Universidad de Valladolid

Facultad de Medicina

Escuela de Ingenierías Industriales

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

PROCESADO AUTOMÁTICO DE LA SEÑAL DE VARIABILIDAD DEL RITMO CARDIACO PARA LA CARACTERIZACIÓN Y MANEJO DE LA EPOC. ESTUDIO DE LA INFLUENCIA DEL NIVEL DE EOSINÓFILOS

Autor:

Yixiao Liang Zhou

Tutores:

Daniel Álvarez González

Tomás Ruiz Albi

Valladolid, Septiembre de 2024

TÍTULO:	**Procesado automático de la señal de variabilidad del ritmo cardiaco para la caracterización y manejo de la EPOC. Estudio de la influencia del nivel de eosinófilos**
AUTOR:	**D. Yixiao Liang Zhou**
TUTORES:	**D. Daniel Álvarez González** **D. Tomás Ruiz Albi**
DEPARTAMENTO:	**Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática** **Departamento de Medicina, Dermatología y Toxicología**

TRIBUNAL

PRESIDENTE:	**D. Jesús Poza Crespo**
SECRETARIO:	**D. Daniel Álvarez González**
VOCAL:	**D. Tomás Ruiz Albi**
SUPLENTE 1:	**D. Roberto Hornero Sánchez**
SUPLENTE 2:	**D. Carlos Gómez Peña**

FECHA:	**Septiembre de 2024**
CALIFICACIÓN:	

A Jin y Jin Yuan,

mis más queridos hermanos.

AGRADECIMIENTOS

Antes de todo, debo agradecer de todo corazón a mis tutores, Dr. Daniel Álvarez González, por la paciencia y el indudable apoyo que me ha brindado durante todo el proyecto y Dr. Tomás Ruiz Albi, por la guía y la ayuda que me ha prestado en todo momento durante mi estancia en el Servicio de Neumología.

Asimismo, también debo transmitir mis agradecimientos a Fernando Moreno, por la ayuda y la compañía prestada durante el proyecto y las prácticas. Agradecer también, a todo el Servicio de Neumología del Hospital Universitario Río Hortega, en especial a las enfermeras y las auxiliares de enfermería, por su amabilidad y colaboración brindada durante toda mi estancia allí.

Paralelamente, agradezco a todos los momentos, conocimientos, amigos, profesores y buenas enseñanzas que me han brindado durante estos cuatro años de carrera, etapa de mi vida que sin duda recordaré con mucho cariño y ternura.

Por último, pero no menos importante, debo de expresar mi más profunda gratitud a mis padres, por la crianza, el apoyo incondicional y la confianza que siempre han tenido en mí. A mis hermanos, unos de mis motivos de orgullo, también por el apoyo que me han brindado.

Resumen

Antecedentes. La Enfermedad Pulmonar Obstructiva Crónica (EPOC) hace referencia a una patología pulmonar caracterizada por la obstrucción de las vías aéreas de forma progresiva e irreversible. Los pacientes de EPOC presentan 3 sintomatologías denominadas "la triada clásica", las cuales son: la bronquiolitis crónica, obstrucción de las vías aéreas y enfisema. Hay genes que predisponen a pacientes a sufrir esta enfermedad, pero la mayoría de los casos ocurren en pacientes que han estado expuestos a factores ambientales nocivos, como el humo de las fábricas, contaminación aérea, sustancias químicas y sobre todo, tabaquismo. Se diagnostica mediante una prueba espirométrica y no existe cura, los tratamientos disponibles van orientados al control y manejo de la enfermedad. La EPOC un gran desafío para la comunidad médica ya que supone una de las causas de mortalidad más importante en los países desarrollados, incluido el nuestro. Con una estimación de 328 millones de afectados mundialmente, es la tercera causa de muerte, sólo por detrás de las cardiopatías isquémicas y los accidentes cerebrovasculares.

Hipótesis y Objetivos. El presente trabajo se ha desarrollado bajo la hipótesis de que el análisis de la señal de variabilidad del ritmo cardiaco (HRV) podría aportar información adicional y relevante para estudiar la asociación entre el nivel de eosinófilos en la sangre con la tasa de exacerbaciones en pacientes con EPOC. El objetivo principal del estudio consistió en caracterizar las señales de variabilidad del ritmo cardiaco de pacientes diagnosticados de EPOC con y sin eosinofilia para determinar los índices autonómicos con mayor nivel de asociación y poder diferenciarlos en dos grupos separados.

Material y Métodos. Para la toma de datos, se reclutaron a pacientes que acudieron de forma consecutiva al Servicio de Neumología del Hospital Universitario Río Hortega de Valladolid, sean pacientes ya previamente diagnosticados con la enfermedad o pacientes de nuevo diagnóstico. Tras determinar la tasa de eosinófilos, se les agrupó en 2 categorías: (i) sujetos de EPOC sin eosinofilia, que podrían considerarse el grupo control, (ii) sujetos de EPOC con eosinofilia. Se les registró a todos la señal de electrocardiograma en reposo durante un mínimo de 5 minutos. La metodología llevada a cabo en este proyecto se dividió en 3 fases: extracción, selección y clasificación de características. En la extracción, se aplicaron métodos de análisis temporal, frecuencial y no lineal a la señal HRV. En la selección de características, se usaron algoritmos genéticos para realizar una criba de todas las variables halladas en la fase previa, para quedarnos únicamente con un subconjunto óptimo de ellas. Por último, la clasificación se realizó a través de una red neuronal de

perceptrón multicapa (MLP), entrenada para identificar los dos grupos de paciente bajo estudio.

Resultados. Un total de 29 pacientes cumplieron los criterios de inclusión-exclusión para la incorporación al estudio. De los 29 sujetos, 27 pasaron a la etapa de modelado de datos, con la siguiente distribución: (i) 21 sujetos controles y (ii) 6 sujetos con eosinofilia. Además de obtener correlaciones leves-moderadas en ciertas relaciones como la tasa de eosinófilos y NN50 o RMSSD, entre otros, se seleccionaron dos conjuntos de características mediante algoritmos genéticos: (i.1) NN50, (ii.1) años fumando, (iii.1) frecuencia respiratoria y (iv.1) test CAT; y el grupo dos siendo: (i.2) NN50, (ii.2) cigarros/día, (iii.2) número de ingresos y (iv.2) FEV1. El subconjunto 1 demostró alcanzar un rendimiento y capacidad de generalización mayor. Tomando estas variables como única entrada a la red neuronal MLP, se alcanzó una precisión del 89%, y en las métricas de clasificación alcanzó una sensibilidad del 67% y una especificidad del 100%.

Conclusiones. El análisis automático de la actividad cardiaca ha demostrado aportar información relevante y complementaria a las variables clínicas tradicionales sobre el papel de los eosinófilos en la EPOC. Los resultados obtenidos sugieren que el empleo de esta información podría tener cierta utilidad en el esclarecimiento de la relación EPOC-eosinofilia, aunque se deben llevar a cabo estudios adicionales para su confirmación.

Palabras Clave

EPOC, eosinofilia, fenotipo EPOC eosinofílico, procesado de señales biomédicas, variabilidad del ritmo cardiaco, aprendizaje computacional, algoritmos genéticos, redes neuronales.

Abstract

Background. The Chronic Obstructive Pulmonary Disease (COPD) refers to a lung pathology characterized by progressive and irreversible airway obstruction. COPD patients present with three symptoms known as "classic triad", which includes chronic bronchiolitis, airway obstruction and emphysema. There are genes that predispose patients to this disease, but most of the cases occurs in patients that have been exposed to harmful environmental factors, such as factory smoke, air pollution, chemical substances, and, above all, smoking. It is diagnosed through a spirometry test and there is no cure for this disease. Available treatments focus on controlling and managing it. COPD poses a great challenge for the medical community as it is one of the leading causes of mortality in developed countries, including ours. With an estimated 328 million people affected worldwide, it is the third leading cause of death, only behind ischemic heart disease and stroke.

Hypothesis and Objectives. The present work has been developed under the main hypothesis that the analysis of the heart rate variability (HRV) signal could provide additional and relevant information to study the association between the blood eosinophil levels and the rate of exacerbations in COPD patients. Accordingly, the main objective of this research was to characterize the HRV signal of COPD patients with and without eosinophilia to determine the autonomic indices with the highest level of association and to differentiate them into two separate groups.

Materials and Methods. For data collection, patients consecutively referred to the Pulmonology Department at Río Hortega University Hospital of Valladolid were recruited, either they were already diagnosed with the disease or newly diagnosed patients. After determining the eosinophil levels, they were grouped into two categories: (i) COPD patients without eosinophilia, which can be considered the control group, (ii) COPD patients with eosinophilia. The resting electrocardiogram signal was recorded for all of them for a minimum of 5 minutes. The methodology carried out in this project is divided into three phases: feature extraction, selection and classification. In the extraction stage, temporal, frequency and nonlinear analysis methods were applied to the HRV signal. For feature selection, Genetic Algorithms were used to filter all the variables found in the previous phase, focusing only on an optimal feature subset. Finally, classification was performed using a multilayer perceptron neural network (MLP), trained to identify the two patient groups under study.

Results. A total of 29 patients met the inclusion-exclusion criteria and were enrolled in the study. Out of the 29 subjects, 27 progressed to the data modelling stage, with the following distribution: (i) 21 control subjects and (ii) 6 subjects with eosinophilia. In addition to obtaining mild-to-moderate correlations in certain relationships, such as eosinophil rate and NN50 or RMSSD, among others, two feature subsets were selected using Genetic Algorithms: the first subset included (i.1) NN50, (ii.1) years of smoking, (iii.1) respiratory rate, (iv.1) CAT test; the second group included (i.2) NN50, (ii.2) cigarettes/day, (iii.2) number of hospital admissions and (iv.2) FEV1. The subset 1 showed higher performance and generalizability. Using these variables as unique input to the MLP neural network, the model reached an accuracy of 89%, with a sensitivity of 67% and a specificity of 100%.

Conclusions. The automatic analysis of cardiac activity has proven to provide relevant and complementary information to conventional clinical variables about the role of eosinophils in COPD. The results suggest that the use of this information could be useful for clarifying the COPD-eosinophilia relationship, although further studies must be carried out to confirm these findings.

Keywords

COPD, eosinophilia, eosinophilic COPD phenotype, biomedical signal processing, heart rate variability, computational learning, genetic algorithms, neural networks.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Glosario de siglas y acrónimos

Acc	Precisión
ACO	Solapamiento asma-EPOC
AOS	Apnea del sueño
ApEn	Entropía aproximada
APEPOC	Asociación de pacientes con EPOC
AR	Modelos autorregresivos
AVNN	Media de intervalos NN
CAT	*COPD Assessment Test*
CEIm	Comité de Ética de la Investigación con Medicamentos
CTM	Medida de la tendencia central
CV	Enfermedades Cardiovasculares
DFT	Transformada de Fourier discreta
DM	Diabetes Mellitus
ECG	Electrocardiograma
EDR	Respiración derivada del electrocardiograma
EOS	Eosinofílico
EPOC	Enfermedad Pulmonar Obstructiva Crónica
FC	Frecuencia cardiaca
FEV_1	Volumen espiratorio forzado en el primer segundo
FEV_1/FVC	Relación entre el volumen espiratorio y capacidad vital forzadas
FFT	Transformada rápida de Fourier
FN	Falsos negativos
FP	Falsos positivos
FR	Frecuencia respiratoria
FVC	Capacidad vital forzada

GA	Algoritmo genético
GOLD	*Global Initiative for Chronic Obstructive Lung Disease*
HF	Banda de alta frecuencia
HRV	Variabilidad del ritmo cardiaco
HTA	Hipertensión arterial
IMC	Índice de masa corporal
LABA	*Long-Acting Beta Agonist*
LAMA	*Long-Acting Muscarinic Agonist*
LF	Banda de baja frecuencia
LR-	Razón de verosimilitud negativa
LR+	Razón de verosimilitud positiva
LZC	Complejidad de Lempel-Ziv
MF	Frecuencia mediana
MLP	Perceptrón multicapa
MMP	Metaloproteinasas de la Matriz extracelular
mMRC	*Modified Medical Research Council Dyspnea Scale*
Nh	Número de nodos para la red neuronal
NHBLI	*National Heart, Lung and Blood Institute*
NN50	Número de intervalos NN adyacentes que difieran más de 50ms
NSA	Nódulo sinoauricular
OCD	Oxigenoterapia domiciliaria
OMS	Organización Mundial de la Salud
PaO_2	Presión parcial de oxígeno
PC	Probabilidad de cruce
P_{HF}	Potencia en frecuencias altas
P_{LF}	Potencia en frecuencias bajas
Pm	Probabilidad de mutación
pNN50	Porcentaje de intervalos NN que difieran más de 50ms

PSD	Densidad espectral de potencia
PT	Potencia Total
P_{VLF}	Potencia en frecuencias muy bajas
RMSSD	Raíz cuadrada de la media de las diferencias en intervalos NN
SABA	*Short-Acting Beta Agonist*
SAMA	*Short-Acting Muscarinic Agonist*
SampEn	Entropía muestral
SDNN	Desviación estándar de intervalos NN
SE	Entropía espectral
Sen	Sensibilidad
SGDM	Descenso de gradiente estocástico con optimizador de momento
SNA	Sistema nervioso autónomo
Sp	Especificidad
TAD	Tensión arterial diastólica
TAS	Tensión arterial sistólica
TINN	Índice triangular de la señal de HRV
TN	Verdaderos negativos
TP	Verdaderos positivos
ULF	Bandas de frecuencia ultra bajas
VLF	Banda de frecuencias muy bajas
VPN	Valor predictivo negativo
VPP	Valor predictivo positivo
α	Regularización para la red neuronal

1. Introducción a la Enfermedad Pulmonar Obstructiva Crónica

La enfermedad pulmonar obstructiva crónica (EPOC) es una patología pulmonar que se caracteriza por la obstrucción de las vías aéreas de forma progresiva e irreversible.

A pesar de que el nombre que se le da procede de los siglos XVII y XVIII, esta enfermedad ya fue descrita en los tiempos de la antigua Grecia, donde el médico Hipócrates describió síntomas similares a la EPOC, relacionándolo con la inhalación de sustancias tóxicas como humo y vapores.

Más tarde, René Théophile introdujo el término de "obstrucción bronquial", y durante la Revolución Industrial se produjo un aumento significativo de incidencias de esta enfermedad, debido a surgimiento de fábricas y la quema masiva de combustibles fósiles.

Sin embargo, no fue hasta mitades del siglo pasado, cuando se acuño el término de "Enfermedad Pulmonar Obstructiva Crónica" para describir a la sintomatología causada por la fusión de lo que antes se consideraban dos enfermedades separadas: bronquitis crónica y enfisema.

Fue también, durante esa época, cuando se identificó el tabaquismo como factor de riesgo para el desarrollo de la enfermedad (1).

La EPOC tiene naturaleza heterogénea y compleja. Su fenotipo y comorbilidades dependen, además, del tipo de tóxicos al que haya estado expuesto el sujeto de forma recurrente (2).

En la mayoría de ocasiones se manifiesta como una enfermedad inflamatoria progresiva de las vías aéreas, alveolos y la microvasculatura que cursan con una limitación al flujo aéreo progresiva e irreversible. Esto es principalmente debido al enfisema, que se debe a una destrucción proteolítica con remodelado de bronquiolos y alveolos.

La mayoría de esta respuesta inflamatoria desmesurada es debido a la inhalación de partículas o gases tóxicos, causando la activación de leucocitos y linfocitos que son los causantes de la destrucción y remodelación de la parénquima pulmonar.

Este proceso causa, de forma irreversible, una pérdida de la elasticidad pulmonar de forma continuada en el tiempo, lo que conlleva a una disminución del volumen

espiratorio forzado en el primer segundo (FEV1), vaciado pulmonar inadecuado durante la espiración e hiperinsuflación pulmonar.

Todo esto se puede resumir en lo que se denomina "La Triada Clásica" que presentan los pacientes con esta enfermedad:

- Bronquiolitis crónica
- Obstrucción de las vías aéreas
- Enfisema

Por último, aunque se ha hecho mucho hincapié en los factores como el tabaquismo, inhalación de gases tóxicos y otros tipos de tóxicos, también son significativos los factores genéticos, la dieta, hiperreactividad bronquial y el sexo de los sujetos.

Por otra parte, la epidemiología de la EPOC no resulta inmediata de cuantificar, ya que dependen de diferentes métodos usados para calcular su prevalencia.

Según la revista *Global Burden of COPD* publicado en 2015 por la *Oficial Journal of the Asian Pacific Society of Respirology,* no siempre se sigue un mismo método para considerar un diagnóstico de la EPOC, ya que algunas veces sólo se valora la función pulmonar sin tener en cuenta la exposición a agentes tóxicos, otras veces no se tiene en cuenta la presencia de la triada clásica que suele presentar la enfermedad.

Se estima que alrededor del 30-40% de las personas fumadoras desarrollarán, a lo largo de su vida, una EPOC, frente a un 10% de los no fumadores debido a otros factores, como los genéticos. Existen un 15-20% (3) de los casos en los que la EPOC va a ir asociada a la exposición ambiental o a la ocupacional.

Actualmente la prevalencia es mayor en hombres que en mujeres, pero se espera que la tendencia cambie en las próximas décadas, ya que el consumo de tabaco en mujeres jóvenes es superior al de hombres (2).

Según la asociación de pacientes con EPOC (APEPOC), la prevalencia de esta enfermedad en España a partir de los 40 años es de 34 casos por cada 1000 habitantes (3.4% de la población), existiendo diferencias significativas entre comunidades, siendo la Comunidad Valenciana la región con maoyr incidencia de todas, con 47.7 casos por cada 1000 habitantes. La incidencia más baja la posee Castilla la Mancha, con tan sólo 16.1 casos por cada 1000 habitantes.

Castilla y León se sitúa en torno a los 32.9 casos por cada 1000 habitantes, por lo que también es de interés el estudio en una comunidad como la nuestra (4).

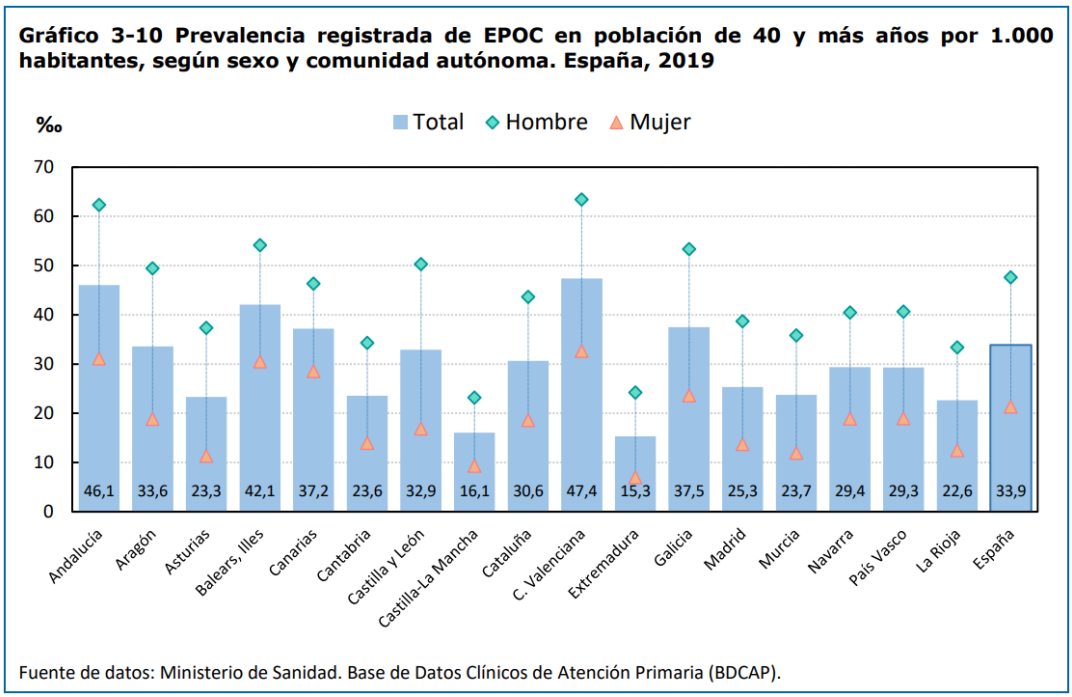

Figura 1. Prevalencia de la EPOC en España según las diferentes CCAA (4).

Con todo, su estimación ronda los 328 millones de personas, mientras que la Organización Mundial de la Salud (OMS) estima que existen 210 millones de personas en el mundo con dicha enfermedad, lo que la convierte en la tercera causa de muerte en el mundo.

1.1 Análisis de datos e inteligencia artificial

En el presente trabajo se realiza un estudio sobre la señal HRV. La elección de la señal se debe a la gran versatilidad, extensa literatura y facilidad con la que se maneja, además, es posible registrarse de forma portátil y cómoda para el paciente.

Se han implementado 3 etapas diferenciadas dentro del procesado para los registros biomédicos: extracción de características, selección de variables y clasificación.

Se extraerán métricas temporales, espectrales y no lineales en la extracción en la primera etapa. En la segunda, se trabajará con los algoritmos genéticos para que nos ayude a elegir las variables más significativas. Por último, como algoritmo de clasificación binaria, se plantea usar una red neuronal, comparándola con un modelo clásico, que será la regresión logística.

1.2 Estructura

La memoria del presente trabajo fin grado (TFG) está dividida en 9 apartados o capítulos, siguiendo los pasos convencionales que se han establecido para llevar a cabo una investigación con señales biomédicas.

El primer capítulo es introductorio, se presenta el problema bajo estudio, la relevancia e importancia del mismo, así como su prevalencia en la población. Dada la alta mortalidad que produce esta enfermedad en nuestro país, es un gran factor que motiva el estudio.

En el segundo apartado se ahonda sobre las características de la enfermedad, describiendo todos los aspectos relevantes para el estudio de la misma.

En el tercer apartado se describe el estado del arte de la señal HRV en el contexto de la EPOC. Se explica brevemente la señal a analizar y variables relevantes.

En el cuarto capítulo se establecen las hipótesis y objetivos del TFG.

En el quinto se describe la población bajo estudio, así como criterios de inclusión y exclusión, clasificación en grupos y el equipamiento utilizado para el registro de la señal HRV.

En el sexto apartado se describe la metodología utilizada para procesar la señal en los tres dominios, la selección de características mediante algoritmos genéticos y la clasificación binaria que se realizará a través de una red neuronal. Finalmente, se describen las métricas que se usarán para medir el rendimiento de la red neuronal.

En el séptimo capítulo se exponen los resultados obtenidos, comenzando por un análisis descriptivo, hasta la extracción de características, selección de variables y, por último, evaluación del clasificador empleado.

La discusión de los resultados se relata en el octavo capítulo, donde se tratarán de justificar los resultados obtenidos, así como las limitaciones del estudio.

Finalmente, en el último capítulo se describirán las conclusiones, destacando los hallazgos encontrados y futuras líneas de investigación.

2. HIPÓTESIS Y OBJETIVOS.

2.1 Hipótesis

El presente trabajo se sustenta sobre varias hipótesis. En primer lugar y desde el punto de vista clínico se busca estudiar las potenciales diferencias entre los fenotipos agudizador y no agudizador de la EPOC haciendo hincapié entre dos grupos de pacientes: (i) pacientes con una tasa normal de eosinófilos en sangre, considerado en la literatura, inferior a 300 células por microlitro (grupo NO eosinofílico) y (ii) pacientes con una tasa anormalmente alta de eosinófilos, cuyo conteo es de mayor o igual a 300 células por microlitro (grupo eosinofílico).

Se espera que los pacientes con el perfil eosinofílico puedan presentar mayor número de exacerbaciones en el contexto de la enfermedad en comparación con aquellos que no presenten esta tasa de eosinófilos elevada.

Por otra parte, se espera que los pacientes del grupo positivo presenten cambios relevantes en su modulación autonómica, por lo que además de los índices de función pulmonar habituales, se busca analizar la HRV, derivada del ECG en los diferentes dominios (tiempo, frecuencial y no lineal).

Se espera que, abordando la caracterización de dichos pacientes, sea posible aportar información útil para detectar los potenciales cambios en la modulación autonómica asociada a este fenotipo eosinofílico, y de ahí, medidas de cuantificación de su magnitud.

2.2 Objetivos

De acuerdo con las hipótesis anteriormente expuestas, el **objetivo principal** de este Trabajo Fin de Grado consiste en analizar y caracterizar las diferencias en la modulación autonómica de pacientes de EPOC con y sin eosinofilia, intentando, a partir de un modelo automático, predecir la pertenencia a uno de estos grupos y el posible efecto de estas células sobre el curso de la enfermedad.

Para lograr este objetivo principal, se han establecido los siguientes **objetivos específicos**:

- Estudiar y describir los antecedentes, variables clínicas, sintomatología y la afectación de la calidad de vida de pacientes con EPOC, así como los efectos de la eosinofilia que se conocen actualmente sobre la enfermedad.

- Caracterizar la modulación automática cardiaca de los dos grupos de pacientes a través del análisis automático de la señal de HRV en los dominios temporal, espectral y no lineal.

- Describir las variables clínicas y derivadas del procesado automático de la señal, estableciendo estudios de correlación y normalidad.

- Hallar posibles correlaciones significativas entre las variables clínicas o de procesado de la señal HRV con la tasa de eosinófilos.

- Desarrollar y validar un modelo automático para la clasificación binaria de pacientes con EPOC en 2 grupos mutuamente excluyentes, implementando una etapa de selección de variables y posteriormente un modelo de regresión logística binaria.

- Establecer una comparación entre la predicción realizada por el modelo automático y un modelo de regresión logística binaria (tradicional).

3. ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA

Pasado el capítulo introductorio, será objetivo de este capítulo el describir de forma exhaustiva la enfermedad propiamente dicha, eso incluye aspectos como la fisiopatología, sintomatología, seguimiento y tratamiento, así como la magnitud del problema.

3.1 Factores de riesgo para la EPOC

La EPOC es una enfermedad causada por una inflamación progresiva y crónica de las vías aéreas, los alveolos y la microvasculatura de estos que es causada, en la mayoría de los casos, por la inhalación de gases tóxicos o que causan inflamación. Lo que resulta, a la larga, en una limitación al flujo aéreo de forma irreversible. Esta limitación es la característica esencial y definitoria de la enfermedad.
En general, los casos de EPOC en países desarrollados se atribuyen al tabaquismo, no obstante, también existen multitud de factores de riesgo que causan el desarrollo de esta patología:

- **Factores genéticos**
  En los últimos años, han surgido múltiples estudios indicando que la EPOC puede tener influencias genéticas. Esto es debido al hecho de que no todos los pacientes fumadores desarrollan la enfermedad, sino que esto sólo ocurren en un 15-20% de ellos.
  El flujo espiratorio forzado en el primer segundo (FEV1) puede estar influenciado por uno o varios genes. Existen candidatos que han demostrado tener influencia, aunque sus efectos no están del todo dilucidados actualmente (2).
  En la Tabla 1 se muestran los principales genes relacionados con la predisposición a desarrollar EPOC.

Tabla 1. Genes que posiblemente influyan en el desarrollo de la EPOC (2).

Genes candidatos a influir en el desarrollo de la EPOC
• **α1 antritripsina (AAT)**
• **α1-antiquimiotripsina (AACT)**
• **Hidrolasa epóxida micosomal (EPHX)**
• **Glutation-S-transferasa (GSTs)**
• **Hemooxigenasa-1 (EPHx)**
• **Factor de necrosis tumoral (TNF- α)**
• **Regulador de transmembrana de fibrosis quística (CFRT)**

- **Factores dietéticos**

  El exceso de ciertas vitaminas antioxidantes como la A, C y E se asocian a mayor riesgo de EPOC. Dietas bajas en carbohidratos con ácidos grasos insaturados, encontradas típicamente en el pescado y aceite de oliva cuyas propiedades pueden ayudar a disminuir la producción de CO_2 se consideran beneficiosas, así como dietas con alto contenido en vegetales (2).
- **Atopia e hiperreactividad bronquial**

  Se ha reportado que la mortalidad por EPOC es mayor en aquellos individuos con atopia más grave y con mayor hiperreactividad de las vías aéreas. También se ha identificado una asociación importante con la eosinofilia, de la cual se tratará en profundidad más adelante.
- **Sexo**

  En general, la prevalencia de la EPOC en mujeres es mayor que en hombres, pero es discutida su influencia ya que también podría ser que las mujeres sean más sensibles a los efectos del tabaco que los hombres.
- **Factores ambientales**

  Se denomina factor ambiental aquellos aspectos que no dependen exclusivamente del individuo, a excepción quizás del tabaquismo, podemos encontrar los siguientes (2,5):
  - **Tabaquismo**, uno de los factores prevenibles más importantes causantes de la EPOC. Las partículas del humo del tabaco inspiradas por el paciente causan inflamación y destrucción de la parénquima pulmonar, lo que conlleva al desarrollo de la bronquitis y enfisema, que son dos de los tres pilares fundamentales de la enfermedad.

    Si bien es cierto que no todos los fumadores la desarrollan, hay un porcentaje significativo que sí lo hacen, combinado con el gran número de fumadores en los países desarrollados, hace de la

EPOC, una enfermedad muy estrechamente relacionada con el consumo del tabaco y prevalente en países como el nuestro.

- **Contaminación atmosférica**, en especial la aspiración del dióxido de azufre, humo negro, material particulado, humos o gases tóxicos, producidos por la quema de combustibles como la leña o carbón pueden ser causantes de la enfermedad.
- **Químicos, polvo o gases en el ambiente laboral**, lo que se conoce como **EPOC ocupacional**. Se sabe que el polvo del carbón, sílex, cuarzo, vapores de isocianato y disolventes son un factor de riesgo para desarrollar la EPOC.
- **Infecciones respiratorias**, sobre todo las infecciones víricas latentes provocados por un adenovirus, pueden causar inflamación en el pulmón.

3.2 Fisiopatología de la EPOC

La EPOC es el resultado de la conjunción de diferentes, pero no incorreladas, patologías. Tradicionalmente se considera la tríada clásica conformada por (2):

- Bronquitis crónica
- Obstrucción de las vías aéreas
- Enfisema

La bronquiolitis crónica es una patología causada por la inflamación de los bronquios, lo que hace disminuir su calibre y por ende el flujo de aire que entra y sale de ellos. Simultáneamente, aumenta la secreción mucosa en dichos conductos debido a un mayor número de células caliciformes y glándulas submucosas agrandadas debido a la irritación constante a la que se les somete. Estas glándulas submucosas producen una secreción mayor que las células caliciformes. Sin embargo, no sólo es mucosa lo que se segrega, sino también mucinas, las cuales espesan el moco, estrechan las vías aéreas pequeñas y aceleran el deterioro de la función pulmonar, contribuyendo a la magnificación de la obstrucción.

Se considerará esta patología como crónica cuando existe tos y expectoración durante la mayoría de los días en un periodo de 3 meses al año, durante 2 años consecutivos. El exceso de mucosidad que presentan los pacientes se elimina a través de la tos, que suele ser peor al despertarse por la mañana. El esputo puede ser amarillo o verde y contener trazas de sangre.
Dado que la enfermedad es crónica y progresiva, la tos se vuelve ineficaz poco a poco y el exceso de mucosa acaba por obstruir gran parte de las vías aéreas. Se

sabe que una bronquiolitis crónica grave da lugar a un mayor número de exacerbaciones.

Por otra parte, el enfisema pulmonar es una patología que cursa con destrucción de la pared alveolar pudiendo dar lugar a fibrosis, donde también se produce el agrandamiento de los espacios aéreos distales a los bronquiolos terminales.

De ahí que los pacientes se hayan clasificado tradicionalmente como *Blue Bloaters*, refiriéndose a pacientes con sobrepeso, bronquitis crónica e hipoxemia, y *Pink puffers*, para pacientes de bajo peso, enfisema y con niveles de oxígeno normales en reposo (6), que se muestran en la Figura 2.

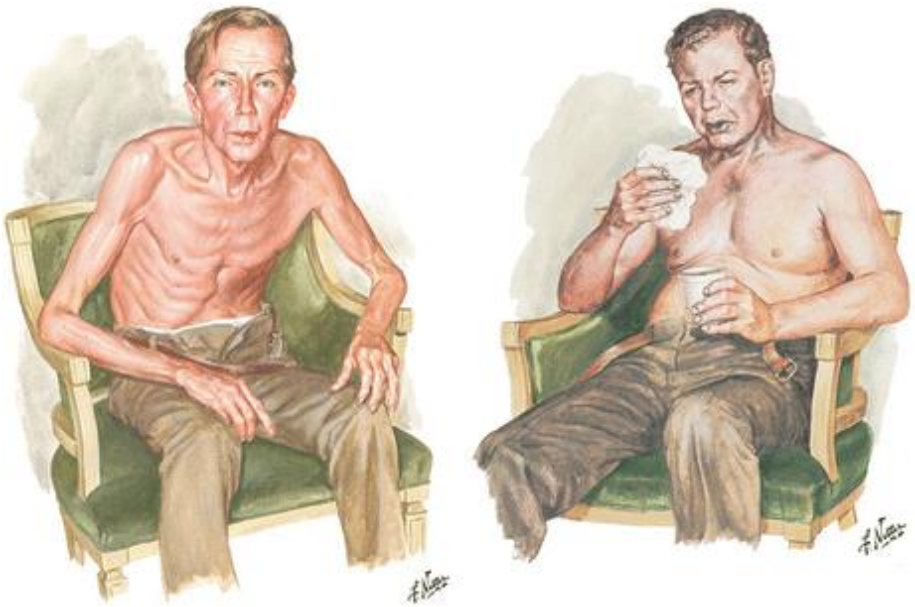

Figura 2. Los fenotipos tradicionales Blue Bloater (derecha) y Pink Puffer (izquierda) (7).

Se teoriza que en los pulmones existe un equilibrio entre la síntesis y degradación de elastina, que se ve alterada en el enfisema pulmonar, principalmente debido a la inhalación de agentes tóxicos como el humo del tabaco y contaminantes.

Sumado a lo anterior, la existencia de la alfa 1-antitripsina, una proteína cuya función es la de inhibidor de proteasa sérico, protege a los tejidos de la acción de las proteasas segregadas por las células inflamatorias. Aquellas personas que presenten un déficit genético de esta proteína tienen mayor probabilidad de desarrollar el enfisema pulmonar (2).

La EPOC se evidencia clínicamente con polipnea, taquipnea y disminución del murmullo vesicular al auscultarse. El enfisema causa además disnea, escasa expectoración, alteraciones en la difusión de gases evidenciado por la prueba del carbono, resistencia aérea ocasionalmente elevada e hiperinsuflación (5,8).

Todos estos mecanismos pueden ser explicados cuando consideramos la inhalación de gases o partículas tóxicas de forma consistente.

Estas partículas, generalmente inorgánicas, provocan una respuesta inflamatoria inespecífica por parte de los leucocitos, que liberan una serie de moléculas intracelulares como las citocinas, factor necrótico tumoral alfa e interleucinas 1 y 8. Los factores quimiotácticos cumplen la función de reclutar a los macrófagos, linfocitos y neutrófilos para que viajen al lugar de inflamación. Una vez allí, las células liberarán una serie de proteasas que, junto con las especies reactivas al oxígeno, afectarán a las metaloproteinasas de la matriz (MMP) y a la elastasa de los neutrófilos. Lo que provoca daños y alteraciones en el tejido conjuntivo. Se ha teorizado la posibilidad de que el humo del tabaco pueda dañar la célula epitelial bronquial y con el tiempo, se pueden generar agregados linfoides junto a los pequeños conductos que exacerban todavía más la patología.

La conjunción de estas características llevan a teorizar que la EPOC pueda tener carácter autoinmune (2).

Asimismo, los pacientes de EPOC también pueden clasificarse según sus fenotipos, ya que muchos de ellos no mueren por el EPOC en sí, si no por comorbilidades o enfermedades adjuntas como las cardiovasculares o cáncer.

Por lo general, los fenotipos de la EPOC incluyen (9):

- **Fenotipo de bronquitis crónica**: Se caracterizan por una inflamación bronquial crónica. Conducen, muy a menudo, a moco y expectoración. Se estima que alrededor del 50-70% de los pacientes con EPOC presentan este fenotipo.
- **Fenotipo enfisema**: Se caracteriza por la destrucción de la parénquima pulmonar, siendo especialmente afectados los alvéolos. Conduce a una capacidad reducida de respiración.
  Se estima que el 15-50% de los pacientes con EPOC lo presentan (9).
- **Fenotipo mixto**: Los pacientes presentan una mezcla de los dos anteriores. Se estima entre el 20-30% de los casos.
- **Fenotipo eosinofílico**: El cual es objeto de nuestro estudio. Estos pacientes se caracterizan por un aumento en los niveles de eosinófilos en sangre y tejido pulmonar, lo que se asocia a una mayor respuesta a los corticoides y una mayor susceptibilidad a las exacerbaciones. Además, pueden estar relacionadas con la inflamación de las vías aéreas, lo que contribuye en la obstrucción del flujo aéreo y el empeoramiento de los síntomas respiratorios asociados a la EPOC. Algunos estudios estiman que el 20-30% (9) de los pacientes con la enfermedad puedan presentar un fenotipo eosinofílico, aunque ésta es difícil de estimar debido al tratamiento por corticoides.
- **Otros fenotipos:** En ellos pueden incluirse los asociados a la exposición al humo de biomasa, edad, factores genéticos y medioambientales.

3.3 Diagnóstico de la EPOC.

La enfermedad se diagnostica principalmente mediante una prueba espirométrica (7,8).

En esta prueba, se le pide al paciente que realice 4 acciones:

- Respirar de forma normal
- Vaciar el pulmón
- Insuflarse con todo el aire posible
- Expulsar el aire insuflado con la mayor velocidad posible

Este procedimiento mide la capacidad funcional pulmonar y son muchas otras variables las que pueden obtenerse a partir de esta prueba, pero los índices espirométricos son especialmente relevantes para el diagnóstico de esta enfermedad:

- $FEV_{1,}$ nos indica el volumen de aire absoluto que el sujeto es capaz de espirar en el primer segundo. Interesa tanto el pre como post-broncodilatador.
- FEV_1/FVC, indica el porcentaje del volumen respecto al total espirado en el primer segundo.

Los sujetos que presenten un FEV_1 post-broncodilatador menor del 80% junto con un FEV_1/FVC menor del 0.7, son consideradas patológicas. Se interpretan clínicamente como obstructiva, y por tanto, compatible con la EPOC. La interpretación espirométrica se muestra en la Figura 3.

Esta prueba es ambulatoria y no invasiva, pero requiere de la colaboración del paciente para que los resultados sean concluyentes y/o manejables.

La exploración física también tiene su importancia a la hora del diagnóstico, pero cuando se advierten estas anomalías, la enfermedad suele estar relativamente avanzada.

Las radiografías y la tomografía computarizada de tórax tienen su importancia a la hora de confirmar el diagnóstico del enfisema y la bronquitis crónica, pero, al igual que la exploración física, los pulmones no suelen mostrar anomalías hasta que el paciente empeora (10).

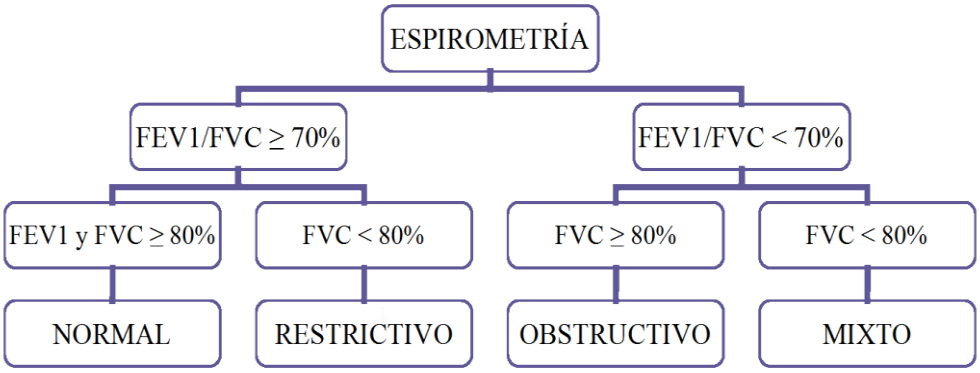

Figura 3. Interpretación clínica de la espirometría (10).

3.4 El papel de las exacerbaciones en la EPOC.

Las exacerbaciones o agudizaciones en el contexto de la EPOC, se definen como un empeoramiento repentino de los signos y síntomas que conllevan, siendo éstos generalmente tos, aumento de esputo y disnea. Un signo relativamente común es el del atrapamiento aéreo, lo que dificulta una exhalación completa. En muchas ocasiones, una agudización de la EPOC conlleva a un cambio de medicamento (11).

Una causa común de las exacerbaciones es por una infección vírica, como podría ser un resfriado común, por lo que es frecuente tener más pacientes afectados por reagudizados en invierno en comparación con cualquier otra época del año. Los esputos de estos pacientes durante estas exacerbaciones pueden ser amarillos o verdes.

Las infecciones no son la única fuente de posibles exacerbaciones, sino que también pueden influir los alérgenos, exposición a aire contaminado, humo de tabaco o de incendios, entre otros.

Las recidivas de la enfermedad son muy relevantes, pues son un indicador determinante de su progresión y el deterioro de la función pulmonar. Influyen de forma importante en la mortalidad del paciente, así como una disminución de la calidad de vida del mismo.

La condición del paciente no sólo es dependiente de su estado en la EPOC, si no que también influyen de forma significativa las comorbilidades que pueda presentar, como la insuficiencia cardiaca, diabetes mellitus, hipertensión, anemia y trastornos psicológicos, entre otros (11).

Estas exacerbaciones son más frecuentes a medida que avanza la enfermedad. Durante las primeras etapas de la enfermedad las reagudizaciones suelen ser leves o moderadas, pudiendo tratarse habitualmente de forma ambulatoria,

mientras que las severas, que se manifiestan en las fases más tardías de la enfermedad pueden conllevar a patologías asociadas como la insuficiencia respiratoria, disminuyendo drásticamente las probabilidades de supervivencia del paciente y necesitando, de forma más frecuente, los ingresos hospitalarios.

En general, el pronóstico de los pacientes de EPOC con exacerbaciones frecuentes suele ser malo, ya que indica gravedad en la patología y mayor decaimiento en las funciones pulmonares. Es más grave todavía, si tenemos en cuenta las comorbilidades, que no sólo incrementan la mortalidad de los sujetos si no que también agranda el gasto del sistema sanitario de los países con una incidencia considerable de EPOC (12).

Por último, es interesante destacar la relación entre eosinofilia y las recidivas de la EPOC. Se sabe que existen relaciones entre los eosinófilos y otras patologías pulmonares como el asma, evidenciado en los esputos. Si bien es cierto que todavía no se ha observado una relación evidente, se ha visto que en determinadas cohortes aproximadamente el 20% de los pacientes tienen un número de eosinófilos mayor de 300 cels/µL (2).

3.5 Gravedad de la EPOC según la clasificación GOLD.

Las siglas GOLD se refieren al nombre de una organización auspiciada por la OMS y el NHLBI (*National Heart, Lung and Blood Institute*) fundada en 1997 con el propósito de concienciar y mejorar la atención de la EPOC. Desde 2001 publica las guías y recomendaciones que actualizan de forma periódica con evidencias nuevas.

La forma de evaluar y clasificar a los pacientes es, además de las pruebas diagnósticas clásicas como pueden ser la espirometría y el valor FEV1/FVC, los cuestionarios de síntomas como el CAT, que es el Test de evaluación de la EPOC (*COPD Assesment Test*) y el mMRC, que se trata de un test para evaluar la disnea (*modified Medical Research Council*) (13).

Los datos recabados en base a los tests y pruebas diagnósticas anteriormente mencionadas se usan para establecer una clasificación de la severidad de la EPOC.

En la Figura 4 se muestra la clasificación GOLD para la gravedad de la EPOC, donde va del menos al más severo de los casos:

Figura 4. Clasificación de la severidad de la EPOC según la guía GOLD (14).

La guía también cataloga a los pacientes en cuatro categorías basados en la evaluación sintomática, limitación al flujo aéreo e historial de reagudizaciones.

Con todo, es una clasificación relativamente completa que actualmente se considera apropiada para el diagnóstico, clasificación y tratamiento en atención primaria.

3.6 Tratamiento de la EPOC.

La EPOC, como se ha descrito a lo largo de este último punto, es una enfermedad multifactorial y heterogénea, cuyo fenotipo y comorbilidades dependen del individuo y de las sustancias tóxicas a las que haya estado sometido. Por su naturaleza crónica y progresiva, sumado al daño irreversible que causa en la parénquima pulmonar, no hay una cura propiamente dicha, sino que más bien se habla del control de la enfermedad. Los tratamientos existentes van orientados al manejo de los síntomas y a la mejora de la calidad de vida de los pacientes.

Uno de los tratamientos más inmediatos y eficientes es el abandono del hábito tabáquico y reducción a la exposición de sustancias tóxicas.

En la figura 5 se muestran valores de la espirometría forzada en el primer segundo para pacientes que nunca han fumado, han dejado de fumar a distintas edades y aquellos que no han abandonado el hábito tabáquico.

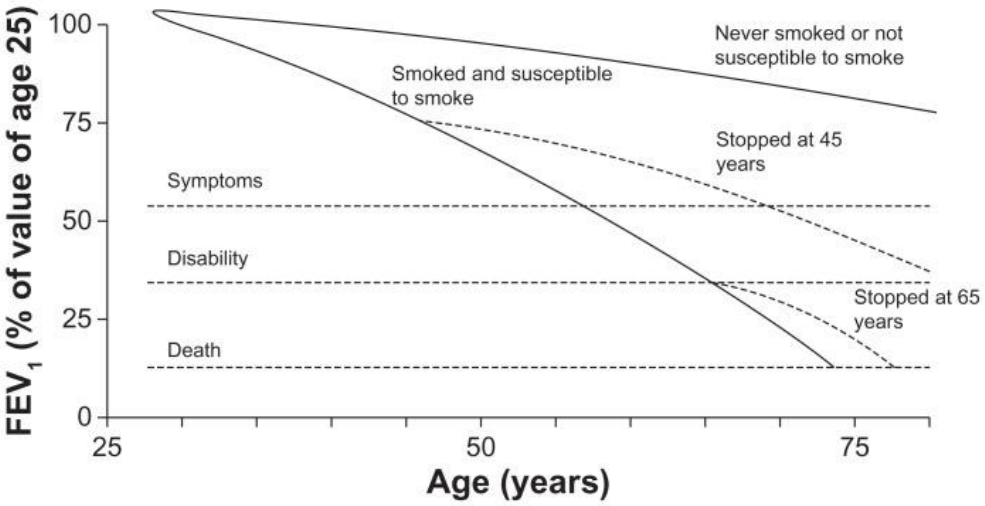

Figura 5. Comparativa de FEV_1 entre pacientes de EPOC que han abandonado el tabaco y los que no (15).

Los medicamentos recetados más frecuentemente son aquellos que estabilizan la enfermedad y previenen exacerbaciones. Se utilizan 3 grupos, principalmente: **broncodilatadores, corticoides y oxigenoterapia** (16):

- Los **broncodilatadores** se refieren a un amplio espectro de fármacos cuyo principio activo son los agonistas de los β2-adrenérgicos, como podría ser el Salbutamol o el Salmeterol, o anticolinérgicos, como podría ser el Ipratropio.
  En cuanto a su mecanismo de acción, los agonistas de β2-adrenérgicos se unen a los receptores del músculo liso en los bronquiolos causando su relajación, lo que aumenta el calibre de los mismos y mejora el flujo o paso del aire.
  Los anticolinérgicos, por otra parte, son fármacos que inhiben la acetilcolina (neurotransmisor en el Sistema Nervioso Central y Periférico). Hay varias categorías de anticolinérgicos, pero en el caso de la EPOC, se refiere concretamente a los antimuscarínicos, que bloquean los receptores muscarínicos de la acetilcolina reduciendo la secreción de moco en las vías respiratorias.

  Los broncodilatadores pueden ser de corta y larga duración:
  - SABA (*Short-Acting Beta-2 Agonists*).
  - SAMA (*Short-Acting Muscarinic Antagonists*).
  - LABA (*Long-Acting Beta-2 Agonists*).
  - LAMA (*Long-Acting Muscarinic Antagonist*).

En general, los de corta duración hacen efecto durante 4 a 6 horas, mientras que los de larga duración pueden durar de 12 a 24 horas.
Los antimuscarínicos y β2-adrenérgicos se usan habitualmente a la vez, constituyendo el primer escalón terapéutico para pacientes que presenten sintomatología (16).

- Por otro lado, también está indicado el tratamiento con **corticoides inhalados**, cuyo efecto es antinflamatorio. Estos están indicados para pacientes con agudizaciones frecuentes, con historial de hospitalización previa y/o pacientes con eosinofilia. Se utilizan de esta forma ya que han demostrado ser capaces de mejorar la calidad de vida de los pacientes, función pulmonar y reducir las exacerbaciones.

  Los corticoides suelen usarse junto a los broncodilatadores, ya sea con antimuscarínicos (SAMA/LAMA) o β2-adrenérgicos (SABA/LABA), denominándose doble terapia, o si se usan los 3 de forma simultánea (Corticoides + SAMA/LAMA + SABA/LABA) se denominaría triple terapia.

- Por último, también hay que destacar la **Oxigenoterapia domiciliaria (OCD)**, que está indicado para pacientes de EPOC con niveles de oxígeno en la sangre bajos (PaO_2 < 55 mmHg) o que también presenten comorbilidades como poliglobulia, hipertensión pulmonar, insuficiencia cardiaca, *cor pulmonare*, entre otros.
  El oxígeno se debe administrar, al menos, entre 15 y 18 horas al día (o la máxima posible, en su defecto). El paciente no debe de permanecer sin oxígeno durante más de 3 horas, especialmente en aquellos que tengan unos niveles de oxígeno bajos (8,16).

Se pueden considerar, aunque ya no de forma principal, otros medicamentos como antibióticos, vacunas o cirugía para casos más extremos. Además del tratamiento farmacológico o quirúrgico, es importante considerar de igual modo cambios en el estilo de vida y rehabilitación pulmonar.

3.7 EPOC y su relación con la eosinofilia

Como ya se ha expuesto con anterioridad, el proceso inflamatorio es clave en el desarrollo de esta enfermedad. De forma sintetizada, el humo del tabaco, otras sustancias tóxicas o incluso infecciones pueden desencadenar una inflamación en las vías aéreas. Este proceso, sobrevenido de forma recurrente, puede provocar la destrucción de la parénquima pulmonar dando lugar al enfisema pulmonar y la bronquitis crónica, constituyendo un mecanismo, que en casos de normalidad ocurre de forma puntual, crónico.

La inflamación crónica está caracterizada por un aumento de una serie de células, factores quimiotácitos y quimiocinas, entre los que encontramos:

- Macrófagos activados, reclutados gracias a los mediadores quimiotácitos liberados por las células epiteliales y macrófagos del tejido pulmonar, en concreto los bronquios y bronquiolos.
- Neutrófilos, junto con los macrófagos, pueden hallarse en el esputo.
- Linfocitos, $CD8^+$ y $CD4^+$, en menor medida.
- Quimiocinas como CXCL9, CXCL10, CXCL11, que se expresan en quimiocinas unidas a los linfocitos $CD8^+$ y $CD4^+$.
- Células linfoides innatas, como ILC1 e ILC3 (9).

Se cree, con gran probabilidad, que éstas son las células que contribuyen a la persistencia inflamatoria, incluso cuando los pacientes han dejado de fumar.

Se sabe con certeza la existencia de las células, quimiocinas y factores quimiotácitos anteriormente mencionadas, éstas juegan un papel fundamental en el desarrollo de la EPOC convencional, pero el número de eosinófilos suele estar muy reducido, entonces ¿en cómo es posible el estudio de la relación EPOC-eosinofilia?

Hay dos razones principales:

1- Existen pacientes de EPOC con fenotipo eosinofílico (EOS).
2- Existen pacientes Asma-EPOC (fenotipo mixto).

Los eosinófilos son leucocitos granulocitos que se forman en la médula ósea y que contienen multitud de proteínas granulares (de ahí su aspecto granulado al microscopio) con funciones proinflamatorias, siendo una de las patogénesis principales de las alergias. Suelen presentarse en un porcentaje pequeño en la sangre periférica, aunque con gran varianza durante el tiempo.
Se considera EPOC de fenotipo agudizador eosinofílico cuando el recuento de éstas células en sangre sea igual o mayor de 300 eosinófilos/µL y con un historial

de agudizaciones de repetición (2 o más exacerbaciones ambulatorias, una o más exacerbación grave que requieran de ingreso hospitalario) (17).
Otras fuentes indican una concentración eosinofílica más baja, con 279 eosinófilos/µL (18).
Existe, además, un fenotipo no agudizador eosinofílico, que tiene lugar cuando el nivel de éstas células está por debajo de 300 eosinófilos/µL, pero no las consideraremos objeto de estudio en el presente TFG.
La razón de que algunos pacientes tengan un mayor nivel de estas células en sangre parece tener una base genética. En un trabajo desarrollado por Soler Cataluña et al., se observaron, para 2 grupos de pacientes, una "firma genética" de Th2, la cual se caracteriza por una mayor concentración de eosinófilos tanto en las vías aéreas como en la sangre, junto con una pérdida de función pulmonar acelerada y mayor reversibilidad en la prueba broncodilatadora (19).

El otro "fenotipo" interesante a destacar sería el *asthma-COPDoverlap* (ACO), que se aplica para pacientes de EPOC con rasgos asmáticos, como podría ser una prueba broncodilatadora muy positiva y eosinofilia sanguínea. O simplemente aplicado para pacientes con EPOC y que hayan tenido historial de asma o alergias. Sin embargo, actualmente, no se considera un fenotipo propiamente dicho, ya que la prueba broncodilatadora positiva casi nunca se encuentra de forma aislada y tiene escaso valor diagnóstico, sumado a que los pacientes asmáticos presentan diferencias en la respuesta al tratamiento con corticoides (9,17).

Volviendo al fenotipo eosinofílico, se han observado mejorías con el tratamiento con corticoides (junto con broncodilatadores), reduciendo exacerbaciones, aumentando el volumen espirado en el primer segundo (FEV1) y calidad de vida (9). Otros estudios han confirmado la disminución de eosinófilos en sangre para ciertos pacientes a los que se les ha prescrito triple terapia (9).
La importancia de los eosinófilos radica en la posibilidad de provocar mayor número de exacerbaciones en aquellos individuos que presenten el fenotipo agudizador eosinofílico.
No obstante, la interacción eosinófilos-EPOC no está totalmente dilucidado por lo que no se puede afirmar con absoluta certeza las relaciones anteriormente mencionadas.
Será objetivo del presente TFG el intentar esclarecer, aunque sea de forma somera, la relación subyacente entre este leucocito y la EPOC.

4. ESTADO DEL ARTE EN EL CONTEXTO EPOC-HRV.

La señal biomédica que se analizará será principalmente la señal de variabilidad del ritmo cardiaco (*Heart Rate Variability – HRV*), que es una señal derivada del ECG ya que el objetivo del TFG consiste en analizar si existen diferencias en cuanto a esa señal para pacientes de EPOC-eosinofílicos y no eosinofílicos.

La HRV se trata de una señal derivada del electrocardiograma que mide la variación del intervalo pico-a-pico, típicamente entre picos R-R de latidos consecutivos.
Dado que la señal captura fluctuaciones entre latidos cardiacos consecutivos, es un buen indicador de la actividad autónoma del sistema nervioso, en concreto ambos sistemas simpático y parasimpático.

Estudios previos han analizado la relación entre la señal de variabilidad cardiaca y la EPOC, revelando que, la naturaleza crónica e inflamatoria de la enfermedad, puede influenciar al sistema nervioso autónomo (SNA). Esto puede dar lugar a alteraciones patológicas en el equilibrio simpático-vagal, que se puede traducir en una sobreexcitación del sistema nervioso simpático, lo que contribuye a una mayor frecuencia cardiaca y una respuesta más extremada frente al estrés, y/o una reducción de la actividad vagal. Con esto, la reducción de la frecuencia cardiaca y relajación se ven disminuidos, lo que contribuye al agravamiento de los síntomas de la enfermedad y aumenta el riesgo de eventos cardiovasculares (20).
Estas alteraciones pueden ser identificadas de forma no invasiva utilizando la señal HRV, pues las fluctuaciones de los intervalos R-R están relacionadas directamente con el nódulo sinoauricular (NSA) y es considerado como un biomarcador de salud, de forma que mayor variabilidad en la señal HRV indica mejor adaptación por parte del sistema cardiovascular a aspectos como el ejercicio físico o el estrés, mientras que una señal de menor variabilidad puede indicar mayor predisposición a patologías cardiovasculares y comorbilidades (20).

En los estudios que analizan la variabilidad cardiaca con EPOC, se utilizan índices temporales, frecuenciales y no lineales de esta señal biomédica para comparar entre pacientes que padezcan de la enfermedad y aquellos que no.

Entre los parámetros temporales, se han visto diferencias principalmente en la **SDNN** (desviación estándar de los intervalos NN) y **RMSSD** (raíz cuadrada de la media de las diferencias sucesivas de intervalos NN) donde en los pacientes con EPOC se ven más disminuidos.

También se ha podido observar una disminución similar en otros parámetros temporales, como el **TINN** (índice de SDNN), **NN50** y **pNN50** (número y porcentaje de intervalos NN que difieren en un 50%) (21).

En los parámetros frecuenciales, toma especial importancia en la potencia total (TP), y las diferentes bandas de frecuencia:

- **Alta frecuencia (HF)**, entre 0.15-0.40 Hz.
- **Baja frecuencia (LF)**, entre 0.04-0.15 Hz.
- **Muy baja frecuencia (VLF)**, entre 0.0033-0.04 Hz.
- **Ultra-Baja frecuencia (ULF)**, menor de 0.003 Hz.

Igualmente, también resulta interesante en el contexto de la HRV estudiar la relación entre las bandas **LF** y **HF**, conocido como balance simpático-vagal (22,23). Todas estas bandas reflejan diversas funciones del SNA y la activación de ambas ramas simpática y parasimpática. De forma resumida (20):

- **La potencia total (PT)** refleja la influencia combinada de ambos sistemas sobre el ritmo cardiaco.
- **La banda HF** refleja la banda respiratoria y el control parasimpático sobre el ritmo cardiaco cuando ésta se encuentra bajo una arritmia sinusal respiratoria.
- **La banda LF** refleja la modulación simpática y el reflejo barorreceptor bajo respiración normal.
- **La banda VLF y ULF** son comprendidos en menor medida, pero se cree que están relacionados con mecanismos como la termorregulación, regulación hormonal y ritmos circadianos.
- **La relación LF/HF** es un indicador del equilibrio entre la actividad simpática y parasimpática. Un aumento de la relación puede indicar un predominio simpático, mientras que una disminución indicaría lo contrario.

En el contexto de la EPOC, los parámetros frecuenciales muestran resultados dispares ya que, mientras en ciertos estudios se han reportado una mayor PT y HF en pacientes con fenotipos de exacerbación aguda, otros han mostrado una menor PT.
En la banda LF, parece haber mayor consenso en que los pacientes de EPOC presentan valores más bajos.
Por último, en las ratios de potencia (LF/HF), se han reportado también resultados dispares entre estudios (24).
Por todas estas discrepancias en la mayoría de los estudios relacionados con EPOC, se considera que no hay una evidencia consistente en la utilización de índices frecuenciales en comparación con las del dominio temporal.

5. SUJETOS Y SEÑALES.

En esta sección se procederá a describir la población de estudio y variables adquiridas para esta investigación. Adicionalmente, también se hará una descripción de los distintos tipos de señales que se hayan analizado.

Como primer apartado, se expondrán los criterios fijados a la hora de incluir o excluir individuos de la población a estudiar, presentando además otros datos de carácter clínico y registros biomédicos.

5.1 Diseño del estudio. Criterios de inclusión y exclusión.

En primer lugar, la investigación llevada a cabo es de naturaleza prospectivo, longitudinal y observacional. En el estudio se analizan pacientes con diagnóstico de EPOC confirmado, de acuerdo con su historia clínica, durante un periodo estipulado. No se interviene ni manipulan las variables del estudio, limitándose a observar correlaciones que puedan existir según los datos y variables recogidas.

La población de estudio está formada por pacientes con EPOC que hayan acudido en el Servicio de Neumología del Hospital Universitario Río Hortega de Valladolid, entre febrero y mayo de 2024, de acuerdo a los siguientes criterios de inclusión:

- Diagnóstico confirmado de EPOC, principalmente a través de una prueba espirométrica post-broncodilatadora, midiendo la FEV_1 y la relación FEV_1/FVC.
- Ambos sexos
- Edad comprendida entre los 18 y 80 años.

Además, se crean dos grupos de pacientes balanceados basados en un único criterio, el contaje de eosinófilos en sangre. Por tanto, se definen dos grupos de pacientes bajo estudio, de la siguiente forma:

- Grupo 0 (clase negativa: No eosinofilia): Pacientes con diagnóstico clínico de EPOC, con un contaje de eosinófilos inferior a 300 cels/µL.
- Grupo 1 (clase positiva: Eosinofilia): Pacientes con diagnóstico clínico de EPOC, con un contaje de eosinófilos igual o superior a 300 cels/µL.

Como criterios de exclusión aplicables en los dos grupos de estudio, se ha fijado únicamente que el origen de la EPOC no se deba a ningún factor genético u ocupacional.

Por último, se solicita la firma del consentimiento informado a cada paciente, comunicando los detalles de la investigación a través de una hoja de información. Este protocolo de estudio ha sido aprobado por el Comité de Ética de la Investigación con Medicamentos (CEIm) del Hospital Universitario Río Hortega de Valladolid (Referencia 23-PI198).

5.2 Procedimiento para la recopilación de datos.

Las condiciones de toma de registro fueron en un entorno controlado como es la consulta de Neumología. El paciente se encontraba sentado, en reposo a la toma el ECG, procediendo después, a la toma de la tensión y realización de los test CAT y mmRC. La duración media del procedimiento fue de 12 a 15 minutos, con 5 a 9 minutos para el ECG y el resto para las demás pruebas.

La información adquirida se detalla en los próximos apartados:

5.2.1 Variables clínicas

La primera información tomada del paciente fueron las variables clínicas, la mayoría extraídas de la historia clínica electrónica, aunque también se interactúo con el paciente para que éste completase los cuestionarios propuestos en el protocolo del estudio.

Variables sociodemográficas: aquí se incluyen variables como el peso, altura, edad, sexo, presiones sistólica y diastólica, frecuencia cardiaca y respiratoria.

Comorbilidades: se codificó la presencia o ausencia de otras enfermedades como hipertensión arterial (HTA), diabetes mellitus (DM), enfermedades cardiovasculares (CV) o apnea del sueño (AOS).

Variables relacionadas con la EPOC: si el paciente es fumador o no (1/0) o exfumador (2), años fumando, número de agudizaciones en el último año, agudizaciones que hayan requerido de ingreso hospitalario, variables espirométricas (FEV_1 y FEV_1/FVC).

Test de evaluación del paciente: el test mMRC (*modified Medical Research Council*) y el test CAT (*COPD Assessment Test*).

- La escala mMRC: es una escala utilizada para evaluar la disnea en pacientes con patologías respiratorias crónicas, como es la EPOC. Clasifica el grado de disnea en función del impedimento o la dificultad que supone realizar las actividades diarias del paciente.
  La escala consta de 5 grados, que van del 0 al 4 y de forma ascendente (25)
  Grado 0, no presenta disnea excepto durante el ejercicio intenso.
  Grado 1, siente disnea al caminar deprisa en llano o pendiente.

Grado 2, necesita caminar más despacio que las personas de su misma edad por la disnea o detenerse puntualmente para respirar en terreno llano.
Grado 3, se detiene a respirar tras caminar unos 100 m o algunos minutos en llano.
Grado 4, tiene dificultad para salir de casa o experimenta disnea al realizar tareas como vestirse o desvestirse.

- El test CAT: es un cuestionario diseñado específicamente para evaluar el impacto de la EPOC en la vida diaria de los pacientes, evaluando los síntomas y el impacto de la enfermedad en la calidad de vida. Ha demostrado ser útil para evaluar la prognosis de la enfermedad y la respuesta al tratamiento.
  Valora 8 ítems relativos a la enfermedad, las cuales son la tos, flema, presión en el pecho, dificultad para subir cuestas o escaleras (disnea), limitaciones en la actividad doméstica, confianza para salir de casa, calidad de sueño, energía o sensación de fatiga.

  Cada pregunta se puntúa de 0 a 5, de forma ascendente, por lo que la suma total deriva en una puntuación de 0 a 40: Impacto bajo (1-10 puntos), impacto medio (11-20 puntos), impacto alto (21-30 puntos), impacto muy alto (31-40 puntos) (25).

5.2.2 Señal de variabilidad del ritmo cardiaco

La señal HRV es una señal derivada del ECG. Como es ampliamente conocido, el electrocardiograma es una señal biomédica que detalla la actividad eléctrica del corazón, donde cada uno de sus formas de onda (onda P, complejo QRS y onda T) indican la despolarización (y repolarización) de las cámaras cardiacas, hasta cumplir un ciclo cardiaco completo.

Este voltaje es medible en la superficie corporal y el ECG es una herramienta diagnóstica de gran utilidad para evaluar el estado de salud del corazón y detectar problemas de conducción del mismo. Algunos entre los que se incluyen: arritmias, bloqueos coronarios, alteraciones electrolíticas, entre otros. También se puede utilizar para determinar la eficacia de los tratamientos cardiacos. Es una de las herramientas básicas para la monitorización del paciente en estado crítico (26).

La HRV mide la variación de tiempo entre intervalos sucesivos del corazón, analizando la fluctuando entre intervalos, típicamente R-R, consecutivos. Una de sus funcionalidades principales es la de la evaluación de la actividad del

sistema nervioso autónomo, que está estrechamente relacionada con la frecuencia y variabilidad cardiaca.

Es una señal que mide dos tiempos, el primero es el tiempo entre latidos consecutivos y el segundo es el momento en el que sucede esa medición, un ejemplo gráfico de cómo se obtiene se muestra en la Figura 6.

Dado el alto grado de información que provee y la facilidad con la que se obtiene, es muy útil analizarlo en una gran cantidad de patologías, como es el caso de la EPOC.

Diferentes indican una disfunción del sistema nervioso autónomo por varios factores relacionados con la EPOC, como la inflamación crónica, hipoxia o el estrés respiratorio, lo que conlleva a una alteración de equilibrio en los sistemas simpático y parasimpático y regulación autónoma cardiaca. Además, también se ha considerado una herramienta útil para el seguimiento de la enfermedad, riesgo cardiovascular y evaluación del riesgo cardiovascular (21,28).

El dispositivo utilizado para realizar los registros de ECG fue el Biosignalsplux (Plux Wireless Biosignals S.A., Lisboa, Portugal) con sensor y canal específico de 3 derivaciones para realizar el ECG. El programa permite varias frecuencias de muestreo. Se ha optado por usar una frecuencia de muestreo de 300 Hz.

Los electrodos, proporcionados por el servicio de neumología del Hospital Universitario Río Hortega, se colocaron en las dos clavículas y en la cadera izquierda del paciente.

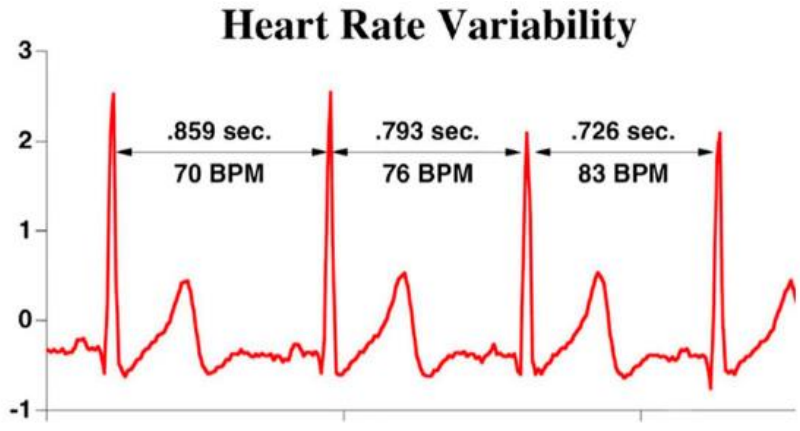

Figura 6. Proceso de obtención de la señal de variabilidad cardiaca (27).

Por último, una de las funcionalidades que provee el fabricante, es un análisis del ECG para obtener la señal HRV, junto con otras funcionalidades (*heart rate variability*) del software Opensignals (r)evolution, el dispositivo utilizado se muestra en la Figura 7:

Figura 7. Dispositivo Biosignalsplux (Plus Wireless Biosignals S.A., Lisboa Portugal)

6. METODOLOGÍA

En este capítulo se expondrán los métodos, técnicas y algoritmos empleados en el desarrollo del estudio.
En primer lugar, se describirán los procedimientos que se han utilizado para la extracción de variables de la señal bajo estudio, realizando una caracterización en el dominio del tiempo, frecuencial y no lineal. Tras la extracción de dichas características, se detalla el proceso empleado para la selección de características óptimas, basado en algoritmos genéticos (GAs). Finalmente, se presentará el algoritmo de clasificación binaria para identificar los dos grupos de pacientes bajo estudio (eosinofílico y no eosinofílico) para, describir los test estadísticos empleados en el análisis descriptivo de las variables y las métricas utilizadas para la evaluación del rendimiento diagnóstico del clasificador.

6.1 Preprocesado de la señal HRV.

La señal HRV que se obtiene del programa anteriormente mencionado no ha pasado por una etapa de preprocesado y a pesar de tener un algoritmo de detección de intervalos RR, sigue proporcionando datos crudos con potenciales pérdidas de latidos (falsos negativos) o latidos extra (falsos positivos), por lo que es necesaria una etapa en la cual se eliminen los artefactos y los valores que no se consideren dentro de la normalidad.

La duración estipulada inicialmente para las señales había sido de cinco minutos, pero ésta se flexibilizó posteriormente para realizar una etapa de *data augmentation* e incrementar el tamaño de la base datos, especialmente el de la clase minoritaria (pacientes con eosinofilia).

El primer paso fue la inspección individual de cada registro para eliminar e interpolar las muestras que fueran fisiológicamente incompatibles, dando lugar a tres condiciones de intervención:

- Señales cuyo valor del intervalo RR sea inferior a 0.33 segundos (180 bpm) (29). Estos valores son imposibles para una persona que se encuentra en reposo y/o que no se encuentre en taquicardia. Por lo que consideraremos que el algoritmo ha detectado un falso latido. La muestra será eliminada y el valor asignado a dicha muestra se asignará a la muestra consecutiva siempre y cuando ésta sea válida.

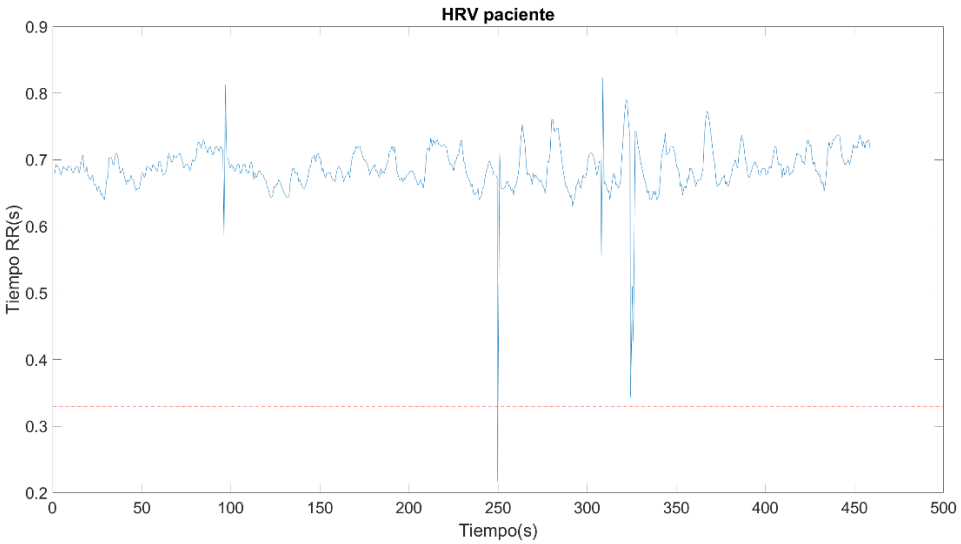

Figura 8. HRV con un elemento menor de 0.33s

- Señales cuyos valores sean superiores a 1.5 segundos, lo que indicaría una frecuencia cardiaca inferior a 40 bpm, algo que es muy inusual para una persona en reposo que no experimenta bradicardia, por lo que consideraremos que el algoritmo de detección ha perdido un latido válido (29).

  Se procederá a la eliminación de dicho valor para sustituirlo por otro que se asigne mediante interpolación lineal. Si existen valores superiores a 1.5 consecutivos, entonces se procederá a la asignación de los mismos mediante interpolación cúbica. Consideraremos que son consecutivos cuando existan 3 o más valores seguidos.

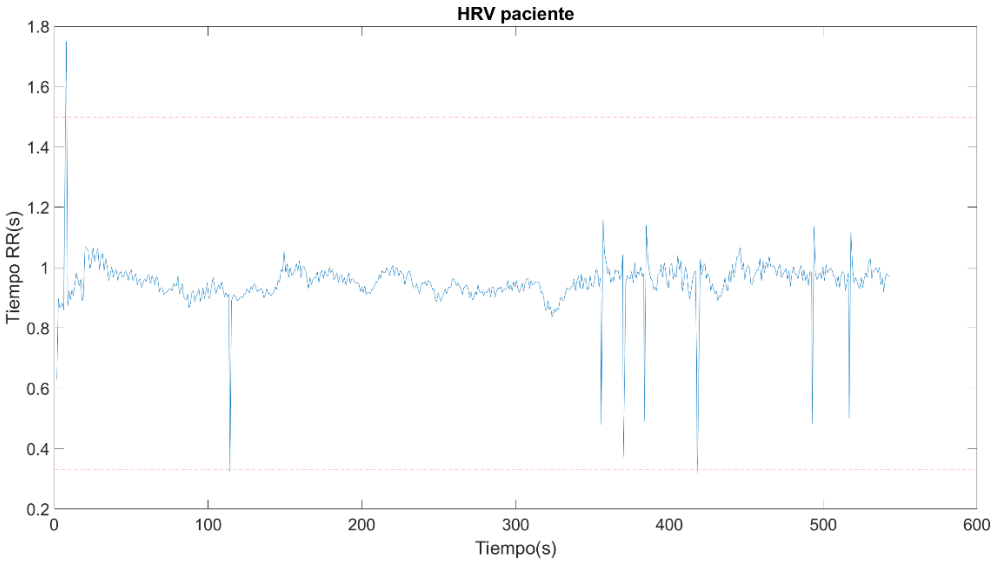

Figura 9. HRV con un elemento mayor a 1.5s.

- Señales con valores consecutivos sean mayores de 0.66 segundos, ya que podría indicar una transición brusca o irregularidad no fisiológica, como podrían ser errores a la toma del ECG o a la hora de realizar la medición (29).
  En ese caso también se realizará una interpolación lineal, si no existen valores consecutivos, o cúbica, en caso de sí haberlos.

Tras la implementación correcta de estas condiciones, se esperaría obtener señales sin muestras fisiológicamente inválidas. No obstante, esto no ocurrió ya que, a la hora de revisar los registros, todavía existían valores cercanos a 0.33s y a 1.5s, las condiciones impuestas por la literatura. Esto indica que todavía subyacen latidos falsos o latidos que no hubieran sido detectados por el algoritmo.

Se optó, entonces, por implementar una segunda fase de preprocesado adicional que lo perfeccionase, utilizando una media móvil a través de una ventana deslizante.

Una media móvil es un algoritmo utilizado para procesar datos secuenciales o series temporales. En este contexto, se analiza sólo un subconjunto de muestras del registro HRV a la vez, creando una media para ese subconjunto. Este subconjunto no es estático, ya que se desplaza de una muestra en una muestra hasta recorrer todo el registro, lo que significa que se puede examinar una porción de muestras cada vez (con cada subconjunto) y decidir en base a esa información.

Primero, se define un tamaño de ventana, en el caso del HRV, se ha visto que una ventana de 30 segundos es óptima para el análisis (30–32).

Dada la cantidad de artefactos que existían en ciertos registros, se decidió definir un intervalo de tiempo, entre diez y sesenta segundos para que se detecten la mayor parte de los espurios u *outliers*. Una vez definida la ventana, que se podría considerar como una serie temporal o un vector, se hace un promedio y se calcula la desviación estándar para todos los elementos dentro de ese intervalo.

En la Figura 10 se muestra un ejemplo gráfico de cómo se realiza la ventana deslizante para una serie temporal cualesquiera.

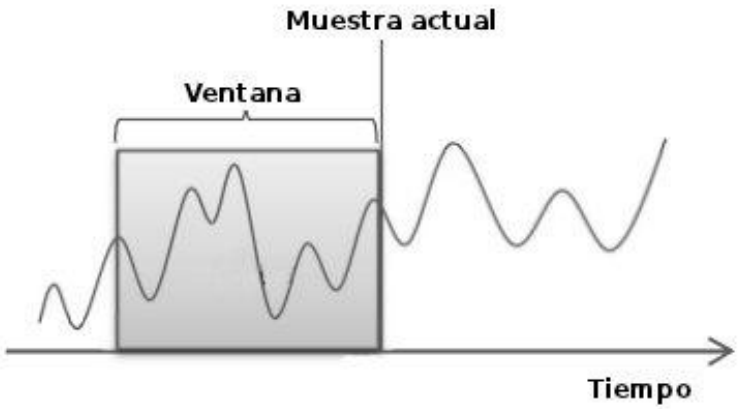

Figura 10. Ejemplo gráfico de una ventana deslizante (33).

Este intervalo es deslizante, lo que significa que con cada iteración del bucle o del proceso, se abandona el primer elemento del intervalo y se coge el siguiente elemento, creándose así una nueva ventana hasta que se consiga recorrer todos los elementos del registro:

$$y_k = \frac{1}{w}\sum_{i=0}^{w-1} x_{k+i} \tag{6.1}$$

Se procede, por último, a comparar elemento por elemento de la señal, siempre que esté dentro de la ventana definida. Si el valor de ese elemento es inferior o superior a la media ± 1.96 veces la desviación estándar, se considera ese elemento como artefacto y se procederá a la eliminación e interpolación de dicho valor.

Se considera esta última condición, ya que contemplamos que los datos tienen una distribución gaussiana y establecemos un intervalo de confianza del 95%, lo que significa que, con una probabilidad del 95%, se espera que se encuentren los verdaderos valores de los datos, o que la media poblacional de la muestra sí es representativa.

$$IC = \bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \tag{6.2}$$

- Donde $\bar{X}$ es la media muestral.
- σ es la desviación estándar de los datos.
- n es el tamaño de la muestra.

Una vez obtenidos los valores *outlier*, se procede a su eliminación y la realización de una interpolación lineal o cúbica, según las situaciones comentadas con anterioridad.
En la figura 11 se muestra un registro HRV que ha sido preprocesada:

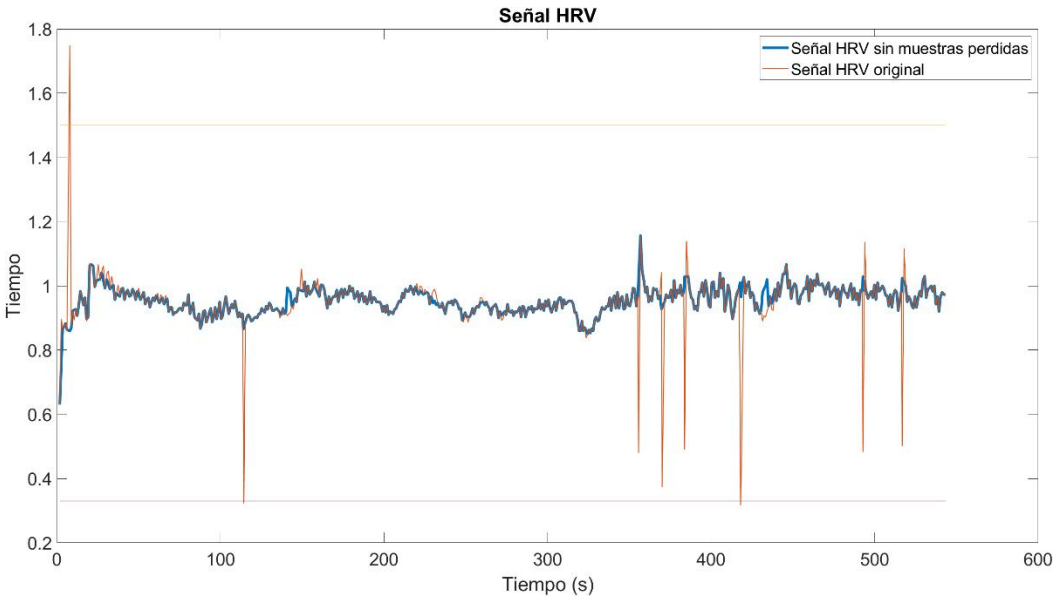

Figura 11. HRV completa preprocesada, este proceso ocurre varias veces.

6.2 Análisis temporal de la señal HRV

También es interesante estudiar el registro HRV en el dominio temporal. Los parámetros que se extraen son de la estadística clásica, como son la media o la desviación típica. En total se han elegido 6 parámetros y se extraen directamente sobre la HRV preprocesada:

Average NN Interval (AVNN): Se define como la media del registro HRV completo.

$$\overline{NN} = \frac{1}{K}\sum_{i=1}^{K} NN_i \tag{6.3}$$

Standard Deviation of NN intervals (SDNN): Se define como la desviación típica de la señal HRV en el registro completo.

$$SDNN = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(NN_n - \overline{NN})^2} \tag{6.4}$$

Root Mean Square of the Differences of Successive NN Intervals (RMSSD): Se define como el valor cuadrático medio de las diferencias entre valores consecutivos de la señal completa HRV.

$$RMSSD = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K-1}(NN_{i+1} - NN_i)^2} \tag{6.5}$$

NN50: Número de intervalos NN adyacentes que difieran por más de 50ms. Ésta requiere de un registro de duración mínima de 2 min (21).

Percentage of NN50 (pNN50): Se define como el porcentaje de intervalos NN adyacentes que difieran por más de 50ms con respecto al número total de muestras del registros. También requiere de un registro mínimo de 2 minutos. Está estrechamente relacionada con la actividad del sistema nervioso parasimpático (SNP) y posee correlación con la RMSSD y la banda HF (21).

$$pNN50 = \frac{NN50}{N-1} \cdot 100\% \tag{6.6}$$

HRV Triangular Index (TINN): Se define como una medida geométrica que se obtiene mediante el cálculo integral de la densidad del histograma del intervalo RR. Se necesita un registro de, típicamente, 5 minutos para poder calcular esta métrica (21).

$$HTI = \frac{N}{h_{\max}} \tag{6.7}$$

donde *N* es el número total de intervalos RR en el histograma y h_{max} es la altura máxima del histograma.

6.3 Estimación de la densidad espectral de potencia

Como muchas de las señales que se manejan en el ámbito biomédico, es posible realizar un análisis en el dominio espectral. La densidad espectral de potencia, o PSD, es una técnica que se usa para analizar la distribución de potencia de una señal en este último dominio.

Es posible emplear distintas técnicas para obtener la PSD, de las más comunes son la Transformada Rápida de Fourier (FFT) y Modelado Autorregresivo (*Autorregresive Modeling,* AR) (34)mbién son frecuentes las técnicas basadas en la transformada discreta de Fourier (*Discrete Fourier Transform*, DFT) como la transformada Wavelet y otras variantes, siendo especialmente destacada el Periodograma de Welch basado en la FFT. Este algoritmo separa los componentes armónicos de la HRV para describirlos, a posteriori, en el dominio espectral, lo que hace posible una asignación de potencia a las diferentes bandas, atribuidas a las diversas funciones fisiológicas (35).

En el presente TFG se trabajó con la transformada de Welch a la hora de obtener la estimación de la densidad espectral de potencia, ya que cuenta con múltiples ventajas, como una reducción del sesgo y varianza, mayor resolución frecuencial, buena supresión de ruido y menor carga computacional (21,28).

Este método divide la señal en ventanas para calcular el periodograma modificado de una de ellas para luego promediarlos y obtener la PSD estimada.

Existe, sin embargo, un último problema con la extracción de la PSD para la señal HRV y es que, esta señal no está equiespaciada pues no es posible un muestreo uniforme. Hay que recordar que la señal HRV involucra dos vectores de tiempos, una indicando cada intervalo RR y otra indicando el tiempo de registro, que se corresponde con el instante en el que se produce cada latido y, por tanto, no es un muestreo uniforme.

Conseguiremos un muestreo homogéneo si realizamos una interpolación mediante *splines* cúbicos con una frecuencia de 3.41 Hz (29). Se usará únicamente para extraer las características espectrales de la HRV, por las condiciones de uniformidad que asume el análisis espectral. Al obtener el registro interpolado, es necesario definir los parámetros del método de Welch para estimar la densidad espectral de potencia. Para ello, se empleó ventanas de 30 muestras y, como la frecuencia de muestro es 3.41 Hz, equivale a 10.23 segundos de la señal por cada ventana. También se utilizó un solapamiento del 50% y 128 puntos para la DFT.

Como se expuso con anterioridad, se extrajeron características sobre ciertas bandas de frecuencia de interés (22).

- Banda de muy bajas frecuencias: (**VLF**), comprendida entre 0.003 y 0.04Hz
- Banda de bajas frecuencias: (**LF**), comprendida entre 0.04 y 0.15Hz
- Banda de altas frecuencias: (**HF**), comprendida entre 0.15 y 0.4Hz.

Además de dichas potencias, también se calculó la relación de la potencia absoluta entre la banda LF y HF, lo que se conoce como balance simpático-vagal. En la Figura 12 se muestra la estimación de la densidad espectral de potencia a través del método de Welch sobre uno de los registros que hemos obtenido:

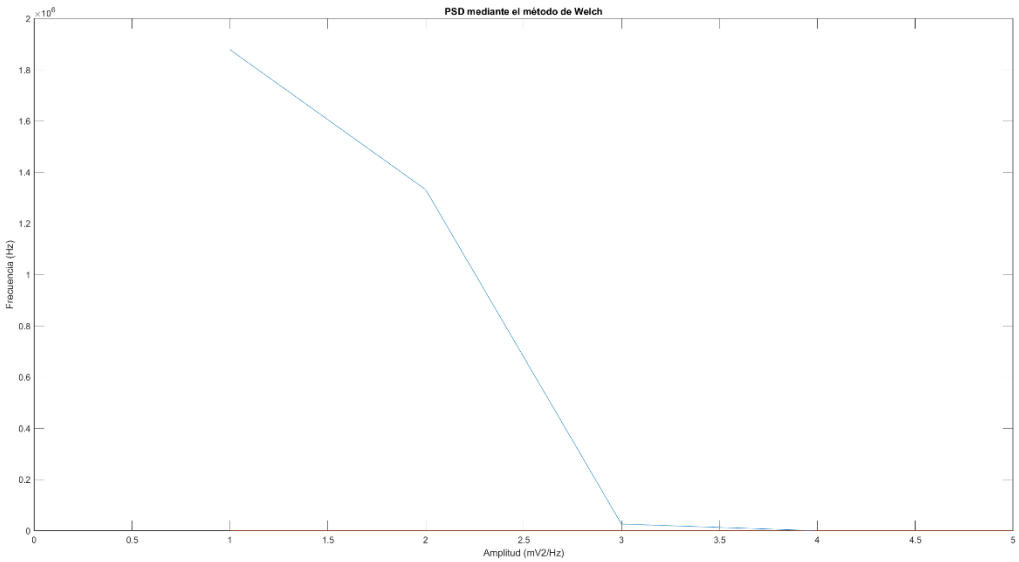

Figura 12. PSD de un HRV a partir del método de Welch

Método de Welch

A continuación, se detalla el algoritmo para la obtención de la PSD con el método de Welch (36):

1. La serie temporal, *x*, se divide en M segmentos o *batches* que pueden estar solapados (D) y de longitud *L*, lo que da a lugar a K segmentos que recorren toda la secuencia inicial. El grado de solapamiento se ha fijado con anterioridad.

$$\begin{aligned} X_1(j) &= X(j) && j = 0, \dots, L-1 \\ X_2(j) &= X(j+D) && j = 0, \dots, L-1 \\ X_k(j) &= X(j + (K(-1) \cdot D && j = 0, \dots, L-1 \end{aligned} \tag{6.8}$$

2. Para cada segmento, desde *j* hasta *L-1*, se computa una ventana de la misma longitud que los segmentos señal $\big(W(j)\big)$, éstos se usan para

mitigar efectos de discontinuidades en los extremos. Se forman entonces, las secuencias $X_1(j)W(j), \dots, X_k(j)W(j)$ sobre las que se calcula la transformada discreta de Fourier.
En la Figura 13 se muestra un ejemplo gráfico del método.

$$A_k\,(n) = \frac{1}{2}\sum_{j=0}^{L-1} e^{-\frac{2K_i jn}{L}} \tag{6.9}$$

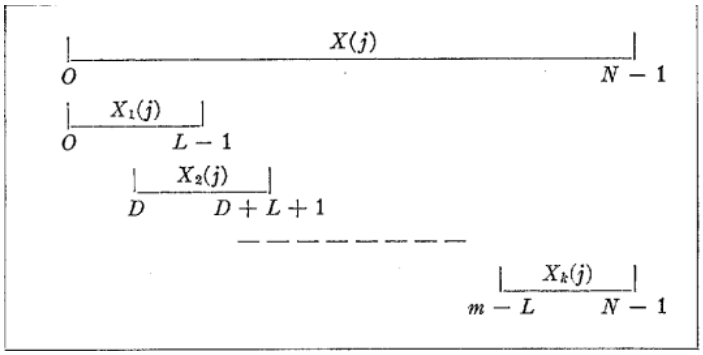

Figura 13. Segmentación de la secuencia temporal original (37).

3. Cada uno de los periodogramas se han obtenido de la siguiente forma:

$$P_k\,(f_n) = \frac{1}{W}|A_k(n)|^2 \tag{6.10}$$

$donde$

$$W = \frac{1}{L}\sum_{k=0}^{L-1} w^2(j)\ , \qquad f_n = \frac{n}{L} \tag{6.11}$$

4. Por último, la PSD se obtiene con el promedio de todos los periodogramas:

$$S_x(f_n) = \frac{1}{K}\sum_{K=1}^{K} P_k(f_n) \tag{6.12}$$

En la Figura 14 se muestran las ventanas posibles para diferentes usos, todas ellas dependen de la forma de los lóbulos centrales y laterales:

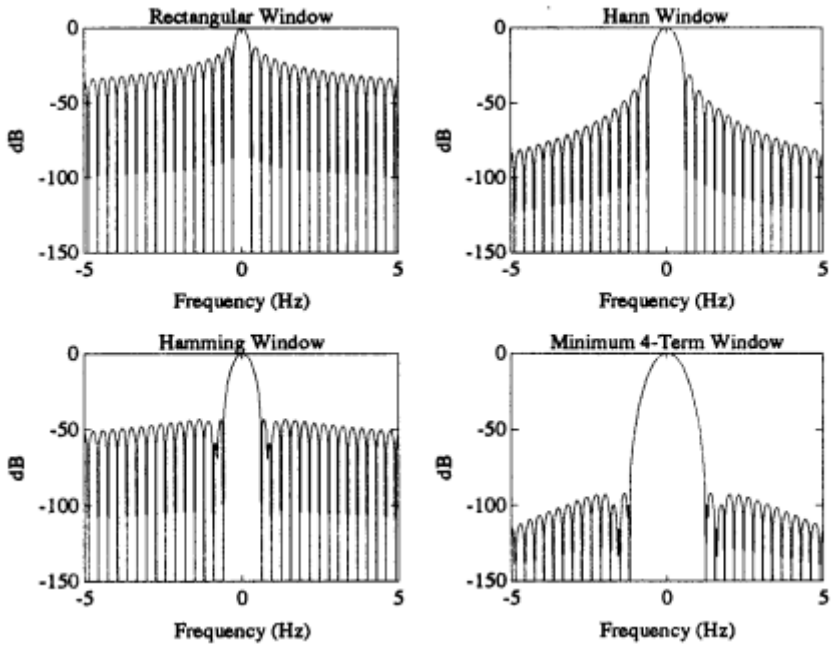

Figura 14. Posibles ventanas para definir en el método de Welch (36).

El consenso considera que el lóbulo principal es la resolución de la transformada discreta de Fourier, mientras que los lóbulos laterales definen cuánto una onda sinusoidal con una frecuencia fuera del lóbulo principal pueda contaminar la medición de frecuencias dentro del lóbulo principal. Por el principio de incertidumbre en la transformada de Fourier, no es posible tener lóbulos laterales más pequeños (y por tanto, menos afectación de ruido) conservando el ancho de ventana principal. Es decir, no es posible obtener simultáneamente el lóbulo principal estrecho y los lóbulos laterales pequeños (36).

A continuación, se detallarán las características extraídas de la PSD para parametrizar el contenido espectral de la señal HRV**:**

<u>Potencia Total</u>

La potencia total (PT) de una señal en el dominio espectral, se refiere a la suma de la potencia contenida en todas las frecuencias de la señal. Es un parámetro descriptivo muy común en este ámbito.

Para obtenerlo, se debe calcular el área total bajo la curva de la PSD, es decir, una integral o suma de coeficientes hasta fs/2.

$$PT = \int_0^{\frac{fs}{2}} PSD(f)\, df \qquad (6.13)$$

$$En\ caso\ de\ no\ ser\ infinitesimal:$$

$$PT = \sum_{f=0}^{fs/2} PSD(f_j) \qquad j = 1,2,\dots,fs/2 \tag{6.14}$$

Frecuencia mediana

La frecuencia mediana (MF), es otro de los parámetros muy comunes para describir una señal en el ámbito frecuencial.

Se define como la componente espectral donde se alcanza la mitad de toda la potencia (PT) de la señal (38), es decir, es la frecuencia que divide el espectro de potencia en 2 partes iguales, donde cada mitad está por encima o por debajo de ella (39).

$$MF = \frac{1}{2}\int_0^{fmax} PSD(f)df \tag{6.15}$$

$$En\ caso\ de\ no\ ser\ infinitesimal$$

$$MF = 0.5 \sum_{fj=0}^{\frac{fs}{2}} PSD(f_j) \tag{6.16}$$

siendo *fs* la frecuencia de muestro y f_j valor de frecuencia en cada punto.

Entropía espectral

La entropía espectral (SE) cuantifica la irregularidad o complejidad en el dominio frecuencial de una señal. Está relacionada con la entropía de Shannon (40).

Dada su definición, un valor alto de entropía espectral indica una distribución frecuencial más uniforme y, por tanto, mayor incertidumbre en el espectro. En cambio, un valor bajo sugiere que la potencia está más concentrada en ciertas frecuencias, lo que da a entender menor irregularidad en la señal.

Hay que recordar que la concentración de la potencia en ciertas frecuencias indica, en el dominio del tiempo, una o varias componentes periódicas, mientras que la presencia de homogeneidad (valores más altos de SE) indican la ausencia de periodicidad dominante (39).

$$p_j = \frac{PSD(f_j)}{\sum_{f_j=0Hz}^{0.5fs} PSD(f_j)} \tag{6.17}$$

siendo P_j el valor de la PSD normalizado.

Por último, la SE se calcula mediante la siguiente expresión:

$$SE = -\sum p_j \ln(p_j) \tag{6.18}$$

6.4 Análisis no lineal de la señal HRV

Existe cierta parte de información que el análisis temporal y espectral no son capaces de capturar, ya que muchas señales biológicas no son estrictamente lineales y/o estacionarias (40). Por eso es necesario la utilización de otras métricas y métodos que nos permitan capturar y describirlas. Entre ellas, destacan los métodos no lineales, dada la naturaleza no lineal de muchos procesos biológicos.

Entropía muestral

La entropía muestral (SampEn) es una métrica no lineal originada por la modificación de una métrica "hermana", la entropía aproximada (ApEn). Se utiliza para evaluar la complejidad de series temporales cuyo origen es fisiológico.

La SampEn se define como el logaritmo negativo de la probabilidad condicional de que dos secuencias sean similares, dentro de una tolerancia *r*. Se establece una ventana, de *m* puntos adyacentes y que esa probabilidad condicional de que sean similares lo sigan siendo cuando se incremente dicha ventana a una serie de *m+1* puntos (41).

SampEn y ApEn son dos de las métricas no lineales más utilizadas en las últimas décadas para cuantificar la regularidad de datos biológico. Si bien es cierto que la estimación proporcionada por ambos es muy similar (la regularidad), SampEn es más consistente, ya que requiere de series más cortas para converger y menos sesgada que la otra ApEn, incluso para series relativamente cortas (42).

Para calcular la entropía muestral, se siguen los siguientes pasos:

1. Se obtienen los segmentos de longitud *m* y *m+1* de un vector o serie temporal de longitud *x*. Además, se elimina la autocomparación, que era un factor de sesgo de la ApEn (43):

$$C_i^m(r) = \frac{número\ de\ j\ tal\ que\ d[x(i), x(j)] \leq r}{N - m + 1}, (i \neq j) \tag{6.19}$$

   Se define:

$$C^m(r) = (N - m + 1)^{-1} \sum_{i=1}^{N-m+1} C_i^m(r) \quad (6.20)$$

2. Por último, SampEn se define como el logaritmo negativo de la relación entre la probabilidad de $C^m(r)\ y\ C^{m+1}(r)$.

$$SampEn\ (m, r, N) = -ln \frac{C^m(r)}{C^{m+1}(r)} \quad (6.21)$$

La ventana de tolerancia *r,* se suele tomar como una proporción (0.10, 0.15, 0.20, o 0.25) de la desviación estándar de la serie original.

Medida de la tendencia central

La medida de la tendencia central (CTM) es un parámetro cuantitativo que mide la variabilidad de una serie temporal, en este caso de los sucesivos intervalos RR. Se usa para clasificar y diferenciar patologías como insuficiencia cardiaca congestiva en una persona sana, epilepsia, afectaciones neurológicas, diagnóstico de enfermedades de la arteria coronaria, entre otras (44).

Está basada en el cálculo de diagramas de dispersión, los cuales representan diferencias de segundo orden, formadas a partir del desplazamiento temporal de la señal original x[n] (44).

La CTM se calcula seleccionando una región circular de un radio *r* alrededor del origen del diagrama de dispersión. A continuación, se contabiliza el número de puntos que se hallen dentro del círculo, dividiendo finalmente entre el número total de puntos, lo que hace que la CTM siempre esté en un rango de [0, 1].

La medida de la tendencia central se define como (44):

$$CTM = \frac{1}{N-2} \sum_{i=1}^{N-2} \delta(d_i) \quad (6.22)$$

$se\ define\ \delta(d_i)\ como$:

$$\delta(d_i) = \begin{cases} 1, si\ \left[\left(x(i+2) - x(i+1)\right)^2 + \left(x(i+1) - x(i)\right)^2\right]^{\frac{1}{2}} < r \\ 0, en\ caso\ contrario \end{cases} \quad (6.23)$$

Complejidad de Lempel-Ziv

La complejidad Lempel-Ziv (LZC) es un método no paramétrico utilizado para evaluar la complejidad de un algoritmo. La medición se asocia al número de las distintas sub-cadenas o tramas y la tasa de reaparición, es decir, refleja el aumento gradual de patrones a lo largo de cada secuencia (45).

En el análisis de las señales biomédicas, las sub-tramas o cadenas suelen ser secuencias binarias o ternarias seguidas o alternadas entre ellas. La asignación de uno de estos dos valores de determina mediante el uso de umbrales. Para las secuencias binarias se utilizará un umbral, mientras que si es ternaria se utilizarán 2 umbrales (46).

En el cálculo de la complejidad Lempel-Ziv, se define primeramente una letra del alfabeto, A, el conjunto de símbolos que componen la secuencia (si es binaria y es la letra A, entonces sería {0, 1}). Se define a continuación una secuencia finita *S*, donde cada subconjunto que se pueda obtener de esta secuencia *S* pertenezca estrictamente a A, el vocabulario de las secuencias de S $(v(S))$ son todos los subconjuntos que se puedan obtener de ella.

Por ejemplo, si existe una secuencia *P* que se defina de la siguiente forma:

$$P = s(1), s(2), \dots, s(n) \tag{6.24}$$

siendo

$$s(i) = \begin{cases} 0, si\ x(i) < Td \\ 1, en\ caso\ contrario \end{cases} \tag{6.25}$$

siendo Td el umbral empleado para la binarización, que suele ser la mediana de la serie temporal.

1. Definamos 2 subsecuencias *P* y *Q,* que podrían ser 2 cadenas de letras, *SQ* se definiría como la concatenación de ambas dos, y *SQπ* se definiría como la concatenación de ambas dos eliminando el último carácter de la secuencia (45).

$$S = s_1, s_2, \dots, s_r \qquad Q = s_{r+1} \tag{6.26}$$

$$v(SQ\pi) = SQ - s_{r+1} \tag{6.27}$$

2. Asumiendo, como punto de partida, que la complejidad *c(n)* es 1. Al añadir una nueva muestra si no añade información al contenido de la secuencia, entonces dejaremos S sin cambios, y añadimos a Q un nuevo carácter s_{r+2}. Si al añadir este nuevo carácter Q sigue perteneciendo al

conjunto $v(SQ\pi)$, entonces no consideramos que la complejidad haya aumentado.
La adición de caracteres continúa hasta que Q no pertenezca al conjunto $v(SQ\pi)$. Cuando esto ocurre, se aumenta la complejidad en 1 y volvemos a concatenar SQ (45).

$$Q = s_{r+2}, se\ verifica\ si\ Q \in v(SQ\pi) \tag{6.28}$$

$$c(n) = \begin{cases} c(n)+1, si\ no\ se\ verifica \\ c(n), en\ caso\ contrario \end{cases} \tag{6.29}$$

3. Se repite el procedimiento hasta que Q vuelva a contener el último carácter de la secuencia. En ese momento, el número de sub-tramas diferentes de S equivaldría a $c(n)$, es decir, la medida de la complejidad. Es importante recordar que esta medida refleja la tasa de aparición de nuevos patrones con el aumento de la longitud de la secuencia.
   Es necesario, por ello, realizar una normalización, ya que $c(n)$ depende de la longitud de la serie temporal (41,45).

$$LZC = \frac{c(n)}{b(n)} \tag{6.30}$$

donde:

$$\lim_{n\to\infty} c(n) = b(n) = \frac{n}{\log_\alpha(n)} \tag{6.31}$$

Al aplicar esta normalización, el valor de LZC estará acotado entre 0 y 1, con 1 representando la mayor complejidad (ya que tiene más subsecuencias) y 0 representando nula complejidad según esta métrica.

6.5 Aumento de datos. *Data Augmentation.*

El proceso de *data augmentation* se define como una técnica utilizada para generar nuevos ejemplos de datos para el entrenamiento de un modelo. Es útil para obtener más información a partir de una base de datos más (o relativamente) pequeña.

Generalmente se utiliza en el contexto de procesado o clasificación de imágenes, como imágenes de tomografía computacional y rayos X, donde se generan imágenes artificiales para ayudar al modelo a entrenar y generalizar mejor (47).

Mientras que, en el campo de la clasificación de imágenes hay muchas técnicas a la hora de generar datos nuevos (como la rotación, cambio de escala, inversiones, entre otras), en el caso de los datos clínicos es algo más difícil, pues suelen ser datos cuya naturaleza que no pueden ser modificadas a libertad sin comprometer la imparcialidad del estudio.

En el caso del presente TFG, realizaremos *data augmentation* combinando datos existentes:

1. Obtener 2 registros de HRV a partir de 1. Aquellos registros de duración mayor que 6 minutos, se divididos en 2 de 4 minutos. De esta forma "dos" observaciones (pacientes) a partir de 1, aceptando un solapamiento de máximo 2 minutos (50%) entre las 2 muestras.
   Las variables de análisis temporal, espectral y no lineal no serán iguales entre ambos sujetos.
2. Las demás variables sociodemográficas, clínicas y relacionadas con la EPOC se duplican.

No obstante, no se realizará esta técnica sobre todo el conjunto de datos, si no sólo aquellos que pertenezcan al conjunto de eosinofilia positiva, pues es el grupo minoritario a la hora de entrenar el modelo.

En la Figura 15, se muestra un diagrama de flujo en el proceso para el duplicado de datos:

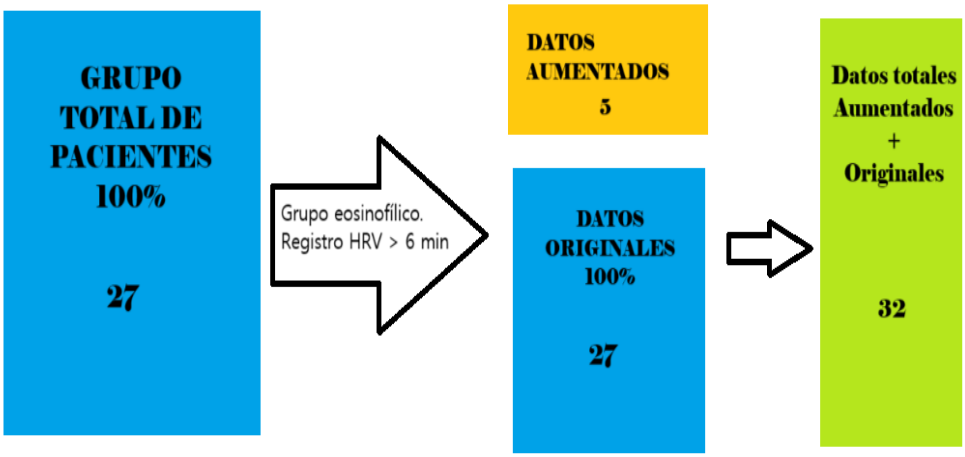

Figura 15. Proceso de duplicado de datos.

6.6 Algoritmos genéticos para la selección de variables.

Este es el último paso antes de pasar al entrenamiento del modelo predictivo, dada la numerosidad de variables definidas en el protocolo del estudio, las de naturaleza clínica y derivadas de señales, la posibilidad de que al introducir todas ellas en el modelo y que éste no pueda converger es grande. Paralelamente, el coste computacional puede ser excepcionalmente alto. Es por ello que se realiza una selección de las variables más destacadas para optimizar el modelo, reduciendo el coste computacional, tiempo de entrenamiento y facilita la comprensión de datos (48).

Los algoritmos genéticos (GA), son una técnica basada en la teoría de evolución propuesta por Charles Darwin. Siguiendo esta teoría, las especies más aptas sobreviven al proceso evolutivo y las más débiles o inadaptadas mueren y no consiguen pasar sus características a la descendencia. Poniéndolo en palabras de selección de variables, la idea es muy similar: las variables más "fuertes" sobreviven a todas (o gran parte) de las iteraciones del proceso y las otras son eliminadas (49).

La reducción de información como entrada al modelo presenta numerosas ventajas para el rendimiento predictivo, lo que hace que la red entrene mucho mejor (50,51):

- Elimina información redundante proveniente de variables que están altamente correlacionadas entre ellas.
- Identificar características que se complementen.
- Reduce el *overfitting*.

Por otra parte, este algoritmo también tiene sus propias limitaciones como:

- En ciertas ocasiones, los algoritmos genéticos tienden a converger hacia óptimos locales o incluso puntos arbitrarios en lugar del óptimo global del problema. Esto ocurre, en muchas ocasiones, debido a la propia función *fitness*, que se podría solucionar si se escoge una función diferente, aumentando la tasa de mutación o haciendo uso de técnicas para mantener una población diversa de soluciones (52).
- También resulta difícil, en ciertos casos, resolver problemas cuya única función *fitness* sea la de un resultado binario (0/1, éxito/fracaso), pues la convergencia es más difícil cuando no hay pasos intermedios que proporcionen una guía (52).

El método parte definiendo parámetros que recuerdan al proceso evolutivo, como podría ser la probabilidad de mutación, cruce, tamaño de población, número de generaciones, entre otras.

- El tamaño de población se define como el número de potenciales soluciones, en este TFG se ha fijado un valor entre 15 y 25.

- La población inicial se define como el conjunto de partida, se trata de una serie de soluciones/individuos seleccionados de forma pseudo-aleatoria. En este TFG, dado el número de pacientes y el número de variables, se ha partido de 2 soluciones iniciales y el algoritmo explorará convenientemente todo el espacio de búsqueda de soluciones potenciales.

- La Probabilidad de cruce (P_c) se define como la probabilidad de que dos individuos se recombinen para generar un nuevo individuo hijo o nueva solución potencial. En este trabajo el valor es de 0.8, típicamente tomado en la literatura (53).

- La Probabilidad de mutación (P_m) se define como la probabilidad de que un individuo se vea alterado de forma aleatoria durante la reproducción de los individuos. Esto permite explorar nuevas soluciones en el espacio de búsqueda, pues introduce variaciones aleatorias. Una tasa de mutación alta favorece una mayor exploración de nuevas combinaciones o individuos en el espacio de búsqueda.

- El criterio de parada define cuándo debe de terminar el algoritmo. En este trabajo se ha fijado a un valor 100 iteraciones, que es común en la literatura (53).

El problema principal del estudio es el tamaño de la población, es demasiado pequeña en comparación con el número de características extraídas y pequeña en general. Es por ello que no podemos obtener subconjuntos individuales de entrenamiento, validación y test que sean representativos, así que optaremos por una partición utilizando estrategias *hold-out* y validación cruzada para llevar a cabo todas las pruebas y obtener unas variables válidas.

En primer lugar, se divide a la población total en dos conjuntos, uno de entrenamiento que representa el 70% de la población total y el 30% restante pasará a formar parte del conjunto test. Haciendo uso de las estrategias anteriormente mencionadas, obtendremos 23 pacientes en el conjunto de entrenamiento y 9 en el conjunto test.

Es importante, en este punto, resaltar dos aspectos esenciales:

1- La partición debe de mantener la proporcionalidad de los grupos, esto es, que se debe mantener la misma proporción de pacientes eosinofílicos en ambas subdivisiones. Siendo 3 de 9 en el conjunto test y 8 de 23 en el conjunto de entrenamiento (33% en ambos grupos).
2- Los individuos del conjunto test, tanto EOS positivos como EOS negativos, deben de provenir de la toma de datos reales, es decir, no se pueden introducir en el conjunto test individuos obtenidos a partir de *data augmentation*. Pues éstos derivan directamente del conjunto de entrenamiento (sus variables sociodemográficas y clínicas son iguales a las del entrenamiento, aunque las derivadas del análisis temporal, espectral y no lineal no lo sean), lo que podría conllevar a sesgos indeseados si se introducen en el conjunto test.

Una vez realizada esta primera partición, se procederá a particionar el conjunto de entrenamiento una segunda vez, ya que necesitamos un conjunto de validación para el algoritmo genético.

El conjunto de validación proviene del conjunto de entrenamiento, realizándose un *Hold-Out* 2 con la misma proporcionalidad (70-30%), dejando 17 individuos en el conjunto de entrenamiento y 6 en el conjunto de validación. La partición también se realiza con la misma proporcionalidad de los grupos eosinofílicos.

El conjunto de test que supone el 30% de los pacientes no se utiliza hasta obtener el modelo final, y se usa exclusivamente para verificar si el modelo es generalizable o no.

El conjunto restante será utilizado para seleccionar las características óptimas mientras que para la optimización de hiperparámetros en la red neuronal se utilizará todo el conjunto de entrenamiento, no siendo necesario un conjunto de validación independiente, aunque se aplicará validación cruzada para minimizar los sesgos.

En la figura 16 se muestra un diagrama de flujo, detallando el proceso de partición del *dataset.*

Adentrándonos un poco más en el método, definiremos a los algoritmos genéticos como un método heurístico aplicable a un amplio espectro de problemas de optimización el cual supone un punto de partida que son un conjunto de soluciones candidatas, las cuales se irán optimizando (54).

El *dataset* dado podría ser un vector, una matriz o una cadena de *strings*, el algoritmo se encargaría de encontrar, para una instancia $\overrightarrow{x} \in \mathbb{R}^k$, un nuevo subespacio $\mathbb{R}^l$ proveniente de $\mathbb{R}^k$ (siendo $l \leq k$) mientras se mantiene un rendimiento comparable (o mejor) en la instancia del espacio de búsqueda original $\mathbb{R}^k$. También es posible que sean símbolos que aparezcan en forma de lista.

El GA empieza con una población inicial aleatoria la cual, a través de operadores genéticos, aplicando cruce (Pc) y mutación (Pm) a los individuos seleccionados para que se "reproduzcan", tiene su semejanza a la selección natural en tanto a que los individuos elegidos que tienen mejor nivel de precisión o rendimiento son aquellos que más probabilidades tienen de sobrevivir a la siguiente generación (propagarse) (55).

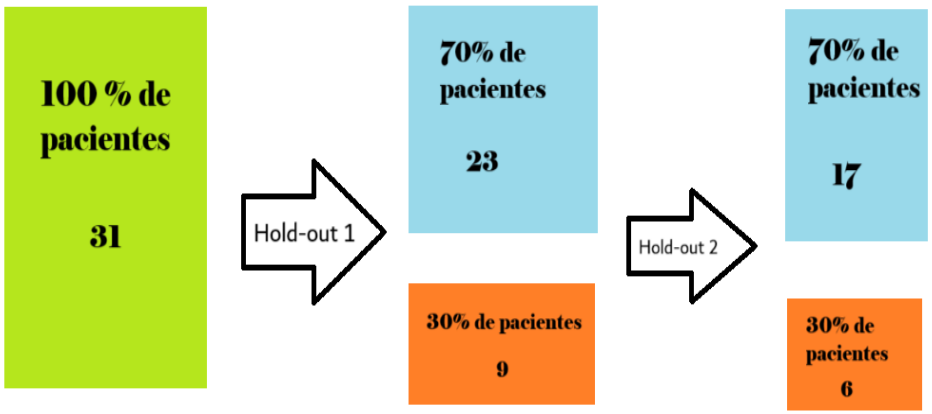

Figura 16. Proceso de partición de los individuos

Como el algoritmo trabaja sobre cadenas binarias de la misma longitud para todos los individuos del conjunto de soluciones potenciales, el tamaño de la población es fijo con el paso de las generaciones. Ambos valores, tamaño de población y el número de generaciones, dependen de la complejidad del problema bajo estudio. Se definen las siguientes etapas (54):

1. Inicializar la población de soluciones potenciales.
2. Evaluar el ajuste (*fitness*) para cada uno de los individuos que formen el conjunto de la población.
3. Aplicar una estrategia para seleccionar a los padres (*parents*).
4. Realizar operaciones genéticas (cruce y mutación) para que se "reproduzcan" y así obtener la descendencia a partir de los padres.
5. Con la descendencia es posible crear un nuevo conjunto de soluciones potenciales.
6. Repetir el proceso, con excepción del primero pues ya no es necesario introducir una población inicial. Hasta que se cumpla alguna condición de parada.

Encontrar un subconjunto óptimo de individuos para un *dataset* de alta dimensionalidad requiere ejecutar una gran cantidad de evaluaciones de la función de ajuste, incluso para poblaciones pequeñas. Lo que hace de la técnica computacionalmente muy costosa. Por otro laod, esta función de ajuste es independiente de la estructura del algoritmo genético (55).

Nuestra función de ajuste o *fitness*, que será la encargada de evaluar el rendimiento de la población de soluciones potenciales, se calcula en términos de la precisión (*accuracy)* del modelo predictivo construido con las variables identificadas por la solución potencial bajo evaluación. Dado que solo queremos evaluar la afectación o diferencia que produce el hecho de estar en una clase u otra, será de clasificación binaria.

Como función *fitness*, en el presente TFG se ha empleado un modelo de regresión logística que se encargará de transformar las variables predichas en 0s y 1s, dando lugar a un vector de predicción (*Ypred*) binario, donde 0 es la no pertenencia al grupo eosinofílico y el 1 es la pertenencia al padecimiento de esta condición (eosinofilia), de la misma longitud que la dimensión del espacio de búsqueda.

Figura 17 muestra un diagrama de flujo de búsqueda de soluciones potenciales mediante un algoritmo genético. En este TFG no se ha empleado "elitismo", esto es, mantener fijo a los individuos con mejores precisiones.

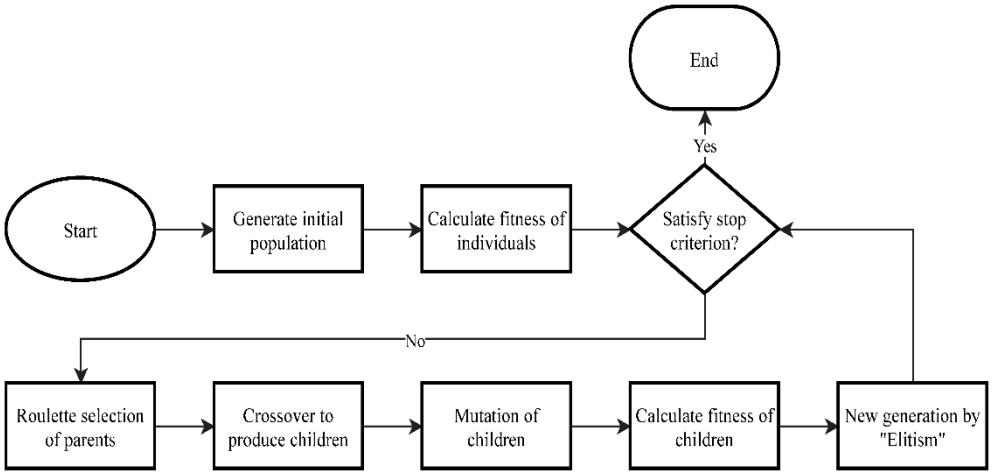

Figura 17. Diagrama de flujo de un GA (56).

6.7 Red neuronal, perceptrón multicapa.

Una red neuronal perceptrón multicapa (MLP) es una red neuronal artificial basadA en la forma que el cerebro humano organiza y procesa la información, hecho que sucede gracias a la interconexión neuronal.

La red estará compuesta por diferentes capas de diferentes niveles, entrada, capas ocultas y salida, todas interconectadas entre sí mediante unidades llamadas perceptrones.

Figura 18 muestra la comparación de un modelo de neuronas biológico con uno artificial.

La utilidad de redes neuronales artificiales reside en resolver problemas complejos cuya linealidad no está asegurada, como podría ser procesado de imagen, o de lenguaje natural (57). Cada perceptrón estará conectado con otros de diferentes capas. Cada una tiene una función de activación y están asociados a pesos adaptativos, siguiendo la siguiente expresión (58):

$$z_j = g_{act}(\sum_{i=1}^{d} w_{ij} \cdot x_i + b_j) \tag{6.32}$$

donde:

- x_i se refiere a cada una de las componentes del vector de características de entrada.
- d es la dimensión del vector de entrada o, lo que es lo mismo, el número de características.
- g_{act} es el valor de activación.
- b_j es el sesgo.

- w_{ij} son los pesos que caracterizan la interconexión del nodo con el resto de nodos.

La función de activación utilizada comúnmente es la función sigmoidal. No obstante, para nuestro problema de clasificación binaria en concreto, haremos uso de la función *softmax*. Similar a la sigmoidal, se podría considerar como la versión normalizada de ésta y permite describir una distribución probabilística (58).

En la red MLP, los valores de las variables se introducen en los nodos de la primera capa. La salida se calcula según la ecuación descrita con anterioridad y esta misma salida se convierte en una nueva entrada para la siguiente capa, la cual hace lo mismo. Este proceso se repite hasta que se haya llegado a la capa de salida.

El valor de salida es conocido y se calcula mediante la siguiente expresión:

$$y_k = f_k(x, w) = g_l \left\{ \sum_{j=1}^{N_h} w_{jk} \cdot g_t \left(\sum_{i=1}^{d} w_{ij} \cdot x_i + b_j \right) + b_k \right\}, \qquad k = 1, \dots, N_o \tag{6.33}$$

donde:

- d es el número de características del vector de entrada.
- N_h es el número de nodos en la capa oculta.
- w_{jk} son los pesos que conectan la salida de los nodos de la capa oculta con los nodos de la capa de salida.
- w_{ij} son los pesos adaptativos que conectan la característica i del vector de entrada con el nodo h_j de la capa oculta.
- b_j es el sesgo del nodo j en la capa oculta.
- g_t es la función de activación en la capa oculta.
- g_l es la función de activación en la capa de salida.
- N_o es el número de nodos en la capa de salida, que corresponde al número de clases a clasificar.

La red neuronal que se empleará en el presente TFG guarda similitudes con el que se presenta a continuación, pero en vez de tener dos capas ocultas, solo poseerá una.

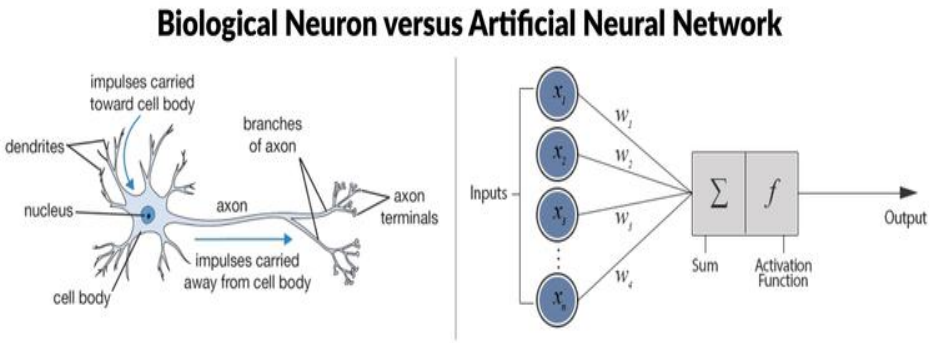

Figura 18. Símil de una red neuronal con red neuronal artificial (57).

Figura 19 presenta un modelo de MLP utilizado típicamente en la literatura.

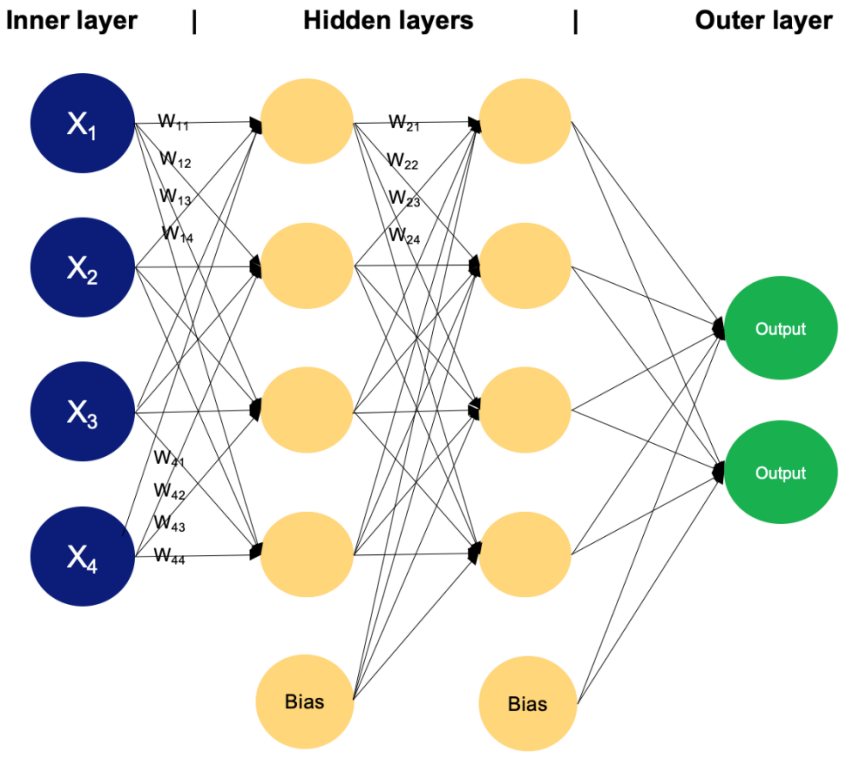

Figura 19. Ejemplo de una MLP común con una capa de entrada, nodos en capas oculta y 2 de salida (59).

En el presente trabajo se han utilizado dos nodos en la capa de salida, uno para cada uno de los dos grupos de pacientes que ya se han mencionado:

(i) Grupo de pacientes EPOC sin eosinofilia, codificado comúnmente con un 0. Se podría considerar el grupo control contra el que comparar.

(ii) Grupo de pacientes EPOC con eosinofilia, codificado comúnmente con un 1.

La red neuronal debe aprender los pesos óptimos w_{ij} a través de un proceso de aprendizaje. Dado que partimos desde una red neuronal nuevamente definida y no hemos tomado ninguna red neuronal ya entrenada, los pesos se deben de inicializar aleatoriamente para luego ajustarse al conjunto de entrenamiento, que viene definido por el par (x_n, t_n). Se corresponden al patrón de entrada y de salida, respectivamente.

La función de densidad probabilística tiene la siguiente expresión (58):

$$p(x.t) = p(t|x) \cdot p(x) \tag{6.34}$$

donde:

- $p(t|x)$ es la densidad de probabilidad de t condicionado a x.
- $p(x)$ es la densidad de probabilidad de x.

El objetivo de las MLPs es predecir el valor real t a partir de las características o parámetros proporcionados por el vector x. Como la clasificación binaria es la que se realiza en este TFG, t representará al grupo que pertenece el paciente [0 o 1].

Los pesos se han optimizado utilizando *stochastic gradiente descent with momentum optimizer* (SGDM), indicando 1 *epoch*, 1 *minibatch*, 2 capas de salida y valores óptimos de Nh y α que se han obtenido utilizando el conjunto de entrenamiento (Hold-Out 1), es decir, 23 pacientes y una estrategia de *Leave-One-Out cross-validation.*

Este último método consiste en entrenar la red con todos los individuos a excepción de 1. El individuo apartado servirá como validación. Este proceso implica entrenar tantas veces como individuos haya en la matriz de entrenamiento:

- Como, en nuestra matriz de entrenamiento tenemos 23 pacientes, para la primera iteración de entrenamiento usaremos 22 pacientes [paciente 2 al 23], el primer individuo se usará para validarlo.

- En la siguiente iteración usaremos los pacientes 1 y del 3 al 23, dejando el paciente número 2 para la validación, así hasta recorrer todos los pacientes.
- Como sólo validamos con 1 individuo en cada iteración, al repetir el proceso con todo el conjunto de entrenamiento se obtiene una matriz de confusión en el que están todos los individuos evaluados. De esta matriz de confusión se deriva la métrica que guía el proceso de optimización de los hiperparámetros. En este TFG se ha empleado la tasa global de aciertos (accuracy).

Se obtiene la precisión para cada par de valores nodos-regularización (Nh-α) y se usa, para el modelo final, el que maximice la métrica accuracy.

Los valores de Nh evaluados son de 1 nodo hasta 10 nodos. Los valores de α evaluados son: 0.01, 0.1, 1, 10 y 100.

6.8 Modelo de referencia. Regresión Logística.

El modelo de regresión logística binaria es una técnica estadística utilizada para resolver problemas de clasificación binaria, donde se predice la probabilidad de pertenencia de un elemento a una de las 2 clases que se definen, es decir, se usa para variables categóricas. Esto no sólo se usa para 2 clases, si no que puede usarse para el número de clases cualesquiera (regresión logística multinominal) (60).

La entrada de características x, es un conjunto de variables o características que no tienen por qué ser categóricas. La probabilidad, p, describe la posibilidad de que una instancia pertenezca a la clase positiva. Si ésta es mayor o igual que 0.5, generalmente se considera como clase 1, y si es menor, entonces será 0.

El modelo se obtiene a base de lo que cada ensayo Bernoulli (distribución de probabilidad binomial) y el conjunto de variables explicativas puedan informar acerca de la probabilidad final (60).

$$p_i = E\left(\frac{Y_i}{n_i}\middle| X_i\right) \tag{6.35}$$

Para que pueda suceder la transformación de dicha probabilidad, es necesaria una función logística, *logit*, que se define como:

$$p(X) = \frac{1}{1 + e^{-f(x)}} \tag{6.36}$$

donde $p(X)$ es la probabilidad de pertenencia a la clase positiva (presencia de la condición bajo estudio: 1) dada la entrada X, y $f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ es la combinación lineal de las características X con sus correspondientes coeficientes β.

En la Figura 20 se muestra una función de transformación logit característica que permite interpretar su salida en términos de probabilidad:

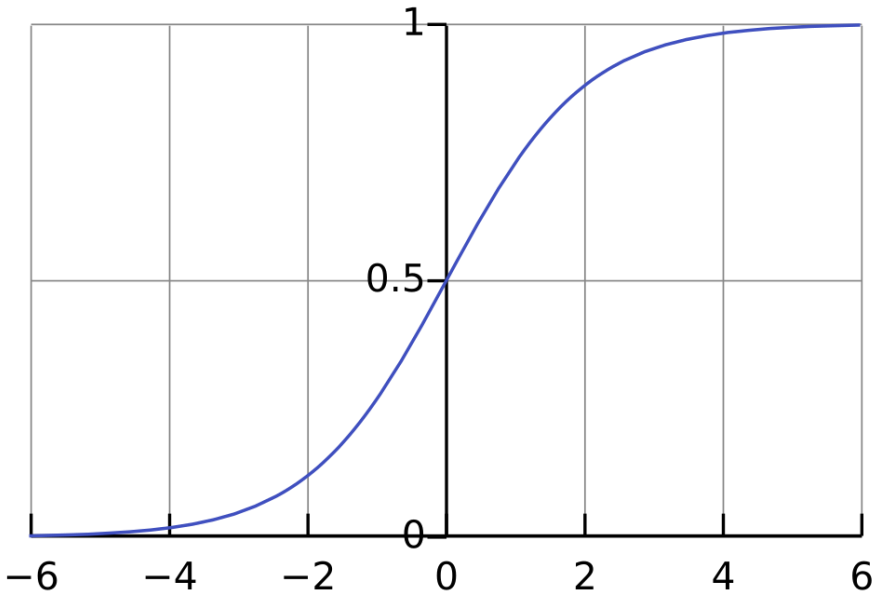

Figura 20. Transformación logit característica (59).

Es un modelo simple, eficiente y fácil de interpretar, y ha sido el modelo preferido en la medicina tradicionalmente. No obstante, tiene una serie de desventajas importantes, como la asunción de un modelo lineal, algo que, como ya hemos visto, no es la regla en cuanto a señales biomédicas se refiere.

En el presente TFG se empleará la regresión logística como *benchmark* o modelo de referencia al que comparar la eficacia de la red neuronal, que es el enfoque principal propuesto para detectar la condición de eosinofilia.

6.9 Evaluación del clasificador binario

Se realizará una comparación de los dos grupos de pacientes bajo estudio, construyendo las matrices de confusión correspondientes con sus respectivos parámetros:

- **Verdaderos positivos (TP):** Son los casos en los que el modelo predice la clase positiva de forma correcta. Lo que viene a ser lo mismo, tanto la clase predicha como la real son positivas y coinciden.

- **Falsos negativos (FN):** Son los casos en los que el modelo predice la clase negativa, pero en realidad es positiva. El modelo falla en identificar una clase positiva.

- **Falsos positivos (FP):** Son los casos en los que el modelo predice la clase positiva, pero en realidad es negativa. El modelo falla en identificar una clase negativa.

- **Verdaderos negativos (TN):** Son los casos en los que el modelo predice correctamente la clase negativa. La clase predicha y la real son negativas y coinciden.

A partir de estos índices obtenidos sobre la matriz de confusión, es posible calcular las diferentes métricas (61):

- **Precisión (Acc):** Cuantifica la tasa global de aciertos, que se representa como la proporción de los datos etiquetados correctamente entre el número total de casos examinados.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.37}$$

- **Sensibilidad (Sen):** Mide la proporción de casos positivos identificados de forma correcta:

$$Sen = \frac{TP}{TP + FN} \tag{6.38}$$

- **Especificidad (Sp)**: Mide la proporción de casos negativos que se han clasificado de forma correcta:

(6.39)

$$Sp = \frac{TN}{TN + FP}$$

- **Valor predictivo positivo (VPP):** Mide la proporción de sujetos clasificados como positivos por el modelo y que realmente lo son de acuerdo con el *GOLD standard*:

$$VPP = \frac{TP}{TP + FP} \tag{6.40}$$

- **Valor predictivo negativo (VPN):** Mide la probabilidad de que un caso clasificado negativamente lo sea realmente lo sea de acuerdo con el *GOLD standard*:

$$VPN = \frac{TN}{TN + FN} \tag{6.41}$$

- **Razón de verosimilitud positiva (LR+):** La razón de verosimilitud positiva mide cuánto más probable es que un resultado positivo proceda de un individuo que lo sea realmente, comparado con uno que no lo sea:

$$LR_{+} = \frac{Sen}{1 - Sp} \tag{6.42}$$

- **Razón de verosimilitud negativa (LR-):** La razón de verosimilitud negativa mide cuánto más probable es que un resultado negativo provenga de un individuo que presente la condición (sea 1) comparado con uno que no la presenta (sea 0):

$$LR_{-} = \frac{1 - Sen}{Sp} \tag{6.43}$$

Se realizará una comparación de los dos grupos de pacientes bajo estudio, construyendo las matrices de confusión correspondientes con sus respectivos parámetros:

Tabla 2. *Matriz de confusión al uso.*

DATOS (CLASE Y PREDICCIÓN)		**Clases predichas**	
		1	**0**
Clases reales	**1**	TP	FN
	0	FP	TN

6.10 Evaluación de la correlación entre variables con el conteo de eosinófilos

Por último, para el análisis descriptivo de los datos clínicos y sociodemográficos, se han definido y expresado en términos de mediana y rango intercuartil (q1 y q3), pues no se presentan propiedades gaussianas. Los tests estadísticos utilizados para la evaluación de diferencias significativas entre los dos grupos de pacientes bajo estudio (EOS vs No EOS) han sido:

- *Mann-Whitney*, si las variables son cuantitativas.
- *Chi2*, si las variables son categóricas.

Se analizaron 5 correlaciones:

1- Correlación entre los índices de modulación cardiaca derivados de la señal de HRV con la tasa de eosinófilos.
2- Correlación entre los índices de modulación cardiaca derivados de la señal de HRV con el número de exacerbaciones-
3- Correlación entre los índices de modulación cardiaca derivados de la señal de HRV con el número de exacerbaciones que hayan resultado en hospitalización.
4- Correlación entre la tasa de eosinófilos y el número de exacerbaciones.
5- Correlación entre la tasa de eosinófilos y el número de exacerbaciones que hayan resultado en hospitalización.

Para ello, se ha empleado la prueba de *Spearman*, si la variable es cuantitativa y el test de *Fisher*, si es categórica. Son medidas no paramétricas que evalúan la relación monotónica entre dos variables, esto es, si una variable tienda a aumentar o disminuir según la otra.

Todo ello fijando un umbral de probabilidad $\alpha = 0.05$ para considerar significación estadística.

7. RESULTADOS

Una vez realizado el reclutamiento de pacientes según los criterios de inclusión y exclusión para completar la base datos, se realizó un preprocesado de la señal a tratar, eliminando *outliers* y suavizando la señal, de tal modo que quedase libre de artefactos. A continuación, se extrajeron un conjunto de características a la señal estudiada, la HRV en el dominio temporal, espectral y no lineal, así como información sobre variables clínicas.

Con todos estos datos, a continuación, se presentan los resultados alcanzados y se evaluará la potencial existencia de diferencias significativas entre los dos grupos bajo estudio. También se mostrará la evaluación realizada por la red neuronal y las variables escogidas finalmente por los GAs.

7.1 Población bajo estudio

Un total de 29 pacientes cumplieron los requisitos de inclusión y firmaron el consentimiento informado. Dos de ellos fueron descartados por problemas con la adquisición y almacenamiento de registros. No se descartó a ningún paciente por exceso de artefactos, *outliers* ni por incompatibilidad en la historia, obteniendo un total de 27 pacientes finales.

La Figura 21 muestra un diagrama de flujo entre los pacientes que cumplieron los requisitos de inclusión y los que finalmente cuya señal de HRV fue incluida en el estudio.

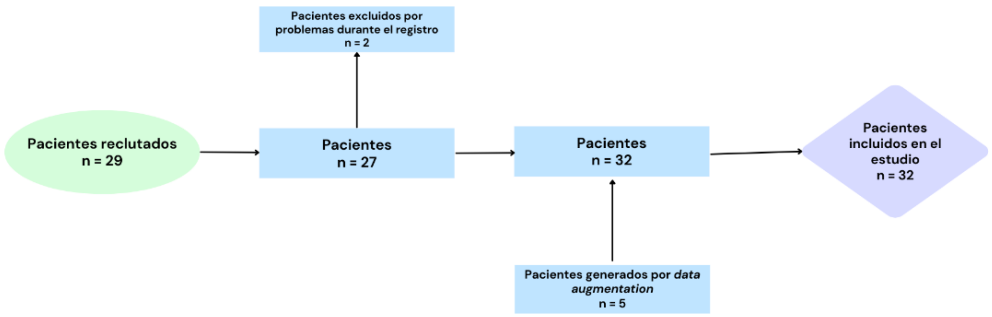

Figura 21. Diagrama de flujo de pacientes que forman parte de la población bajo estudio.

7.2 Análisis descriptivo de la población bajo estudio.

En la Tabla 3 se muestran las características sociodemográficas y antecedentes de la población bajo estudio, mientras que en la Tabla 4 se describen los signos vitales tomados en la consulta de Neumología durante el registro del ECG, así como variables de caracterización de la EPOC.

Es de denotar, en la tabla 3, que existen diferencias significativas en la edad.

Tabla 3. Mediana, cuartiles 1,3 (Q1, Q3) y p-valor de las características sociodemográficas y clínicas para los dos grupos bajo estudio.

Variables	***Pacientes n = 27***	***Grupo 0 n = 21***	***Grupo 1 n = 6***	***p-valor***
Edad	67.70 [65.20, 71.750]	67.20 [65.20, 70.80]	71.95 [51.20, 76.10]	$1.2 \cdot 10^{-5}$
Sexo (masculino)	11 (40.70%)	7 (38.88%)	4 (44.44%)	0.4479
IMC	23.03 (19, 26.08)	22.22 (18.92, 25.58)	25.21 (21.41, 35.56)	0.1705
FC	79 (60.25, 87.5)	79 (60.75, 84.5)	81.50 (55, 91)	0.8154
Enfermedades	13 (48.1%)	8 (38.09%)	5 (83.33%)	0.6946
Nº agudizaciones	0 (0, 1)	1 (0, 1)	0 (0, 1)	0.6468

Los datos se expresan en forma de mediana y los cuartiles 1,3 (Q1, Q3) para las variables continuas y en términos de número y porcentaje para las variables categóricas. IMC: Índice de masa corporal. FC: Frecuencia cardiaca.

Tabla 4. Mediana, Q1, Q3 y p-valor para los signos vitales y variables de caracterización de la EPOC para los dos grupos bajo estudio.

Variables	***Pacientes n = 27***	***Grupo 0 n = 21***	***Grupo 1 n = 6***	***p-valor***
FC	79 (60.25, 87.5)	79 (60.75, 84.5)	81.50 (55, 91)	0.8154
FR	20 (16, 23)	20 (20,24)	14 (12, 16)	0.0043
TAS	135 (125.25, 153.25)	133 (125.75, 146.5)	148.5 (123, 159)	0.4479
TAD	81 (74, 89.25)	82 (72.25, 90.25)	80 (77, 81)	0.7482
mMRC	1 (1, 3.75)	1 (1, 4)	1 (0,2)	0.217
CAT	12 (6.25, 16.75)	12 (7,17.25)	8.5 (3, 15)	0.3652
FEV1/FVC	52.51 (42.73, 65.02)	52.51 (44.24, 63.82)	53.765 (40.82, 66.82)	0.9303

Los datos se expresan en forma de mediana cuartiles Q1, Q3 para las variables continuas. FC: Frecuencia cardiaca; FR: Frecuencia respiratoria; TAS: Tensión arterial sistólica; TAD: Tensión arterial diastólica; mMRC: Test disnea mMMRC; CAT: Test de calidad de vida CAT; FEV_1/FVC: Indice de Tiffeneau.

En la tabla anterior se detallan las variables estrechamente relacionadas a la EPOC. Se ha observado un valor estadísticamente significativo para la frecuencia respiratoria.

7.3 Caracterización de la señal de variabilidad de la frecuencia cardiaca.

En la Tabla 5 se muestran las variables temporales extraídas del procesado de la señal biomédica, mientras que en la Tabla 6 se muestran las características espectrales y no lineales extraídas de la misma.

Tabla 5. Mediana, cuartiles Q1 y Q3 y p-valor de las características temporales de la señal de HRV para los dos grupos bajo estudio.

Variables	***Pacientes n = 27***	***Grupo 0 n = 21***	***Grupo 1 n = 6***	***p-valor***
AVNN	0.7230 (0.6882, 0.9407)	0.7320 (0.6963, 0.9407)	0.7192 (0.5899, 1.0099)	0.746
SDNN	0.0219 (0.015, 0.027)	0.0219 (0.0157, 0.0277)	0.0193 (0.0125, 0.0266)	0.5403
RMSSD	0.0148 (0.0107, 0.0174)	0.0148 (0.011, 0.017)	0.0131 (0.0077, 0.0175)	0.4311
NN50	3 (1, 8.5)	3 (1, 9.5)	1.5 (0, 3)	0.1957
pNN50	0.5848 (0.1481, 1.9797)	0.8757 (0.1527, 2.2868)	0.5070 (0, 0.8152)	0.4297
TINN	0.2048 (0.1172, 0.4114)	0.2048 (0.1175, 0.4411)	0.2376 (0.0786, 0.3256)	1

Los datos se expresan en forma de mediana y cuartiles Q1 y Q3. AVNN: media de intervalos NN; SDNN: desviación estándar de los intervalos NN; RMSDD: raíz cuadrada de la media de las diferencias en los intervalos NN; NN50: número de intervalos NN adyacentes que difieran más de 50ms; pNN50: porcentaje de intervalos NN que difieran más de 50ms entre el número de muestras; TINN: índice triangular de la señal HRV.

Tabla 6. Mediana, cuartiles Q1, Q3 y p-valor de las características espectrales y no lineares de la señal de HRV para los dos grupos bajo estudio.

Variables	***Pacientes n = 27***	***Grupo 0 n = 21***	***Grupo 1 n = 6***	***p-valor***
P_T	19.61 (17.74, 34.02)	19.61 (18.16, 33.53)	19.42 (13.01, 38.16)	0.7046
P_{VLF}	48.95 (47.97, 49.83)	48.95 (48.17, 49.89)	48.78 (47.02, 49.80)	0.7046
P_{LF}	0.9654 (0.9554, 0.97)	0.9636 (0.957, 0.969)	0.9676 (0.9534, 0.9708)	0.7046
P_{HF}	0.0346 (0.03, 0.0446)	0.0364 (0.0308, 0.043)	0.0324 (0.0292, 0.0466)	0.7046
$P_{LF/HF}$	27.90 (21.45, 32.32)	26.44 (22.30, 31.44)	29.88 (20.47, 33.26)	0.7046
SE	0.7267 (0.7259, 0.7288)	0.7267 (0.7261, 0.7291)	0.7261 (0.7252, 0.7284)	0.3977
SampEn	2.785 (2.31, 3.53)	2.81 (2.32, 3.54)	2.636 (1.91, 3.505)	0.4311
CTM	0.9744 (0.9323, 0.984)	0.9744 (0.914, 0.985)	0.9749 (0.9606, 0.9809)	0.9767
LZC	0.6562 (0.5749, 0.7898)	0.6622 (0.592, 0.798)	0.606 (0.54, 0.6875)	0.2319

Los datos se expresan en forma de mediana y cuartiles Q1 y Q3. PT: potencia total; PVLF: potencia en la banda de VLF; PLF: potencia en la banda de LF; PHF: potencia en la banda de HF; PLF/HF: relación de potencia absoluta entre la banda de LF y de HF; SE: entropía espectral; SampEn: entropía muestral; CTM: medida de la tendencia central; LZC: complejidad Lempel-Ziv.

7.4 Correlación de las distintas métricas con la eosinofilia.

En la Tabla 7 se muestran las métricas halladas por la prueba de *Spearman* por correlaciones entre las características de modulación autonómica y el conteo de eosinófilos, mientras que la Tabla 8 se muestra la correlación de las características de la modulación autonómica con el número de exacerbaciones.

Tabla 7. Valores rho y p-val obtenidos de la correlación de Spearman entre características extraídas de la señal de HRV con la tasa de eosinófilos

Correlación señal-EOS	rho	pval
AVNN	NaN	NaN
SDNN	-0.204	0.264
RMSSD	-0.353	**0.048**
NN50	-0.360	**0.043**
pNN50	-0.468	**0.007**
TINN	-0.397	**0.025**
PT	-0.030	0.265
MF	-0.190	0.298
SE	NaN	NaN
VLF	-0.234	0.198
LF	-0.129	0.480
HF	0.267	0.139
Relación HF/LF	-0.267	0.139
SampEn	0.267	0.139
CTM	-0.289	0.109
LZC	0.065	0.725

AVNN: media de intervalos NN; SDNN: desviación estándar de los intervalos NN; RMSDD: raíz cuadrada de la media de las diferencias en los intervalos NN; NN50: número de intervalos NN adyacentes que difieran más de 50ms; pNN50: porcentaje de intervalos NN que difieran más de 50ms entre el número de muestras; TINN: índice triangular de la señal HRV; PT: potencia total; PVLF: potencia en la banda de VLF; PLF: potencia en la banda de LF; PHF: potencia en la banda de HF; PLF/HF: relación de potencia absoluta entre la banda de LF y de HF; SE: entropía espectral; SampEn: entropía muestral; CTM: medida de la tendencia central; LZC: complejidad Lempel-Ziv.

Tabla 8. Valores rho y p-val obtenidos de la correlación de Spearman entre las características extraídas de la señal HRV con el número de exacerbaciones.

Correlación señal-exacerbación	rho	pval
AVNN	NaN	NaN
SDNN	0.055	0.766
RMSSD	0.053	0.775
NN50	-0.171	0.350
pNN50	-0.171	0.349
TINN	-0.202	0.268
PT	-0.025	0.892
MF	0.055	0.766
SE	NaN	NaN
VLF	-0.253	0.163
LF	0.015	0.934
HF	0.238	0.190
Relación HF/LF	-0.238	0.190
SampEn	0.238	0.190
CTM	-0.063	0.731
LZC	0.413	**0.019**

AVNN: media de intervalos NN; SDNN: desviación estándar de los intervalos NN; RMSDD: raíz cuadrada de la media de las diferencias en los intervalos NN; NN50: número de intervalos NN adyacentes que difieran más de 50ms; pNN50: porcentaje de intervalos NN que difieran más de 50ms entre el número de muestras; TINN: índice triangular de la señal HRV; PT: potencia total; PVLF: potencia en la banda de VLF; PLF: potencia en la banda de LF; PHF: potencia en la banda de HF; PLF/HF: relación de potencia absoluta entre la banda de LF y de HF; SE: entropía espectral; SampEn: entropía muestral; CTM: medida de la tendencia central; LZC: complejidad Lempel-Ziv.

En la Tabla 9 se muestra la correlación encontrada mediante el método de *Spearman* entre las características de la modulación autonómica y el número de exacerbaciones que resulten en hospitalización.

Tabla 9. Valores rho y p-val obtenidos de la correlación de Spearman entre las características extraídas de la señal HRV con el número de exacerbaciones que resulten en hospitalización.

Correlación señal-exacerbación que resulten en hospitalización	rho	pval
AVNN	NaN	NaN
SDNN	0.023	0.903
RMSSD	0.039	0.831
NN50	-0.141	0.442
pNN50	-0.071	0.701
TINN	-0.059	0.750
PT	0.073	0.693
MF	0.023	0.903
SE	NaN	NaN
VLF	-0.240	0.185
LF	0.058	0.753
HF	0.179	0.326
Relación HF/LF	-0.179	0.326
SampEn	0.179	0.326
CTM	-0.080	0.662
LZC	0.209	0.252

AVNN: media de intervalos NN; SDNN: desviación estándar de los intervalos NN; RMSDD: raíz cuadrada de la media de las diferencias en los intervalos NN; NN50: número de intervalos NN adyacentes que difieran más de 50ms; pNN50: porcentaje de intervalos NN que difieran más de 50ms entre el número de muestras; TINN: índice triangular de la señal HRV; PT: potencia total; PVLF: potencia en la banda de VLF; PLF: potencia en la banda de LF; PHF: potencia en la banda de HF; PLF/HF: relación de potencia absoluta entre la banda de LF y de HF; SE: entropía espectral; SampEn: entropía muestral; CTM: medida de la tendencia central; LZC: complejidad Lempel-Ziv.

Las Tablas 10 y 11 muestran la correlación de la tasa de eosinófilos con el número de exacerbaciones y el número exacerbaciones que resulten en hospitalización, respectivamente.

Tabla 10. Valores rho y p-val obtenidos de la correlación de Spearman entre la tasa de eosinófilos y el número de exacerbaciones.

Correlación tasa eosinófilos-exacerbaciones	rho	Pval
	-0.043	0.815

Tabla 11. Valores rho y p-val obtenidos de la correlación de Spearman entre la tasa de eosinófilos y el número de exacerbaciones que resulten en hospitalización.

Correlación tasa eosinófilos-exacerbaciones que resulten en hospitalización	**rho**	**p-val**
	0.197	0.279

Se puede observar una leve correlación inversa entre la *RMSSD* y *NN50* con la tasa absoluta de eosinófilos, donde, si ésta aumenta, los índices autonómicos tenderán a disminuir levemente.

Con el parámetro *pNN50* esta correlación inversa es más fuerte, aproximándose a -0.5. Además, la tasa de eosinófilos tiene correlación positiva con el parámetro TINN, aunque moderada.

Por otra parte, se ha observado una correlación positiva moderada entre la LZC y el número de exacerbaciones.

Como no hay ninguna variación perfectamente positiva o negativa, no se puede concluir que una variable aumente o disminuya de forma consistente según la otra, aunque se pueden observar ciertas tendencias.

7.5 Selección de características

Dada la poca cantidad de pacientes en la población final bajo estudio, las ejecuciones del algoritmo genético no convergen a variables que superen el umbral del 50% de precisión (métrica empleada como valor *fitness* para guiar la selección de variables) y muchas veces las variables importantes se difuminan entre las redundantes. Por motivos de colinealidad entre algunas de las variables propuestas, se ha eliminado de la selección la medida de: *reagudizaciones que resulten en ingreso*, quedándonos con únicamente *número de agudizaciones*.

En las Figuras 22 a 24, se muestra el histograma con la distribución del número de veces que cada variable del conjunto de partida fue seleccionada para formar parte del modelo óptimo final tras 100 ejecuciones del GA.

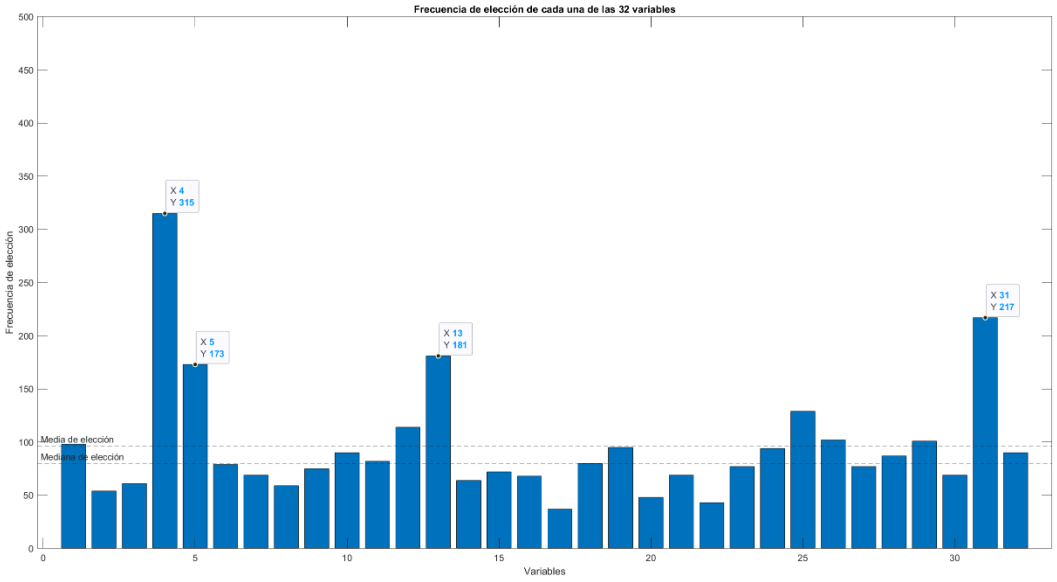

Figura 22. Variables escogidas por los algoritmos genéticos, primera ejecución.

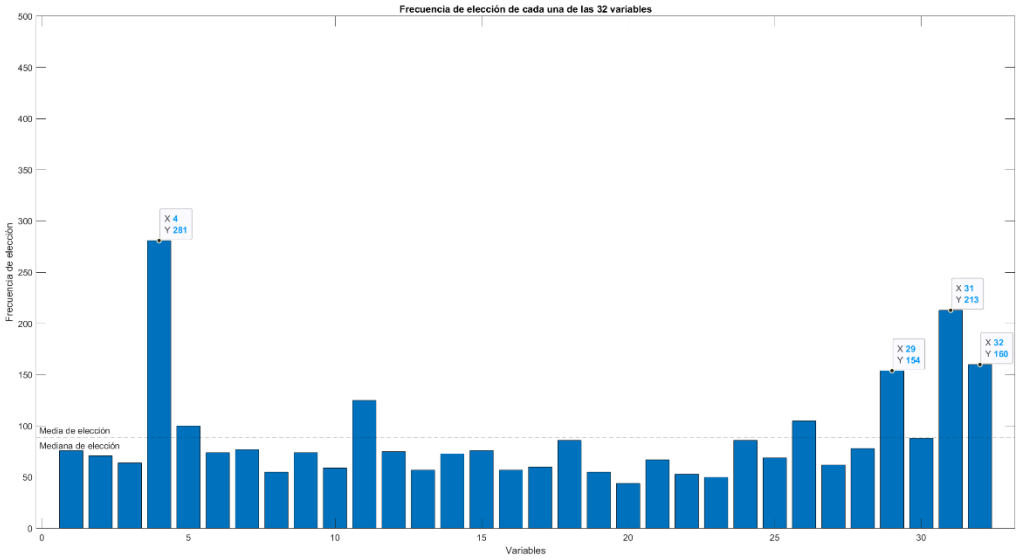

Figura 23. Variables escogidas por los algoritmos genéticos, segunda ejecución.

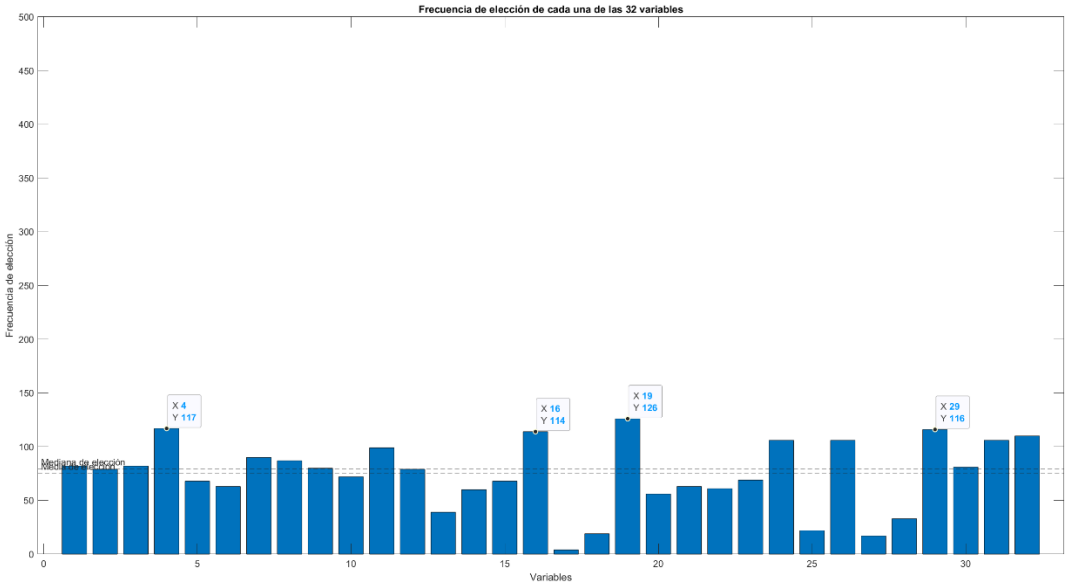

Figura 24. Variables escogidas por los algoritmos genéticos, tercera ejecución.

Con esta selección, es posible observar cierta tendencia de elección de variables, sobre todo la número 4, que se refiere al *NN50*, aunque el *número de reagudizaciones* también es destacable. Escogeremos únicamente 4 variables para evitar el problema del *course of dimensionality*, fenómeno que hace que los datos se vuelvan dispersos y no permitiría que la red neuronal entrene de forma correcta.

Como tenemos únicamente a 31 pacientes, el número máximo de variables de entrada al modelo se ha fijado en 4 (10% de 31, redondeado para completar un número entero de variables).

Por último, como los parámetros son de distinta naturaleza (*NN50, CAT, Frecuencia cardiaca, número de agudizaciones…*), y para aumentar nuestras opciones, haremos dos selecciones de variables. Esto es, elegir 2 sets de 4 variables más escogidas siendo éstas:

- Opción 1: *NN50, Años fumando, Frecuencia respiratoria y CAT.*
- Opción 2: *NN50, Cigarros/día, Número de ingresos y FEV1.*

Es interesante destacar que en ambos subconjuntos se incluyen de forma conjunta, variables de procesado de señal, clínicas y de caracterización de EPOC. Esto es, que tienen representación en dicho subconjunto diferentes enfoques de caracterización de los pacientes, lo que sugiere su complementariada.

En las Tablas 12 y 13 se mostrarán la media y desviación estándar de los dos subconjuntos:

Tabla 12. Media y desviación estándar del primer subconjunto de variables

EOS	NN50	Años fumando	Frecuencia respiratoria	Test CAT
0	*9.200* *12.451*	*42.80* *13.586*	*21.600* *5.193*	*13.80* *5.966*
1	*0.750* *1.388*	*41* *17.328*	*14.50* *2.976*	*9.625* *6.653*

Se muestra la media y desviación estándar para el primer subconjunto de variables, tanto para el grupo positivo como para el grupo negativo. NN50: número de intervalos NN adyacentes que difieran por más de 50ms; Test CAT: test de calidad de vida CAT.

Tabla 13. Media y desviación estándar del segundo subconjunto de variables

EOS	NN50	Cigarros al día	Número de ingresos	FEV1
0	*9.200* *12.451*	*25* *9.819*	*0.266* *0.457*	*1.325* *0.564*
1	*0.750* *1.388*	*25.50* *11.250*	*0.625* *0.9161*	*1.625* *0.509*

Se muestra la media y desviación estándar para el primer subconjunto de variables, tanto para el grupo positivo como para el grupo negativo. NN50: número de intervalos NN adyacentes que difieran por más de 50ms; FEV1: volumen espiratorio forzado en el primer segundo.

Una vez definidos estos dos subconjuntos, procederemos a entrenar la red neuronal.

7.6 Red neuronal perceptrón multicapa (MLP)

Elección de los hiperparámetros óptimos.

Una vez obtenida las variables por parte de los algoritmos genéticos, determinaremos los parámetros óptimos para definir la red neuronal final, la que entrenaría con todo el set de entrenamiento y clasificaría con el set de test que hemos reservado para este momento.

Los dos únicos hiperparámetros a optimizar serán el número de nodos, *Nh* y la regularización de la red, *alpha.* Para cada uno de los subconjuntos de variables óptimos propuestos, es necesario evaluar con cuál de los dos se obtiene la precisión más alta para problema de clasificación binaria propuesto en el presente trabajo.

Es necesario resaltar la necesidad de determinar hiperparámetros óptimos de forma diferenciada para los dos subconjuntos de variables que han mostrado tener resultados aceptables anteriormente.

En las Figuras 25 y 26 se muestra la evolución de la métrica de rendimiento para cada combinación de hiperparámetros y subconjunto de variables bajo estudio:

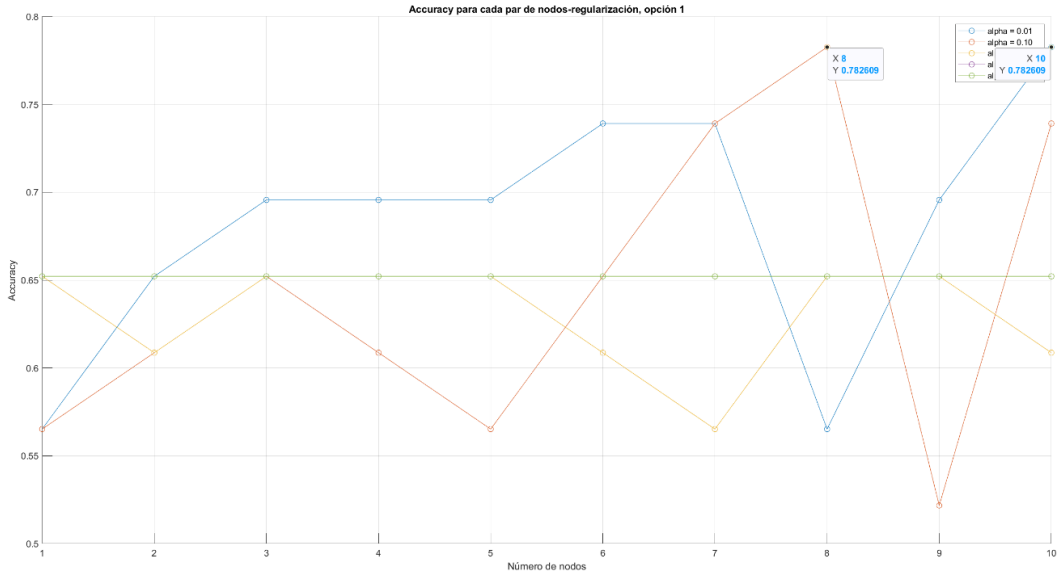

Figura 25. *Accuracy alcanzada por cada par nodo-regularización para el subconjunto 1*

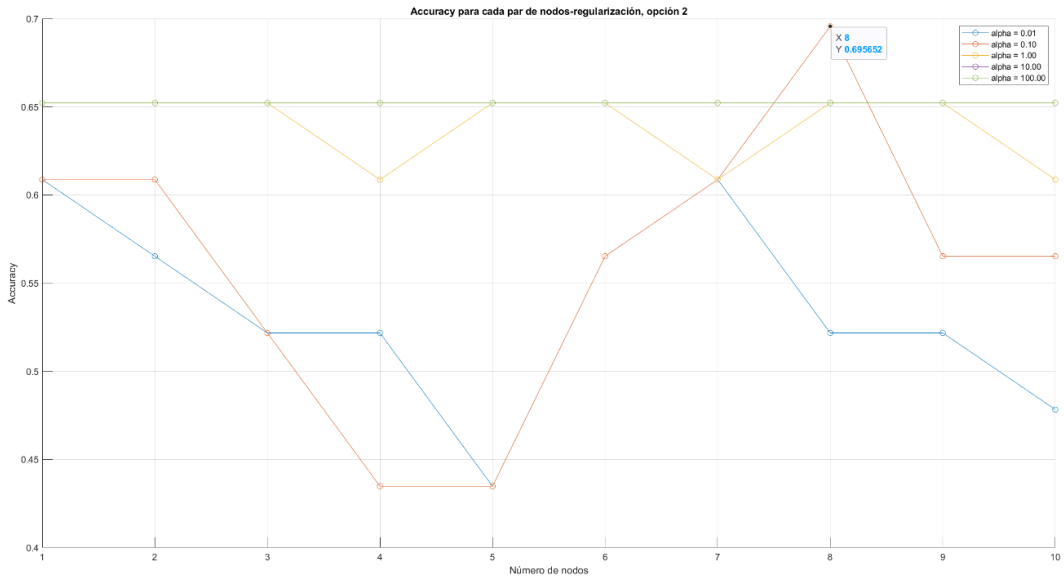

Figura 26. Accuracy alcanzada por cada par nodo-regularización para el subconjunto 2

Para el subconjunto de variables 1, se alcanzó una precisión máxima de 78.26% para el valor de *Nh* = 8 u 10 y *alpha* = 0.10.

Por otra parte, para el subconjunto de variables 2, se alcanzó una precisión máxima de 73.9% para un valor de *Nh* = 8 y *alpha* = 0.10.

7.7 Evaluación de la capacidad de predicción del clasificador – Red neuronal

Tras haber establecido los subconjuntos de entrada (variables) e hiperparámetros óptimos, se procedió a entrenar el modelo final de red neuronal manteniendo todos los parámetros constantes y usando los hiperparámetros *Nh* y *alpha* óptimos que determinados en el apartado anterior.

En las Tablas 14 a 15 se muestran las matrices de confusión obtenidas tras evaluar el modelo de red neuronal en el conjunto de variables 1 y 2 en los pacientes del conjunto independiente de test.

Tabla 14. Matriz de confusión para clasificación binaria en el subconjunto de variables 1.

DATOS (CLASE Y PREDICCIÓN)		**Clases predichas**	
		1	***0***
Clases reales	***1***	**2**	**1**
	0	**0**	**6**

En la Tabla 15 se resume el modelo rendimiento del modelo predictivo de la red neuronal en este conjunto de test para el subconjunto óptimo de variables 1.

Tabla 16 muestra la matriz de confusión para el subconjunto de variables 2.

Tabla 15. *Métricas de rendimiento del modelo predictivo para el subconjunto de variables 1.*

Acc	***Se***	***Sp***	***VPP***	***VPN***	***LR+***	***LR-***
89%	67%	100%	100%	85.71%	Inf	0.333

Tabla 16. Matriz de confusión para clasificación binaria en el subconjunto de variables 2

DATOS (CLASE Y PREDICCIÓN)		**Clases predichas**	
		1	***0***
Clases reales	***1***	**6**	**0**
	0	**3**	**0**

En la Tabla 17 se resume el modelo rendimiento del modelo predictivo de la red neuronal en este conjunto de test para el subconjunto óptimo de variables 2.

Tabla 17. Métricas de rendimiento del modelo predictivo para el subconjunto de variables 2.

Acc	*Se*	*Sp*	*VPP*	*VPN*	*LR+*	*LR-*
67%	0%	100%	NaN	67%	NaN	1

En las dos matrices de confusión se puede observar una diferencia evidente, las variables escogidas en la opción 1 clasifican con una buena precisión, tanto el grupo 0 y 1. Se podría argumentar que, dado un *dataset* más grande, ésta podría logar mayor precisión y seguramente capacidad de generalización.

El conjunto de variables número 2, a pesar de haber logrado una precisión decente a la hora de escoger los hiperparámetros de la red neuronal, no ha sido capaz de clasificar de forma eficaz el conjunto test, de hecho, ha clasificado a todos los pacientes en la clase positiva (eosinofilia). Esto demuestra que no ha sido capaz de entrenarse de forma correcta.

7.8 Evaluación de la capacidad de predicción del clasificador – Regresión Logística

En las Tablas 18 y 19 se muestras la matriz de confusión y las métricas de rendimiento, respectivamente, obtenidas empleando la regresión logística. Es necesario destacar que las dos opciones de variables óptimas (1 y 2) han generado exactamente la misma matriz de confusión, por lo que no se duplican las tablas.

Tabla 18. Matriz de confusión lograda por el modelo de regresión logística, ambas opciones.

DATOS (CLASE Y PREDICCIÓN)		**Clases predichas**	
		1	***0***
Clases reales	***1***	**0**	**3**
	0	**0**	**6**

Tabla 19. Métricas de rendimiento logradas por el modelo de regresión logística, ambas opciones.

Acc	*Se*	*Sp*	*VPP*	*VPN*	*LR+*	*LR-*
67%	**0%**	**100%**	**NaN**	**67%**	**NaN**	**1**

Como podemos observar a la vista de los resultados, el modelo obtenido con la regresión logística no consigue aprender las características generales del problema bajo estudio, a pesar de alcanzar una precisión del 67%, ya que el predictor asigna la clase negativa (no eosinofilia) a todos los pacientes.

8. DISCUSIÓN

En el presente TFG se analizó el registro HRV derivado del ECG realizados a pacientes diagnosticas de EPOC en el Servicio de Neumología, con el objetivo de establecer una correlación EOS-exacerbación entre pacientes con EPOC-eosinofilia y EPOC-no eosinofilia. Se comparó, además, el rendimiento de un modelo predictivo para la clasificación binaria en contraposición a la regresión logística.
Para construir el modelo de red neuronal propuesto, se implementaron las etapas comunes en el campo de diseño de modelos predictivos: (i) extracción de características; (ii) selección de variables; (iii) evaluación de la capacidad clasificadora.

En este penúltimo apartado se analizan los resultados obtenidos, examinando primeramente el análisis descriptivo de las características provenientes del análisis de señales (tiempo, espectral y no lineal), así como las variables clínicas. Se justifica la presencia en el modelo final de las variables seleccionadas por los algoritmos genéticos y, tras ello, se analiza la capacidad predictiva del modelo de red neuronal comparándola con la regresión logística junto con un breve comentario sobre las correlaciones halladas. Finalmente, se enumeran las limitaciones del trabajo.

8.1 Extracción de características y análisis estadístico

Respecto a las variables clínicas, es de esperar alguna diferencia en cuanto a la presencia de comorbilidades entre ambos grupos, pues las personas con EPOC tienen mayor riesgo de desarrollar enfermedades cardiovasculares y predisposición de diabetes por la inflamación sistémica y el estrés oxidativo al que somete al cuerpo la propia enfermedad (62). Sin embargo, no se han encontrado diferencias significativas en ambos grupos. A pesar de ello, es destacable que el grupo eosinofílico presenta estas comorbilidades en un porcentaje mayor que el grupo de pacientes sin eosinofilia.

Lo mismo ocurre con el número de agudizaciones, pues como se ha comentado en otros apartados, hay estudios que sugieren una relación entre la presencia de eosinófilos y las exacerbaciones. Sin embargo, tampoco se han observado diferencias signficativas.

Únicamente se han observado diferencias significativas en el grupo de edad, donde la mediana de edad en el grupo eosinofílico es mayor. Esto se podría explicar ya que, con el envejecimiento y la degeneración orgánica, los mecanismos antinflamatorios y antioxidantes fallan o son menos efectivos. Este

efecto junto a la naturaleza de la enfermedad puede causar un mayor daño y riesgo para personas de avanzada edad. Además, los eosinófilos pueden verse aumentados en las vías respiratorias por un sistema inmune y mecanismos de protección menos eficientes (62).

Por otra parte, y adentrándonos más en las variables más específicas de la EPOC, se esperarían encontrar diferencias significativas en las tensiones sistólica y diastólica por la presencia de comorbilidades (entre ellas las enfermedades cardiovasculares), los tests CAT y mMRC. Esto es porque el primero indica el grado de disnea y la discapacidad que ésta provoca, lo que la convierte en un indicador de exacerbación, mientras que el segundo evalúa la calidad de vida de los pacientes. Estas tampoco presentan una diferencia significativa, siendo muy similar en ambos grupos.

La única variable que presenta una diferencia significativa es la frecuencia respiratoria. Esta variable suele estar aumentada, por norma general en los pacientes de EPOC, pues se podría considerar respuesta compensatoria ante la disminución de la elasticidad pulmonar a la hora de respirar.

Es posible que los pacientes de EPOC eosinofílico al tener una inflamación mayor tengan la función pulmonar más degenerada, impidiendo a los pacientes respirar y expandir de forma correcta los pulmones. Esto daría lugar a una frecuencia respiratoria disminuida, ya que a los pacientes les cuesta mucho respirar, y las veces que lo hacen, lo hacen de forma profunda (63).

Por último, observando las variables extraídas de la señal HRV, no se encontraron diferencias estadísticamente significativas entre los dos grupos de pacientes para cada una de ellas de forma individual.

Sí que se esperaba encontrar alguna tendencia significativa, como en la *AVNN, SDNN o RMSSD*, todas características temporales y según la literatura deberían de verse disminuidos, caso que no se ha cumplido esta vez. Sin embargo, es destacable que la *NN50* haya formado parte de las variables incluidas en el modelo óptimo final, lo que sugiere su complementariedad con el resto de variables clínicas seleccionadas.

En cuanto a las características frecuenciales, la literatura no los abala ya que diferentes estudios no reportan diferencias significativas o eran contradictorias, aspecto que ocurre en este proyecto también.

Las métricas no lineales tampoco mostraron diferencias significativas entre ambos grupos.

8.2 Selección y clasificación de características

Las variables seleccionadas automáticamente por el algoritmo genético dentro del subconjunto 1 fueron el único conjunto de variables que ha conseguido un entrenamiento en la red neuronal y, por ende, predecir de forma efectiva.

La primera variable seleccionada, es NN50, variable no directa de la EPOC, pero que puede mostrar información sobre el estado del corazón. Aquellos con una EPOC más grave pueden tener la variable más disminuido por el aumento de la hipoxemia.

La segunda variable son los *años fumando*, cuya correlación con las enfermedades respiratorias queda ya demostrada en una variedad de estudios. El humo del tabaco es perjudicial para el pulmón y las vías respiratorias, causando inflamación y respuesta inmune crónica. Los pacientes que hayan fumado durante más tiempo tendrán el sistema respiratorio más degenerado.

La tercera variable es la *frecuencia respiratoria*. Como se ha explicado con anterioridad, una de las razones por las que existen diferencias significativas puede deberse a la pérdida de elasticidad pulmonar, haciendo más difícil la respiración. Los pacientes que estén más graves optarán por respirar de forma menos frecuente y más hondo.

Por último, la variable *CAT*, que es un evaluador de la afectación de la EPOC a la calidad de vida del paciente y discapacidad asociada. Está también relacionada la probabilidad de exacerbaciones.

Tras la selección de estas características se analizó la capacidad diagnóstica de la red neuronal, aunque alcanzó una precisión global del 89%, es importante mencionar la limitada generalización y potencia estadística de los resultados obtenidos, pues el conjunto de test consta sólo de 6 individuos.

Lo que sí se puede destacar, no obstante, es que la capacidad diagnóstica de la red neuronal, que es ampliamente superior a la de una regresión logística, enfoque tradicional que no ha conseguido clasificar de forma efectiva.

Este hecho sugiere dos interpretaciones:

- La relación de variables y la eosinofilia puede no ser lineal, requiriendo modelos no lineales para hallar la relación entre la eosinofilia y el resto del conjunto.
- Existe una ganancia predictiva en el uso de las redes neuronales sobre la regresión logística (al menos en este caso), pues la red neuronal sí es capaz de clasificar a los pacientes y la regresión no.

8.3 Correlación entre diferentes variables

Si bien es cierto que no se han encontrado diferencias estadísticas significativas entre ambos grupos (eosinofílico y no eosinofílico) para las variables de análisis de señal, sí se han podido observar correlaciones significativas (leves y moderadas) entre la tasa de eosinófilos y las variables temporales, tal y como parecían indicar las referencias bibliográficas. También se ha podido observar una correlación moderada positiva entre LZC y el número de exacerbaciones.

Las relaciones más a destacar son:

- La *pNN50* con una relación inversa del 46%. Es una de las métricas que definen la variabilidad cardiaca, indicaría de forma moderada que, a una mayor tasa de eosinófilos, menor es la *pNN50*, lo que conlleva una disfunción autonómica y mayor riesgo de complicaciones cardiovasculares.
- La LZC, con una relación positiva del 41%. Es otra de las métricas que extraídas del procesado de la señal de HRV. Esto puede indicar que, en pacientes con mayor tendencia a reagudizaciones, los patrones de ritmo cardiaco sean menos predecibles y erráticas (de ahí el aumento de la LZC), aumentando la variabilidad del SNA.

8.4 Limitaciones del estudio

Una vez analizado y discutido los resultados del presente TFG, es de importancia destacar las limitaciones de este, pues serán necesarias a la hora de interpretar y derivar conclusiones de los resultados que se han alcanzado.

La limitación más relevante es el número de sujetos, son demasiado pocos (27 tomados, 32 tras *data augmentation*) como para que sean representativos de la población a estudiar. Esto provoca directamente que las métricas estadísticas no sean representativas, donde una variable que podría tener significación a la hora de comparar entre grupos de acuerdo la literatura, no lo sea, pues sólo tenemos 6 pacientes eosinófilos de 27.

El tamaño reducido de la muestra también podría afectar a la selección de variables, al tener más variables que pacientes, influye en el diseño el algoritmo genético. Ninguna de las elegidas supera el umbral del 50% tras las repeticiones realizadas. Además, también puede provocar que la elección no sea la óptima, de nuevo, porque los datos no son representativos.

El diseño de la red neuronal ha sido muy simple para adaptarse a la cantidad de datos, con una única capa oculta, dando lugar a 2 consecuencias:

- La red podría tener cierta tendencia al infra-ajuste (*underfitting*) porque es demasiado simple, no tiene las suficientes capas para abordar el problema
- La red entrenada alcanza unas métricas de rendimiento con gran potencial de mejora, ya sea por simplicidad de la red o ya sea por falta de datos.

Además, también dificulta sacar conclusiones sobre correlaciones entre variables, dado que los grupos están tan desbalanceados que las correlaciones se difuminan o no son muy concluyentes, como en este caso.

Por otra parte, como todos los sujetos proceden del servicio de Neumología, incluir a todos los sujetos en la categoría EPOC (con o sin eosinofilia) no implica que no hayan podido sufrir otras patologías que interfieran con las variables del estudio y que no hayan sido identificadas. Sería interesante probar el modelo con pacientes de otros servicios o población general.

Relacionado con el servicio, todos los registros se realizaron con el mismo dispositivo. Sería interesante probar el modelo con los datos de otros equipos de adquisición para ver si es generalizable, si el equipo tiene algún error/sesgo leve, entre otros efectos.

Respecto a los registros recogidos, hubiera sido mejor registrar una duración más larga (8-10 minutos) e igual para todos los pacientes, pues paliaría (levemente) el problema de insuficiencia de datos además de permitir el empleo de otras técnicas.

Finalmente, en la selección de características, hubiera sido ventajoso usar *bootstrapping* para el remuestreo de instancias, así, se hubiera podido asignar medidas de precisión a los estimadores estadísticos utilizados, además de poder validar mejor el modelo con un conjunto de datos tan pequeño como el empleado en el presente TFG.

9. CONCLUSIONES

La Enfermedad Pulmonar Obstructiva Crónica, EPOC, es actualmente una de las enfermedades más prevalentes y mortales en los países desarrollados, tanto por ser una enfermedad que no es curable actualmente, como por tener estrecha relación con el consumo del tabaco. La EPOC supone, desde hace años, una gran carga asistencial para el sistema de salud público español. Por esa razón, este TFG se ha centrado en el análisis de la HRV y la relación EPOC-eosinofilia, con el objetivo de encontrar una correlación concluyente entre ambas dos, así como comparar el uso de modelos matemáticos convencionales como lo es la regresión logística binaria, contra otros más nuevos como una red neuronal.

Para llevar a cabo la investigación, fue necesario realizar registros de ECG a pacientes del servicio de Neumología del Hospital Universitario Río Hortega de Valladolid, de la cual se han analizado a 27 pacientes subdividiéndolos en 2 grupos: 0 para el EPOC sin eosinofilia, 1 para el EPOC con el conteo de eosinófilos por encima del umbral de normalidad.

La metodología consistió en 3 etapas fundamentales: extracción de características con las técnicas anteriormente descritas, selección de variables y clasificación de pacientes, utilizando técnicas temporales, espectrales y no lineales para la extracción.
Se usaron GAs para la selección de características y clasificación binaria con una red neuronal vs regresión logística.

Finalmente, los resultados se verían comparados con otros estudios que apliquen metodologías similares, pero debido a la falta de los mismos, no ha sido posible realizar una comparativa directa, lo que sugiere la novedad del trabajo.

9.1 Contribuciones del estudio

Las principales contribuciones al campo de la investigación en este ámbito han sido:

- Análisis y comparativa binaria directa entre pacientes de EPOC sin eosinofilia y EPOC con eosinofilia, observándose ciertas diferencias estadísticamente significativas entre sujetos de los dos grupos, algo que no ha sido estudiado en profundidad hasta el momento de la realización de este TFG.

- Uso de parámetros del HRV no tan extendidos (*NN50, pNN50 y* TINN) y encontrando correlaciones moderadas entre una de ellas y la tasa de eosinófilos.
- Implementación de una red neuronal artificial para la clasificación binaria utilizando escasos datos en el contexto de la patología y alcanzado una presión moderada-alta en uno de los modelos.
- Identificación de un subconjunto de variables óptimas heterogéneas, esto es, que son tanto de origen clínico como procedentes del análisis de señales. Esta complementariedad no abordada aún en el contexto EPOC-eosinofilia podría suponer un avance en la literatura para futuros estudios.

9.2 Líneas futuras de investigación

De acuerdo con las conclusiones, a continuación se exponen brevemente una serie de propuestas para futuras líneas de investigación que podrían mejorar sustancialmente el presente estudio.
En primer lugar, el tamaño de la base datos es el aspecto más esencial por mejorar, pues en este estudio se han manejado un número de sujetos excesivamente escaso, tanto que, si llegara a disminuirse aún más, se verían comprometidas las premisas del mismo o incluso el proyecto entero. Se propondría una distribución de grupos mucho más balanceada y más sujetos en general para una mayor generalización, variables y modelos más efectivos.

En segundo lugar, dado que sólo se analizó la señal HRV, también sería interesante analizar las señales de oximetría y acoplamiento cardiorrespiratorio (CRC) para tener una visión mucho más completa. Además, podría ser efectivo no considerar de nuevo las variables espectrales de la HRV, pues tanto en la literatura como en este estudio no han sido relevantes y buscar, en su lugar, otras métricas que fuesen más prometedoras. También sería interesante utilizar otros métodos de estimación espectral como la Wavelet o STFT para obtener la PSD.

En tercer lugar, podría mejorar la caracterización de la modulación autonómica cardíaca de los pacientes el uniformizar los tiempos de registro (incrementándolos) para un mejor preprocesado, generalización y aplicación de *data augmentation* en caso de ser necesario.

Finalmente, sería interesante evaluar la capacidad predictiva de otros enfoques de *machine learning,* predictivo como Gradient Boosting Machines (GBM) para mejorar la precisión.

9.3 Conclusiones principales del estudio

Las resultados alcanzados permiten obtener las siguientes conclusiones:

- La caracterización de la HRV en los 3 dominios (tiempo, frecuencia, no lineal) no mostró ningún índice significativo de forma individual.
- Existen diferencias significativas entre los grupos EPOC sin eosinofilia y EPOC eosinófilo para variables sociodemográficas (Edad) y clínicas (Frecuencia respiratoria), aún con la limitación del tamaño de la muestra.
- Existe una correlación entre características clínicas y del dominio del tiempo con la tasa de eosinófilos, incluso con la limitación del *dataset*.
- Las redes neuronales artificiales han demostrado ser útiles para la identificación de eosinofilia en pacientes de EPOC, siendo superiores a los modelos tradicionales de clasificación y alcanzando una precisión del 89% para variables mixtas.

Tras analizar las conclusiones obtenidas y compararlas con otros resultados publicados en la literatura, se puede concluir que el presente proyecto aporta información útil a tener en cuenta para futuros trabajos centrados en EPOC-eosinofilia.

ANEXOS

A continuación, se recoge el código desarrollado para el preprocesado y la extracción de variables. También se adjunta el código utilizado para programar la red neuronal. Se adjunta el código del *script* principal y después las funciones que son llamadas.

ANEXO 1. Código para el preprocesado y extracción de variables de la señal de HRV

```
clearvars
close all
clc

mypath_hrv = 'C:\Users\10518\Desktop\TFG\ECG_Bucle\';
myfile_hrv = dir([mypath_hrv '*.csv']);

window = 30*3.41;
noverlap = 0.5*window;
nfft = 128;
fs = 3.41;
tam = 0;
radio_ctm = 2.0092;
m = 1;

filename = "BBDD.xlsx";

for Npac = 1:length(myfile_hrv)
    HRV = readmatrix([mypath_hrv myfile_hrv(Npac).name]);
    HRV_f = HRV(:,1:2);
    [hrv, t] = elimina_interpola(HRV_f);
    RR = [hrv, t];
    radio_se = 0.1*std(RR);
    % Análisis en frecuencia:
    window = 30*3.41;
    noverlap = 0.5*window;
    nfft = 128;
    fs = 3.41;
    [pxx, f] = pwelch(RR,window,noverlap,nfft,fs);
    [pxx, f] = pwelch(RR,window,noverlap,nfft,fs);
    [v(1), v(2), v(3)] = procesado_tiempo(RR);
    [v(4), v(5), v(6)] = procesado_tiempo_2(RR);
    [v(7), v(8), v(9)] = analisis_frecuencia(pxx, f);
    [v(10), v(11), v(12), v(13)] = analisis_frecuencia_2(pxx, f);
    [v(14), v(15), v(16)] = analisis_no_lineal(RR, tam, m, radio_se,
radio_ctm);
    n = num2str(Npac+1);
    RA = "B";
    RB = "Q";
    rango1 = append(RA, n);
    rango2 = append(RB, n);
    rango = append(rango1,":",rango2);
    %Se escriben todas las variables la matriz deseada.
    writematrix(v, filename, "Range", rango);
end
```

Función: eliminar_interpola

```
function [HRV,t, window_size] = elimina_interpola(vector)
%Se preprocesa la señal para que elimine los valores menores de 0.33s,
%mayores a 1.5s y cuyos valores consecutivos sean mayores que 0.66s.
HRV = vector(:,2);
t = vector(:,1);
HRV_orig = HRV;
t_orig = t;

%Eliminamos muestras perdidas

% 1. Eliminamos las muestras menores de 0.33
pos_33 = find(HRV <= 0.33);

while ~isempty(pos_33)
    HRV_n = [(HRV(1:pos_33(1)-1))' (HRV(pos_33(1)+1:end))']'; %Asignamos un
HRV nuevo que no contenga el 0.33, es decir, el valor anterior y el
siguiente.
    HRV_n(pos_33(1)) = HRV_n(pos_33(1)) + HRV(pos_33(1)+1); %Para la posición
en la que existe la condición, le sumamos el siguiente.
    t_n = [(t(1:pos_33(1)-1))' (t(pos_33(1)+1:end))']'; %Se actualiza el
vector de tiempos.

    HRV = HRV_n;
    t = t_n;
    pos_33 = find(HRV <= 0.33);
end
% 2. Interpolamos las muestras mayores de 1.5
pos_150 = find(HRV >= 1.50);
pos_150_consec = detecta_consecutivos(pos_150);

if ~isempty(pos_150_consec)
    for i = 1:length(pos_150_consec)
        pos_150 = setdiff(pos_150, pos_150_consec{i}); %Eliminamos la
intersección de vectores
        interpol_150_consec = interp1(t, HRV, pos_150_consec{i}, "cubic");
        HRV(pos_150_consec{i}) = interpol_150_consec;
    end
end
if pos_150
    HRV = interpolacion_manual(t, HRV, pos_150);
end

% 3. Interpolamos las muestras cuyo valor consecutivo sea mayor de 0.66
%OBJETIVO, SI INTERP1 es NaN, identificar los puntos NaN y realizar la
%interpolación manual para esos NaN.
vc_06 = abs(diff(HRV));
pos_66 = find(vc_06 >= 0.66);
consec_66 = detecta_consecutivos(pos_66);
if ~isempty(consec_66)
    for i = 1:length(consec_66)
        pos_66 = setdiff(pos_66, consec_66{i});
        interpol_3d = interp1(t, HRV, consec_66{i}, "cubic");
        HRV(consec_66{i}) = interpol_3d;
    end
else
    HRV = interpolacion_manual(t, HRV, pos_66);
end
```

```
% 4. Interpolamos las muestras cercanas al valor 0.33 y 1.5, para esto
% usaremos una ventana deslizante que hemos definido en una función
% anterior.
%Usaremos una ventana de 10s, aunque habría que revisar la bibliografía
%para escoger una mejor ventana.
Vc_vent = [t, HRV];
seg = maximiza_seg(Vc_vent);
[outliers, window_size] = Ventana_Des(Vc_vent, seg);
consec = detecta_consecutivos(outliers);

%Interpolamos, si existen posiciones a eliminar, y son consecutivas, se
%eliminan las consecutivas de las posiciones originales y se interpolan
%linealmente. Luego, para las posiciones consecutivas utilizamos
%interpolación cúbica.
%Adicionalmente, nos aseguramos de que también se hagan interpolaciones
%aunque no existan posiciones consecutivas.
%Obviamente, si no existen posiciones a eliminar e interpolar, no se hace
%nada.
if ~isempty(consec)
    for i = 1:length(consec)
    outliers = setdiff(outliers, consec{i}); %Eliminamos la intersección de
vectores
    end
end
HRV = interpolacion_manual(t, HRV, outliers);
if ~isempty(consec)
    for i = 1:length(consec)
        interpol_3d = interp1(t, HRV, consec{i}, "cubic");
        HRV(consec{i}) = interpol_3d;
    end
end

%Sacar las posiciones en las que dif sea 1, e interpolar cúbicamente para
%todos los que sean 1, el anterior y el posterior



 %%% REPRESENTACIÓN DE LA SEÑAL HRV ORIGINAL Y DE LA SEÑAL HRV PROCESADA
%%%
 figure(2)
 plot(t, HRV, 'LineWidth', 3)
 hold on
 plot(t_orig, HRV_orig, 'LineWidth', 1)
 plot(t, 1.5*ones(1, length(t)), '-', t, 0.33*ones(1, length(t)))
 title('Señal HRV')
 legend('Señal HRV sin muestras perdidas', 'Señal HRV original')
 xlabel('Tiempo (s)')
 ylabel('Tiempo')
 hold off
```

Función: Ventana_Des

```
function [indices_outliers, window_size] = Ventana_Des(vector, seg)
%Vetana deslizante
t = vector(:,1);
HRV = vector(:,2);
window_size = 1;
```

```
dif = 0;
x = 1;
while dif <= seg
    diference = t(x+1) - t(x);
    window_size = window_size+1;
    dif = dif+diference;
    x = x+1;
end

% Inicializamos el vector
indices_outliers = [];
media_ventana = zeros(size(HRV));
std_ventana = zeros(size(HRV));

% Calcular la media móvil
for i = 1:length(HRV) - window_size + 1
    window = HRV(i:i+window_size-1);
    media_ventana(i) = mean(window);
    std_ventana(i) = std(window);
end
media_ventana = media_ventana(media_ventana ~= 0);
std_ventana = std_ventana(std_ventana ~= 0);
% Detectar outliers
for i = 1:length(HRV)
    if i < length(HRV) - window_size +1
        if HRV(i) > media_ventana(i)+1.96*std_ventana(i) || HRV(i) <
media_ventana(i) - 1.96*std_ventana(i)
            indices_outliers = [indices_outliers i];
        end
    elseif i >= length(HRV) - window_size +1
        if HRV(i) > media_ventana(end)+1.96*std_ventana(end) || HRV(i) <
media_ventana(end) - 1.96*std_ventana(end)
            indices_outliers = [indices_outliers i];
        end
    end
end
```

Función: Detecta_consecutivos

```
function segmentos = detecta_consecutivos(vector)
    % Detectar segmentos de 3 o más posiciones consecutivas del vector.
    %
    % Parámetros:
    % - vector: Vector de entrada (posiciones cualesquiera).

    % Encontrar las diferencias entre elementos consecutivos
    diferencias = diff(vector);

    % Inicializar variables. Queremos segmentos de 3 o más.
    inicio_segmento = 1;
    segmentos = {};

    for i = 1:length(diferencias)
        if diferencias(i) ~= 1
            % Si la diferencia no es 1, finalizar el segmento actual
            if i - inicio_segmento + 1 >= 3
                segmentos{end+1} = vector(inicio_segmento:i);
```

```
            end
            % Iniciar un nuevo segmento
            inicio_segmento = i + 1;
        end
    end

    % Añadir el último segmento si tiene 3 o más elementos
    if length(vector) - inicio_segmento + 1 >= 3
        segmentos{end+1} = vector(inicio_segmento:end);
    end
end
```

Función: interpolación_manual

```
function HRV = interpolacion_manual(t, HRV, pos)
    for i = 1:length(pos)
        idx = pos(i);

        % Verificar límites del índice
        if idx <= 1
            idx = 2; % Evitar el primer elemento
        elseif idx >= length(t)
            idx = length(t) - 1; % Evitar el último elemento
        end

        x1 = t(idx-1);
        x2 = t(idx+1);
        y1 = HRV(idx-1);
        y2 = HRV(idx+1);
        x_intrp = t(idx);
        y = y1 + (y2-y1)*(x_intrp-x1) / (x2-x1);
        % Asignamos el valor
        HRV(idx) = y;
    end
end
```

Función: procesado_tiempo

```
function [AVNN, SDNN, RMSSD] = procesado_tiempo(vector)
 % AVNN: el promedio del intervalo de pulso a pulso, siendo una estimación
 % global del período entre latidos (inverso de la frecuencia del pulso).
 AVNN = mean(vector);

 % SDNN: la desviación estándar de los intervalos de pulso a pulso, que
 % cuantifica el grado de variabilidad.
 SDNN = std(vector);

 % RMSSD: la raíz cuadrada media de las diferencias sucesivas de intervalos
 % de pulso a pulso (RMSSD), que representa la actividad vagal.
 diferencias = diff(vector);
 RMSSD = rms(diferencias);
end
```

Función: procesado_tiempo_2

```
function [NN50, pNN50, HRV_triangular_index] = procesado_tiempo_2(vector)

%NN50: El número de intervalos adyacentes que varían por más de 50 ms
diferencias = abs(diff(vector));
xNN50 = diferencias(diferencias > 0.050);
NN50 = length(xNN50);

%pNN50: El número de intervalos adyacentes que varían por más de 50ms,
%expresado en porcentaje.
pNN50 = (NN50/length(vector))*100;

%La triangular index se trata de una medida geométrica que calcula la
%integral de la densidad de un histograma de intervalos RR divididos por su
%altura. Básicamente, es calcular el área bajo un triángulo del grueso del
%histograma.
%Típicamente, usaremos un ancho para cada banda de 7.8125, pues es un valor
%estándar en estudios de HRV relacionado con el muestreo de los ECG.
bin_width = 0.0078125;
hist_edges = 0:bin_width:max(vector)+bin_width;
[n, edges] = histcounts(vector, hist_edges);

%Calcular la base
base_triangle = sum(n > 0);
max_frequency = max(n);
HRV_triangular_index = base_triangle / max_frequency;
```

Función: análisis_frecuencia

```
function [PT, MF, SE] = analisis_frecuencia(pxx,f)
% Cálculo de todas las métricas relevantes para el análisis ESPECTRAL.
% La potencia total se define como la suma de los coeficientes de la PSD en
todo el espectro.
PT = sum(pxx);
% FRECUENCIA MEDIANA
% Buscamos la componente frecuencial (Fm) en la que se alcanza la mitad de
%la potencia total de la PSD (PT/2).
PSD_acumulado = cumsum(pxx);
PSD_acumulado_norm = PSD_acumulado / PSD_acumulado(end);

ind_mediana = find(PSD_acumulado_norm >= 0.5, 1, "first");
MF = f(ind_mediana);

% ENTROPÍA ESPECTRAL
% La SE se calcula a partir de uns PSD normalizada, que se convierte en una
% función de densidad de la probabilidad.
fdp = pxx/PT;

% Cálculo de la entropía espectral.
SE = (-sum(fdp.*log(fdp)));
end
```

Función: análisis_frecuencia_2

```
function [VLF, LF, HF, rel_HLF] = analisis_frecuencia_2(pxx,f)
    VLF = sum(pxx(find(f >= 0.003 & f<0.04)));
    LF = sum(pxx(find(f >=0.04 & f<0.15)));
    HF = sum(pxx(find(f>= 0.15 & f <0.40)));
    %Relación LF/HF
    rel_HLF = LF/HF;

    %Lo normalizamos
    norm = LF+HF;
    VLF = VLF/norm;
    LF = LF/norm;
    HF = HF/norm;
end
```

Función: análisis no_lineal

```
function [SampEn, CTM, LZC] = analisis_no_lineal(vector, tam, m, radio_se,
radio_ctm)
%SampEn: Entropía muestral es una métrica que se utiliza para cuantificar
%la complejidad y regularidad de una señal.
SampEn = f_calculaSampEn(vector, m, radio_se);
%CTM: Medida que resume una distribución de datos mediante un valor que
%representa el punto central o típico de esos datos.
CTM = f_calcula_CTM(vector, radio_ctm);
%LZC: Medida del grado de "aleatoriedad" de una secuencia de datos, cuántas
%veces aparecen nuevas subsecuencias en una cadena de datos.
umbral = median(vector);
num_simbolos = 2;
LZC = lzcomplexity_tramas(vector, umbral, num_simbolos, tam);
```

Función: f_calculaSampEn

```
function value = f_calculaSampEn(signal, m, r, dist_type)
    % Error detection and defaults
    if nargin < 3, error('Not enough parameters.'); end
    if nargin < 4
        dist_type = 'chebychev';
        fprintf('[WARNING] Using default distance method: chebychev.\n');
    end
    if ~isvector(signal)
        error('The signal parameter must be a vector.');
    end
    if ~ischar(dist_type)
        error('Distance must be a string.');
    end
    if m > length(signal)
        error('Embedding dimension must be smaller than the signal length
(m<N).');
    end

    % Useful parameters
    signal = signal(:)';
    N = length(signal); % Signal length
    sigma = std(signal); % Standard deviation
```

```
    % Create the matrix of matches
    matches = NaN(m+1,N);
    for i = 1:1:m+1
        matches(i,1:N+1-i) = signal(i:end);
    end
    matches = matches';
    % Check the matches for m
    d_m = pdist(matches(:,1:m), dist_type);
    if isempty(d_m)
        % If B = 0, SampEn is not defined: no regularity detected
        % Note: Upper bound is returned
        value = Inf;
    else
        % Check the matches for m+1
        d_m1 = pdist(matches(:,1:m+1), dist_type);

        % Compute A and B
        % Note: logical operations over NaN values are always 0
        B = sum(d_m <= r*sigma);
        A = sum(d_m1 <= r*sigma);
        % Sample entropy value
        % Note: norm. comes from [nchoosek(N-m+1,2)/nchoosek(N-m,2)]
        value = -log((A/B)*((N-m+1)/(N-m-1)));
    end
    % If A=0 or B=0, SampEn would return an infinite value. However, the
    % lowest non-zero conditional probability that SampEn should
    % report is A/B = 2/[(N-m-1)(N-m)]
    if isinf(value)
        % Note: SampEn has the following limits:
        % - Lower bound: 0
        % - Upper bound: log(N-m)+log(N-m-1)-log(2)
        value = -log(2/((N-m-1)*(N-m)));
    end
end
```

Función: f_calcula_CTM

```
function CTM=Calcula_CTM(vector, R)
% Función que calcula la Medida de la Tendencia Central (CTM). La CTM es
% un parámetro que mide el grado de variabilidad de una serie temporal,
% haciendo uso para su cálculo de diagramas de dispersión los cuales
% representan diferencias de segundo orden. Estas están formadas a partir
% de del desplazamiento temporal de la señal original.
% La CTM se calcula seleccionando una región circular de radio r,
% alrededor del origen, contando el número de puntos que caen dentro del
% radio y dividiendo por el número total de puntos. Esto implica que el
% valor se encuentre siempre dentro del rango de 0 a 1.
%
% Argumentos de entrada:
% - registro: serie de datos de entrada de la que estimaremos su CTM.
% - radio: valor del radio que definirá la región circular en la que se
% van a contar los puntos que caen dentro.
%
% Argumentos de salida:
% - CTM: valor de la CTM para el registro introducido.
%
% Programado por: María Fernández Vaquerizo
%
```

```
% Última actualización: 30/03/2023
x = zscore(vector);
sum = 0;
for i = 1:length(x)-2
    if ((x(i+2)-x(i+1))^2 + (x(i+1)-x(i))^2)^(1/2) < R
        sum = sum + 1;
    end
end
CTM = 1/(length(x)-2)*sum;
end
```

Función: lzcomplexity_tramas

```
function LZcomplexity = lzcomplexity_tramas(registro,
umbral_db,num_simbolos,tam)
% Función que calcula la complejidad de Lempel-Ziv (Lempel-Ziv
% complexity, LZC) mediante el algoritmo propuesto en A. Lempel and J.
% Ziv, "On the complexity of finite sequences," IEEE Transactions on
% Information Theory, vol. IT-22, pp. 75-81, 1976.
%
% El LZC es una medida no paramétrica de quantificación de la complejidad
% en series de datos finitas. LZC está directamente relacionada con el
% número de cadenas (subsecuencias) diferentes y su repetición a lo largo
% de una secuencia de datos. Para detectar las diferentes subsecuencias,
% se convierte la señal original en una secuencia binaria de símbolos
% mediante la comparación con un umbral, habitualmente la media o la
% mediana de la serie de datos a analizar.
% La serie de entrada es dividida en segmentos, de forma que se obtiene
% un valor de LZC para cada segmento. Finalmente se realiza un
% promediado para obtener un único valor de LZC para la serie.
%
% Argumentos de entrada:
%
% - registro: Serie de datos de entrada de la que estimaremos su LZC.
% - umbral_db: Umbral empleado en la conversión binaria de la serie.
% Tomará los valores 'media' o 'mediana'.
% - num_simbolos: Número de símbolos empleados en la conversión binaria
% de la serie de datos original.
% - tam: Número de muestras de los segmentos en que se dividirá
% la serie de datos original. Si tam = 0, se aplicará el
% algoritmo sobre la serie original sin segmentar.
%
% Variables de salida:
% - LZcomplexity: Vector de valores de LZC para cada segmento
% del registro de SpO2. Este vector tiene tamaño 1xk,
% donde k es el número de segmentos de tamaño 'tam'
% muestras en que se posible dividir el registro de
% SpO2
%
% Grupo de Ingeniería Biomédica
% http://www.gib.tel.uva.es
% Universidad de Valladolid
%
% Programado por: Daniel Abásolo Baz.
% Modificado por: Daniel Álvarez González.
%
% Última actualización: 29/11/2007
%
```

```
% LZC se puede calcular sobre el registro completo (tam=0) o
% sobre el registro dividido en tramas de tamaño 'tam' muestras.
% Calculamos el número de veces a aplicar el algoritmo en la señal
% correspondiente viendo en cuántas tramas no solapadas de longitud
% 'tam' es posible dividir el registro de SpO2 original

if tam==0 % Si tam = 0 el algoritmo se calcula sobre el registro completo
    n_tramas=1; % El algoritmo se aplica una vez n_tramas=1
    tam=length(registro);
else % El algoritmo se aplica n_tramas veces
    n_tramas=floor(length(registro)/tam);
end
% Calculamos LZC para cada trama
n_tramas
for l=1:n_tramas
    %Extraemos la trama correspondiente del registro original
    trama = registro((1+(l-1)*tam):(l*tam));
    % Número de muestras de la trama.
    n=length(trama);

    % Transformación de la trama en una secuencia binaria
    trama=transforma_binaria(trama,num_simbolos,umbral_db);

    if num_simbolos == 2
        b=n/log2(n);
    else
        b=n/(log(n)/log(3));
    end
    % Inicializamos las variables empleadas en el cálculo la complejidad LZ.
    c = 1; % Valor inicial del contador de complejidad.
    S = trama(1); % Inicialización de la subsecuencia S.
    Q = trama(2); % Inicialización de la subsecuencia Q.

    for i = 2:tam
        % Concatenamos ambas subsecuencias.
        SQ = [S,Q];
        % Eliminamos el último caracter de la subsecuencia resultado de la
        % concatenación.
        SQ_pi = [SQ(1:(length(SQ)-1))];

        % Comprobamos si la subsecuencia Q se encuentra contenida en SQ_pi
        % Para encontrar coincidencias dentro de la secuencia de símbolos
        % bajo estudio, se considera que el contendio de la secuencia es
        % una cadena de caracteres y se utiliza una función típica de
        % comparación de cadenas, muy común en las librerías básicas de
        % muchos lenguajes de programación.
        indice = findstr(Q,SQ_pi); % Nos da los índices en los que Q empieza
dentro de SQ_pi

        if length(indice) == 0
            % Q no se encuentra al inspeccionar SQ_pi: Q es una secuencia
            % nueva
            c = c+1; %Incrementamos el contador de complejidad.
            if (i+1) > tam %Si llegamos al final de la serie hemos terminado
                break;
            else      %Si no hemos llegado al final concatenamos las
subsecuencias S y Q
                S = [S, Q]; %Formamos una nueva subsecuencia S
                Q = trama(i+1); %Actualizamos la subsecuencia Q
```

```
            end
        else
            %Q forma parte del SQ_pi.
            if (i+1) > tam   %Si llegamos al final de la serie hemos
terminado.
                break;
            else
                Q = [Q, trama(i+1)]; %Extendemos la subsecuencia Q.
            end
        end
    end
    %Normalizamos el contador de complejidad c de tal forma que 0 <= c/b.
    % <= 1
    LZcomplexity(l) = c/b;
end
```

Función: transforma_binaria

```
function s = transforma_binaria(serie, num_simbolos, umbral)
% Esta función transforma una señal temporal en una serie compuesta por un
% número de símbolos determinado de naturaleza binaria.
%
% Argumentos de entrada:
% - serie: Señal temporal a transformar
% - num_simbolos: Número de símbolos a utilizar en la transformación (2 o 3)
% - umbral: Umbral de decisión a aplicar en la transformación.
% Será una cadena de caracteres que indique si el
% umbral es la 'media' o la 'mediana' de la serie
% temporal bajo estudio.
%
% Argumentos de salida:
% - s: Señal transformada. Es la secuencia de símbolos
% binarios
%
% Grupo de Ingeniería Biomédica
% http://www.gib.tel.uva.es
% Universidad de Valladolid
%
% Programado por: Alicia Rodrigo de Diego y Jose Victor Marcos Martin.
% Modificado por: Daniel Álvarez González.
%
% Última actualización: 11/12/2007
    if num_simbolos ==2
        if strcmp(umbral,"mediana")
            mediana=median(serie);
            for i = 1:1:length(serie)
                if serie(i) < mediana
                    s(i) = 0;
                else
                    s(i) = 1;
                end
            end
        else
            media = mean(serie);
            for i = 1:1:length(serie)
                if serie(i) < media
                    s(i) = 0;
```

```
                else
                    s(i) = 1;
                end
            end
        end
    else
        mediana = median(serie);
        mx = abs(max(serie));
        mn = abs(min(serie));

        td1 = mediana-mn/16;
        td2 = mediana+mx/16;

        for i = 1:1:length(serie)
            if serie(i) <= td1
                s(i) = 0;
            elseif serie(i) < td2
                    s(i) = 1;
            else
                s(i) = 2;
            end
        end
    end
end
```

ANEXO 2. Código para la selección de variables mediante algoritmos genéticos

```
mypath_MD = 'C:\Users\10518\Desktop\TFG\Funciones_Matlab\';
myfile_MD = dir([mypath_MD '*.xlsx']);
MD = readmatrix("MD.xlsx");
MD = MD(:,2:end);
MD(:,24) = [];
MD(19,29) = median(MD(:,29),"omitnan");
MD(19,30) = mean(MD(:,30),"omitnan");
%
EOS = readmatrix("EOS_real.xlsx");
Y = EOS(:,1);
c = cvpartition(Y, 'HoldOut', 0.3, 'Stratify', true);
ind_train = training (c);
% ind_test = test(c);
%El ind_test está modificado manualmente sobre el primer índice creado para
%que los que se incluyan en el test sean únicamente reales.
ind_test = [0; 0; 0; 0; 0; 1; 1; 0; 0; 0; 0; 0; 0; 0; 1; 1; 0; 0; 0; 1; 0; 0;
1; 0; 1; 0; 1; 1; 0; 0; 0; 0;];
ind_test = logical(ind_test);
MDtrain = MD(ind_test == 0, :); %Esto lo aplico a MD, quiero todos los que
tengan el índice en train = 1, serán los que conformen MDtrain
Ytrain = Y(ind_test == 0);
MDtest = MD(ind_test == 1, :); %Esto se puede hacer así porque ind_test es
complementario de ind_train.
Ytest = Y(ind_test == 1);

%% Partición 2 del conjunto de test
c_ga = cvpartition(Ytrain, "HoldOut", 0.3, "Stratify", true);
ind_train_ga = training (c_ga);
ind_test_ga = test(c_ga);
MDtrain_ga = MDtrain(ind_train_ga == 1, :); %Esto lo aplico a MD, quiero
todos los que tengan el índice en train = 1, serán los que conformen MDtrain
```

```
Ytrain_ga = Ytrain(ind_train_ga == 1);
MDtest_ga = MDtrain(ind_train_ga == 0, :); %Esto se puede hacer así porque
ind_test es complementario de ind_train.
Ytest_ga = Ytrain(ind_train_ga == 0);
Psize = ceil(15 + rand*10);
[xopt,Acc_opt,dim_opt,Acc_mean,scores] = mi_ga(MDtrain_ga,Ytrain_ga,
MDtest_ga, Ytest_ga, Psize,0.8,0.01,100);
Var_opt = zeros(100, 32);
frec_var = zeros(10,32);
for i = 1:size(frec_var,1)
    for j = 1:length(Var_opt)
        [xopt] = mi_ga(MDtrain_ga,Ytrain_ga, MDtest_ga, Ytest_ga,
Psize,0.8,0.01,100);
        Var_opt(j,:) = xopt;
    end
    frec_var(i,:) = sum(Var_opt)
    i
end
frec_var_2 = sum(frec_var)
media_frec = ceil(mean(frec_var_2));
median_frec = median(frec_var_2);
figure
bar(frec_var_2);
xlabel("Variables")
ylabel("Frecuencia de elección")
title("Frecuencia de elección de cada una de las 32 variables");
hold on
yline(media_frec, '--', 'Media de elección', 'LabelHorizontalAlignment',
'left');
yline(median_frec, "--", "Mediana de elección", "LabelHorizontalAlignment",
"left");
yline(500, "--", "Límite del 50%", "LabelHorizontalAlignment", "left");
hold off
```

Función: mi_ga

```
function [xopt,Acc_opt,dim_opt,Acc_mean,scores] =
mi_ga(MDtrain_ga,Ytrain_ga,MDtest_ga, Ytest_ga, PopulationSize,Pc, Pm, NGen)
    [nf,nvars]=size(MDtrain_ga);
    %Se sobreentiende que MD e Y son train para el ga, nunca se ha de usar el
    %test y el otro hasta el final.
    fhandle_pobini = @(GenomeLength, FitnessFcn, options)mi_pobini(nf, nvars,
FitnessFcn, options, PopulationSize); %PopulationSize es el número de filas
de MD
    %aquí creo la función handle, no la pob_ini
    MDtrain_std = zscore(MDtrain_ga);
    [rowtest, coltest] = size(MDtest_ga);
    MDtest_std = zeros(rowtest, coltest);
    for i = 1:nvars
        meanMDtrain = mean(MDtrain_std(:,i));
        stdMDtrain = std(MDtrain_std(:,i));
        MDtest_std(:,i) = (MDtest_ga(:,i) - meanMDtrain) / stdMDtrain;
    end
    fhandle_fit = @(x)mi_fitfun(x, MDtrain_std, Ytrain_ga, MDtest_std,
Ytest_ga);%Después del arroba se ponen los obligatorios

    opciones=optimoptions('ga',...
        'CreationFcn',fhandle_pobini,...
```

```
        'CrossoverFraction',Pc,...
        'CrossoverFcn',@crossoversinglepoint,...
        'Generations', NGen,...
        'PopulationType','bitstring',...
        'MutationFcn',{@mutationuniform, Pm},...
        'PlotFcns',{@gaplotbestf},...
        'PlotInterval',1,...
        'PopulationSize',PopulationSize,...
        'SelectionFcn', @selectionroulette)


[xopt,Acc_opt,exitflag,output,population,scores]=ga(fhandle_fit,nvars,[],[],[
],[],[],[],[],[],opciones);
    dim_opt=sum(xopt);
    Acc_mean = mean(scores);
```

Función: mi_pobini

```
function SP= mi_pobini(nf, nvars, FitnessFcn, options, PopulationSize)
    row = nf;
    col=nvars;
    SP = zeros(row,col);
    for i = 1:row
        indices=randperm(col,2);
        SP(i,indices)= 1;
    end
end
```

Función: mi_fitfun

```
function Acc = mi_fitfun(x, MDtrain_std, Ytrain, MDtest_std, Ytest)
    if all(x == 0)
        Acc = Inf;  % Penaliza si no selecciona ninguna característica
    else
        B = glmfit(MDtrain_std(:, x == 1), Ytrain, 'binomial');
        yprob = glmval(B, MDtest_std(:, x == 1),'logit');
        yclass = (yprob>=0.5)*1;
        Acc = -sum(yclass == Ytest) / length(Ytest);
        x
    end
end
```

ANEXO 3. Código para la clasificación binaria mediante una red neuronal y comparación con un modelo de regresión logística.

```
%Se estandarizan las matrices de entrenamiento
MDtrain_1 = MDtrain(:,[4,22,26,31]);
MDtrain_std1 = zscore(MDtrain_1);
MDtrain_2 = MDtrain(:,[4, 21, 24, 29]);
MDtrain_std2 = zscore(MDtrain_2);
alpha = [0.01, 0.1, 1, 10, 100];
Nh = 1:10;
all_Ypred = zeros(size(MDtrain,1),1);
all_Ypred = categorical(all_Ypred);
Acc = zeros(10,5);
%Convertimos Ytrain en categorical para que funcione en la red neuronal
Y_cat = categorical(Ytrain);

%Primer se hallan los hiperparámetros óptimos
for i = 1:length(alpha)
    lambda = alpha(i);
    for j = 1:length(Nh)
        Nh_val = Nh(j);
        %Definimos la red neuronal con el número de neuronas
        layers = [
            featureInputLayer(size(MDtrain_std2,2))  %Capa de entrada con
número de características
            fullyConnectedLayer(Nh_val)         %Capa oculta con Nh_val
            reluLayer                           %Activación ReLU
            fullyConnectedLayer(2)              %Capa salida 2 clases
(binaria)
            softmaxLayer                        %Capa softmax para salida
            classificationLayer                 %Capa de clasificación
            ];
        % Configurar las opciones de entrenamiento, incluyendo la
regularización (L2 regularization)
        options = trainingOptions('sgdm', ...
            'MaxEpochs', 1, ...
            "MiniBatchSize",1, ...
            "L2Regularization",lambda, ...
            "Shuffle","never", ...
            "Verbose",false, ...
            "Plots","none");
        %Entrenamiento con Leave-One-Out
        for k = 1:size(MDtrain_2,1)
            MD_LOO = MDtrain_std2([1:k-1, k+1:end], :);
            Y_LOO = Y_cat([1:k-1, k+1:end]);
            MD_val = MDtrain_std2(k, :);
            Y_val = Y_cat(k,:);
            %Entrenar la red con el conjunto
            net_i = trainNetwork(MD_LOO, Y_LOO, layers, options);
            %Predecir para el LOO
            Y_pred = classify(net_i, MD_val)
            %Clasificación

            % Y_pred_class = round(Y_pred);
            %Guardamos todas las predicciones
            all_Ypred(k) = Y_pred;
            j
        end
```

```
        C = confusionmat(Y_cat,all_Ypred)
        Acc(j,i) = (C(1)+C(4))/(sum(sum(C)));
        disp(['Regularización: ', num2str(lambda), ', Nodos: ',
num2str(Nh_val)]);
    end
end
```

Se define el Código utilizado para la representación

```
%% Para la representación del par nodo-regularización óptimos

Acc_1 = readmatrix("Acc.xlsx");
Acc_2 = readmatrix("Acc_md2.xlsx");

% Graficar las series
figure(1);
hold on;  % Mantener todas las líneas en el mismo gráfico
for i = 1:length(alpha)
    plot(1:10,Acc_1(:,i),'-o', 'DisplayName', sprintf('alpha = %.2f',
alpha(i)));  % Graficar cada columna
end
% Añadir leyenda, etiquetas y título
legend('show');
xlabel('Número de nodos');
ylabel('Accuracy');
title('Accuracy para cada par de nodos-regularización, opción 1');
grid on;
hold off;  % Finalizar la adición de nuevas líneas


% Graficar las series
figure(2);
hold on;  % Mantener todas las líneas en el mismo gráfico
for i = 1:length(alpha)
    plot(1:10,Acc_2(:,i),'-o', 'DisplayName', sprintf('alpha = %.2f',
alpha(i)));  % Graficar cada columna
end
% Añadir leyenda, etiquetas y título
legend('show');
xlabel('Número de nodos');
ylabel('Accuracy');
title('Accuracy para cada par de nodos-regularización, opción 2');
grid on;
hold off;  % Finalizar la adición de nuevas líneas
```

Se define el código utilizado para definir la red neuronal con los hiperparámetros finales

```
%% Definimos las 2 redes neuronales y las pasamos por Ytest
%Se estandarizan las matrices de test con las de entrenamiento.
MDtrain_1 = MDtrain(:,[4,22,26,31]);
MDtrain_std1 = zscore(MDtrain_1);
MDtrain_2 = MDtrain(:,[4, 21, 24, 29]);
MDtrain_std2 = zscore(MDtrain_2);
MDtest_1 = MDtest(:,[4,22,26,31]);
MDtest_2 = MDtest(:,[4,21,24,29]);
```

```
[rowtest, coltest] = size(MDtest_1);
MDtest_std1 = zeros(rowtest, coltest);
MDtest_std2 = zeros(rowtest, coltest);
for i = 1:size(MDtrain_std1,2)
    meanMDtrain = mean(MDtrain_1(:,i));
    stdMDtrain = std(MDtrain_1(:,i));
    MDtest_std1(:,i) = (MDtest_1(:,i) - meanMDtrain) / stdMDtrain;
end
for i = 1:size(MDtrain_std2,2)
    meanMDtrain = mean(MDtrain_2(:,i));
    stdMDtrain = std(MDtrain_2(:,i));
    MDtest_std2(:,i) = (MDtest_2(:,i) - meanMDtrain) / stdMDtrain;
end
%Las que se obtienen por estandarizar
lambda = 0.1;
% lambda2 = 10;
Nh_1 = 8;
Nh_2 = 1;
Ytrain = categorical(Ytrain);
Ytest = categorical(Ytest);
Metricas_1 = zeros(100, 7);
vec_int = zeros(1,7);

% Definir la arquitectura de la red según el entrenamiento
layers_1 = [
    featureInputLayer(size(MDtrain_1,2))                % Capa de entrada
    fullyConnectedLayer(Nh_1)                       % Capa oculta con 7
neuronas
    reluLayer                                       % Función de activación ReLU
    fullyConnectedLayer(2)              % Capa de salida con número de clases
    softmaxLayer                                    % Función de activación
softmax para clasificación
    classificationLayer                             % Capa de clasificación
];
options = trainingOptions('sgdm', ...               % Optimizador sgdm
    'MaxEpochs', 1, ...                            % Número máximo de épocas
    'MiniBatchSize', 1, ...                        % Tamaño del mini-batch
    'L2Regularization',lambda,...
    'Shuffle', 'never', ...                % Mezcla los datos en cada época
    'Plots', 'none', ...
    'Verbose', false);
C_max = [];
for i = 1:size(Metricas_1,1)
    %Entrenamiento
    net_1 = trainNetwork(MDtrain_std1, Ytrain, layers_1, options);
    %Evaluar la red
    Y_pred = classify(net_1, MDtest_std1);
    %Metricas
    C = confusionmat(Ytest,Y_pred);
    vec_int(:,1:7) = Calcula_metricas(C);
    Metricas_1(i,1:7) = vec_int;
end


Metricas_2 = zeros(100, 7);
vec_int = zeros(1,7);
layers_2 = [
    featureInputLayer(size(MDtrain_2,2))                % Capa de entrada
```

```
    fullyConnectedLayer(Nh_2)                    % Capa oculta con 6
neuronas
    reluLayer                                    % Función de activación ReLU
    fullyConnectedLayer(2)              % Capa de salida con número de clases
    softmaxLayer                                 % Función de activación
softmax para clasificación
    classificationLayer                          % Capa de clasificación
];
options_2 = trainingOptions('sgdm', ...          % Optimizador sgdm
    'MaxEpochs', 1, ...                         % Número máximo de épocas
    'MiniBatchSize', 1, ...                     % Tamaño del mini-batch
    'L2Regularization',lambda,...
    'Shuffle', 'never', ...               % Mezcla los datos en cada época
    'Plots', 'none', ...
    'Verbose', false);
for i = 1:size(Metricas_2,1)
    %Entrenamiento
    net_2 = trainNetwork(MDtrain_std2, Ytrain, layers_2, options_2);
    %Evaluar la red
    Y_pred = classify(net_2, MDtest_std2);
    %Metricas
    C = confusionmat(Ytest,Y_pred);
    vec_int(:,1:7) = Calcula_metricas(C);
    Metricas_2(i,1:7) = vec_int;
end
```

Función: Calcula_métricas

```
function Met = Calcula_metricas(C)
    TN = C(1);
    FN = C(2);
    FP = C(3);
    TP = C(4);
    Acc = (TP + TN) / (TP + TN + FP + FN);
    Sen = TP / (TP + FN);
    Sp = TN / (TN + FP);
    VPP = TP / (TP + FP);
    VPN = TN / (TN + FN);
    LR_pos = Sen / (1 - Sp);
    LR_neg = (1 - Sen) / Sp;
    Met = [Acc, Sen, Sp, VPP, VPN, LR_pos, LR_neg];
end
```

Se define el modelo de regresión logística binaria

```
%% Regresión lineal
MDtrain_std1 = zscore(MDtrain_1);
MDtrain_std2 = zscore(MDtrain_2);
[rowtest, coltest] = size(MDtest_1);
MDtest_std1 = zeros(rowtest, coltest);
MDtest_std2 = zeros(rowtest, coltest);
Ytrain = double(Ytrain);
Metricas_RL_1 = zeros(1,7);
Metricas_RL_2 = zeros(1,7);
for i = 1:size(MDtrain_std1,2)
    meanMDtrain = mean(MDtrain_std1(:,i));
    stdMDtrain = std(MDtrain_std1(:,i));
```

```
    MDtest_std1(:,i) = (MDtest_std1(:,i) - meanMDtrain) / stdMDtrain;
end
for i = 1:size(MDtrain_std1,2)
    meanMDtrain = mean(MDtrain_std2(:,i));
    stdMDtrain = std(MDtrain_std2(:,i));
    MDtest_std2(:,i) = (MDtest_std2(:,i) - meanMDtrain) / stdMDtrain;
end

C_reg1 = regresion_logistica(MDtrain_std1,Ytrain,MDtest_std1,Ytest);
C_reg2 = regresion_logistica(MDtrain_std2,Ytrain,MDtest_std2,Ytest);
```

Función: regression_logística

```
function C = regresion_logistica(MDtrain_std,Ytrain,MDtest_std,Ytest)

    B = glmfit(MDtrain_std, Ytrain, "binomial");
    Yprob = glmval(B, MDtest_std, "logit");
    Yclass = (Yprob>=0.5)*1;
    C = confusionmat(Ytest,Yclass);
end
```

BIBLIOGRAFÍA

1. Leovigildo Ginel and Marta Gómez del Valle. “Historia de la enfermedad pulmonar obstructiva crónica (EPOC): desde Hipócrates hasta nuestros días”.

2. Christenson SA, Smith BM, Bafadhel M, Putcha N. Chronic obstructive pulmonary disease. Lancet [Internet]. 2022 Jun 11 [cited 2024 Sep 24];399(10342):2227–42. Available from: https://pubmed.ncbi.nlm.nih.gov/35533707/

3. Lõpez-Campos JL, Tan W, Soriano JB. Global burden of COPD. Respirology [Internet]. 2016 Jan 1 [cited 2024 Sep 22];21(1):14–23. Available from: https://pubmed.ncbi.nlm.nih.gov/26494423/

4. Prevalencia de la EPOC en España - APEPOC - Asociación de Pacientes con EPOC [Internet]. [cited 2024 Sep 17]. Available from: https://www.apepoc.es/actualidad/750-prevalencia-de-la-epoc-en-espana?jjj=1710329615397&jjj=1726587165050

5. Walter RE, Wilk JB, Larson MG, Vasan RS, Keaney JF, Lipinska I, et al. Systemic inflammation and COPD: The Framingham heart study. Chest [Internet]. 2008 Jan 1 [cited 2024 Sep 17];133(1):19–25. Available from: http://journal.chestnet.org/article/S0012369215489536/fulltext

6. Hersh CP, Make BJ, Lynch DA, Barr RG, Bowler RP, Calverley PMA, et al. Non-emphysematous chronic obstructive pulmonary disease is associated with diabetes mellitus. BMC Pulm Med [Internet]. 2014 [cited 2024 Sep 22];14(1). Available from: /pmc/articles/PMC4216374/

7. Definition of Chronic Obstructive Pulmonary Disease (COPD): Is the Latest GOLD Classification of Severity Still Valid? | Thoracic Key [Internet]. [cited 2024 Sep 17]. Available from: https://thoracickey.com/definition-of-chronic-obstructive-pulmonary-disease-copd-is-the-latest-gold-classification-of-severity-still-valid/

9. Barnes PJ. Inflammatory endotypes in COPD. Allergy [Internet]. 2019 Jul 1 [cited 2024 Sep 17];74(7):1249–56. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/all.13760

10. FORMACIÓN MÉDICA ACREDITADA. Conceptos clínicos básicos sobre exploración funcional respiratoria (II) [Internet]. [cited 2024 Sep 17]. Available from: https://corporativa.amyts.es/noticias/noticia/?slug=exploracion-funcional-respiratoria-ii

11. Ko FW, Chan KP, Hui DS, Goddard JR, Shaw JG, Reid DW, et al. Acute exacerbation of COPD. Respirology [Internet]. 2016 Oct 1 [cited 2024 Sep 17];21(7):1152. Available from: /pmc/articles/PMC7169165/

12. Wedzicha JA, Seemungal TA. COPD exacerbations: defining their cause and prevention. Lancet [Internet]. 2007 Sep 1 [cited 2024 Sep 22];370(9589):786. Available from: /pmc/articles/PMC7134993/

13. Llauger Rosselló M, Naberan Toña K. GOLD: estrategia mundial para la enfermedad pulmonar obstructiva crónica (EPOC). Aten Primaria [Internet]. 2003 Sep 30 [cited 2024 Sep 17];32(5):306–10. Available from: https://www.elsevier.es/es-revista-atencion-primaria-27-articulo-gold-estrategia-mundial-enfermedad-pulmonar-13051600

14. Grados de EPOC | Clasificacion GOLD EPOC | Fundación Rene [Internet]. [cited 2024 Sep 17]. Available from: https://www.fundacionrenequinton.org/blog/grados-epoc-pautas-recomendadas/

15. Tantucci C, Modina D. Lung function decline in COPD. Int J Chron Obstruct Pulmon Dis [Internet]. 2012 [cited 2024 Sep 17];7:95. Available from: /pmc/articles/PMC3282601/

16. COPD - Diagnosis and treatment - Mayo Clinic [Internet]. [cited 2024 Sep 17]. Available from: https://www.mayoclinic.org/diseases-conditions/copd/diagnosis-treatment/drc-20353685

17. García Castillo E, Vargas G, García Guerra JA, López-Giraldo A, Alonso Pérez T. [Chronic Obstructive Pulmonary Disease]. Open respiratory archives [Internet]. 2022 Apr 1 [cited 2024 Sep 17];4(2). Available from: https://pubmed.ncbi.nlm.nih.gov/37497315/

18. Enfermedad pulmonar obstructiva crónica DOCUMENTO DE CONSENSO.

19. Soler-Cataluña JJ, Piñera P, Trigueros JA, Calle M, Casanova C, Cosío BG, et al. Spanish COPD Guidelines (GesEPOC) 2021 Update Diagnosis and Treatment of COPD Exacerbation Syndrome. Arch Bronconeumol [Internet]. 2022 Feb 1 [cited 2024 Sep 17];58(2):159–70. Available from: https://pubmed.ncbi.nlm.nih.gov/34172340/

20. Alqahtani JS, Aldhahir AM, Alghamdi SM, Al Ghamdi SS, AlDraiwiesh IA, Alsulayyim AS, et al. A systematic review and meta-analysis of heart rate variability in COPD. Front Cardiovasc Med [Internet]. 2023 [cited 2024 Sep 17];10:1070327. Available from: /pmc/articles/PMC9981678/

21. Shaffer F, Ginsberg JP. An Overview of Heart Rate Variability Metrics and Norms. Front Public Health [Internet]. 2017 Sep 28 [cited 2024 Sep 17];5:258. Available from: /pmc/articles/PMC5624990/

22. Guidelines Heart rate variability Standards of measurement, physiological interpretation, and clinical use.

23. Stein PK, Pu Y. Heart rate variability, sleep and sleep disorders. Sleep Med Rev. 2012 Feb 1;16(1):47–66.

24. Mohammed J, Meeus M, Derom E, Da Silva H, Calders P. Evidence for autonomic function and its influencing factors in subjects with COPD: A systematic review. Vol. 60, Respiratory Care. American Association for Respiratory Care; 2015. p. 1841–51.

25. Jones PW, Harding G, Berry P, Wiklund I, Chen WH, Kline Leidy N. Development and first validation of the COPD Assessment Test. European Respiratory Journal [Internet]. 2009 Sep 1 [cited 2024 Sep 17];34(3):648–54. Available from: https://erj.ersjournals.com/content/34/3/648

26. Monitorización y estudio del paciente en cuidados críticos - Cuidados críticos - Manual MSD versión para profesionales [Internet]. [cited 2024 Sep 17]. Available from: https://www.msdmanuals.com/es-es/professional/cuidados-cr%C3%ADticos/abordaje-del-paciente-con-enfermedad-cr%C3%ADtica/monitorizaci%C3%B3n-y-estudio-del-paciente-en-cuidados-cr%C3%ADticos#Monitorizaci%C3%B3n-card%C3%ADaca_v924337_es

27. Dong JG. The role of heart rate variability in sports physiology (Review). Exp Ther Med. 2016 May 1;11(5):1531–6.

28. McCraty R, Shaffer F. Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-regulatory Capacity, and Health Risk. http://dx.doi.org/107453/gahmj2014073 [Internet]. 2015 Jan 1 [cited 2024 Sep 17];4(1):46–61. Available from: https://journals.sagepub.com/doi/10.7453/gahmj.2014.073

29. Penzel T, Kantelhardt JW, Grote L, Peter JH, Bunde A. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. IEEE Trans Biomed Eng. 2003 Oct 1;50(10):1143–51.

30. Xie T, Ma N. Tracking vigilance fluctuations in real-time: a sliding-window heart rate variability-based machine-learning approach. Sleep [Internet]. 2024 Aug 26 [cited 2024 Sep 17]; Available from: https://dx.doi.org/10.1093/sleep/zsae199

31. Nano M, Fonseca P, Overeem S, Vullings R, Aarts RM. Lying Awake at Night: Cardiac Autonomic Activity in Relation to Sleep Onset and Maintenance. Front Neurosci [Internet]. 2020 Jan 15 [cited 2024 Sep 17];13:499467. Available from: www.frontiersin.org

32. Almeida DLF, Soares FA, Carvalho JLA. A sliding window approach to detrended fluctuation analysis of heart rate variability. Annu Int Conf IEEE Eng Med Biol Soc [Internet]. 2013 [cited 2024 Sep 17];2013:3278–81. Available from: https://pubmed.ncbi.nlm.nih.gov/24110428/

33. (PDF) Algoritmos para la reducción de ruido en señales industriales mediante la Transformada Wavelet Discreta [Internet]. [cited 2024 Sep 17]. Available from: https://www.researchgate.net/publication/308174674_Algoritmos_para_la_reduccion_de_ruido_en_senales_industriales_mediante_la_Transformada_Wavelet_Discreta

34. Schaffer T, Hensel B, Weigand C, Schüttler J, Jeleazcov C. Evaluation of techniques for estimating the power spectral density of RR-intervals under paced respiration conditions. J Clin Monit Comput [Internet]. 2014 Sep 20 [cited 2024 Sep 17];28(5):481–6. Available from: https://pubmed.ncbi.nlm.nih.gov/23508826/

35. Kovács L, Jurkovich V, Bakony M, Szenci O, Póti P, Tå'Zsér J. Welfare implication of measuring heart rate and heart rate variability in dairy cattle: Literature review and conclusions for future research. Animal [Internet]. 2014 Feb [cited 2024 Sep 17];8(2):316–30. Available from: https://pubmed.ncbi.nlm.nih.gov/24308850/

36. PSD Computations Using Welch's Method. 1991.

37. Welch PD. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. IEEE Transactions on Audio and Electroacoustics. 1967;15(2):70–3.

38. Álvarez D, Hornero R, Marcos JV, Wessel N, Penzel T, Glos M, et al. ASSESSMENT OF FEATURE SELECTION AND CLASSIFICATION APPROACHES TO ENHANCE INFORMATION FROM OVERNIGHT OXIMETRY IN THE CONTEXT OF APNEA DIAGNOSIS. https://doi.org/101142/S0129065713500202 [Internet]. 2013 Aug 7 [cited 2024 Sep 17];23(5). Available from: https://www.worldscientific.com/worldscinet/ijns

39. Poza J, Hornero R, Abásolo D, Fernández A, García M. Extraction of spectral based measures from MEG background oscillations in Alzheimer's disease. Med Eng Phys. 2007 Dec 1;29(10):1073–83.

40. Hornero R, Kheirandish-Gozal L, Gutiérrez-Tobal GC, Philby MF, Alonso-Álvarez ML, Alvarez D, et al. Nocturnal Oximetry-based Evaluation of Habitually Snoring Children. Am J Respir Crit Care Med [Internet]. 2017 Dec 15 [cited 2024 Sep 17];196(12):1591–8. Available from: https://pubmed.ncbi.nlm.nih.gov/28759260/

41. Álvarez D, Hornero R, Víctor Marcos J, Delcampo F. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. IEEE Trans Biomed Eng. 2010 Dec;57(12):2816–24.

42. Aktaruzzaman M, Sassi R. Parametric estimation of sample entropy in heart rate variability analysis. Biomed Signal Process Control. 2014 Nov 1;14(1):141–7.

43. Byun S, Kim AY, Jang EH, Kim S, Choi KW, Yu HY, et al. Entropy analysis of heart rate variability and its application to recognize major depressive disorder: A pilot study. Technol Health Care [Internet]. 2019 [cited 2024 Sep 17];27(S1):S407–24. Available from: https://pubmed.ncbi.nlm.nih.gov/31045557/

44. Belfort REAU, Treccossi SPC, Silva JLF, Pillat VG, Freitas CBN, dos Santos L. Extended Central Tendency Measure and difference plot for heart rate variability analysis. Med Eng Phys. 2019 Dec 1;74:33–40.

45. Ferrario M, Signorini MG, Cerutti S. Complexity analysis of 24 hours heart rate variability time series. Conf Proc IEEE Eng Med Biol Soc [Internet]. 2004 [cited 2024 Sep 17];2004:3956–9. Available from: https://pubmed.ncbi.nlm.nih.gov/17271163/

46. Aboy M, Hornero R, Abásolo D, Álvarez D. Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis. IEEE Trans Biomed Eng. 2006 Nov;53(11):2282–8.

47. Hassan H, Ren Z, Zhao H, Huang S, Li D, Xiang S, et al. Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. Vol. 141, Computers in Biology and Medicine. Elsevier Ltd; 2022.

48. Vaquerizo-Villar F, Álvarez D, Kheirandish-Gozal L, Gutiérrez-Tobal GC, Barroso-García V, Crespo A, et al. Utility of bispectrum in the screening of pediatric sleep apnea-hypopnea syndrome using oximetry recordings. Comput Methods Programs Biomed. 2018 Mar 1;156:141–9.

49. Akhter N, Dabhade S, Bansod N, Kale K. Feature selection for heart rate variability based biometric recognition using genetic algorithm. In: Advances in Intelligent Systems and Computing. Springer Verlag; 2016. p. 91–101.

50. GuyonIsabelle, ElisseeffAndré. An introduction to variable and feature selection. The Journal of Machine Learning Research [Internet]. 2003 Mar 1 [cited 2024 Sep 17]; Available from: https://dl.acm.org/doi/10.5555/944919.944968

51. Jacak W, Pröll K, Winkler S. Neural Networks Based Feature Selection in Biological Data Analysis. 2014 [cited 2024 Sep 17];79–94. Available from: https://link.springer.com/chapter/10.1007/978-3-319-01436-4_5

52. Taherdangkoo M, Paziresh M, Yazdi M, Bagheri MH. An efficient algorithm for function optimization: Modified stem cells algorithm. Central European Journal of Engineering [Internet]. 2013 Mar 1 [cited 2024 Sep 17];3(1):36–50. Available from: https://www.degruyter.com/document/doi/10.2478/s13531-012-0047-8/html

53. Álvarez D, Hornero R, Marcos JV, del Campo F. Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis. Med Eng Phys. 2012 Oct 1;34(8):1049–57.

54. Kramer O. Genetic Algorithms. Studies in Computational Intelligence [Internet]. 2017 Jan 1 [cited 2024 Sep 17];679:11–9. Available from: https://link.springer.com/chapter/10.1007/978-3-319-52156-5_2

55. Altarabichi MG, Nowaczyk S, Pashami S, Mashhadi PS. Fast Genetic Algorithm for feature selection — A qualitative approximation approach. Expert Syst Appl. 2023 Jan 1;211:118528.

56. GuyonIsabelle, ElisseeffAndré. An introduction to variable and feature selection. The Journal of Machine Learning Research [Internet]. 2003 Mar 1 [cited 2024 Sep 17];3:1157–82. Available from: https://dl.acm.org/doi/10.5555/944919.944968

57. Rout S, Dwivedi V, Srinivasan B. Numerical Approximation in CFD Problems Using Physics Informed Machine Learning. 2021 Nov 1 [cited 2024 Sep 17]; Available from: https://arxiv.org/abs/2111.02987v1

58. Bishop C. Neural networks for pattern recognition. Choice Reviews Online [Internet]. 1995 Jun 1 [cited 2024 Sep 17];31(10):31-5500-31–5500. Available from: http://choicereviews.org/review/10.5860/CHOICE.31-5500

59. Multilayer Perceptrons in Machine Learning: A Comprehensive Guide | DataCamp [Internet]. [cited 2024 Sep 17]. Available from:

https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning?dc_referrer=https%3A%2F%2Fwww.google.com%2F

60. Agresti A. Categorical Data Analysis. 2002 Jul 3 [cited 2024 Sep 17]; Available from: https://onlinelibrary.wiley.com/doi/book/10.1002/0471249688

61. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009 Jul 1;45(4):427–37.

62. Negewo NA, Gibson PG, McDonald VM. COPD and its comorbidities: Impact, measurement and mechanisms. Respirology [Internet]. 2015 Nov 1 [cited 2024 Sep 17];20(8):1160–71. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/resp.12642

63. Tantucci C, Modina D. Lung function decline in COPD. Int J Chron Obstruct Pulmon Dis [Internet]. 2012 [cited 2024 Sep 17];7:95–9. Available from: https://pubmed.ncbi.nlm.nih.gov/22371650/