Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022

Contents

[Resumen 1](#_Toc98188599)

[Autores 1](#_Toc98188600)

[Objetivos 1](#_Toc98188601)

[General: 1](#_Toc98188602)

[Específicos: 1](#_Toc98188603)

[Información Relevante 2](#_Toc98188604)

[Base de datos 2](#_Toc98188605)

[Descripción del formato 4](#_Toc98188606)

[Algoritmo 4](#_Toc98188607)

[Subir el Algoritmo 4](#_Toc98188608)

[Métrica del desafío 4](#_Toc98188609)

[Descripción de los datos 6](#_Toc98188610)

[Metodología 6](#_Toc98188611)

[Metas (Cronograma) 6](#_Toc98188612)

[Resultados 6](#_Toc98188613)

[Referencias 6](#_Toc98188614)

# Resumen

Las enfermedades

# Autores

Andrés Felipe Romero Gómez,

# Objetivos

## General:

Crear un algoritmo que permita identificar la presencia, la ausencia o los casos poco claros de soplos en las grabaciones de sonidos cardíacos recogidos en múltiples lugares de auscultación del cuerpo utilizando un estetoscopio digital.

## Específicos:

* Realizar una lectura y preprocesamiento de los datos.
* Extraer características en tiempo y frecuencia de las señales de audio.
* Hacer un análisis exploratorio de los datos demográficos y de las características extraídas a las señales de audio.
* Evaluar 2 métodos de aprendizaje automático con las métricas provistas por el challenge. El primer método con algoritmos clásicos de ML, el segundo usando DL (RNN, CNN y Transformers).
* Implementar los modelos.

# Información Relevante

https://moody-challenge.physionet.org/2022/#rules

## Base de datos

Los datos de Challenge contienen una o más grabaciones de sonidos cardíacos de 1568 pacientes, así como información demográfica rutinaria sobre los pacientes de los que se tomaron las grabaciones. Las etiquetas de Challenge indican si un anotador experto detectó la presencia o ausencia de un soplo en un paciente a partir de las grabaciones, o si el anotador no estaba seguro de la presencia o ausencia de un soplo.

Los datos del Reto se recogieron en una población pediátrica durante dos campañas de cribado masivo realizadas en el noreste de Brasil en julio-agosto de 2014 y junio-julio de 2015. La recopilación de datos fue aprobada por la junta de revisión institucional 5192-Complexo Hospitalar HUOC/PROCAPE, a petición del Real Hospital Portugués de Beneficencia en Pernambuco. La población objetivo fueron individuos de 21 años o menos que se presentaron voluntariamente para el cribado con un formulario de consentimiento firmado por los padres o el tutor legal. Todos los participantes completaron un cuestionario sociodemográfico y posteriormente se sometieron a un examen clínico, una evaluación de enfermería y una investigación cardíaca

Cada paciente de los datos del Reto tiene una o más grabaciones de una o más localizaciones de auscultación destacadas: válvula pulmonar (PV), válvula aórtica (AV), válvula mitral (MV), válvula tricúspide (TV) y otras (Phc). Las grabaciones se recogieron de forma secuencial (no simultánea) en diferentes lugares de auscultación utilizando un estetoscopio digital. El número, la ubicación y la duración de los registros varían entre los pacientes.

Diagram

Description automatically generated with low confidence

Table

Description automatically generated

Los datos del Desafío están organizados en tres conjuntos distintos: conjuntos de entrenamiento, de validación y de prueba. Hemos hecho público el 60% del conjunto de datos como conjunto de entrenamiento del Desafío 2022, y hemos conservado el 40% restante como datos ocultos para fines de validación y prueba. Los datos ocultos se utilizarán para evaluar las candidaturas al Desafío 2022 y sólo se harán públicos una vez finalizado el Desafío 2022.

Para crear los conjuntos de entrenamiento, validación y prueba, el conjunto de datos original se dividió por pacientes (ningún paciente pertenecía a varios conjuntos) mediante un muestreo aleatorio estratificado para proporcionar proporciones similares de pacientes con soplos (presentes), pacientes sin soplos (ausentes) y casos desconocidos en los diferentes conjuntos. El conjunto de entrenamiento contiene 3163 grabaciones de 942 pacientes.

El conjunto de entrenamiento público contiene grabaciones de sonidos cardíacos, información demográfica rutinaria, etiquetas (presencia, ausencia o soplos desconocidos), anotaciones de las características de los soplos (ubicación, tiempo, forma, tono, calidad y grado) y segmentaciones de sonidos cardíacos. Los conjuntos privados de validación y prueba sólo contienen grabaciones de sonidos cardíacos e información demográfica.

The participants are welcome and encouraged to use external PCG or audio datasets, including the [2016 PhysioNet Challenge](https://physionet.org/content/challenge-2016/) data [[8](https://moody-challenge.physionet.org/2022/#Clifford16), [9](https://moody-challenge.physionet.org/2022/#Clifford17)] and [PhysioNet EPHNOGRAM](https://physionet.org/content/ephnogram/) dataset [[12](https://moody-challenge.physionet.org/2022/#Kazemnejad2021)] for training their models or for transfer learning

## Descripción del formato

Hay 4 tipos de archivos en el conjunto de entrenamiento. Archivos *.wav* con las señales de audio, archivos *.hea* que describen los archivos *.wav*, los archivos *.tsv* contienen los segmentos respecto al inicio de s1 y s2.

Los conjuntos de datos de validación y de prueba tienen la misma estructura, pero el archivo .txt no proporciona información sobre los soplos, y los archivos de segmentación .tsv no se proporcionan.

Para los archivos *.tsv*, se tiene la siguiente numeración para cada una de las ondas

* The S1 wave is identified by the integer 1.
* The systolic period is identified by the integer 2.
* The S2 wave is identified by the integer 3.
* The diastolic period is identified by the integer 4.
* The unannotated segments of the signal are identified by the integer 0.

## Algoritmo

Para cada paciente (independientemente del número de ubicaciones de registro), su algoritmo debe identificar la etiqueta de clase (presente, ausente, desconocido) así como una puntuación de probabilidad o confianza para cada clase por ID de sujeto. Como ejemplo, suponga que tiene cuatro grabaciones en cuatro lugares del cuerpo, su clasificador necesita analizar esas grabaciones, pero al final debe generar sólo una etiqueta (por ejemplo, presente) por ID de sujeto con la puntuación/probabilidad para todas las clases, que son números entre cero y uno.

### Subir el Algoritmo

Asd

## Métrica del desafío

Para este problema, asumimos que los algoritmos pueden cometer errores de diagnóstico. A la inversa, suponemos que los médicos de cabecera y los especialistas no cometen errores de diagnóstico, pero el cribado de los médicos de cabecera puede ser inseguro/desconocido y requerir que un especialista lo identifique definitivamente como un caso positivo o negativo. Suponemos que estos casos desconocidos se determinan como positivos una fracción α de las veces y negativos una fracción 1-α de las veces.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

## Descripción de los datos

Cada una de las variables del txt se describe en: <https://moody-challenge.physionet.org/2022/data/>, es de importancia:

La etiqueta ID adicional proporciona el otro identificador utilizado por los sujetos que asistieron a las dos campañas de cribado.

Nótese que para los sujetos que asistieron a ambas campañas de cribado, los datos del mismo sujeto (independientemente de la campaña de cribado) se proporcionan sólo en el conjunto de entrenamiento, validación o prueba.

Las muestras de audio adquiridas se segmentaron automáticamente utilizando los tres algoritmos propuestos en [15], [16] y [17]. Estos algoritmos sólo se utilizaron para detectar e identificar los sonidos cardíacos fundamentales (sonidos S1 y S2) y sus correspondientes límites. Dos fisiólogos cardíacos inspeccionaron los resultados de los algoritmos en datos mutuamente excluyentes (ya que cada experto examinó sólo una de las dos campañas). En consecuencia, cada experto analizó las anotaciones automáticas y, cuando el anotador no estaba de acuerdo con las anotaciones automáticas sugeridas, era necesario realizar una anotación manual. En estos casos, se instruyó al anotador para que anotara al menos cinco ciclos cardíacos completos. Las etiquetas de segmentación se mantuvieron para las secciones de las grabaciones de sonidos cardíacos que los fisiólogos cardíacos consideraron de alta calidad y representativas. El resto de la señal puede incluir tanto datos de baja como de alta calidad. De este modo, los usuarios del conjunto de datos pueden optar por utilizar (o no) las ventanas de tiempo sugeridas, en las que se inspeccionó manualmente la calidad de la señal y se validaron las etiquetas automatizadas.

# Metodología

<https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>

<https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/>

<https://github.com/dataprofessor/imbalanced-data/blob/main/imbalanced_learn.ipynb>