de la red se realizó tomando como base los hiperparámetros usados en el trabajo (Krizhevsky *et al.*, 2012) para ser ajustados a nuestro problema específico con una evaluación sistemática de los valores más favorables.

De las 10,200 imágenes se hizo uso del 80 % (8,160 imágenes) para el entrenamiento y 20 % (2040 imágenes) para la validación. Usando la máquina virtual de *Google Colaboratory* de paga, el modelo tomaba en ser entrenado seis segundos por época, tomando así 15 minutos en ser entrenada. Esto fue repetido para las seis representaciones tiempo-frecuencia obtenidas. Dando los resultados de la Tabla 9, tomando en consideración lo que será explicado en el Capítulo 5.4 [1].

Tabla 9. Resultados de la validación del modelo AlexNet

Representación tiempo-frecuencia	Precisión
Espectrograma	82.745 %
Espectrograma escala Mel	81.47 %
WSST	77.745 %
WSST + Espectrograma + Mel	99.95 %
WSST + Espectrograma	99.901 %
WSST + Mel	98.774 %

4.4.2. VGG16

La red VGG16 es una arquitectura usada para ganar la competencia ILSVR (Imagenet) en 2014. Fue introducida por primera vez en el artículo (Simonyan y Zisserman, 2015), y fue usada para el reconocimiento de objetos. La abreviación VGG significa *Visual Geometry Group*, el cuál es un grupo de investigadores de la Universidad de Oxford, así como el 16 significa que es una estructura de 16 capas.

[1] Se ha excluido del entrenamiento la base *E*

Las primeras trece capas de la red VGG16 son convolucionales y las últimas tres capas son densas. La arquitectura del modelo está descrita en la Figura 37.

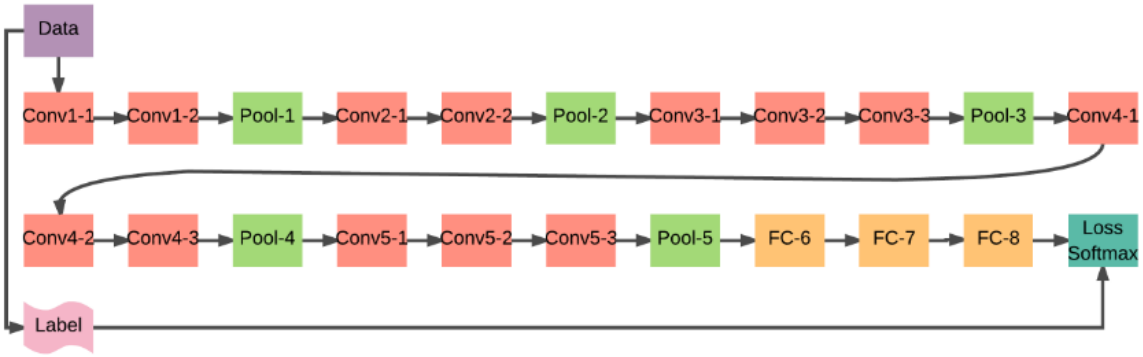


Figura 37. Arquitectura de la red VGG16, imagen tomada de (Qassim *et al.*, 2018).

La composición de cada capa de la estructura de la red usada está detallada en la Figura 40, teniendo un total de 134,267,586 parámetros.

Layer (type)	Output Shape	Param #
conv2d_18 (Conv2D)	(None, 224, 224, 64)	640
conv2d_19 (Conv2D)	(None, 224, 224, 64)	36928
max_pooling2d_8 (MaxPooling2	(None, 112, 112, 64)	0
conv2d_20 (Conv2D)	(None, 112, 112, 128)	73856
conv2d_21 (Conv2D)	(None, 112, 112, 128)	147584
max_pooling2d_9 (MaxPooling2	(None, 56, 56, 128)	0
conv2d_22 (Conv2D)	(None, 56, 56, 256)	295168
conv2d_23 (Conv2D)	(None, 56, 56, 256)	590080
conv2d_24 (Conv2D)	(None, 56, 56, 256)	590080
max_pooling2d_10 (MaxPooling	(None, 28, 28, 256)	0
conv2d_25 (Conv2D)	(None, 28, 28, 512)	1180160
conv2d_26 (Conv2D)	(None, 28, 28, 512)	2359808
conv2d_27 (Conv2D)	(None, 28, 28, 512)	2359808
max_pooling2d_11 (MaxPooling	(None, 14, 14, 512)	0
conv2d_28 (Conv2D)	(None, 14, 14, 512)	2359808
conv2d_29 (Conv2D)	(None, 14, 14, 512)	2359808
conv2d_30 (Conv2D)	(None, 14, 14, 512)	2359808
max_pooling2d_12 (MaxPooling	(None, 7, 7, 512)	0
flatten_2 (Flatten)	(None, 25088)	0
dense_6 (Dense)	(None, 4096)	102764544
dense_7 (Dense)	(None, 4096)	16781312
dense_8 (Dense)	(None, 2)	8194

Figura 38. Estructura de las capas de la red VGG16 implementada.

Las imágenes usadas durante el entrenamiento de esta red son las mismas 10,200 imágenes usadas en el modelo de AlexNet. Los hiperparámetros usados para el entrenamiento de la red fueron mantenidos para tener una mejor comparación del rendimiento de los modelos. Estos parámetros son:

- Optimizador *Stochastic Gradient descent* (SGD)
- Tasa de aprendizaje de 0.008
- Función de pérdida: *categorical_crossentropy*
- 150 épocas
- Un tamaño de lote de 64 muestras

De las 10,200 imágenes se repitió el uso del 80 % (8,160 imágenes) para el entrenamiento y 20 % (2040 imágenes) para la validación. Usando la máquina virtual de *Google Colaboratory* de paga, el modelo tomaba en ser entrenado cincuenta segundos por época, tomando así dos horas en ser entrenada. Esto fue repetido para las seis representaciones tiempo-frecuencia obtenidas. Los resultados de dicha experimentación están en la Tabla 10, tomando en consideración lo que será explicado en el Capítulo 5.4[2].

Tabla 10. Resultados de la validación del modelo VGG16

Representación tiempo-frecuencia	Precisión
Espectrograma	74.607 %
Espectrograma escala Mel	75.49 %
WSST	75 %
WSST + Espectrograma + Mel	99.5 %
WSST + Espectrograma	99.264 %
WSST + Mel	85.539 %

4.4.3. Red Ullah

Esta arquitectura propuesta en el artículo Ullah *et al.* (2020) presentó resultados muy prometedores, está compuesta de cuatro capas convolucionales y dos capas

[2] Se ha excluido del entrenamiento la base *E*

densas es de fácil implementación, así como la comparación en su artículo del modelo propuesto con AlexNet y VGGNet, por lo que fue elegida debido al desempeño mostrado en el artículo.

La arquitectura del modelo está descrita en la Figura 39.

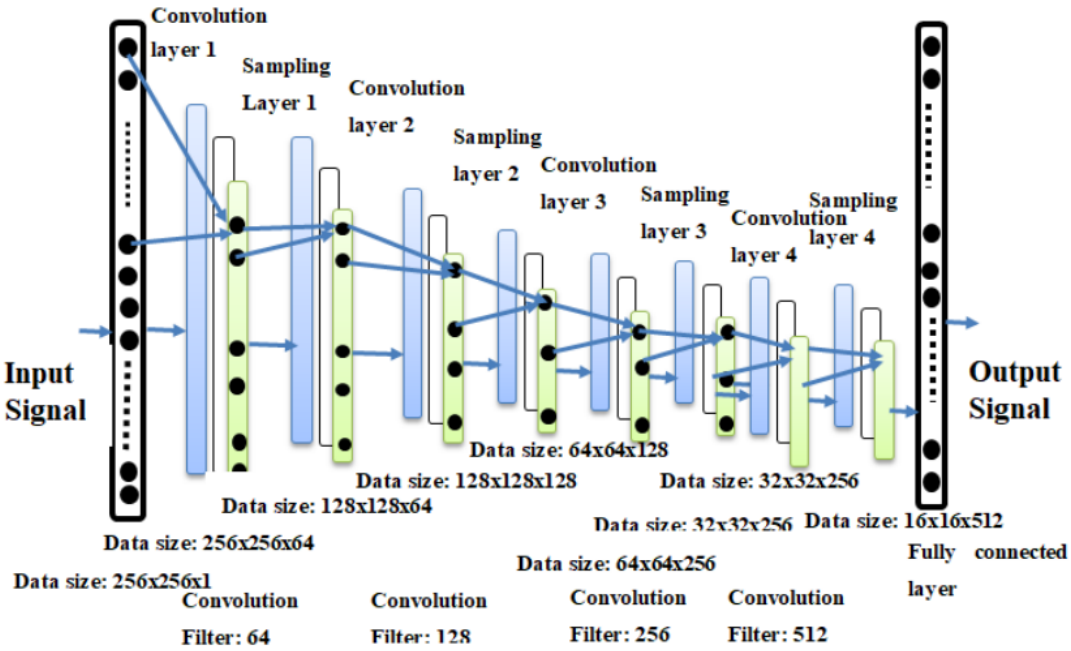


Figura 39. Arquitectura de la red Ullah, imagen tomada de (Ullah *et al.*, 2020).

La estructura de cada capa está detallada en la Figura 36, teniendo un total de 412,603,906 parámetros.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 224, 224, 64)	640
max_pooling2d (MaxPooling2D)	(None, 112, 112, 64)	0
conv2d_1 (Conv2D)	(None, 112, 112, 128)	73856
max_pooling2d_1 (MaxPooling2	(None, 56, 56, 128)	0
conv2d_2 (Conv2D)	(None, 56, 56, 256)	295168
max_pooling2d_2 (MaxPooling2	(None, 28, 28, 256)	0
conv2d_3 (Conv2D)	(None, 28, 28, 512)	1180160
max_pooling2d_3 (MaxPooling2	(None, 14, 14, 512)	0
flatten (Flatten)	(None, 100352)	0
dense (Dense)	(None, 4096)	411045888
dense_1 (Dense)	(None, 2)	8194

Figura 40. Estructura de las capas de la red Ullah implementada.

Nuevamente, las imágenes usadas para el entrenamiento de esta red son las mismas que las usadas en los otros modelos. Los hiperparámetros son los mismos para poder lograr una mejor comparación entre el rendimiento de los tres modelos, siendo los hiperparámetros usados los siguientes:

- Optimizador *Stochastic Gradient descent* (SGD)
- Tasa de aprendizaje de 0.008
- Función de pérdida: *categorical_crossentropy*
- 150 épocas
- Un tamaño de lote de 64 muestras

De las 10,200 imágenes se repitió el uso del 80 % (8,160 imágenes) para el entrenamiento y 20 % (2040 imágenes) para la validación. Usando la máquina virtual de *Google Colaboratory* de paga, el modelo tomaba en ser entrenado cuatro segundos por época, tomando así 10 minutos en ser entrenada. Esto fue repetido para las seis representaciones tiempo-frecuencia obtenidas. Dando los resultados de la Tabla 11, tomando en consideración lo que será explicado en el Capítulo 5.4[3].

Tabla 11. Resultados de la validación del modelo Ullah

Representación tiempo-frecuencia	Precisión
Espectrograma	71.666 %
Espectrograma escala Mel	66.029 %
WSST	70.735 %
WSST + Espectrograma + Mel	99.705 %
WSST + Espectrograma	99.705 %
WSST + Mel	67.05 %

[3]Se ha excluido del entrenamiento la base *E*

Capítulo 5. Análisis de resultados

En este capítulo se mostrarán y analizarán los resultados obtenidos del proceso de clasificación usando las diferentes representaciones tiempo-frecuencia.

Los resultados serán analizados y evaluados usando validación cruzada *(10-fold cross validation)*, así como la matriz de confusión obtenida de la evaluación realizada en el Capítulo 4. Cada representación tiempo-frecuencia será evaluada de forma individual.

Se debe tener en cuenta que durante la fase final del entrenamiento el subconjunto *E* de Clifford *et al.* (2016) no fue tomado en cuenta, esto es debido a que se encontró que inducía una gran merma en los resultados.

5.1. Validación cruzada

La validación cruzada es un método estadístico empleado para determinar el desempeño de un modelo de aprendizaje automático. Es muy usado para comparar y seleccionar un modelo para un problema predictivo, debido a su facilidad de entendimiento, implementación y obtención de resultados, además de que en general tiene un sesgo menor que otros métodos.

El funcionamiento parte de un parámetro llamado *k* que se refiere al número de grupos en los que se dividirá un conjunto de datos (Refaeilzadeh *et al.*, 2016). Por lo que el procedimiento es también llamado *k-fold cross validation*. El procedimiento general es:

1. Mezclar de manera aleatoria el conjunto de datos.
2. Dividir el conjunto de datos en *k* grupos.
3. Por cada grupo:
    a) Tomar un grupo y separarlo para hacer de conjunto de validación.
    b) Tomar el resto de grupos como un conjunto de datos de entrenamiento.

c) Entrenar el modelo con el conjunto de datos de entrenamiento y luego evaluar el modelo con el conjunto de validación.

d) Almacenar la puntuación de la evaluación y desechar el modelo.

4. Determinar el rendimiento del modelo usando las puntuaciones de evaluación obtenidas.

De manera gráfica esto se puede representar con la siguiente figura 41 cuando el valor de *k* es igual a 4.

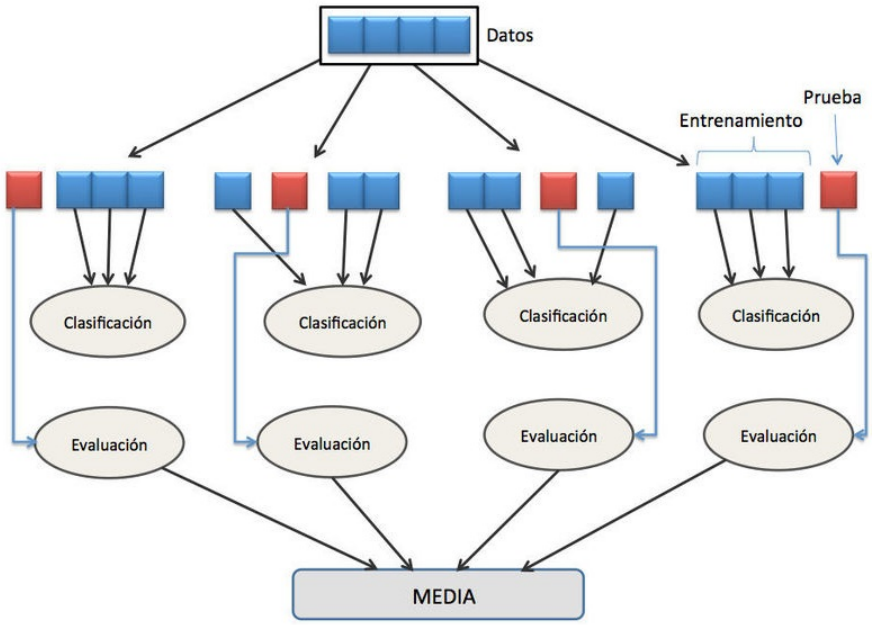


Figura 41. Diagrama de la validación cruzada, tomada de: `https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada`.

La elección correcta del valor *k* es muy importante, pues un valor incorrecto podría resultar en una mala interpretación del rendimiento real del modelo. Los valores más usuales para el valor de *k* son:

- Representativo: El valor de *k* es elegido de tal manera que cada grupo de entrenamiento/validación de las muestras es suficientemente grande para ser estadísticamente representativo del total del conjunto de datos.
- $k = 10$: El valor de *k* es 10, un valor que ha sido encontrado a través de mucha experimentación que tiene muy buenos resultados en la estimación del rendimiento de un modelo (Refaeilzadeh *et al.*, 2016).

- $k = n$: El valor de k es de n, donde n es el tamaño del conjunto de datos para dar oportunidad a cada muestra de ser usado para la validación. Esta aproximación es llamada *leave-one-out cross-validation*.

5.2. Matriz de confusión

La matriz de confusión, también llamada tabla de contingencia, es una herramienta que nos muestra el desempeño de un algoritmo de clasificación, describiendo cómo se distribuyen los valores reales y nuestras predicciones. La Figura 42 detalla la distribución de las etiquetas que serán usadas.

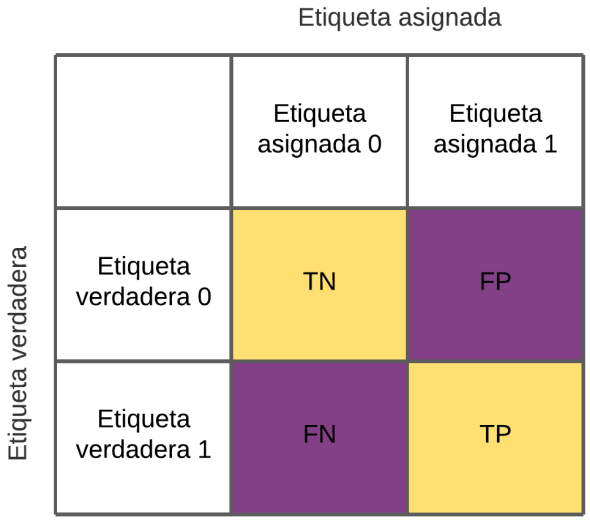


Figura 42. Matriz de confusión.

- Verdadero positivo (TP): predicción correcta positiva.
- Falso positivo (FP): predicción incorrecta positiva.
- Verdadero negativo (TN): predicción correcta negativa.
- Falso negativo (FN): predicción incorrecta negativa.

5.2.1. Métricas de la matriz de confusión

Teniendo en cuenta las partes que componen la matriz de confusión, se pueden obtener diferentes métricas para medir el desempeño de la clasificación (Luque *et al.*,

2019).

La exactitud (*Accuracy*) hace referencia a lo cerca que está el resultado de una medición del valor verdadero. Se representa como la relación entre el número de predicciones correctas y el número total del conjunto de datos, dada por la Ecuación 19. El mejor valor de exactitud es 1, siendo el peor 0.

$$\text{Exactitud (ACC)} = \frac{TP + TN}{TP + FN + TN + FP}. \tag{19}$$

La sensitividad (*Sensitivity*) hace referencia a la habilidad del algoritmo para clasificar casos positivos, dado por la Ecuación 20.El mejor valor de sensibilidad es 1, siendo el peor 0.

$$\text{Sensitividad (SNS)} = \frac{TP}{TP + FN}. \tag{20}$$

La especificidad (*Specificity*) hace referencia a la habilidad del algoritmo para clasificar casos negativos, dado por la Ecuación 21. El mejor valor de especificidad es 1, siendo el peor 0.

$$\text{Especifidad (SPC)} = \frac{TN}{TN + FP}. \tag{21}$$

La precisión (*Precision*) hace referencia a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Se representa como la relación entre el número de predicciones verdaderas y el número total de predicciones verdaderas, dada por la Ecuación 22.El mejor valor de precisión es 1, siendo el peor 0.

$$\text{Precisión (PRC)} = \frac{TP}{TP + FP}. \tag{22}$$

La métrica F1 es una métrica que resume la previsión y sensibilidad en un resultado, dado por la Ecuación 23.

$$\text{Métrica F1} = 2\frac{PRC * SNS}{PRC + SNS}. \tag{23}$$

5.3. Resultados obtenidos incluyendo la base *E*

En un principio se utilizaron sólo las representaciones del Espectrograma, Espectrograma en escala Mel y WSST. Usando 10,200 imágenes de cada representación se obtuvieron los resultados de la Tabla 12:

Tabla 12. Resultados obtenidos usando la base de datos *E*

	AlexNet	VGG16	Ullah
Espectrograma	84.59	77.39	72.63
Espectrograma Mel	83.15	77.44	70.74
WSST	49.31	49.29	49.31
WSST + Espectrograma + Mel	49.25	49.3	49.25
WSST + Espectrograma	49.25	49.25	49.25
WSST + Mel	49.25	49.3	49.25

Estos resultados no se comportaron de la forma esperada, por lo que se procedió a la verificación de los sonidos cardíacos, encontrando así que la base de datos e de Clifford *et al.* (2016) introducía una gran cantidad de inconsistencia en la habilidad del clasificador, y se tomó la decisión de excluirla del entrenamiento hasta que fuera analizada con más detenimiento. El análisis de esta base de datos no fue llevado a cabo de manera exhaustiva debido a limitaciones de tiempo y alcance de este trabajo, pero se encontró que los sonidos de esta base de datos son en extremo ruidosos por lo que dificulta en gran medida la tarea de la clasificación.

Las matrices de confusión de la clasificación incluyendo la base *E* serán expuestos en el Anexo A.

5.4. Resultados obtenidos excluyendo la base *E*

Considerando lo anteriormente explicado se evaluará el desempeño de cada representación tiempo frecuencia y los tres modelos usados para su clasificación de manera individual.

Se debe tener en cuenta que la validación que se utilizó para el modelo VGG16 fue la 5-validación cruzada. Esto es debido al tiempo que requiere el modelo para ser entrenado (gasto computacional) para llevar a cabo la evaluación del modelo. Se determinó que la 5-validación cruzada muestra un comportamiento del rendimiento

similar a la 10-validación cruzada pero con un gasto computacional sustancialmente menor.

5.4.1. Resultados de las imágenes espectrograma

Las representaciones tiempo-frecuencia obtenidas con el espectrograma demuestran que esta representación por sí sola es inadecuada para una tarea de clasificación de sonidos cardíacos de acuerdo a la Tabla 13. El modelo de AlexNet lidera la clasificación con 85.89 %, su matriz de confusión dada por la Figura 43 nos da un mejor entendimiento del desempeño de la clasificación, la cantidad de falsos negativos y falsos positivos es relativamente baja pero notoria.

Tabla 13. Resultados de la 10-validación cruzada de las imágenes espectrograma.

	AlexNet	VGG16	Ullah
Precisión	85.89 (± 0.95)	75.53 (± 0.84)	72.01 (± 1.45)

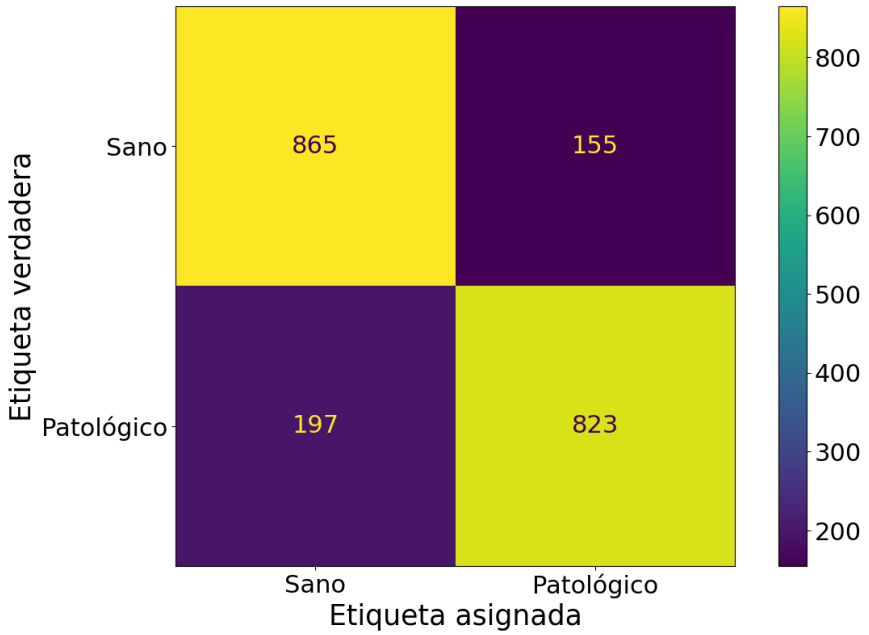


Figura 43. Matriz de confusión de la clasificación usando el modelo de AlexNet con imágenes espectrograma.

El modelo de VGG16 tiene un rendimiento de 75.53 % en su clasificación, lo cuál para una tarea de clasificación relacionado a la salud puede ser visto como relativamente

bajo o poco fiable, la matriz de confusión dada por la Figura 44 nos revela que al igual que el modelo de AlexNet los falsos negativos son mayores que los falsos positivos.

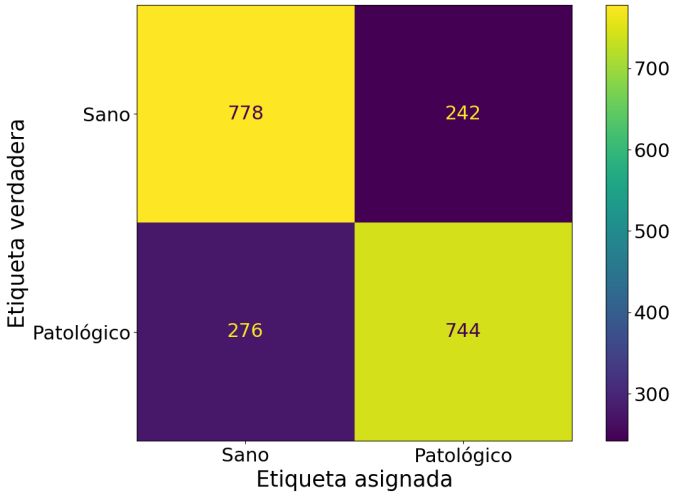


Figura 44. Matriz de confusión de la clasificación usando el modelo de VGG con imágenes espectrograma.

El modelo de Ullah tiene un rendimiento de 72.01 %, siendo el más bajo obtenido usando esta representación tiempo-frecuencia, la matriz de confusión dada por la Figura 45 revela que los falsos positivos son mayores a los falsos negativos y que tiene un mejor rendimiento encontrado casos patológicos que sanos.

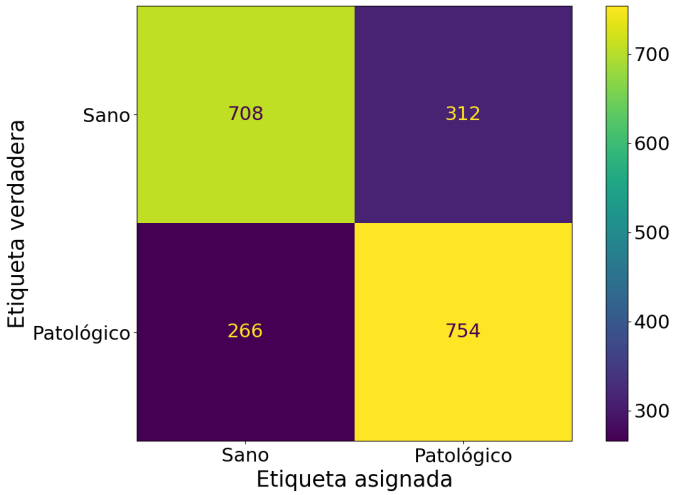


Figura 45. Matriz de confusión de la clasificación usando el modelo de Ullah con imágenes espectrograma.

Tabla 14. Métricas de la matriz de confusión de las imágenes espectrograma

	Exactitud	Precisión	Sensibilidad	Especificidad	F1
AlexNet	0.8275	0.8415	0.8068	0.848	0.8238
VGG16	0.746	0.7545	0.7294	0.7627	0.7417
Ullah	0.7166	0.7073	0.7392	0.6941	0.7229

Las métricas obtenidas de las matrices de confusión de los modelos están resumidas en la Tabla 14. Esto nos ayuda a comparar de manera rápida el rendimiento de los tres modelos evaluados, la ventaja en tareas de clasificación siendo para AlexNet, seguido del modelo de VGG16 y por último el de Ullah. Además de los resultados obtenidos también se tiene que tener en cuenta los gastos computacionales requeridos para la tarea de clasificación, siendo el de AlexNet y Ullah los que requieren de un gasto computacional menor a comparación del modelo VGG16.

5.4.2. Resultados de las imágenes espectrograma en escala Mel

Las representaciones tiempo-frecuencia obtenidas con el espectrograma en escala Mel demuestran que la combinación del espectrograma con el espectrograma en escala Mel no genera una gran mejoría en comparación con el espectrograma por sí solo, de acuerdo a la Tabla 15. El modelo de AlexNet lidera la clasificación con un 85.07 %, la Figura 46 es su matriz de confusión y nos da un mejor entendimiento de la clasificación de los sonidos cardíacos. Se observa también que la cantidad de falsos negativos supera a los falsos positivos, por lo que el modelo tiene preferencia por los sonidos cardíacos sanos.

Tabla 15. Resultados de la 10-validación cruzada de las imágenes espectrograma en escala Mel.

	AlexNet	VGG16	Ullah
Precisión	85.07 (± 1.23)	75.22 (± 1.79)	66.5 (± 3.2)

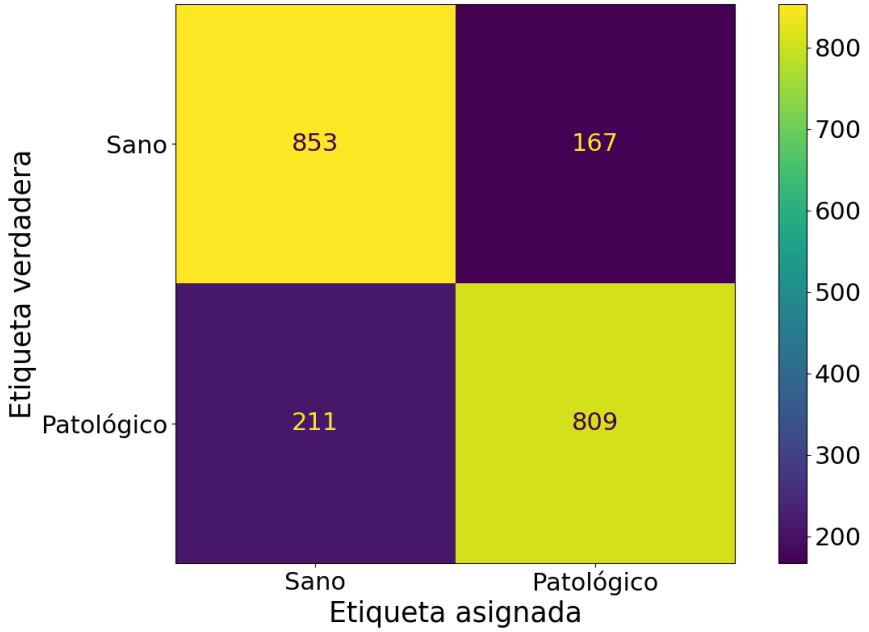


Figura 46. Matriz de confusión de la clasificación usando el modelo de AlexNet con imágenes espectrograma en escala Mel.

El modelo VGG16 tiene una clasificación de 75.22 % de acuerdo a la Tabla 15, lo cual deja ver que no puede ser considerada apta para tareas de clasificación en su estado actual. la matriz de confusión mostrada en la Figura 47 nos muestra la tendencia de estos algoritmos por clasificar de mejor manera los sonidos cardíacos sanos.

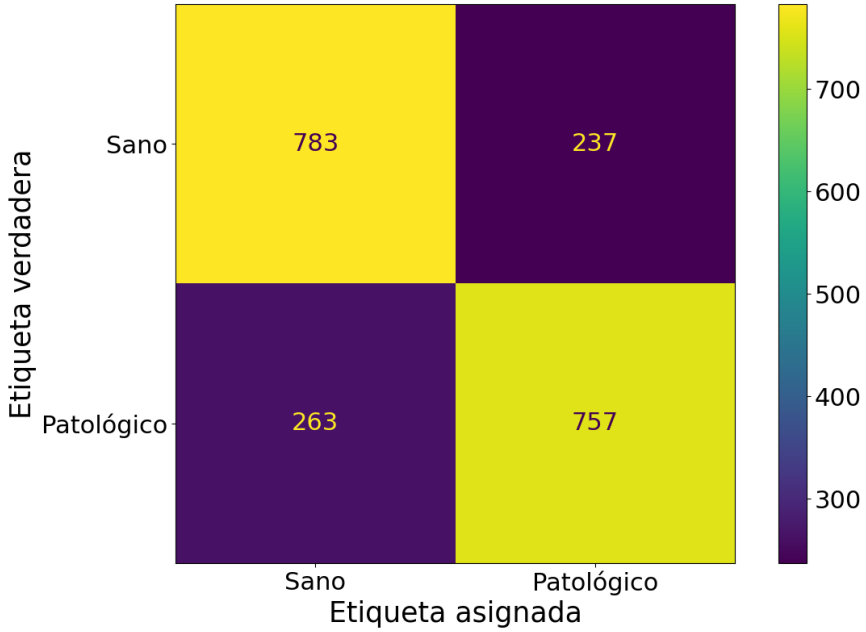


Figura 47. Matriz de confusión de la clasificación usando el modelo de VGG con imágenes espectrograma en escala Mel.

El modelo de Ullah tiene una clasificación de 66.5 % de acuerdo con la Tabla 15, representando el resultado más bajo entre las puntuaciones dadas por todas las configuraciones experimentadas. La matriz de confusión en la Figura 48 respectivamente muestra la mayor cantidad de falsos negativos encontrada entre las configuraciones analizadas.

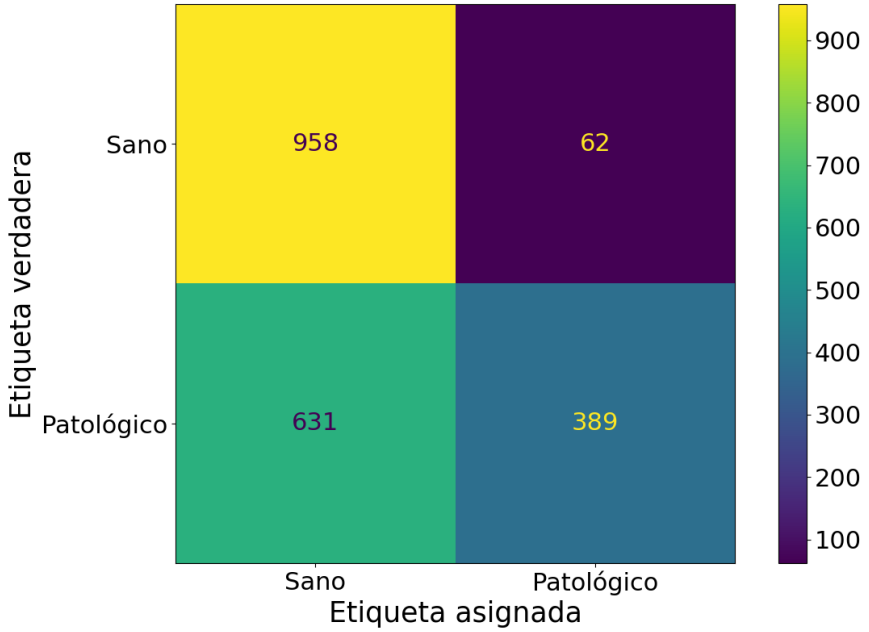


Figura 48. Matriz de confusión de la clasificación usando el modelo de Ullah con imágenes espectrograma en escala Mel.

Tabla 16. Métricas de la matriz de confusión de las imágenes espectrograma en escala Mel

	Exactitud	Precisión	Sensibilidad	Especificidad	F1
AlexNet	0.8147	0.8288	0.7931	0.8362	0.8106
VGG16	0.7549	0.7615	0.7421	0.7676	0.7517
Ullah	0.6602	0.8625	0.3813	0.9392	0.5288

La Tabla 16 nos resume las métricas obtenidas de las matrices de confusión de los modelos evaluados, se observa que la puntuación más alta en Precisión fue alcanzada por el clasificador AlexNet, seguido de VGG16 y por último Ullah. La diferencia entre AlexNet y VGG16 aunque pequeña siendo de seis puntos en su mayoría se tiene que considerar también el gasto computacional requerido para la clasificación, en este rubro el modelo de AlexNet tiene una ventaja considerable pues su gasto computacional es pequeño en comparación al gran gasto que requiere el modelo de VGG16 para ser entrenado.

5.4.3. Resultados de las imágenes WSST

Las representaciones tiempo-frecuencia obtenidas con la WSST demuestran una clasificación poco satisfactoria para este problema. El modelo de AlexNet obtiene una puntuación de 85.07 % de acuerdo a la Tabla 17, su matriz de confusión dada por la Figura 49 se puede apreciar un mayor número de falsos negativos sobre los falsos positivos, siendo estos un número considerable de errores durante la clasificación, así como un mejor desempeño clasificando sonidos sanos sobre los patológicos.

Tabla 17. Resultados de la 10-validación cruzada de las imágenes WSST.

	AlexNet	VGG16	Ullah
Precisión	80.59 (± 1.87)	74.72 (± 0.79)	67.7 (± 3.2)

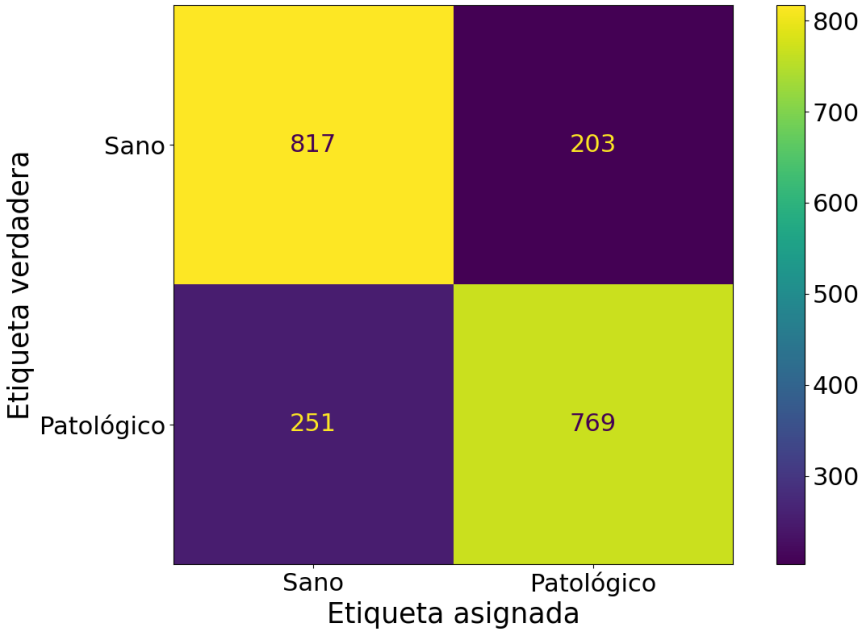


Figura 49. Matriz de confusión de la clasificación usando el modelo de AlexNet con imágenes WSST.

El modelo VGG16 tiene una puntuación de 74.72 % de acuerdo con la Tabla 17, su matriz de confusión dada por la Figura 50 revela que nuevamente los falsos negativos superan a los falsos positivos, resultando en una clasificación inadecuada.

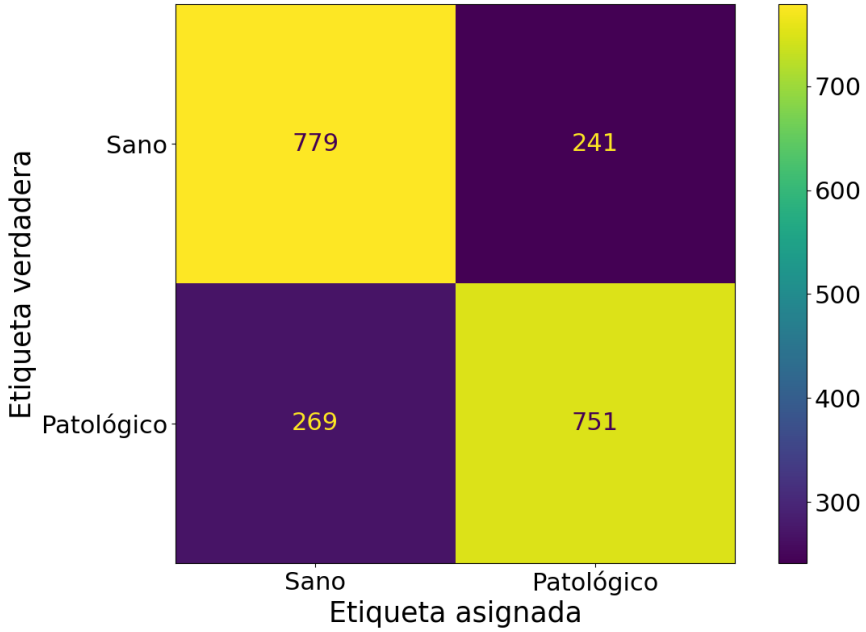


Figura 50. Matriz de confusión de la clasificación usando el modelo de VGG con imágenes WSST.

El modelo Ullah tiene una puntuación de 67.7 % de acuerdo con la Tabla 17, siendo la peor de todas, su matriz de confusión dada por la Figura 51 nos muestra la cantidad alarmante de falsos negativos y falsos positivos que el modelo clasificó.

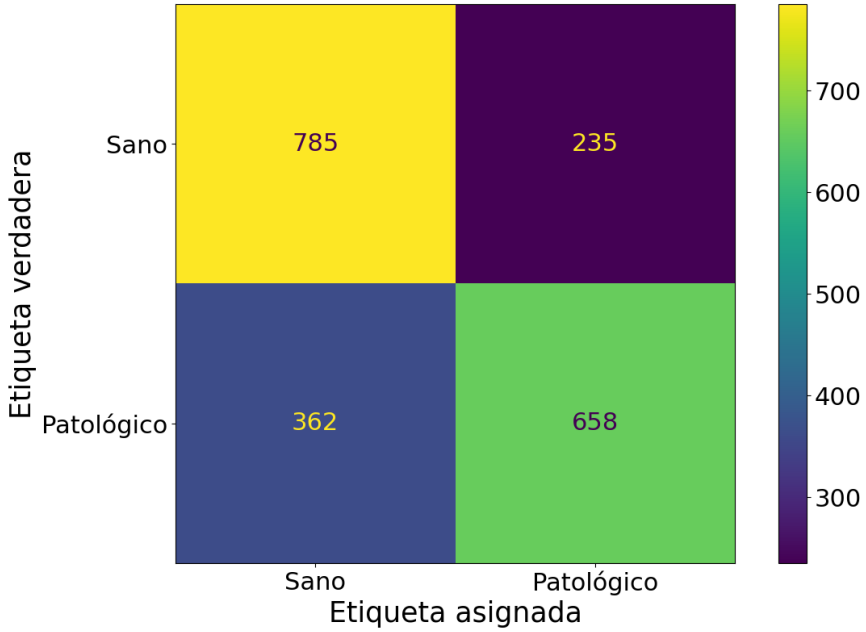


Figura 51. Matriz de confusión de la clasificación usando el modelo de Ullah con imágenes WSST.

Tabla 18. Métricas de la matriz de confusión de las imágenes WSST

	Exactitud	Precisión	Sensibilidad	Especificidad	F1
AlexNet	0.7774	0.7911	0.7539	0.8009	0.772
VGG16	0.75	0.757	0.7362	0.7637	0.7465
Ullah	0.7073	0.7368	0.645	0.7696	0.6879

La Tabla 17 nos muestra que los resultados de los tres modelos son mediocres, siendo que en sólo un rubro se alcanzó el 0.8 siendo el modelo de AlexNet, a esto se le tiene que considerar el gasto computacional utilizado para la realización de la clasificación de los modelos, en los que los modelos de AlexNet y Ullah son los que menor gasto computacional requieren para su entrenamiento, por lo que AlexNet es una opción para clasificación debido resultados y gasto computacional requerido.

5.4.4. Resultados de las imágenes WSST + Espectrograma + Mel

Las representaciones tiempo-frecuencia obtenidas con la combinación de la WSST + Espectrograma + Mel demuestran una clasificación satisfactoria para este problema. El modelo de AlexNet obtiene una puntuación de 99.96 % de acuerdo a la Tabla 19,

su matriz de confusión dada por la Figura 52 nos muestra que el modelo no tuvo dificultades para clasificar los sonidos cardíacos, siendo que tuvo una clasificación impecable con los sonidos sanos y casi impecable con los sonidos patológicos.

Tabla 19. Resultados de la 10-validación cruzada de las imágenes WSST + Espectrograma + Mel.

	AlexNet	VGG16	Ullah
Precisión	99.96 (± 0.06)	98.92 (± 0.41)	99.82 (± 0.09)

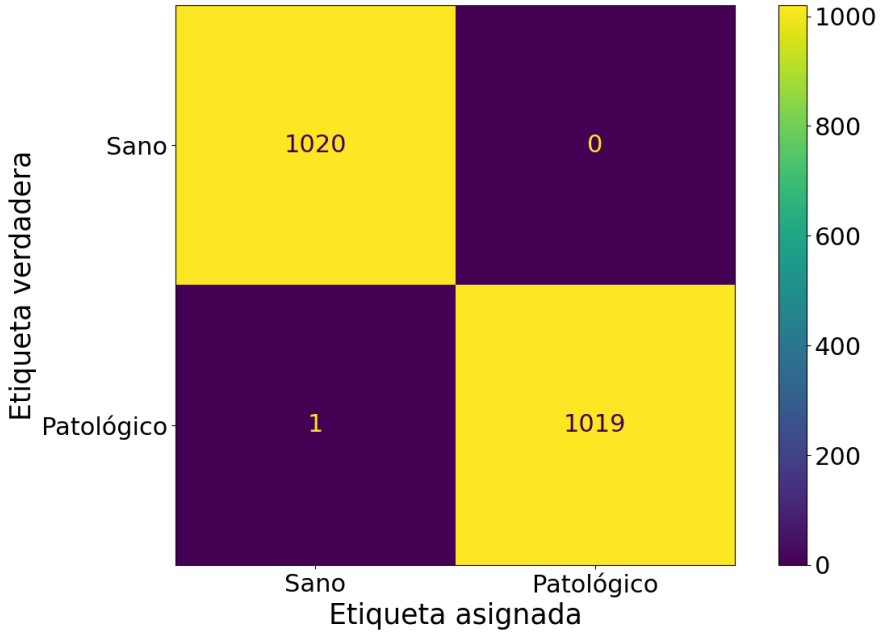


Figura 52. Matriz de confusión de la clasificación usando el modelo de AlexNet con imágenes WSST + Espectrograma + Mel.

El modelo de VGG16 tuvo una puntuación de 98.92 % de acuerdo con la Tabla 19, la Figura 53 muestra su matriz de confusión en la que se muestra el desempeño del modelo, siendo que los falsos negativos y falsos positivos son mínimos en comparación de las clasificaciones verdaderas.

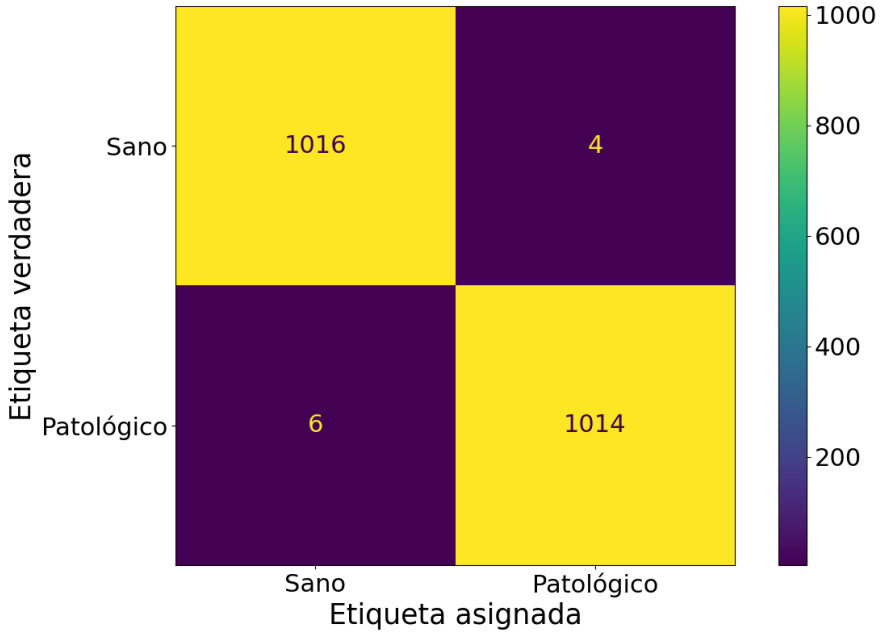


Figura 53. Matriz de confusión de la clasificación usando el modelo de VGG con imágenes WSST + Espectrograma + Mel.

El modelo de Ullah tiene una puntuación de 99.82 % de acuerdo con la Tabla 19, la Figura 54 muestra su matriz de confusión en la que se muestran los errores mínimos que el modelo realizó durante la clasificación de los sonidos cardíacos.

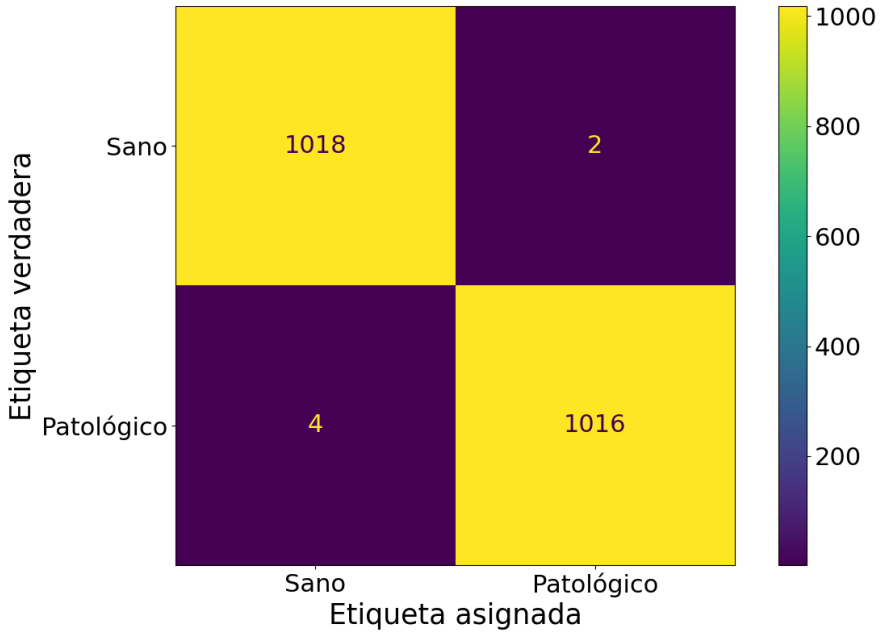


Figura 54. Matriz de confusión de la clasificación usando el modelo de Ullah con imágenes WSST + Espectrograma + Mel.

Tabla 20. Métricas de la matriz de confusión de las imágenes WSST + Espectrograma + Mel

	Exactitud	Precisión	Sensibilidad	Especificidad	F1
AlexNet	0.9995	1	0.999	1	0.9995
VGG16	0.995	0.996	0.9941	0.996	0.995
Ullah	0.997	0.998	0.996	0.998	0.997

La Tabla 20 nos muestra las métricas obtenidas de las matrices de confusión de los modelos analizados, los tres modelos tienen un excelente rendimiento en la clasificación de los sonidos cardíacos, siendo el mejor el modelo de AlexNet, seguido de Ullah y VGG16. Tomando en cuenta el gasto computacional requerido para el entrenamiento de los modelos, AlexNet y Ullah son los más adecuados para realizar tareas de clasificación de sonidos cardíacos usando esta combinación de representaciones tiempo-frecuencia.

5.4.5. Resultados de las imágenes WSST + Espectrograma

Las representaciones tiempo-frecuencia obtenidas con la combinación del Espectrograma + WSST demuestran una clasificación satisfactoria de los sonidos cardíacos.

El modelo de AlexNet obtuvo una clasificación de 99.95 % de acuerdo a la Tabla 21, la Figura 55 muestra que el modelo no tuvo dificultades para realizar la clasificación de los sonidos cardíacos.

Tabla 21. Resultados de la 10-validación cruzada de las imágenes WSST + Espectrograma.

	AlexNet	VGG16	Ullah
Precisión	99.95 (± 0.07)	98.74 (± 0.33)	99.85 (± 0.1)

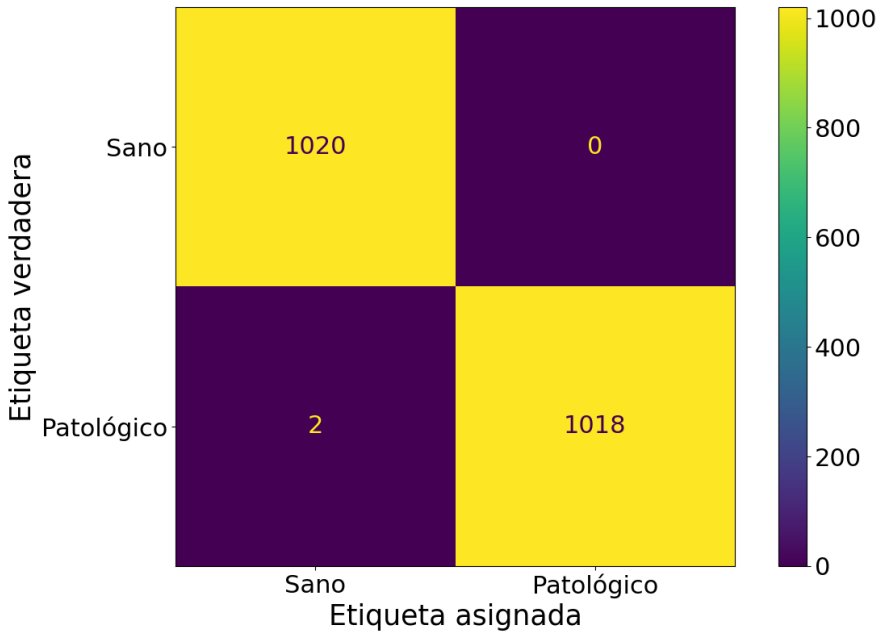


Figura 55. Matriz de confusión de la clasificación usando el modelo de AlexNet con imágenes WSST + Espectrograma.

El modelo VGG16 obtuvo una puntuación de 98.74 % de acuerdo con la Tabla 21, de acuerdo a la matriz de confusión mostrada en la Figura 56 se observa que los falsos positivos y los falsos negativos son menores.

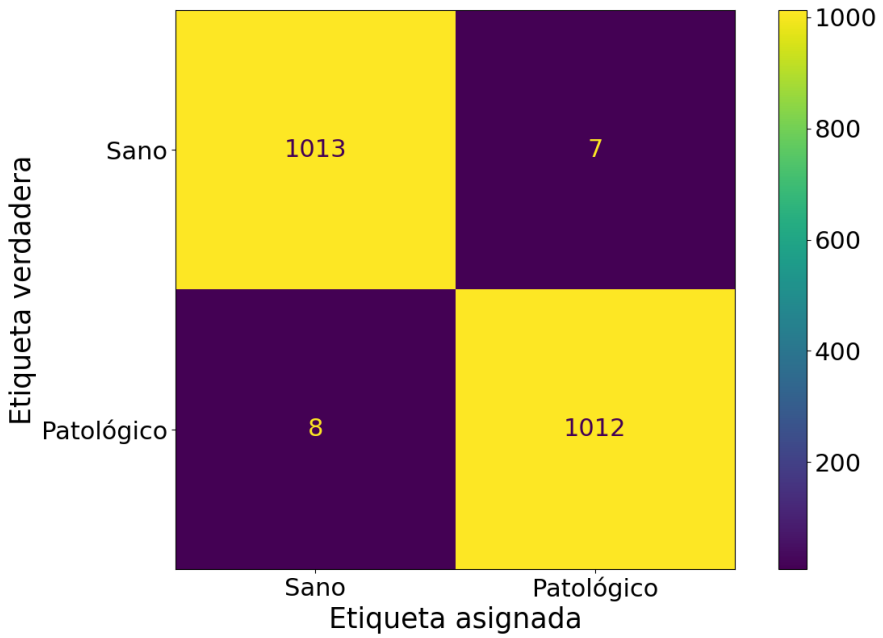


Figura 56. Matriz de confusión de la clasificación usando el modelo de VGG con imágenes WSST + Espectrograma.

El modelo de Ullah obtuvo una puntuación de 99.85 % de acuerdo con la Tabla 21, su matriz de confusión mostrada en la Figura 57 muestra que el modelo realizó una muy buena clasificación de los sonidos cardíacos.

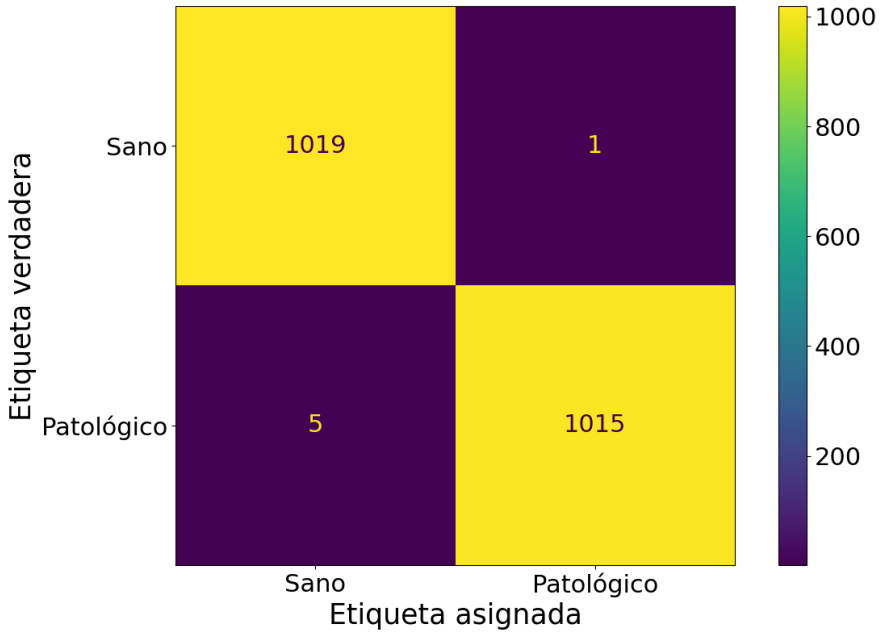


Figura 57. Matriz de confusión de la clasificación usando el modelo de Ullah con imágenes WSST + Espectrograma.

Tabla 22. Métricas de la matriz de confusión de las imágenes WSST + Espectrograma

	Exactitud	Precisión	Sensibilidad	Especificidad	F1
AlexNet	0.999	1	0.998	1	0.999
VGG16	0.9926	0.9931	0.9921	0.9931	0.9926
Ullah	0.997	0.999	0.995	0.999	0.997

La Tabla 22 nos muestra las métricas obtenidas de las matrices de confusión de los modelos analizados, los tres modelos tienen un excelente rendimiento para la clasificación de los sonidos cardíacos, siendo el modelo de AlexNet el que mejor resultados tuvo, seguido de Ullah y el modelo VGG16, usando esta combinación de Espectrograma + WSST se puede observar que la clasificación de los sonidos cardíacos se realiza de manera exitosa de acuerdo a nuestros resultados. Se tiene que considerar también el gasto computacional requerido por los modelos analizados, siendo los modelos de AlexNet y Ullah los que requieren el menor gasto computacional para ser entrenados.

5.4.6. Resultados de las imágenes WSST + Mel

Las representaciones tiempo-frecuencia obtenidas con la combinación del Espectrograma en escala Mel + WSST demuestran tener una buena clasificación en un modelo de red neuronal. El modelo de AlexNet tiene una clasificación de 98.87 % de acuerdo a la Tabla 23, la matriz de confusión mostrada en la Figura 58 muestra un gran rendimiento para la clasificación de los sonidos cardíacos, teniendo pocos falsos positivos y falsos negativos.

Tabla 23. Resultados de la 10-validación cruzada de las imágenes WSST + Mel.

	AlexNet	VGG16	Ullah
Precisión	98.87 (± 1.81)	84.96 (± 0.31)	70.31 (± 2.04)

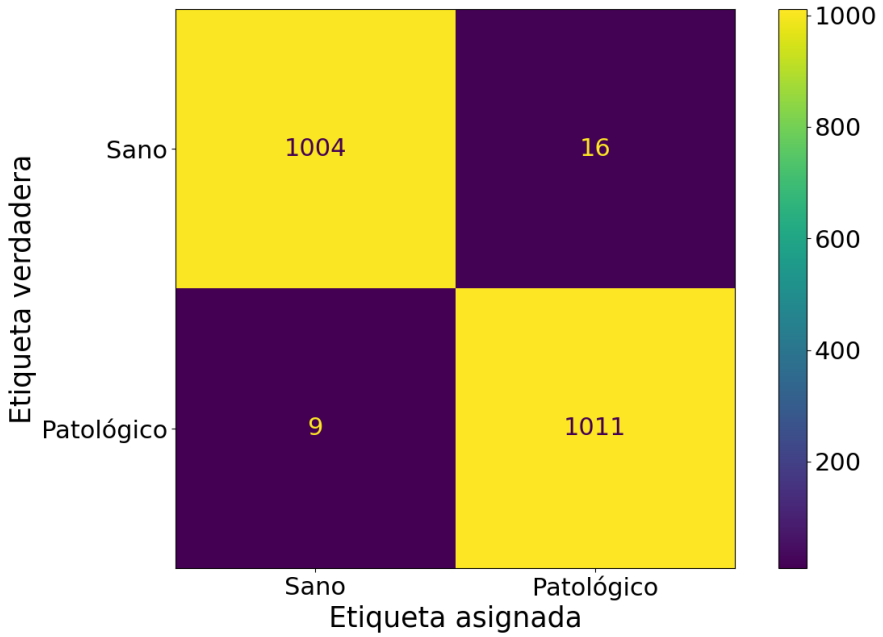


Figura 58. Matriz de confusión de la clasificación usando el modelo de AlexNet con imágenes WSST + Mel.

El modelo de VGG16 obtuvo una puntuación de 84.96 % de acuerdo con la Tabla 23, la Figura 59 muestra su matriz de confusión donde se observa un aumento de los falsos negativos y falsos positivos en comparación al modelo de AlexNet.

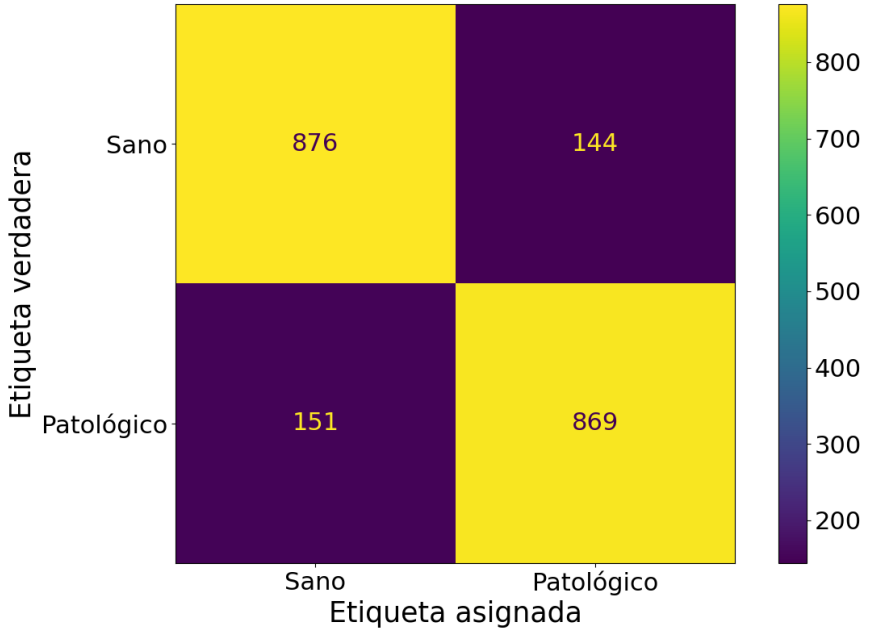


Figura 59. Matriz de confusión de la clasificación usando el modelo de VGG con imágenes WSST + Mel.

El modelo de Ullah obtuvo una puntuación de 70.31 % de acuerdo con la Tabla 23, la Figura 60 muestra el diagrama de confusión en el que se observa que la cantidad de falsos negativos es muy superior a los verdaderos positivos, demostrando que este modelo no pudo generalizar de manera correcta los sonidos cardíacos.

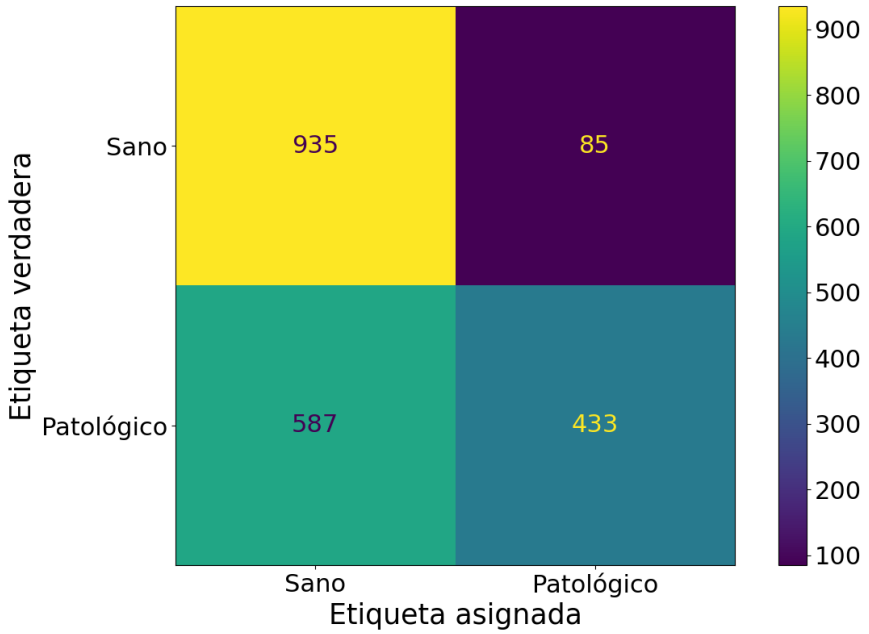


Figura 60. Matriz de confusión de la clasificación usando el modelo de Ullah con imágenes WSST + Mel.

Tabla 24. Métricas de la matriz de confusión de las imágenes WSST + Mel

	Exactitud	Precisión	Sensibilidad	Especificidad	F1
AlexNet	0.9877	0.9844	0.991	0.9843	0.9877
VGG16	0.8553	0.8578	0.8519	0.8588	0.8548
Ullah	0.6705	0.8359	0.4245	0.9166	0.563

La Tabla 24 nos muestra las métricas obtenidas de las matrices de confusión de los modelos analizados, el modelo de AlexNet es el que mejor desempeño tiene de acuerdo a nuestros resultados, seguido del modelo VGG16 que obtuvo resultados aceptables pero no aptos para tareas de clasificación de sonidos cardíacos, el modelo Ullah es el que obtuvo los peores resultados, siendo un modelo que falló para poder generalizar la tarea de la clasificación, su baja sensibilidad es un indicativo de que el algoritmo no es capaz de clasificar casos positivos. Teniendo en cuenta el gasto computacional requerido para el entrenamiento de los modelos, se concluye que para esta combinación de representaciones tiempo-frecuencia se debe hacer uso del modelo de AlexNet para tener una clasificación satisfactoria.

Capítulo 6. Conclusiones

En esta sección se mostrarán las conclusiones generadas durante el desarrollo de la tesis, así como proponer puntos que pudiesen ser investigados en trabajos futuros.

6.1. Sobre los objetivos de la tesis

- El objetivo general de la tesis se considera cumplido, debido a que se ha desarrollado un método de detección de patología en la señal de audio cardíaco usando redes neuronales, obteniendo así una clasificación de 99 % sobre señales patológicas y sanas.
- Se analizaron tres representaciones tiempo-frecuencia para obtener las características de los sonidos cardíacos que se usarían para clasificación, estas características obtenidas de las tres representaciones fueron combinadas entre sí para la creación de seis tipos de representaciones tiempo-frecuencia con el objetivo de encontrar la representación que sea más adecuada para la detección de patologías cardíacas.
- Cada una de las seis representaciones tiempo-frecuencias así como los parámetros usados para la obtención de las mismas fueron descritas con detalle para facilitar la réplica de las mismas si se considerara volver a evaluar el rendimiento del modelo.
- Se analizaron tres modelos de redes neuronales que han demostrado tener un gran rendimiento en tareas de clasificación, estos modelos fueron implementados y modificados para poder ser útiles para este problema de clasificación.
- Los modelos, aunque similares en estructura, tienen un diferente grado de complejidad que influye en el tiempo que tarda cada modelo en poder ser entrenado para realizar una tarea de clasificación.
- Se encontró que los sonidos de la base de datos *E* presentan mucho ruido e interferencias externas que requieren de un mayor preprocesamiento para poder utilizar los sonidos.

- Se ha mostrado los parámetros usados durante el entrenamiento de las redes neuronales, descritos de forma que sean fácilmente repetibles para su posterior replicación.

- De las representaciones tiempo-frecuencia más idóneas para la clasificación se tiene la combinación del Espectrograma + Mel + WSST y la combinación Espectrograma + WSST, que tuvieron un rendimiento en la clasificación de 99.9 % de los sonidos cardíacos patológicos y sanos. El rendimiento de estas representaciones fue consistente en los tres modelos de redes neuronales de acuerdo con la Figura 61, lo cuál deja ver que la mejor combinación para la clasificación de patologías cardíacas es el Espectrograma + WSST siguiendo los parámetros descritos en esta tesis, pues estas dos representaciones igualan el rendimiento de la combinación de las tres representaciones.

- Los resultados demuestran que el preprocesamiento que se realizó sobre los sonidos cardíacos jugaron un papel importante en el rendimiento de los modelos de clasificación. Se observó que un análisis apropiado así como la combinación adecuada de las diferentes representaciones tiempo-frecuencia permiten una clasificación de los fonocardiogramas satisfactoria.

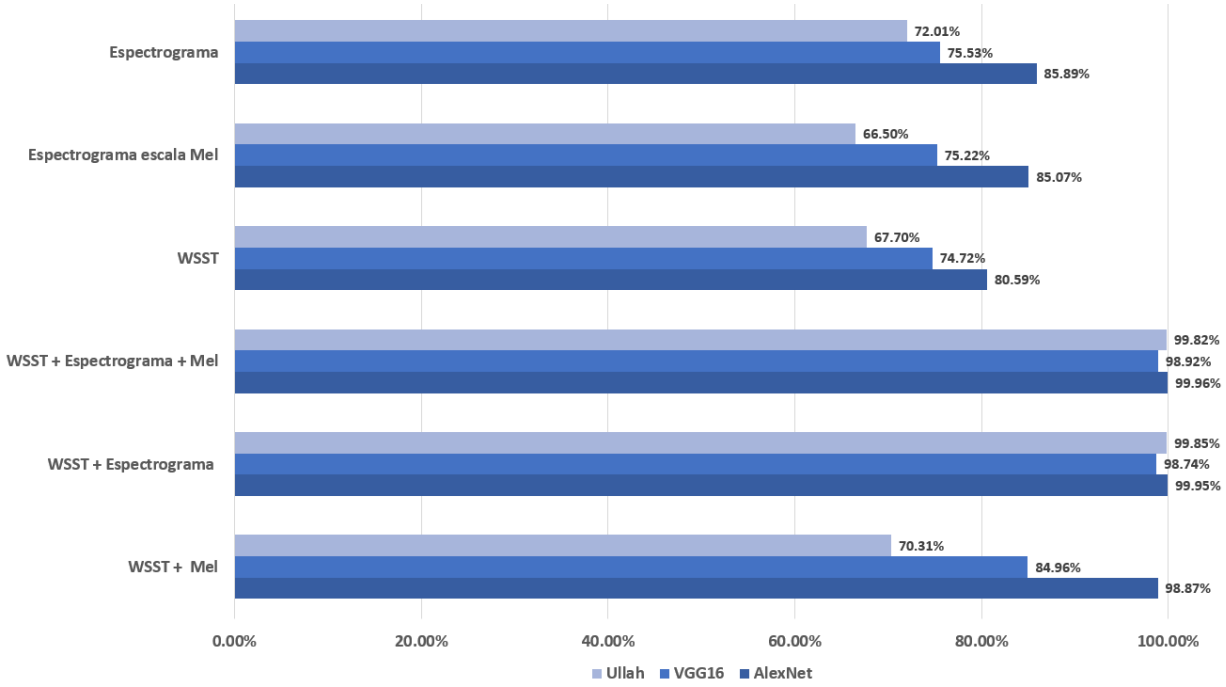


Figura 61. Resultados de la validación cruzada.

6.2. Trabajo futuro

Se mostrarán las recomendaciones para trabajos futuros que puedan surgir en el área de clasificación de sonidos cardíacos usando redes neuronales:

- Utilización de mejor equipo de trabajo para que el número de imágenes con las que los modelos son entrenados aumente, mejorando así su generalización.
- Obtención de más bases de datos de sonidos cardíacos para el entrenamiento de los modelos de redes neuronales.
- Utilización de modelos de redes neuronales convolucionales más avanzados para reducir los gastos computacionales durante el entrenamiento.
- Implementar el modelo entrenado en un ambiente de pruebas reales para verificar el desempeño del modelo en la clasificación de sonidos cardíacos.
- Realizar un preprocesamiento más exhaustivo a la base de datos e, esto con el fin de incrementar la cantidad de muestras disponibles para el entrenamiento de un modelo de clasificación.

Literatura citada

Abbas, A. K. y Bassam, R. (2009). *Phonocardiography Signal Processing*, Vol. 31. pp. 1–189.

Abdollahpur, M., Ghaffari, A., Ghiasi, S., y M. Javad, M. (2017). Detection of pathological heart sounds. *Physiological measurement*, **38**(8).

Addison, P., Walker, J., y Guido, R. (2009). Time - Frequency analysis of biosignals. *IEEE Engineering in Medicine and Biology Magazine*, **28**(5): 14–29.

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning*. Springer. p. 512.

Allen, J. B. (2008). Nonlinear Cochlear Signal Processing and Masking in Speech Perception. En: *Springer Handbooks*. Springer, Berlin, Heidelberg, pp. 27–60.

Antonini, M., Barlaud, M., Mathieu, P., y Daubechies, I. (1992). Image Coding Using Wavelet Transform. *IEEE Transactions on Image Processing*, **1**(2): 205–220.

Azmy, M. M. (2016). Classification of normal and abnormal heart sounds using new mother wavelet and support vector machines. *2015 4th International Conference on Electrical Engineering, ICEE 2015*, pp. 90–92.

Boashash, B. (2015). *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. Número December. pp. 1–1020.

Bozkurt, B., Germanakis, I., y Stylianou, Y. (2018). A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. *Computers in Biology and Medicine*, **100**: 132–143.

Chauhan, S., Wang, P., Sing Lim, C., y Anantharaman, V. (2008). A computer-aided MFCC-based HMM system for automatic auscultation. *Computers in Biology and Medicine*, **38**(2): 221–233.

Clifford, G. D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., y Mark, R. G. (2016). Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. *Computing in Cardiology*, **43**: 609–612.

Clifford, G. D., Liu, C., Moody, B., Millet, J., Schmidt, S., Li, Q., Silva, I., y Mark, R. G. (2017). Recent advances in heart sound analysis. *Physiological Measurement*, **38**(8): E10–E25.

Cruz Ortega, H. A., Calderón Monter, F. X., Cruz Ortega, H. A., y Calderón Monter, F. X. (2016). The heart, normal sounds and murmurs. *Revista de la Facultad de Medicina (México)*, **59**(2): 49–55.

Daubechies, I., Lu, J., y Wu, H. T. (2010). "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool,.Applied and Computational Harmonic Analysis. *Elsevier*, pp. 1–32.

Dirección General de Información de Salud (2020). Sistema de Información de la Secretaria de Salud Recuperado de: http://sinaiscap.salud.gob.mx:8080/DGIS/.

Dwivedi, A. K., Imtiaz, S. A., y Rodriguez-Villegas, E. (2019). Algorithms for automatic analysis and classification of heart sounds-A systematic review. *IEEE Access*, **7**: 8316–8345.

Ghosh, S. K., Tripathy, R. K., Ponnalagu, R. N., y Pachori, R. B. (2019). Automated Detection of Heart Valve Disorders from the PCG Signal Using Time-Frequency Magnitude and Phase Features. *IEEE Sensors Letters*, **3**(12): 0–3.

Ghosh, S. K., Ponnalagu, R. N., Tripathy, R. K., y Acharya, U. R. (2020). Automated detection of heart valve diseases using chirplet transform and multiclass composite classifier with PCG signals. *Computers in Biology and Medicine*, **118**: 103632.

Google (2021). Google Colaboratory. *Consultado el 15 de agosto de 2021, de https://colab.research.google.com*.

Gutiérrez, A. C. (2016). *Segmentación robusta de audio cardíaco mediante análisis tiempo-frecuencia y métodos de optimización*. Tesis de maestría, Centro de Investigación Científica y Educación Superior de Ensenada.

Gutiérrez Uribe, J. I. (2019). *Clasificación del audio cardiaco mediante representación escasa de señales y aprendizaje automático*. Tesis de maestría, Centro de Investigación Científica y Educación Superior de Ensenada.

Harris, F. J. (1978). On the Use of Windows for Harmonic Analysis with Discrete Fourier Transform. *Proceedings of the IEEE*, **66**(Jnauary): 51–83.

Hernández, R. F. I. (2019). *Development of techniques for the analysis and processing of cardiac sound signals Dissertation*. Doctoral thesis, Centro de Investigación Científica y Educación Superior de Ensenada.

Homsi, M. N., Medina, N., Hernandez, M., Quintero, N., Perpinan, G., Quintana, A., y Warrick, P. (2016). Automatic heart sound recording classification using a nested set of ensemble algorithms. *Computing in Cardiology*, **43**: 817–820.

Ibarra, R. F., Alonso, M. A., Villarreal, S., y Nieblas, C. I. (2016). A parametric model for heart sounds. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, **2016-Febru**: 765–769.

Jurado, F. y Saenz, J. R. (2002). Comparison between discrete STFT and wavelets for the analysis of power quality events. *Electric Power Systems Research*, **62**(3): 183–190.

Khan, F. A., Abid, A., y Khan, M. S. (2020). Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features. *Physiological Measurement*, **12**(8): 598–603.

Kiymik, M. K., Güler, I., Dizibüyük, A., y Akin, M. (2005). Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application. *Computers in Biology and Medicine*, **35**(7): 603–616.

Kopparapu, S. K. y Laxminarayana, M. (2010). Choice of Mel filter bank in computing MFCC of a resampled speech. *10th International Conference on Information Sciences, Signal Processing and their Applications, ISSPA 2010*, (Isspa): 121–124.

Krizhevsky, A., Sutskever, I., y E. Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, **25**: 1907–1105.

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, **521**(7553): 436–444.

Luque, A., Carrasco, A., Martín, A., y de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, **91**: 216–231.

Mahnke, C. B. (2009). Automated heartsound analysis/Computer-aided auscultation: A cardiologist's perspective and suggestions for future development. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pp. 3115–3118.

Messer, S. R., Agzarian, J., y Abbott, D. (2001). Optimal wavelet denoising for phonocardiograms. *Microelectronics Journal*, **32**(12): 931–941.

Nieblas, C., Ibarra, R., y Alonso, M. (2014). A Novel Fourth Heart Sounds Segmentation Algorithm Based on Matching Pursuit and Gabor Dictionaries. *Research in Computing Science*, **80**(1): 9–16.

Nieblas, C. I., Alonso, M. A., Conte, R., y Villarreal, S. (2013). High performance heart sound segmentation algorithm based on Matching Pursuit. *2013 IEEE Digital Signal Processing and Signal Processing Education Meeting, DSP/SPE 2013 - Proceedings*, pp. 96–100.

Obaidat, S. . M. (1993). Phonocardiogram signal analysis: Techniques and performance comparison. *J. Med. Eng. Technol*, **17**(6): 221–227.

Organización Mundial de la Salud (2018). Las 10 principales causas de defunción Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Peyré, G. (2020). Mathematics of Neural Networks. *École Normale Supérieure PSL.*, p. 16.

Qassim, H., Verma, A., y Feinzimer, D. (2018). Compressed residual-VGG16 CNN model for big data places image recognition. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*, **2018-Janua**: 169–175.

Reed, T. R., Reed, N. E., y Fritzson, P. (2004). Heart sound analysis for symptom detection and computer-aided diagnosis. *Simulation Modelling Practice and Theory*, **12**(2): 129–146.

Refaeilzadeh, P., Tang, L., Liu, H., Angeles, L., y Scientist, C. D. (2016). *Cross-Validation*. Springer New York. New York, NY, pp. 1—-7.

Roguin, A. (2006). The man behind the stethoscope. *Clinical Medicine and Research*, **4**(3): 230–235.

Rosebrock, A. (2017). *Deep Learning for Computer Vision with Python - Starter*. Pyimageserach, first edit edición. p. 332.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., y Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, **2018-April**: 4779–4783.

Shensa, M. J. (1992). The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms. *IEEE Transactions on Signal Processing*, **40**(10): 2464–2482.

Sigurdsson, S., Petersen, K. B., y Lehn-Schiøler, T. (2006). Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music. *ISMIR 2006 - 7th International Conference on Music Information Retrieval*, (5): 286–289.

Simonyan, K. y Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14.

Smith, J. O. (2011). *Spectral Audio Signal Processing*. W3K Publishing.

Soeta, Y. y Bito, Y. (2015). Detection of features of prosthetic cardiac valve sound by spectrogram analysis. *Applied Acoustics*, **89**: 28–33.

Stevens, S. S. y Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. *American Journal of Psychiatry*, **53**(July): 329–353.

Tary, J. B., Herrera, R. H., y Van Der Baan, M. (2018). Analysis of time-varying signals using continuous wavelet and synchrosqueezed transforms. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **376**(2126).

Ullah, A., Anwar, S. M., Bilal, M., y Mehmood, R. M. (2020). Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation. *Remote Sensing*, **12**(10).

Umesh, S., Cohen, L., y Nelson, D. (2002). Frequency warping and the Mel scale. *IEEE Signal Processing Letters*, **9**(3): 104–107.

Vyas, S., D. PAtil, M., y K. Birajdar, G. (2021). Classification of Heart Sound Signals Using Time-Frequency Image Texture Features. En: *Computational Intelligence and Healthcare Informatics*. John Wiley Sons, Ltd, first edit edición, capítulo 5, pp. 81–101.

Wei, J. (2019). AlexNet: The Architecture that Challenged CNNs. *TowardsDataScience*, (https://towardsdatascience.com/alexnet-the-architecture-that-challenged-cnns-e406d5297951).

Weinhaus, A. J. y Roberts, K. P. (2005). Anatomy of the human heart. *Handbook of Cardiac Anatomy, Physiology, and Devices: 2nd ed*, pp. 59–85.

Wu, G.-d. y Lin, C.-t. (2000). Word Boundary Detection with Mel-Scale Frequency Bank in Noisy Environment. *IEEE Transactions on Speech and Audio Processing*, **8**(5): 541–554.

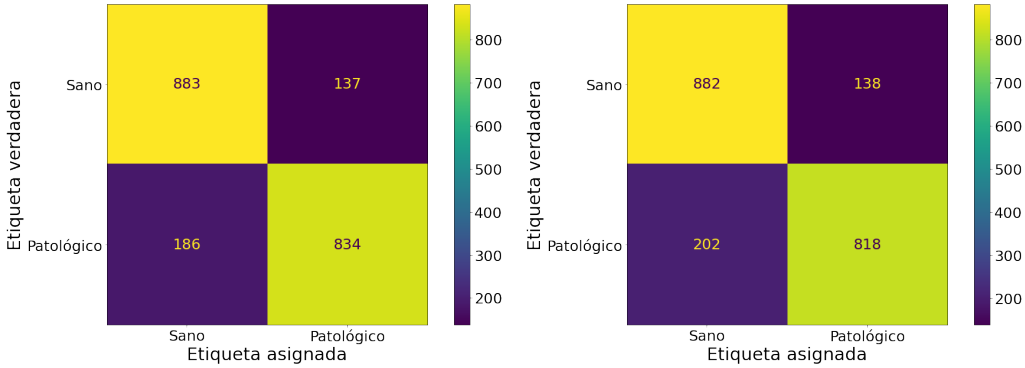
Yaseen, Son, G. Y., y Kwon, S. (2018). Classification of heart sound signal using multiple features. *Applied Sciences (Switzerland)*, **8**(12).

Zhang, W., Han, J., y Deng, S. (2017). Heart sound classification based on scaled spectrogram and tensor decomposition. *Expert Systems with Applications*, **84**: 220–231.

Zhang, X., Durand, L. G., Senhadji, L., Lee, H. C., y Coatrieux, J. L. (1998). Analysis-synthesis of the phonocardiogram based on the matching pursuit method. *IEEE Transactions on Biomedical Engineering*, **45**(8): 962–971.

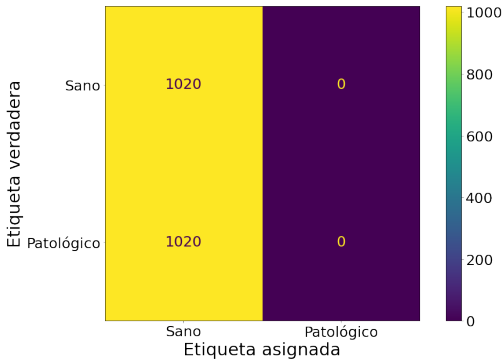
Anexo A

En este Anexo se presentan las matrices de confusión de los diferentes clasificadores usando la base de sonidos *E* para la clasificación. Las Figuras 62, 63 y 64 muestran un particular comportamiento en la clasificación que corresponde a lo establecido en la Tabla 12, las razones por las que se produce este comportamiento no fueron exploradas a profundidad debido a restricciones de tiempo.

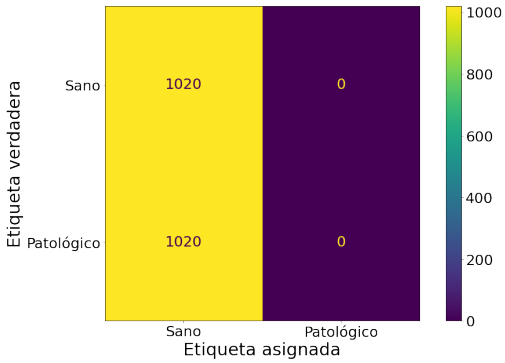


(a) Espectrograma

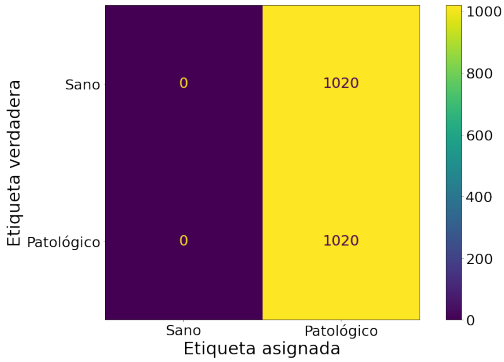
(b) Espectrograma escala Mel



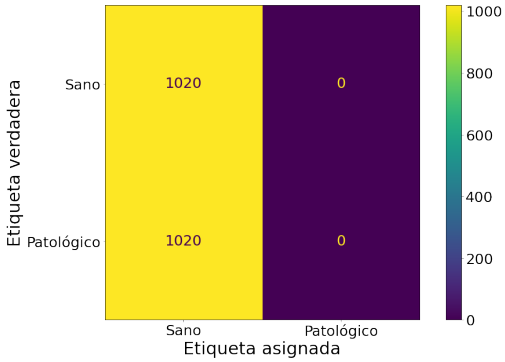
(c) WSST



(d) WSST + Espectrograma + Mel



(e) WSST + Espectrograma



(f) WSST + Mel

Figura 62. Matrices de confusión obtenidos de AlexNet

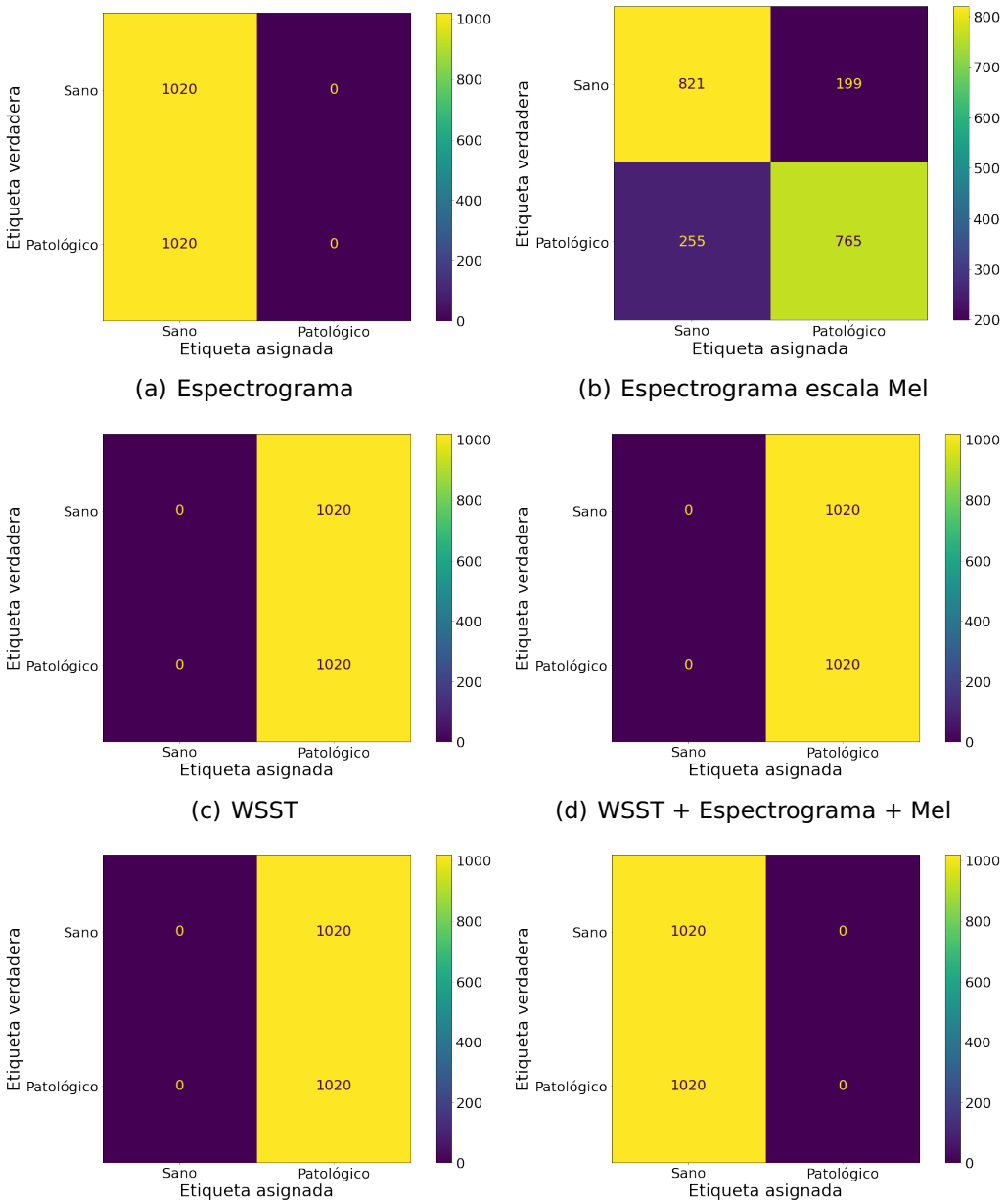
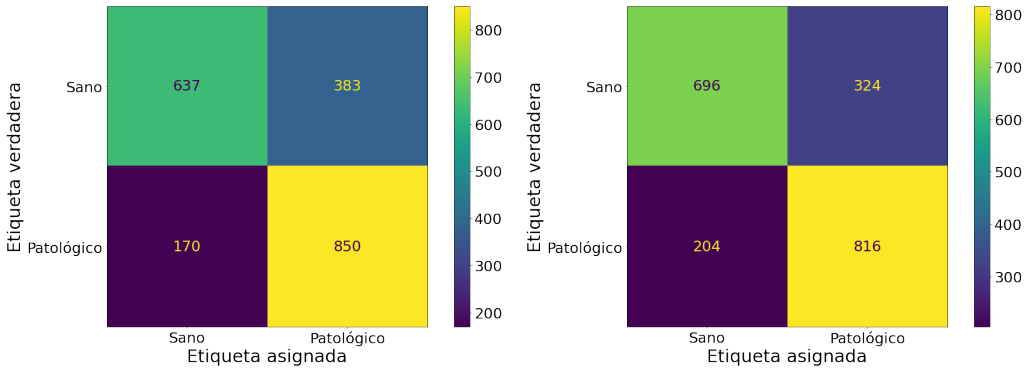
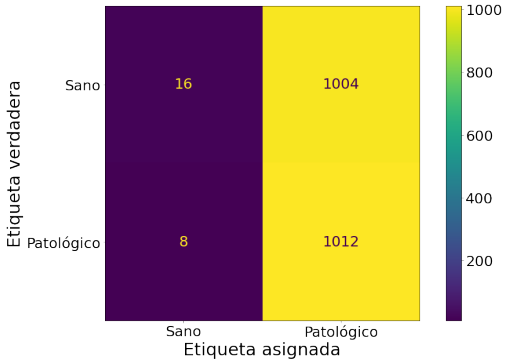


Figura 63. Matrices de confusión obtenidos de VGG16

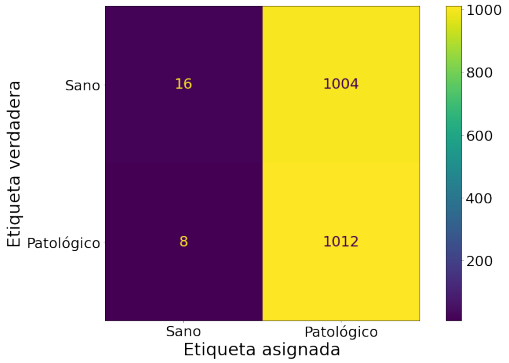


(a) Espectrograma

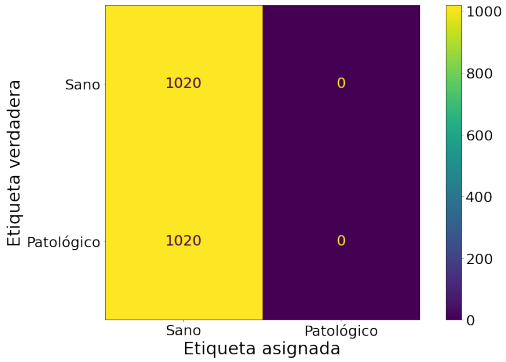
(b) Espectrograma escala Mel



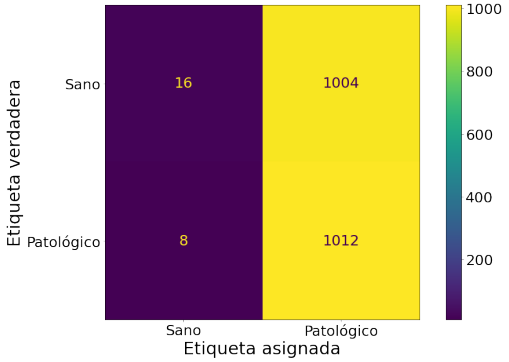
(c) WSST



(d) WSST + Espectrograma + Mel



(e) WSST + Espectrograma



(f) WSST + Mel

Figura 64. Matrices de confusión obtenidos de Ullah