UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIAS EN INTELIGENCIA ARTIFICIAL

“Generación de componentes espectrales sintéticos del electrorretinograma mediante modelos generativos adversarios para predecir factores de riesgo en diabetes tipo 2”

Tesis que como parte de los requisitos para obtener el grado de la Maestría en Ciencias en Inteligencia Artificial

Presenta:
Omar Hernández de los Santos

Dirigido por:
Dra. Stéphanie Colette Thébault

Co-dirigido por:
Dr. Saúl Tovar Arriaga

Dra. Stéphanie Colette Thébault
Presidente

Dr. Saúl Tovar Arriaga
Secretario

Dr. Julio César Muñoz Benítez
Vocal

Dr. Andras Takacs
Suplente

Dr. Juan Manuel Ramos Arreguin
Suplente

Centro Universitario Querétaro, Qro.
Junio, 2025

Abstract

Type 2 Diabetes Mellitus (T2DM) represents a growing global health concern due to its high prevalence and associated complications. The timely identification of modifiable risk factors is essential for the prevention and control of this disease. Recent studies have reported the use of baseline electroretinogram (ERG) signals to predict obesity and metabolic syndrome, which are modifiable risk factors for T2DM. However, a larger, age-unbiased dataset is needed to ensure that the sensitivity and specificity of such predictions are clinically relevant. This study addressed the need to expand and balance a time-series dataset related to baseline retinal function by generating synthetic spectra from baseline ERG signals using Generative Adversarial Networks (GANs). A model named *spectraGAN* was successfully designed for the generation of synthetic spectra. The generated samples were evaluated both qualitatively and quantitatively, confirming their spectral similarity and physiological coherence with real data. Results demonstrated that integrating synthetic data into the control and at-risk categories for T2DM improved prediction performance using a neural network model based on the real Morlet wavelet (ROC-AUC of 0.68). This effect was further enhanced when the dataset was augmented by factors of 10 and 100, achieving average performance metrics ranging from 0.75 to 0.80 and from 0.76 to 0.89, respectively, in comparison with other models. This study highlights the relevance of *spectraGAN* in biomedical research and suggests it as a promising, non-invasive, and accessible alternative for the early detection of modifiable risk factors associated with T2DM.

(**Key words:** Generative adversarial networks, Synthetic spectra, Type 2 Diabetes Mellitus)

Resumen

La Diabetes Mellitus tipo 2 (DM tipo 2) representa una problemática creciente a nivel mundial debido a su alta prevalencia y complicaciones asociadas. La detección oportuna de factores de riesgo modificables es esencial para la prevención y control de esta enfermedad. Se reportó recientemente el uso de señales de electrorretinograma (ERG) basal para predecir la obesidad y el síndrome metabólico que son factores de riesgo modificables de la DM tipo 2, sin embargo, hace falta una base de datos más amplia y sin sesgo relacionado con la edad para que la sensibilidad y especificidad de la predicción sean relevantes para la clínica. Este trabajo abordó la necesidad de ampliar y equilibrar una base de datos de series de tiempo relativas a la función basal de la retina mediante la generación de espectros sintéticos provenientes del ERG basal, utilizando modelos generativos adversarios (GAN). Se diseño exitosamente un modelo llamado *espectraGAN* para la generación de espectros sintéticos. Se evaluaron cualitativa y cuantitativamente las muestras generadas, confirmando su similitud espectral y coherencia fisiológica con los datos orgánicos. Los resultados demostraron que la integración de datos sintéticos en nuestras categorías control y en riesgo de DM tipo 2 mejoró la predicción mediante el modelo basado en redes neuronales con la ondícula Morlet real (ROC-AUC de 0,68). Este efecto mejoró aún más cuando la base de datos se amplió 10 y 100 veces, con métricas en comparativa con otro modelos con un de rendimiento medio entre 0,75 y 0,80 y entre 0,76 y 0,89, respectivamente. Este estudio demuestra la relevancia de *espectraGAN* en la investigación biomédica y sugiere una alternativa prometedora, no invasiva y accesible para detectar tempranamente factores de riesgo modificables asociados con la DM tipo 2.

(**Palabras clave:** Modelos generativos adversarios, Espectros sintéticos, Diabetes Mellitus tipo 2)

Agradecimientos

Quiero expresar mi profundo agradecimiento a todas las personas e instituciones que formaron parte en toda mi trayectoria para alcanzar este logro; sin ellos, esto no habría sido posible:

Principalmente a las tres personas más importantes en mi vida mi mamá, mi hermana y mi hijo. Gracias, mamá por ser siempre mi faro en las peores tormentas, tú amor incondicional y tu apoyo constante me han sostenido en mis peores caídas. Gracias por enseñarme que la zona de confort no es un buen lugar para estar, y por impulsarme a perseguir mis sueños y objetivos con valentía. A mi hermana, gracias por cada consejo, por cada palabra de aliento, por estar siempre presente en cada aventura que decido emprender sin importar la distancia. Eres mi cómplice, mi soporte y mi amiga incondicional. Y a ti hijo, mi motor y mi impulso. Desde el primer instante que te sostuve en mi brazos y abriste esos ojitos para conocerme, disté un nuevo sentido a mi vida. Eres mi razón, mi fuerza y mi inspiración diaria.

A mi directora y co-director de tesis, Stéphanie y Julio, gracias por acompañarme en este largo camino para obtener los objetivos planteados. Agradezco profundamente su paciencia, su sabiduría y cada uno de sus valiosos consejos, que han sido fundamentales en mi formación y desarrollo profesional.

A mis amigos y compañeros de la maestría que formaron parte de esta increíble experiencia, gracias por compartir cada uno sus conocimientos y experiencias especialmente a Sheila, Luis, Oscar, Felipe, Roberto y Aldo.

A la Universidad Autónoma de Querétaro, así como al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONACYT) por la beca de investigación que financió este proyecto.

Al instituto de neurobiología de la Universidad Nacional Autónoma de México por la colaboración, trabajo en equipo y recursos para poder llevar a cabo los experimentos de este proyecto.

Índice general

Índice de figuras

Índice de tablas

Nomenclaturas

Abreviaciones

DM	Diabetes Mellitus
DM tipo2	Diabetes Mellitus Tipo 2
SM	Síndrome Metabólico
HbA1c	Niveles séricos de glucosa y hemoglobina A1c
LDL	Lipoproteínas de baja densidad
HDL	Lipoproteínas de alta densidad
ML	Aprendizaje automático
DL	Aprendizaje profundo
IMC	Índice de Masa Corporal
ERG	Electrorretinografía
EEG	Electroencefalograma
EMG	Electromiograma
PPG	Fotopletismograma
GAN	Redes Generativos Adversarias
RF	Bosques aleatorios
BRB	barrera hematorretiniana
LG	Regresion logística
GRS	Puntuación de Riesgo Genético
PCA	Análisis de Componentes Principales
SMOTE	Técnica de Submuestreo de Minorías Sintéticas

LSTM	Red de memoria a corto-largo plazo
SVM	Maquina de Soporte Vectorial
CNN	Redes Neuronales Convolucionales
RNN	Redes Neuronales Recurrentes
ANN	Red Neuronal Artificial
ACC	Exactitud
MAE	Error Absoluto Medio
MSE	Error Cuadrático Medio
FID	Frechet Inception Distance
PRD	Porcentaje de Diferencia Media Cuadrática
RMSE	Raíz del Error Cuadrático Medio
DWT	Wavelet Discreta
t-SNE	t-Distributed Stochastic Neighbor Embedding

Capítulo 1

Introducción

La DM sigue presentando un crecimiento significativo a nivel global. En 2019, se estimaba que 463 millones de personas vivían con esta enfermedad, y se proyecta que esta cifra podría alcanzar los 589 millones en 2030 y llegar a 750 millones para el año 2045 [1]. De acuerdo con la Encuesta Nacional de Salud y Nutrición en México, se encontró que el número de adultos con diagnóstico médico de DM entre 20 y 79 años es de 14.1 millones (10.3 % de la población), y se estima que para 2045 incremente a 21.2 millones [2].

La DM es una enfermedad sistémica crónica caracterizada por niveles elevados de glucosa en la sangre, asociadas a múltiples complicaciones vasculares, tanto microvasculares como macrovasculares [3]. La DM tipo 2 representa aproximadamente el 90 % de los casos y se caracteriza por hiperglucemia, resistencia insulínica y deterioro en la secreción de insulina [4]. Los factores de riesgo asociados a la DM tipo 2 incluyen hábitos del estilo de vida (como la inactividad física, alimentación poco saludable y la obesidad abdominal), factores demográficos y antecedentes (como la edad avanzada y los antecedentes familiares de diabetes), y factores clínicos y de salud (como la prediabetes y la hipertensión) [5]. El conjunto formado por proaterogenia, dislipidemia, presión arterial elevada, niveles elevados de azúcar en la sangre y niveles abdominales de obesidad, se conoce como síndrome metabólico (SM), y es el principal factor de riesgo para la DM tipo 2 [6].

Tradicionalmente, la diabetes se diagnostica mediante variables obtenidas por pruebas invasivas de laboratorio en ayunas, como HbA1c y los niveles séricos de glucosa, mientras que el SM se diagnostica midiendo elevados niveles séricos de triglicéridos y colesterol, y bajos niveles de HDL, entre otros. Por esta razón, surge la necesidad creciente de desarrollar metodologías no invasivas que permitan determinar los factores de riesgo modificables de la DM tipo 2 [6] - [10].

En este contexto, el ML y el DL, han emergido como herramientas poderosas para predecir, clasificar, identificar patrones, analizar y evaluar la importancia relativa de los diferentes factores de riesgo en la DM tipo 2 [6] - [10]. La gran mayoría de estos modelos se centran en atributos de registros históricos que incluyen el IMC, porcentaje de masa corporal, tipo de diabetes, genero, historial familiar, presión arterial, condiciones sociodemográficas y pruebas de laboratorio [6] - [10]. Recientes avances en tecnologías diagnósticas, especialmente en imagenología ocular, han introducido a las imágenes de retina como una fuente de información relevante para predecir de manera no invasiva el riesgo cardiovascular, estrechamente relacionado con la obesidad [11], [12]. Asimismo, la oftalmoscopia indirecta, fotografía de fondo de ojo, tomografía de coherencia óptica, angiografía por tomografía de coherencia óptica o con fluoresceína/Angiografía con indocianina y la ecografía ocular han demostrado ser efectivas para evaluar indirectamente factores de riesgo asociados con la obesidad y sus complicaciones asociadas, ofreciendo ventajas significativas en términos de precisión diagnóstica en comparación con métodos clínicos tradicionales o pruebas séricas invasivas [12], [13]. En particular, la implementación de algoritmos de aprendizaje supervisado en la imagenología ocular ha revolucionado la capacidad para analizar e interpretar automáticamente pruebas diagnósticas, ayudando a solventar la escasez de personal calificado para dichas tareas [11]. Sin embargo, limitaciones importantes a todos estos acercamientos incluyen la privacidad de los datos (los fondos de ojo son únicos a cada individuo), costo de los estudios, disponibilidad de los equipos, recolección invasiva y/o incómoda de las variables séricas ya que requiere de ayuno, largos tiempos de adquisición, lo que dificulta su implementación rutinaria en contextos clínicos, notablemente en países como México [11], [14].

Una alternativa innovadora para resolver esta problemática es el ERG, una técnica no invasiva que mide los cambios de potencial eléctrico en las células de la retina en respuesta a un estímulo fótico, sus parámetros cuantitativos y objetivos sirven para el diagnóstico de muchas enfermedades oculares [15] y ha demostrado potencial para evaluar condiciones como la obesidad [16]. Yapici y col. [17] aprovecharon las transformadas de onda continua del ERG para procesar respuestas de los bastones y conos de la retina mediante redes neuronales artificiales optimizadas con enjambre de partículas, creando un modelo predictivo eficaz del grado de obesidad y demostrando alteraciones significativas en estas señales en

individuos con obesidad. Sin embargo, estas características aún parecen carecer de la practicidad necesaria para ser adoptadas en la práctica clínica ya que el protocolo de registro ERG para obtener la respuesta de los conos, es tedioso, por ejemplo, periodos de espera de al menos 35 minutos para la adaptación a la obscuridad y a la luz, respectivamente, además del tiempo de registro [16].

Nuestro grupo de trabajo demostró que el ERG basal, que registra la actividad eléctrica en condiciones de estimulación luminosa constante y sólo tarda unos cinco minutos, puede ser analizado mediante transformadas de tipo ondúlelas (o Wavelet en inglés) [18]. Este análisis permite obtener el espectro de frecuencia y la potencia relativa de las oscilaciones que conforman el ERG [19]. Este enfoque se utiliza con el objetivo de entrenar un modelo de diagnóstico predictivo para factores de riesgo de la DM tipo 2, como el sobrepeso, obesidad y SM [18]

Por lo que se ha introducido el uso del ERG basal, una modalidad más practica y rápida que mide la actividad eléctrica basal de la retina sin necesidad de un flash de luz [20]. Debido a que se modifica en ciertas condiciones patológicas [20], [21], se ha desarrollado un modelo de diagnóstico predictivo basado en bosques aleatorios, el cual fue entrenado utilizando componentes espectrales de señales de ERG no evocado [22], para detectar los factores de riesgos modificables asociados con la DM tipo 2, como el sobrepeso, la obesidad y el SM [20]. Sin embargo, su reciente incorporación implica escasez de datos disponibles, creando desafíos como el desbalance de clases y sesgos potenciales en el entrenamiento de modelos predictivos [23]. Estos inconvenientes son particularmente relevantes debido a que la prevalencia de la DM tipo 2 varía significativamente según los grupos etarios. Por ejemplo, en México, la prevalencia es mucho mayor en adultos de 60 a 69 años (25.8 %) que en jóvenes de 20 a 29 años (0.6 %) [2], lo que podría afectar negativamente la precisión de los modelos de aprendizaje supervisado.

Las técnicas de generación sintética de datos, particularmente los GANs, han surgido como una solución efectiva para abordar estos dos desafíos [24]. La mayoría de los estudios en este campo se enfocan en imágenes [25], muy poco abordan la generación de señales biomédicas [26] - [31], y ninguno específicamente en el área del ERG. Esta tesis tiene como objetivo generar componen espectrales sintéticos provenientes del ERG basal mediante

GANs para evaluar su utilidad en la predicción de factores de riesgo modificables asociados a la DM tipo 2.

1.1 Descripción del problema

La atención especializada para la DM tipo 2 en México enfrenta un reto considerable debido a la insuficiente cantidad de endocrinólogos, que alcanza 0.8 especialistas por cada 100,000 habitantes, cifra muy por debajo del promedio internacional [32]. Esta escasez se ve agravada debido a la concentración en zonas urbanas, lo que limita significativamente el acceso adecuado a atención médica especializada para gran parte de la población, especialmente en zonas rurales [33]. En consecuencia, se producen diagnósticos tardíos, manejo subóptimo de la enfermedad y retrasos significativos en la introducción de tratamientos, lo que resulta en mayores complicaciones y un control ineficaz de la DM tipo 2 [34].

Actualmente, el diagnóstico de la DM tipo 2 se basa en varias pruebas estándar que miden los niveles de glucosa en sangre. Las principales incluyen la prueba de glucosa en ayunas, la prueba de tolerancia a la glucosa oral y la prueba de HbA1c [6] - [10]. Estas pruebas requieren acceso de laboratorios clínicos y personal capacitado para su correcta interpretación. Sin embargo, en regiones remotas o con recursos limitados, la disponibilidad de estos servicios puede ser escasa, dificultando la detección temprana de la enfermedad [34]. Además, la DM tipo 2 puede desarrollarse de manera| asintomática, es decir silenciosamente para el enfermo, durante años, lo que retrasa su diagnóstico hasta que aparecen complicaciones significativas [35]. Por lo tanto, es esencial desarrollar métodos de detección fiables, más accesibles y prácticos que permitan identificar la DM tipo 2 en etapas iniciales, especialmente en comunidades con acceso limitado a servicios de salud.

En este contexto, hallazgos recientes de nuestro grupo demuestran que, mediante señales de ERG, es posible predecir factores de riesgo para el desarrollo de DM tipo 2. Aunque dicha enfermedad es reversible, esto implica esfuerzos titánicos constantes y un seguimiento asimismo constante [36]. En contraste, el sobrepeso y la obesidad pueden ser revertidos más rápidamente que la DM tipo 2, pero también requieren modificaciones

profundas y sostenidas en el estilo de vida, que muchas veces resultan difíciles de mantener en el tiempo [37]. Las señales de ERG basal tienen un valor predictivo robusto, alcanzando un área bajo la curva ROC-AUC de 0.926 [21], [22]. Además, el desarrollo tecnológico actual permite registrar estas señales de forma no invasiva, con equipos portátiles (tamaño de un joystick) que son tres veces más económicos que los microscopios para Fundus (350,000 mxn versus 1,000,000 mxn para los equipos de ERG de mesa) [38].

No obstante, existe un desafío significativo relacionado con la escasez y el desbalance en los datos disponibles para desarrollar modelos predictivos robustos [39]. Este desbalance está influenciado por la distribución desigual de la prevalencia de la DM tipo 2 en diferentes grupos etarios, afectando negativamente el desempeño de los modelos predictivos [21]. Por esta razón, se vuelve necesario implementar técnicas avanzadas como los modelos GANs, específicamente entrenados con espectros provenientes de señales de ERG basal, para generar datos sintéticos que permitan equilibrar las bases de datos existentes y mejorar la capacidad de predicción de factores de riesgo modificables en la DM tipo 2.

1.2 Justificación

La DM tipo 2 suele desarrollarse silenciosamente, un gran porcentaje de los casos ya presenta complicaciones severas al momento del diagnóstico debido a la ausencia de síntomas tempranos [40]. Esta situación resalta la importancia crítica de identificar oportunamente los factores de riesgo modificables para esa enfermedad a fin de implementar medidas preventivas que reduzcan su incidencia y desarrollo, generando así un impacto positivo a la salud pública y la economía [41], [42].

La limitada infraestructura hospitalaria, la escasez de equipos especializados y la baja inversión en salud pública, que representa uno de los porcentajes más bajos del Producto Interno Bruto, acompañan/se añaden a un déficit considerable de endocrinólogos en México [43]. Por lo tanto, el diagnóstico asistido por algoritmos de aprendizaje supervisado emerge como una solución prometedora que podría mejorar significativamente la eficiencia y calidad del diagnóstico temprano de la DM tipo 2, especialmente en zonas rurales con acceso limitado a especialistas [41].

Los métodos diagnósticos tradicionales para evaluar factores de riesgo como obesidad, alteraciones lipídicas y resistencia a la insulina suelen ser invasivos y laboriosos [44], [45]. Por ello, métodos no invasivos y cuantitativos como el ERG, especialmente en su modalidad basal que mide la actividad eléctrica de la retina sin necesidad de flashes luminosos, ofrecen alternativas atractivas para predecir el riesgo metabólico y cardiovascular [46], [47]. Nuestro grupo de trabajo demostró que cinco minutos de ERG basal, analizados mediante transformadas de tipo onduletas (o Wavelet en inglés) que permiten obtener el espectro de frecuencia y la potencia relativa de las oscilaciones que conforman el registro [48], pueden ser utilizadas para entrenar un modelo que predice factores de riesgo de la DM tipo 2, como el sobrepeso, obesidad y SM con un rendimiento prometedor [22]. Sin embargo, el conjunto de datos (más de 700 participantes) es sesgado: la edad de los participantes metabólicamente sanos es mucho menor a la de los enfermos (**Figura 1.1**).

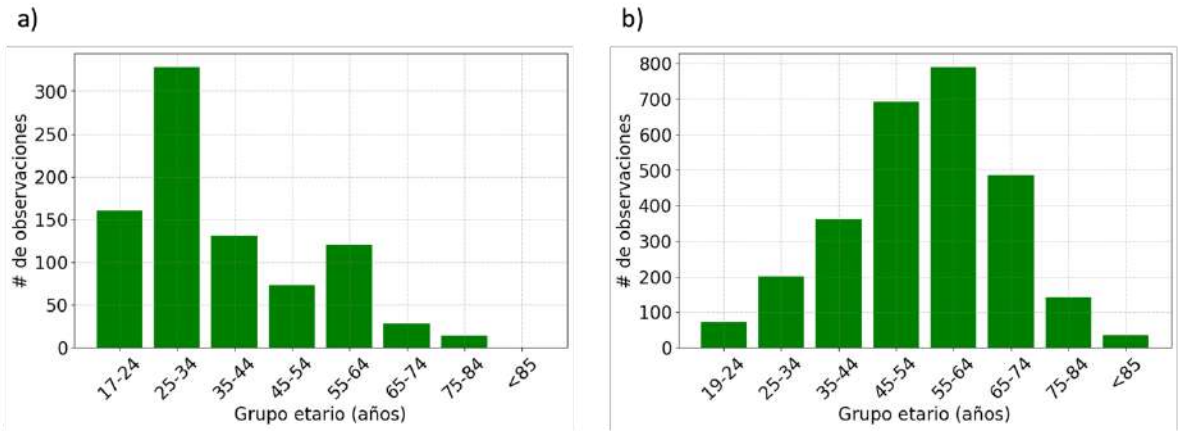


Figura 1.1: *Distribución de los participantes por grupo etarios.*

Se confirmó que este sesgo perjudica a nuestro modelo, ya que, con sólo datos de edad y sexo, es capaz de predecir casos enfermos:

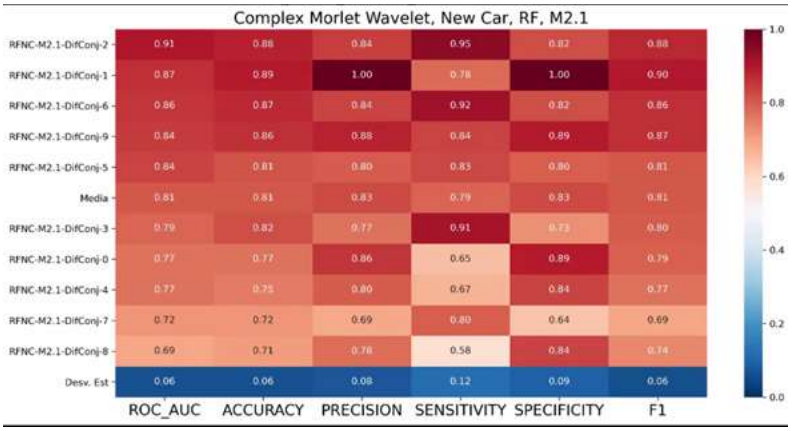


Figura 1.2: *Resultados de predicción de casos enfermos vs. sanos alimentando el algoritmo RF con género y edad de los casos; cada fila corresponde a una semilla diferente con el rendimiento promedio en la sexta fila y la desviación estándar (Desv. Est) de las métricas en la última fila.*

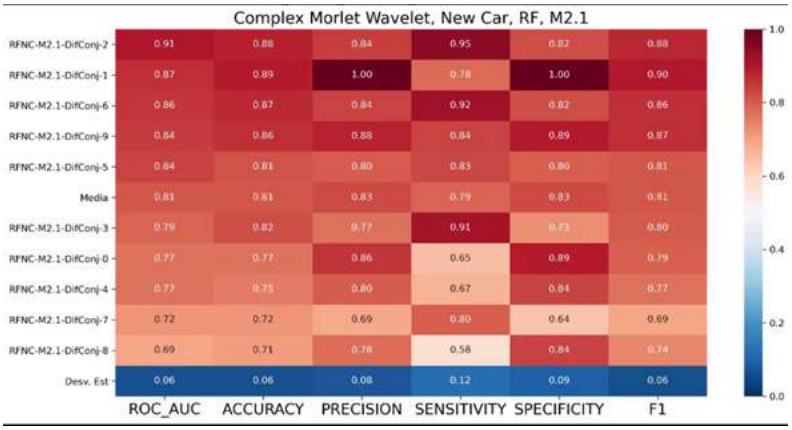


Figura 1.3: *Resultados de predicción de casos enfermos vs. sanos alimentando el RF con género y series de tiempo correspondiente al ERG basal de los casos; cada fila corresponde a una semilla diferente con el rendimiento promedio en la sexta fila y la desviación estándar (Desv. Est) de las métricas en la última fila.*

Es para solventar este problema de desbalance que afecta nuestro modelo en su entrenamiento y pone en tela de juicio los resultados que se propone el desarrollo de un conjunto de datos artificiales a semejanza de los componentes espectrales de los ERG no evocados mediante un modelo GAN. También servirán para ampliar la base de datos ya que, para los algoritmos de aprendizaje, cuanto mayor sea la cantidad de datos disponibles para su entrenamiento y validación, se espera que su desempeño sea mejor.

1.3 Hipótesis

El diseño e implementación de un GAN para generar una base de datos de componentes espectrales del ERG no evocado permite mejorar el desempeño de un modelo de diagnóstico predictivo para factores de riesgo de la DM tipo 2.

1.4 Objetivo General

Diseñar un catálogo de muestras sintéticas provenientes de casos humanos mediante un modelo generativo adversario para mejorar el desempeño de un modelo de diagnóstico predictivo para factores de riesgo de la DM tipo 2.

1.5 Objetivos específicos

- Identificar algoritmos utilizados para generación de componentes espectrales mediante GAN.
- Diseñar y estabilizar un modelo para generar componentes espectrales de tipo electrorretinograma basal con GAN provenientes de humanos.
- Evaluar cuantitativa y cualitativamente la calidad de las muestras de los componentes espectrales sintéticas producidas por los modelos GAN.
- Disponer de un conjunto de datos para mejorar el entrenamiento, prueba y validación de un modelo existente para detectar factores de riesgo modificables para la DM tipo 2.

Capítulo 2

Marco teórico

2.1 Estructura del ojo humano

El ojo humano mide entre 22 y 27 mm de diámetro anteroposterior y entre 69 y 85 mm de circunferencia. La estructura del ojo humano, como se muestra en la **Figura 2.1**, consta de tres capas principales: (1) la externa (córnea y esclera); (2) la intermedia (iris, cuerpo ciliar y coroides) y (3) la interna o retina. La pupila, regulada por el iris, ajusta la cantidad de la luz que ingresa al ojo, mientras que el cristalino enfoca esta luz hacia la mácula [49]. La coroides aporta nutrientes mediante una red vascular, y el vítreo, estructura gelatinosa transparente detrás del cristalino, mantiene la forma interna del ojo [50].

La retina, capa neural sensible a la luz, contiene células neuronales (conos para visión a color y nítida, bastones para visión en condiciones de poca luz), células gliales (astrocitos, microglía y células de Müller) y capilares intrarretinianos. El reflejo de la luz por la materia estimula los fotorreceptores (conos y bastones), generando señales bioeléctricas que se transmiten a través de las células bipolares, horizontales, amacrinas y ganglionares, hasta las fibras del nervio óptico y la corteza cerebral [15], [51]. Por último, podemos mencionar a los pericitos, células contráctiles que se encuentran alrededor de las células endoteliales de los capilares intrarretinianos y participan en BRB.

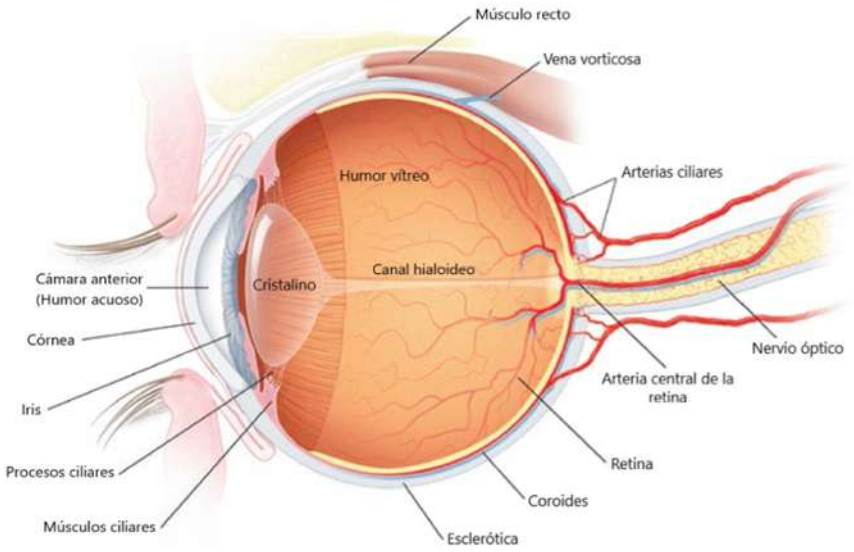


Figura 2.1: *Estructura del ojo humano [15].*

En la retina, se procesa mucho más que sólo detectar el reflejo de la luz por la materia. Se codifica su color, contraste, desplazamiento, dirección y velocidad de la información visual, entre otros. Estructuralmente está formada por diez capas (**Figura 2.2**): (1) epitelio pigmentario, que absorbe luz y recicla fotopigmentos; (2) fotorreceptores (conos y bastones), encargados de la fototransducción; (3) membrana limitante externa, barrera interna; (4) capa nuclear externa, núcleos y bastones; (5) capa plexiforme externa, conexiones entre fotorreceptores, células bipolares y horizontales; (6) capa nuclear interna, cuerpos celulares con núcleos de las neuronas retinianas de 2do orden e interneuronas (células horizontales y amacrinas); (7) capa plexiforme interna, conexiones sinápticas entre células bipolares, amacrinas y ganglionares; (8) capa de células ganglionares, cuyos axones forman la (9) capa de fibras nerviosas, que convergen para formar el nervio óptico y (10) membrana limitante interna, cuyas funciones incluyen el desarrollo de la capa de fibras nerviosas en la embriogénesis, la barrera selectiva de nutrientes, factor de crecimiento y proyecciones de rayos UV [51], [52].

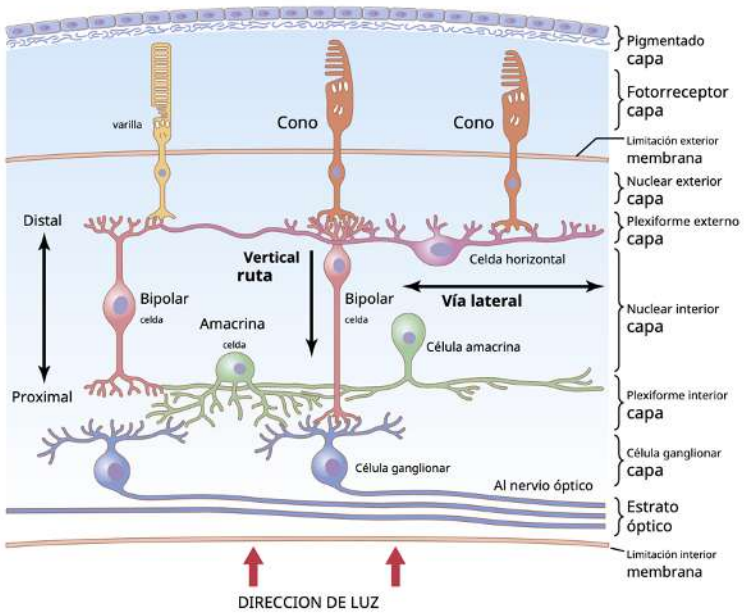


Figura 2.2: *Componentes estructurales de la retina [51].*

Existen dos tipos de fotorreceptores: los conos (visión diurna, alta resolución espacial, percepción de color mediante conopsina) y los bastones (visión nocturna o escotópica, rodopsina, sensibles a poca luz), cuya activación por la información luminosa les hace liberar glutamato, transmitiendo información hacia las células ganglionares a través de las células bipolares. La retina ofrece una serie de superposiciones visuales compuestas por imágenes de contraste positivo y negativo, es decir con refuerzo de contraste, forma, dirección, movimiento y color. En resumen, la retina codifica las señales luminosas y la corteza visual las decodifica para que nos hagan sentido [50].

Mediante una lámpara de hendidura o algún dispositivo para la exploración de las estructuras de la parte posterior del ojo, como se muestra en la **Figura 2.3**, se puede observar la retina en el fondo del ojo, y más particularmente la mácula y la fóvea, que son las porciones de la retina que contienen pigmento xantófilo y que son responsable de la visión central, detallada y aguda; la fóvea mide unos 1.5 mm de diámetro. De la pupila parten los principales vasos sanguíneos que se encargan de nutrir, junto con los vasos sanguíneos de la coroides, la retina neurosensorial.

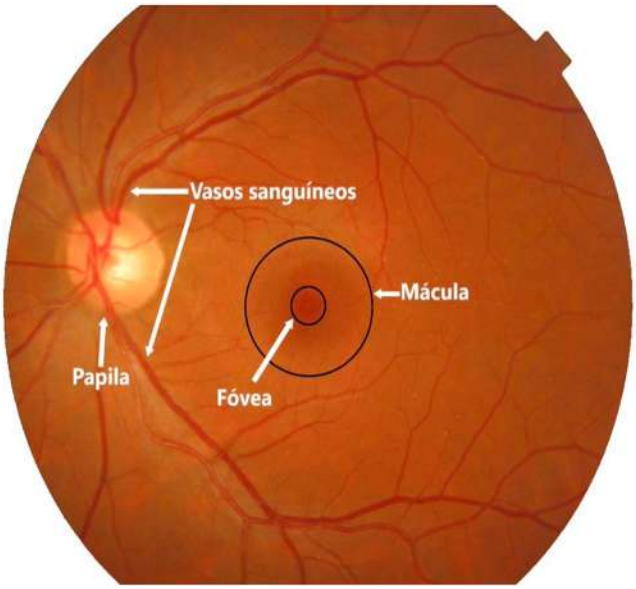


Figura 2.3: *Imagen de ojo en la que se muestra la papila, vasos sanguíneos, la fóvea y la mácula [51].*

En condiciones fisiológicas, diferentes mecanismos mantienen a la retina en un estado transparente relativamente deshidratado, lo que es esencial para su función visual adecuada [53]. La BRB desempeña un papel crucial en este proceso, controlando directamente la entrada de líquidos y moléculas desde la circulación sanguínea sistémica hacia la retina [54]. El componente interno de BRB está constituido por las uniones estrechas entre las células endoteliales de los vasos retinianos, apoyadas por los pericitos, pies astrocíticos y células de Müller. Estas células no solo cubren parcialmente las células endoteliales, sino que también interactúan activamente con otras células retinianas para mantener la homeostasis y función retinal [52].

Durante el desarrollo vascular retinal, la formación de estas uniones estrechas depende fundamentalmente de las interacciones entre las células endoteliales y otras células como los pericitos y astrocitos, destacando la relevancia crítica de la unidad neurovascular (UNV) en las etapas iniciales de la formación de la BRB interna [54]. El término "unidad neurovascular" hace referencia al acoplamiento funcional e interdependencia entre neuronas, células gliales y vasculatura especializada del sistema nervioso central, incluyendo la retina [52]. En este contexto, las células de la UNV mantienen una comunicación estrecha que asegura la integridad de la BRB y regulan dinámicamente el flujo sanguíneo para satisfacer las demandas metabólicas cambiantes del tejido retiniano [52].

En situaciones patológicas como la DM tipo 2, la integridad de la BRB es comprometida debido al aumento crónico de glucosa en sangre, niveles elevados de ácidos grasos libres presentes en condiciones de obesidad y a la presencia de citocinas proinflamatorias [54], entre otros.

2.2 Diabetes Mellitus

La DM abarca un conjunto diverso de trastornos que tienen en común niveles de glucosa en la sangre, también conocido como hiperglucemia, junto con otras alteraciones metabólicas. Es una condición seria y crónica que ocurre cuando el cuerpo no es capaz de producir insulina, la genera en cantidades insuficientes o no la utiliza de forma adecuada. La insulina es una hormona fabricada por el páncreas que permite la entrada de glucosa en las células para ser usada como fuente de energía. Además, esta hormona desempeña un papel clave en el metabolismo de grasas y proteínas. Si hay un déficit de insulina sin control, varios órganos del cuerpo pueden verse afectados, originando problemas de salud como enfermedades cardiovasculares, daño en los nervios (neuropatías), enfermedades renales (nefropatía) y complicaciones oculares como la retinopatía [51].

Clasificación de los tipos de DM:

- DM tipo 1

  Esta forma de diabetes se origina por una respuesta autoinmune en la cual el sistema inmune ataca por error a las células beta (β) del páncreas, encargadas de producir insulina. Como resultado, estas células dejan de funcionar o producen una cantidad insuficiente de insulina, lo que lleva un aumento en los niveles de glucosa en sangre [55]:

  - DM tipo 1A o autoinmune: Hay una destrucción específica de las células (β) pancreáticas mediada por linfocitos [51].
  - DM tipo 1 B o idiopática: es una forma menos comprendida con causas, evolución y pronósticos poco definidos. Se ha observado que en estos casos hay una disminución en la producción de insulina desde el inicio, con tendencia a desarrollar cetosis o cetoacidosis [51].

- DM tipo 2

  En esta variante, se observa una hiperglucemia acompañada por una producción inadecuada de insulina y resistencia a la misma. Como consecuencia, la glucosa no puede ingresar de manera eficiente a las células y se acumula en la sangre. El páncreas intenta compensar aumentando la producción de insulina, pero esto puede agravar el problema de incrementar aún más los niveles de azúcar. A lo largo del tiempo, la capacidad del cuerpo para secretar insulina disminuye, especialmente ante demandas prolongadas, haciendo que algunos pacientes con DM tipo 2 lleguen a requerir tratamiento con insulina. Diversos estudios han evidenciado que el riesgo de desarrollar esta condición es mayor en personas con antecedentes familiares, estilo de vida sedentario, sobrepeso u obesidad [37].

A nivel mundial, la DM se considera una epidemia. Actualmente, se estima que hay 537 millones de adultos entre 20 y 79 años que padecen esta enfermedad, lo que representa un 10.5 % de la población mundial en este grupo de edad. Se proyecta que esta cifra aumente a 643 millones (11.3 %) para 2030 y 783millones (12.2 %) para el 2045. En la **Figura 2.4**, presenta un mapa con la distribución estimada de casos de DM en adultos (de entre 20 a 79 años) para el año 2021.

En México, se calcula que adultos entre 20 y 79 años con DM es de 14.1 millones, y se prevé que esta cifra se eleve a 21.2 millones para el año 2045. Este incremento de casos se atribuye a factores como el envejecimiento de la población, al desarrollo económico, la urbanización, y el cambio en el estilo de vida hacia hábitos menos saludables, con menor actividad física y mayor consumo de alimentos relacionados con el desarrollo de obesidad [1].

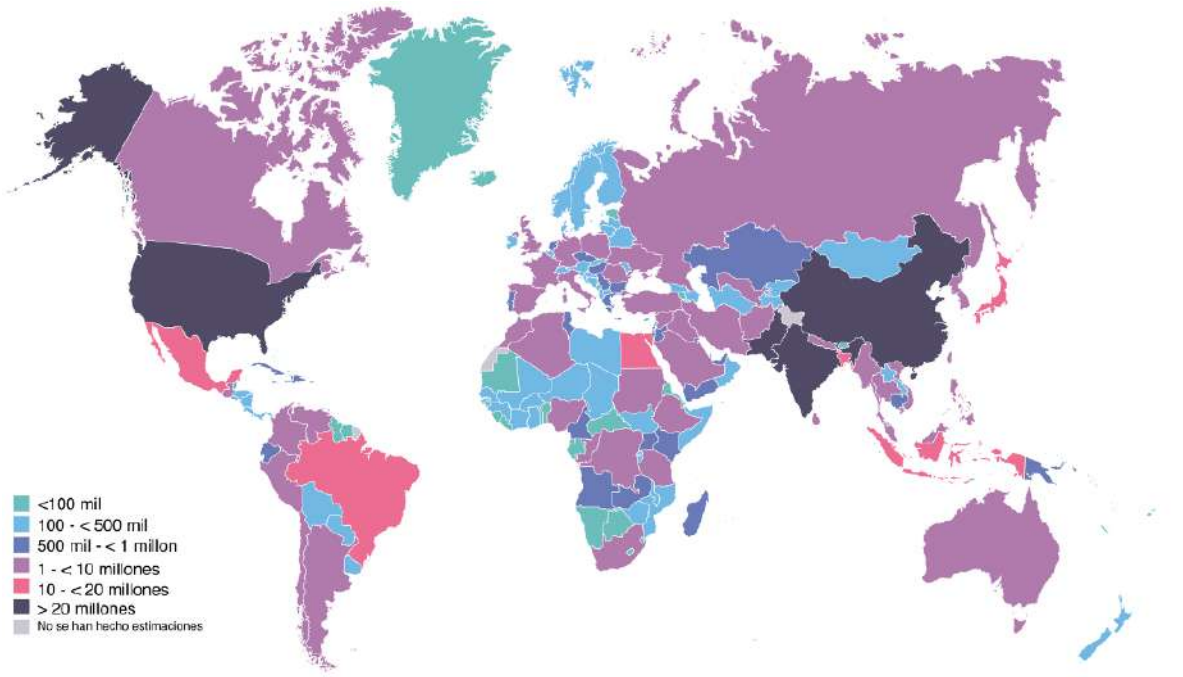


Figura 2.4: *Distribución de la cantidad de adultos (entre 20 y 79 años) con diabetes calculada para el 2021 a nivel mundial [1].*

2.3 Electrorretinograma

El ERG se concibe clásicamente como el registro de la actividad eléctrica de la retina en respuesta a un estímulo luminoso. La retina, es estimulada con breves destellos de luz, controlados por un fotoestimulador, responde con cambios de potencial entre el electrodo de registro (colocado en la superficie de la córnea, conjuntiva o piel del párpado inferior) y el electrodo de referencia colocado en otra parte del cuerpo (usualmente en la sien, la frente, o el lóbulo de la oreja) o no (integrado en electrodo de registro) que tienen una secuencia temporal propia.

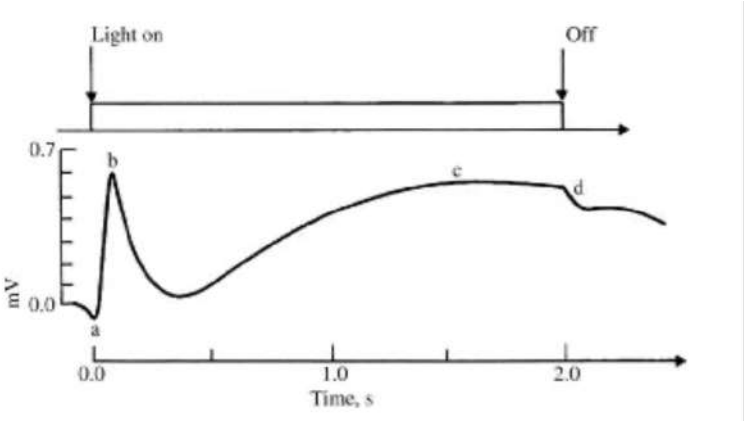


Figura 2.5: *Trazo de ERG (mV,s) en respuesta de un flash de luz prolongado (2 s) en el humano [56] .*

En la **Figura 2.5** se muestra la respuesta típica del ERG en respuesta a un estímulo de luz blanca de 2 s. Las cuatro componentes de esta respuesta son las ondas a, b, c y d; son comunes en la mayoría de los vertebrados, incluyendo los humanos. La onda a es negativa, se llama también potencial receptor temprano, refleja la suma de las hiperpolarizaciones que ocurren en los fotorreceptores en respuesta a la isomerización de los fotopigmentos por la luz reflejada por la materia. El segundo componente (onda b) sigue a la onda a con una latencia de 1 a 5 ms y es conocido como el potencial receptor tardío. Es generado por la actividad de las células bipolares y ganglionares de la retina. La onda c es generada por el epitelio pigmentario y la onda d es predominantemente el resultado de la respuesta de despolarización de la célula bipolar OFF a la terminación del estímulo luminoso. De acuerdo con las condiciones en las que se efectúe el registro del ERG, es posible separar la respuesta de los conos de la respuesta de los bastones. Usualmente, se considera la respuesta promedio a varias estímulos luminosos [50], [56]:

Se han creado varios protocolos de estimulación luminosa en condiciones de luz ambiental distinta para separar la respuesta de las neuronas de la retina como el:

- ERG fotópico: Mide la respuesta de los conos.
- ERG escotópico: Mide la respuesta de los bastones.
- ERG mesópico: Mide la respuesta sumatoria de los conos y bastones.

En recientes estudios se ha introducido a un protocolo de ERG basal, y se encontrado que las respuestas del ERG basal pueden verse afectadas por la obesidad u otras

complicaciones visuales relacionadas con la DM [57], [58]. Para utilizarlas en un diagnóstico, estas señales, que son complejas y no estacionarias, se preprocesan mediante un proceso de transformación matemática que permite pasar del dominio del tiempo al dominio de las frecuencias. Notablemente, se utilizan transformadas discretas de tipo wavelet (DWT por sus siglas en inglés), que descomponen las señales en componentes de frecuencia específicas. A partir de estas descomposiciones, se extraen características estadísticas claves para detectar cambios patológicos. Por ejemplo, las variaciones en la amplitud de los coeficientes obtenidos mediante la descomposición wavelet pueden ser indicativas de daño en estructuras retinianas. Estas características pueden alimentar modelos de aprendizaje automático o sistemas de clasificación para detectar automáticamente patrones específicos asociados con la DM.

2.4 Redes Generativas adversarias

Uno de los principales desafíos del DL es la predicción de alguna condición, lo cual implica el uso de un modelo predictivo. Esto requiere de un conjunto de datos de entrenamiento que se utiliza para entrenar un modelo, compuesto por múltiples ejemplos, llamados muestras, cada uno con una variable de entrada X y etiquetas de clase de salida Y. El modelo se entrena mostrando muestras de entradas, haciendo que prediga la salida y corrigiendo el modelo para que las salidas se parezcan más a la esperada. Se conoce esto como aprendizaje supervisado, porque hay un resultado real esperado con el que se compara la predicción. Algunos modelos de aprendizaje supervisado incluyen clasificación y regresión, como son los LG y RF. Existe otro paradigma de aprendizaje en el que el modelo sólo recibe las variables de entrada X y el problema no tiene variables de salida Y. El modelo se construye extrayendo o resumiendo los patrones de los datos de entrada, sin corrección del modelo, ya que esté no predice nada. Esta falta de corrección suele denominarse aprendizaje no supervisado. Algunos problemas de aprendizaje no supervisado incluyen la agrupación y el modelo generativo como por ejemplo los *K-means* y las GANs [59].

Con el aprendizaje supervisado, se puede desarrollar un modelo para predecir una etiqueta de clase dado un ejemplo de variables de entrada. Esta tarea de modelo predictivo se denomina clasificación. Tradicionalmente, la clasificación también se denomina modelo discriminativo. Esto se debe a que, un modelo debe discriminar ejemplos de variables de

entrada entre clases y debe elegir o tomar una decisión sobre la clase a la que pertenece una muestra dada. Alternativamente, los modelos no supervisados que resumen la distribución de las variables de entrada pueden utilizarse para crear o generar nuevos ejemplos con la distribución de entrada. Como tal, estos tipos de modelos se denominan modelos generativos. Por ejemplo, una sola variable puede tener una distribución de datos conocida, como una distribución gaussiana. Un modelo generativo puede ser capaz de resumir suficientemente esta distribución de datos y utilizarla para generar nuevas muestras que se ajusten de forma plausible a la distribución de las variables de entrada [60].

2.4.1 El modelo generador

El modelo generador parte de un vector aleatorio de longitud fija como entrada y produce una muestra que pertenece al domino de los datos. Este vector se selecciona de manera aleatoria a partir de una distribución gaussiana y se emplea como punto de partida o fuente de ruido para el proceso de generación. Un vez completado el entrenamiento, los elementos dentro de este espacio vectorial multidimensional se corresponden con puntos del dominio del problema, permitiendo construir una representación comprimida de la distribucio1n original de los datos. A este espacio se le conoce como espacio latente, el cual es un espacio vectorial definido por variables latentes [60].

Con frecuencia, estas variables latentes se entienden como una forma de representar o comprimir la informacio1n contenida en una distribución de datos. En otras palabras, el espacio latente ayuda a capturar representaciones de alto nivel de los datos originales. En el caso de las GANs, el generador transforma vectores del espacio latente en puntos del espacio de datos de salida. Así, como los vectores extraídos del espacio latente se utilizan como alimentación para el generador, el cual produce nuevas muestras que simulan los datos reales [60].

2.4.2 El modelo discriminador

El modelo discriminador toma un ejemplo del dominio del problema como entrada (real o generado) y predice una etiqueta de clase binaria de real o falso (generado). El ejemplo real proviene del conjunto de entrenamiento, el ejemplo generado es una salida del modelo generador. El discriminador es comúnmente un modelo de clasificación normal [59].

Las GANs son un modelo generativo basado en aprendizaje profundo. En términos más generales, las GANs son una arquitectura de modelo para entrenar un modelo generativo, y lo más habitual es utilizar modelos de aprendizaje profundo en esta arquitectura. La primera arquitectura GAN se describió por Ian Goodfellow y col. en el 2014 [24].

La arquitectura GAN se conforma por dos modelos entrenados simultáneamente: un modelo generador para generar nuevas muestras y un modelo discriminador para clasificar si las muestras generadas son reales (del dominio) o falsas (generadas por el modelo generador). Esta técnica ha permitido a los ordenadores generar datos realistas utilizando no una sino dos redes neuronales distintas. Las GANs no fueron el primer programa informático utilizado para generar datos, pero sus resultados y su versatilidad los distinguen de otros. El término adversario hace referencia a la dinámica de juego y competición entre los dos modelos que constituyen el marco de la GAN: el generador y el discriminador. El objetivo del generador es producir muestras que capturen las características del conjunto de datos de entrenamiento, hasta el punto de que las muestras que genera sean indistinguibles de los datos de entrenamiento. El generador aprende a través de la información que recibe de la clasificación del discriminador. El objetivo del discriminador es determinar si una muestra en concreto es real (procedente del conjunto de datos de entrenamiento) o falsa (creada por el generador) [60].

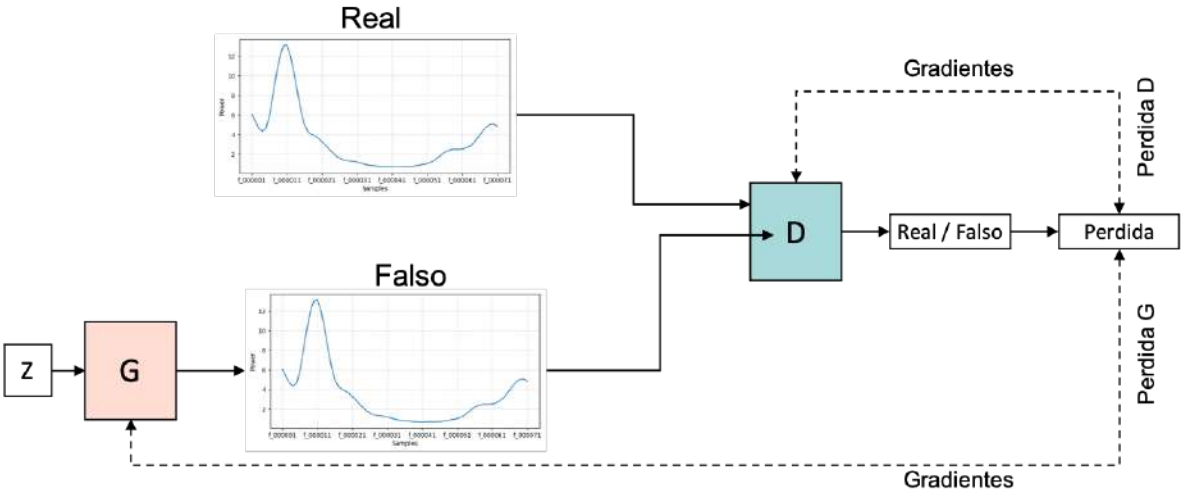


Figura 2.6: *Esquema resumiendo la arquitectura del modelo GAN. D, modelo discriminador y G, modelo generador con un ejemplo de espectro de potencia calculado a partir de un ERG basal en humano.*

Los modelos, el generador y el discriminador, se entrenan de manera simultánea. El generador crea un conjunto de muestras que, junto con datos reales del dominio, son entregados al discriminador para ser etiquetados como muestras reales o falsas [59]. Luego, el discriminador se ajusta con base en su capacidad para diferenciar entre muestras verdaderas y falsas en la siguiente iteración. Lo más relevante es que el generador se modifica [24]. La suma cero significa que cuando el discriminador identifica con éxito muestras reales y falsas, es recompensado y no es necesario cambiar los parámetros del modelo, mientras que el generador es penalizado con grandes actualizaciones de los parámetros del modelo cuando falla. Alternativamente, cuando el generador engaña al discriminador, es recompensado y no es necesario cambiar los parámetros del modelo, pero el discriminador es penalizado y los parámetros de su modelo aumentan [60]. La GAN alcanza un equilibrio cuando se cumplen las siguientes condiciones: el generador produce muestras falsas que son indistinguibles del conjunto de datos de entrenamiento y el discriminador puede, en el mejor de los casos, adivinar aleatoriamente si una muestra es real o falso (es decir, adivinar al 50 % si es real). Con el equilibrio alcanzado, se dice que la GAN ha convergido. En la práctica, es imposible el equilibrio perfecto para las GANs, debido a las inmensas complejidades involucradas en alcanzar la convergencia. Actualmente, la convergencia de GAN sigue siendo una de las preguntas abiertas más importantes en la investigación sobre GAN [24], [59], [60].

Capítulo 3

Antecedentes

3.1 Prevalencia y factores de riesgo de la DM

La prevalencia de diabetes ha mostrado un crecimiento sostenido y alarmante a nivel mundial. En 2021 se estimaba que 537 millos de adultos vivían con diabetes, y se proyecta que esta cifra alcanzará los 783 millones para el año 2045 [3], [61]. En México, el incremente también ha sido constante: la proporción de adultos con diagnósticos médico previo de

diabetes fue de 5.8 % en 2000, 7.0 % en 2006, 9.2 % en 2012 y alcanzó 12.6 % en 2022, con un 5.8 % adicional estimado con diabetes no diagnosticada [61]. En cuanto a la DM tipo 2 específicamente, se observó mayor en mujeres (13.6 %) que en hombres (11.3 %) en 2022. Además, la frecuencia aumenta con la edad: afecta al 2.2 % de los adultos entre 20 y 39 años, el 14.9 % entre los 40 y 59 años, y al 30.3 % en mayores de 60 años [61], [62].

Diversos factores se han vinculado con un mayor riesgo de desarrollar diabetes, entre ellos una alimentación poco saludable, el sobrepeso, la obesidad, y la edad. Además, factores de tipo ambiental y conductual, como nivel educativo, los ingresos económicos, la urbanización, el acceso a servicios médicos y los hábitos de vida, como factores que incrementan dicho riesgo [7], [63]. Gaytán-Hernández y col. [63] señalaron que el porcentaje de diabéticos aumentó sistemáticamente con la edad: de 1.69 % en diabéticos entre 21 y 30 años a 20.69 % en el de 61 años y más. Asimismo, se observó que los habitantes de zonas urbanas presentan una mayor tasa de diabetes en comparación con quienes residen en áreas rurales [7], [9], [64], [65].

En los últimos años, el uso de sistemas de aprendizaje automatizado ha cobrado gran relevancia en la identificación y predicción del riesgo de desarrollar DM. Diversas investigaciones han demostrado que los algoritmos de ML y DL pueden mejorar significativamente la precisión en la detección temprana de esta enfermedad, permitiendo una mejor toma de decisiones clínicas y estratégicas preventivas más eficaces. Mujumdar y col. [66] desarrollaron un modelo de predicción basado en DL con el objetivo de optimizar la clasificación de la diabetes basándose en la inclusión de factores como la glucosa, IMC, edad y la insulina. Obteniendo una precisión del 96 %. Por su parte, Metsker y col. [67] se centraron en la identificación de factores de riesgo aplicados a registros médicos, incluyendo niveles elevados de nefropatía y retinopatía, alteraciones del volumen plaquetario medio, el recuento de glóbulos rojos, hemoglobina y niveles de glucosa en orina, mediante algoritmos de ML, identificando que los niveles de neutrófilos, la concentración de glucosa en sangre y la edad avanzada son variables con alta correlación para el desarrollo de la enfermedad. De manera similar, Faruque y col. [35] estudiaron la predicción de la DM y el análisis de sus factores de riesgo como edad, genero, diabetes, hipertensión, señalando que la presión arterial y las complicaciones renales existe una mayor correlación para el desarrollo de la

enfermedad. Asimismo, Al-Sari y col. [68] desarrollaron algoritmos de ML enfocados en la predicción de la progresión de la diabetes hacia complicaciones más graves, utilizando registros clínicos para modelar el riesgo de padecer enfermedades como la nefropatía diabética. Su estudio destacó que la albúmina en orina, la tasa de filtración glomerular estimada y la hemoglobina glucosilada son factores críticos en la evaluación del riesgo de complicaciones microvasculares en casos diabéticos.

Por otra parte, Tigga y col. [8] evaluaron algoritmos de ML con el objetivo de predecir el riesgo de DM tipo 2, encontrando que variables como la edad, antecedentes familiares de la enfermedad, uso regular de medicamentos y diabetes gestacional desempeñan un papel crucial en la evolución de la enfermedad y sus posibles complicaciones. De manera complementaria, Wang y col. [69] investigaron el impacto de la incorporación de la GRS en modelos de predicción basados en ML. Sus resultados indicaron que la inclusión de esta variable mejoró el rendimiento predictivo de los modelos, considerando como variables predictoras la glucosa en ayunas, los niveles séricos de triglicéridos y colesterol HDL, hipertensión, historial parental, actividad física y el GRS.

Por otro lado, Deberneh y col. [10] desarrollaron un modelo de ML con el propósito de predecir la aparición de DM tipo 2 en el año siguiente, basándose en datos del año actual. Dicho modelo permitió la clasificación de los casos en tres categorías: normales, prediabéticos y diabéticos. Para su desarrollo, se identificaron como factores de riesgo las variables tales como glucosa en ayunas, HbA1c, triglicéridos, gamma-GTP, edad, ácido úrico, genero, tabaquismo, actividad física y antecedentes familiares.

En un análisis más amplio, Fregoso-Aparicio y col. [70] llevaron a cabo una revisión sistemática en la que analizaron 90 estudios enfocados en modelos predictivos de DM utilizando ML y DL. A través de su estudio, los autores concluyeron que no existe un conjunto de características universalmente óptimo para la predicción de la enfermedad, pues la cantidad y elección de los factores de riesgo dependen de la complejidad del modelo. Sin embargo, identificaron que el uso de datos relaciones con el estilo de vida, factores socioeconómicos y datos clínicos tienden a mejorar el rendimiento de los modelos. Uno de los principales desafíos detectados en estos estudios es el desbalance de clases, problema que puede afectar la precisión y generalización de los modelos predictivos. Para abordar estas

limitaciones, la técnica más utilizada ha sido SMOTE, la cual permite generar muestras sintéticas a partir de datos minoritarios con el fin de equilibrar las clases [71]. Asimismo, otro problema recurrente en la predicción de la DM tipo 2 es la presencia de valores faltantes con los conjuntos de datos, lo cual se ha tratado principalmente mediante algoritmos de imputación o, en la mayoría de los casos, mediante la eliminación de registros con datos faltantes [72]. En cuanto a los modelos empleados, aquellos basados en RF han demostrado un desempeño notable, alcanzando una precisión del 90 %, un área bajo la curva ROC (AUC) del 90 % y una sensibilidad del 80 % [70]. No obstante, algunos estudios no detallan de manera explícita el manejo de los valores faltantes ni las estrategias aplicadas para abordar el desbalance de clases. Adicionalmente, muchos de los conjuntos de datos utilizados son relativamente pequeños, como el caso de la base datos que es más reportada con 800 registros y 10 atributos extraídos de registros de salud electrónicos, lo que puede comprometer la capacidad de generalización de los modelos.

Con el propósito de superar algunas de estas limitaciones, han surgido nuevas estrategias basadas en modelos generativos, particularmente el uso de GAN. Estas técnicas han demostrado ser efectivas para abordar el desbalance de clases y la imputación de datos al generar muestras sintéticas con una distribución más realista en comparación de métodos tradicionales antes referidos. Por ejemplo, Jaiswal y col. [73] propusieron el uso de una GAN para la generación de datos sintéticos a partir de la base de datos Pima Indian Diabetes (PID) y la posterior clasificación mediante un LSTM por sus siglas en inglés. Los resultados obtenidos indicaron que la combinación de datos reales y sintéticos permitió mejorar la presión del modelo, alcanzando un 97 % de exactitud. De manera similar, Boughareb y col. [74] presentaron un sistema de detección y predicción de diabetes basado en DL, en el que utilizaron GANs para la expansión del conjunto de datos de entrenamiento en una ANN. Sus resultados experimentales mostraron que el modelo alcanzó una exactitud del 94 % y una precisión del 95 %.

Asimismo, Chushing-Muzo y col. [75] exploraron un enfoque basado en datos clínicos para predecir el riesgo de enfermedad cardiovascular a 10 años en casos con DM tipo 1. En su estudio, implementaron GANs con el objetivo de aumentar el conjunto de datos original para mejorar el rendimiento de los modelos ML. Entre los factores de riesgo

considerados en el estudio incluyeron edad, genero, duración de la DM, presión arterial sistólica, colesterol LDL, HbA1c, tasa de filtración glomerular estimada, ejercicio, tabaquismo, y los niveles de albumina en sangre.

Autor	Modelo IA	Métricas	Variables		Datos
			Invasiva	No invasiva	
[7]	LG Decision Tree	ACC = 78.26	Glucosa Insulina	Embarazos IMC Historial Familiar Edad Presión Arterial	PIMA Indian by National Institute of Diabetes: 286 casos diabéticos
[66]	SVM RF Decision Tree Extra Tree Ada Boost Algorithm Perceptron Linear Gaussian Naïve Bayes	ACC = 96%	Nivel de glucosa Presión Arterial Insulina	Edad IMC Diabético o No	PIMA Data: Contiene 800 registros y 10 atributos
[8]	LG K-Nearest Neighbors SVM Gaussian Naïve Bayes RF	ACC = 94.1%	Concentración de glucosa en plasma Resultados de prueba de laboratorio a partir de muestras de sangre	Edad Género Historial familiar Presión Arterial Actividad física IMC Tabaquismo Consumo de alcohol	PIMA Indian by National Institute of Diabetes
[67]	Artificial Neural Network SVM Decision Tree LG	Precision = 83.28 % Recall = 81.52% F1 score = 80.64% ACC = 82.61% AUC = 0.89	Resultados de prueba de laboratorio a partir de muestras de sangre	Edad Género Retinopatía (Identificadas a partir de diagnósticos en registros médicos)	Sistema de información médica del centro Médico Especializado Almozov: 238,590 registros de laboratorio de 5846 casos con diabetes
[35]	SVM Naive Bayes K-Nearest Neighbor Decision Tree	Precision = 72% Recall = 74% F1 score = 72 ACC = 73.5% AUC = 0.69	Problemas renales	Edad Género Peso Dieta Poliuria Diabetes Hipertensión Presión Arterial	Centro médico de Chittagong (MCC): 200 casos con 16 atributos relacionados con la diabetes mellitus
[68]	RF	AUC = 0.98 ACC = 98 %	Colesterol HDL Colesterol LDL	Edad IMC	Seteno Diabetes Center

		Precision = 99 % Recall = 92 % F-score = 96 %	HbA1c Creatinina sérica albuminaría Variables moleculares (omics)	Presión Arterial Historial de enfermedad cardiovascular	Copenhagen (SDCC): 648 casos que incluyen 17 variables clínicas
[69]	Cox regression Models ANN RF Gradient boosting machine	AUC = 0.885	Colesterol HDL Colesterol LDL Glucosa plasmática en ayunas Triglicéridos GRS	Edad Género Circunferencia de la cintura Historial familiar	Henan Rural Cohort Study: 5,712 individuos entre 19 y 79 años. 324 casos desarrollaron DM tipo 2.
[10]	LG SVM Random Forest XGBoost Métodos de ensamble	ACC = 73 %	Resultados de prueba de laboratorio a partir de muestras de sangre	Datos de registros médicos	Datos de registros médicos electrónicos de un hospital del corea del sur: un total de 253,395 casos
[70]	LG SVM RF Penalized Likelihood Methods	ACC = 0.9 AUC = 0.90 Sensitivity = 0.8	Resultados de prueba de laboratorio a partir de muestras de sangre	Registros de salud electrónicos	PIMA Data: Contiene 800 registros y 10 atributos Base de datos propias
[73]	GAN LSTM	ACC = 92% Precision = 97% AUC = 0.92		Concentración de Glucosa en plasma IMC Grosor de piel Presión arterial diastólica Insulina Historial familiar diabetes Edad Embarazos	Pima Indian Diabetes (PID): 768 muestras con 8 atributos. 268 con diabetes y 500 sin diabetes
[74]	GAN-ANN	Precision = 94 %		Concentración de Glucosa en plasma IMC Grasa corporal de piel Presión arterial diastólica Insulina Historial familiar diabetes Edad Embarazos	Pima Indian Diabetes (PID): 768 muestras con 8 atributos. 268 con diabetes y 500 sin diabetes
[75]	K-Nearest Neighbors Decision Tree	MSE = 0.0088 MAE = 0.017	Colesterol HDL Colesterol LDL HbA1c	Edad Género	Steno Diabetes Center Copenhagen: 677

RF MLP	Duración de la diabetes Tabaquismo Ejercicio Presión arterial sistólica	casos donde incluyen 10 atributos

Tabla 1:*Comparativa de modelos de DL y ML aplicados a la predicción de DM tipo 1 y DM tipo 2.*

3.2 GANs en series de tiempo

En el ámbito del procesamiento de señales biomédicas, las GANs han demostrado ser altamente eficaces en la generación y mejora de datos sintéticos, facilitando la clasificación y el modelado de patrones complejos en registros fisiológicos. Por ejemplo, Galony y col. [76] propusieron un método para mejorar la clasificación de señales de ERG mediante la generación de datos sintéticos utilizando GAN. Para ello emplearon la base de datos MIT-BIH Arrhythmia, la cual contiene señales de ECG de casos con diversas condiciones cardíacas. Su estudio demostró que el uso de datos sintéticos para aumentar el conjunto de entrenamiento de un modelo LSTM mejora significativamente el rendimiento de clasificación. De manera similar Fariha-Hossain y col. [77] utilizaron la misma base de datos para desarrollar un nuevo modelo GAN llamado ECG-adv-GAN diseñado específicamente para generar señales de ECG sintéticas con la capacidad de engañar a clasificadores de arritmias cardíacas, lo que resalta el potencial de estas redes en la evaluación de la robustez de modelos de clasificación.

En un enfoque más avanzado, Xiaomin y col. [78] presentaron una arquitectura GAN basada en Transformers para la aumentación de datos en series de tiempo. Su estudio abarcó cuatro conjuntos de datos diferentes: Simulated Senusoidal Waves, UniMIB Human Activity Recognition, PTB diagnostic ECG y MIT-BIH Arrhytmia, en los cuales generaron datos sintéticos para equilibrar la distribución de clases y mejorar el rendimiento de los modelos de clasificación. Por otro lado, Festag y col. [79] exploraron el uso de GANs para la predicción e imputación de valores faltantes en series de tiempo. Utilizando la base de datos de señales ECG de MIT-BIH Arrhythmia, demostraron que las GANs pueden reconstruir datos faltantes con alta precisión, lo que es especialmente útil en escenarios donde la calidad de los registros biomédicos puede verse afectada por ruido o interrupciones en la captura de los datos.

Además de las aplicaciones biomédicas, las GANs han sido empleadas en otros dominios de series de tiempo. Por ejemplo, Shunjian y col. [80] mejoraron el modelado de series de tiempo en datos de precios de Bitcoin, utilizando GANs para simplificar el entrenamiento y generar mejores predicciones en comparación con otros enfoques tradicionales. Esta aplicación es especialmente relevante en el ámbito financiero, donde las fluctuaciones de los precios pueden ser difíciles de modelar con precisión debido a su naturaleza estocástica.

Finalmente, Matti y col. [81] propusieron un método novedoso para entrenar redes neuronales con ECG sintéticos generados dinámicamente mediante GANs. A diferencia de otros enfoques donde utilizan conjuntos de datos estáticos de señales generadas previamente, su propuesta consiste en generar datos sintéticos durante el proceso de entrenamiento, lo que permite una mayor variabilidad en las muestras y una mejora en la capacidad de generalización del modelo. Este enfoque fue probado en cuatro diferentes conjuntos de datos relacionados con ECG, logrando mejoras significativas en la clasificación. La diferencia con respecto a los otros es que se generan muestras sintéticas de manera dinámica, en lugar de usar un conjunto de datos estáticos, lo generaron durante el entrenamiento, aplicándolo a cuatro diferentes conjuntos de datos relacionados con señales de ECG.

Autor	**Modelo IA**	**Métricas**	**Variables**		**Datos**
			Invasiva	**No invasiva**	
[76]	DCGAN	**Quantitative**: AUC **Qualitative**: Visual Inspection		Señales de ECG	MIT-BIH Arrhythmia
[77]	ECG-ADV-GAN Conditional Generative Adversarial Network (CCGAN)	**Quantitative**: Discriminative Score Modified wavelet coherence MSE **Qualitative**: PCA T-SNE		Señales de ECG	Physionet MIT-BIH Arrhythmia
[78]	TTS-CGAN	**Quantitative**: MSE Structural similarity Index (SSIM)		Señales de ECG y NIRS	Cuatro conjuntos de datos: Simulated Sinusoidal Waves

		Cross-correlation coefficient Normalized Mean Squared Error (NRMSE) **Qualitative**: Visual Inspection		UniMiB Human Activity Recognition PTB Diagnostic ECG database MIT-BIH Arrhythmia Database
[79]	cGAN	**Quantitative**: MSE RMSE SSIM Correlation **Qualitative**: Visual Inspection Human Objective	Señales de ECG	MIT-BIH Arrhythmia
[80]	Wassertein GAN	**Quantitative**: Marginal Distribution Difference of Lag-1 distribution R2 obtained from TSTR Sig-W distance **Qualitative**: Visual Inspection	Datos financieros	Datos de indices bursatil SPX y DJI, datos de precios de Bitcoin
[81]	LSTM	**Quantitative**: ROC-AUC Recall Precision F1-score	Señales de ECG	Se utilizaron cuatro diferentes conjunto de datos relacionados con ECG: Glasgow University ECG database MIT-BIH normal sinus Rhythm database MIT-BIH noise stress test database Computing in Cardiology 2017 single atrial fibrillation database

Tabla 2: *Resumen de antecedentes de GANs en series de tiempo.*

3.3 Generación de espectros mediante GANs

En el campo del análisis espectral, las GANs han sido exploradas para superar la escasez de datos etiquetados, mejorar la calidad de los espectros generados y facilitar la clasificación de muestras en distintos escenarios científicos y médicos. El uso de GANs en este ámbito resulta particularmente atractivo debido a la dificultad de obtener espectros experimentales con alta precisión, ya que muchas técnicas requieren equipos costosos, largas sesiones de adquisición de datos y procesamientos complejos. Por ejemplo, Audebert y col. [82] presentaron un método basado en GANs para la síntesis de muestras hiperespectrales, abordando la escasez de datos etiquetados en este campo. Su modelo fue validado mediante análisis estadístico, reducción de dimensionalidad con PCA y clasificación utilizando un SVM. Por otro lado, Smith y col. [83] propusieron un modelo para la generación de datos sintéticos de series de tiempo unidimensionales, mediante la arquitectura TSGAN, la cual emplea dos redes: la primera convierte los vectores latentes en imágenes de espectrogramas, y la segunda genera las series de tiempo condicionadas por estas imágenes.

Por otra parte, Pavlou y col. [84] desarrollaron un enfoque basado en GANs para la generación de espectros sintéticos de huesos sanos y osteoporóticos, utilizando un conjunto de datos relativamente pequeño que contenía 90 espectros de huesos normales y 72 de huesos osteoporóticos. Esta metodología permitió aumentar la cantidad de datos disponibles, facilitando el entrenamiento de modelos de clasificación de dominio. Finalmente, Hazra y col. [85] introdujeron SynSigGAN, un modelo de GAN diseñado para generar señales sintéticas como ECG, EEG, EMG y PPG. Este enfoque incluyó una etapa de preprocesamiento de señales, donde se aplicaron transformaciones como DWT e inversa DTW (IDWT), además de técnicas de umbralización. Aunque el modelo logró generar señales con características similares a las reales, se identificó como limitante que los datos generados se encuentran en un rango normalizado y no como la naturaleza de los datos.

Uno de los aspectos más relevantes en la generación de espectros mediante GANs es la evaluación de la calidad de los datos sintéticos. Los estudios en este ámbito han adoptado estrategias de evaluación cuantitativa y cualitativa que dependen del dominio de los datos con los que se trabaja. Por ejemplo, en la síntesis de espectros, se han utilizado análisis de desviación estándar y PCA para verificar la similitud de los datos reales. En la generación de

espectros Raman, la evaluación se ha basado en métricas como Discret Frechet Inception Distance (DFID) y variabilidad del conjunto de datos. En el caso de señales biomédicas, se han implementado métricas como coeficiente de correlación de Pearson, RMSE, MAE y FD. Esta diversidad de enfoques refleja la necesidad de desarrollar criterios de evaluación más estandarizados y específicos para la validación de espectros sintéticos generados mediante GANs.

Autor	Modelo IA	Métricas	Variables		Datos
			Invasiva	No invasiva	
[82]	Wassertein GAN	Grafican la media y la desviación estándar del espectro **Quantitative**: Accuracy **Qualitative**: PCA 2D Support Vector Machine (SVM)		Sensores hiperespectrales. Se realizan mediciones de reflectancia espectral de forma remota	Pavia University: 103 bandasespectrales Pavia Center: 224 bandas espectrales
[83]	TSGAN	**Quantitative**: Frechet Inception Score (FID) **Qualitative**: Utilizan la clasificación como criterio de evaluación (CNN inception V3)		Los conjuntos incluyen señales de sensores, espectrogramas, dispositivos de movimiento, datos médicos.	University of California Riverside: 70 conjuntos de datos de series de tiempo unidimensionales.
[84]	GAN	Evalúan la variabilidad del conjunto de datos mediante PCA **Quantitative**: Discret Frechet Inception Distance (DFID)		Espectroscopía Raman	Espectros Raman de huesos de conejo: 90 espectros de huesos normales y 72 de hueso
[85]	SynSigGAN	Quantitative: Pearson Correlation Coefficiente Root Mean Square Error (RMSE) Percent Root Mean Square Difference (PRD) Mean absolute Error (MAE) Frechet Distance		Señales biomédicas utilizadas: ECG, EEG, EMG, PPG	ECG: Base de datos MIT-BIH Arrhytmia EEG: Base de datos Siena Scalp EEG EMG: Base de datos Sleep-EDF PPG: Base de datos BIDMC PPG and respiration

Tabla 3: *Resumen de antecedentes generación de espectros mediante GANs.*

A pesar del avance en el uso de GAN en distintos dominios, la generación de espectros sintéticos sigue siendo un área poco explorada, con desafíos significativos en la evaluación de la calidad de los datos generados. El presente trabajo busca ampliar el estado del arte en esta área. Dado que las métricas de evaluación dependen en gran medida del dominio de los datos, un objetivo clave de esta investigación fue adaptar criterios de validación que permitieron evaluar con mayor precisión la calidad de los datos sintéticos. Aunque la mejora de la predicción de interés se considera la validación definitiva, la validación de la calidad de los datos sintéticos producidos antes de este paso ayuda a optimizar el coste computacional.

Capítulo 4

Metodología

La metodología utilizada en el desarrollo del proyecto *espectraGAN*, fue estructurada en cuatro etapas principales, incluyendo una *primera* etapa de procesamiento de los espectros de potencia derivados de las señales ERG con una fase de exploración, de identificación de valores atípicos (outliers) y análisis detallado de los espectros para comprender su morfología y definir la manera óptima de prepararlos para el entrenamiento del modelo. La *segunda* etapa consistió en la generación de espectros sintéticos con el modelo denominado *espectraGAN*, donde se muestra la implementación del modelo con la incorporación de todos los factores necesarios para asegurar un entrenamiento eficiente. Se analizaron las funciones de pérdida en diferentes escenarios, evaluando tanto los resultados exitosos como los no exitosos. Además, se llevó a cabo la generación de los espectros sintéticos, y se implementó un filtro para asegurar que estas muestras provienen de la misma distribución que los datos orgánicos. En la *tercera* fase, se realizó la evaluación estadística de los espectros generados por *espectraGAN*. Para ello, se realizaron comparativas visuales de los espectros sintéticos frente a los orgánicos con el fin de asegurar la fiabilidad de los espectros generados. Así mismo, se llevó a cabo la evaluación cualitativa y cuantitativa de los espectros sintéticos. El proceso culminó en la *cuarta* etapa, del balance de clases, que resultó en la generación de una base de datos sintética equilibrada y lista para su posterior uso. Respecto a ello, aunque no forman parte como tal del presente proyecto, se comparten en un apartado final de la Discusión sobre la relevancia del trabajo, los resultados del impacto de la base de datos extendida sobre la predicción de casos enfermos y sanos. La **Figura 4.1** ilustra de manera esquemática cada una de estas cuatro etapas.

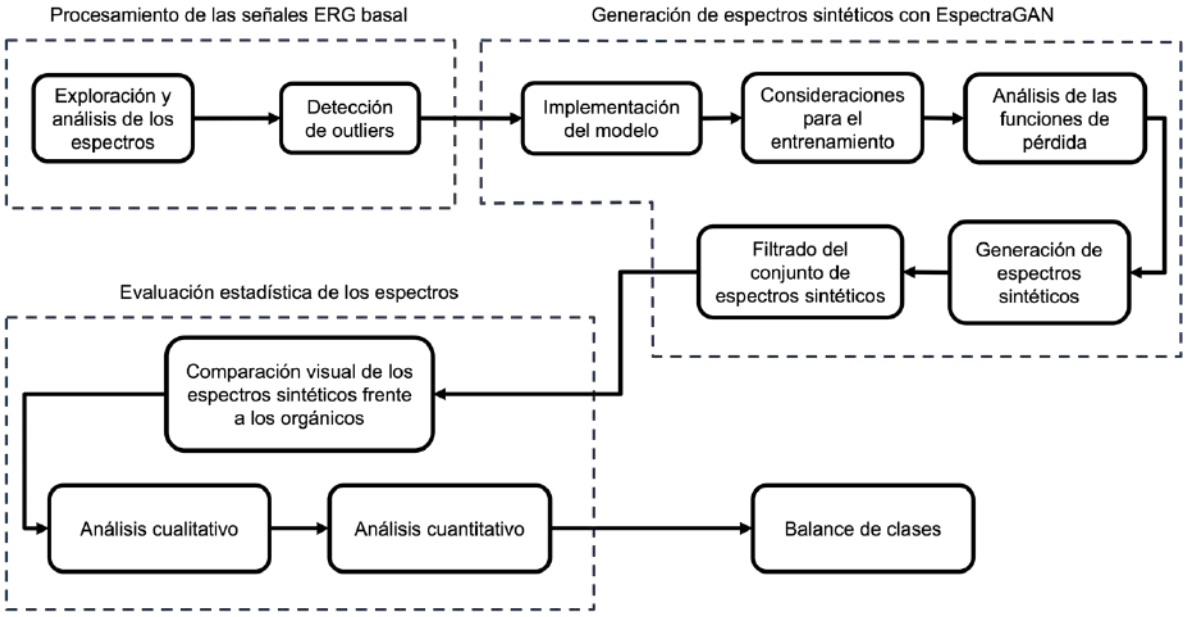


Figura 4.1: *Metodología implementada para el desarrollo de la tesis.*

4.1 Procesamiento de las señales del ERG basal

Se utilizó una base de datos de ERG basales en humanos [18], la cual contó con la aprobación de tres comités de ética: Instituto Mexicano de Oftalmología (IMO), el Comité Nacional de Ética (CONBIOÉTICA-09-CEI-006-20170306), y el Comité de Investigación de la Asociación Para Evitar la Ceguera (APEC, 17 CI 09 003 142). Todos los participantes firmaron un consentimiento informado. Los procedimientos realizados cumplieron con las normas éticas establecida en la declaración de Helsinki [86].

Los ERG basales se obtuvieron con diferentes electrorretinógrafos, pero bajo las mismas condiciones de registro (condición de luz diurna, 400 lux, 5 minutos de registro en ausencia de flash de luz). Se filtraron entre 0.3 Hz y 1 KHz con la finalidad de tener acceso a la mayor información posible y no generar un sesgo *a priori*. Se dividieron en segmentos consecutivos de 60 s para aumentar el número de muestras. Para el procesamiento de las señales se aplicó la transformada Wavelet de tipo Morlet real ya que la señal es discontinua [18]. Las resoluciones temporales y espectrales fueron de 0.01 y 0.05 Hz, respectivamente [18], [87].

4.1.1 Exploración y análisis de los espectros

El proceso de exploración y análisis de los espectros comenzó con una revisión exhaustiva de la base de datos que contenía dichos espectros. La base de datos utilizada para entrenar el modelo *espectraGAN* está conformada por 639 casos, que fueron etiquetados como sanos o enfermos tal como se describió en [18]. Brevemente, se especifica que la categoría *"control"* corresponde a sujetos sin sobrepeso, obesidad, prediabetes, desorden lipídico, hipertensión arterial, diabetes o enfermedad ocular, mientras que la categoría *"enfermos"* incluye a casos con sobrepeso, obesidad, síndrome metabólico sin diabetes o diabetes sin retinopatía diabética (para más detalles, ver [18]). Cabe añadir que la segmentación de 60 s permitió generar 3,665 espectros (854 control y 2,811 enfermos) y que una vez asignado a una categoría (entrenamiento o prueba), los diferentes segmentos de un mismo paciente fueron asignados a la misma categoría.

Inicialmente, se identificaron las distribuciones de cada clase y el número de observaciones en cada grupo etario para comprender cómo estaban organizados los datos y establecer un marco adecuado para su interpretación. Los conjuntos de datos se dividieron en dos grupos de entrenamiento: uno para el grupo *control* y otro para el grupo *enfermos*, dado que el análisis se realizó por separado para cada conjunto. Para cada grupo se aseguró que la base de datos estuviera limpia: se eliminaron valores repetidos o nulos, y se excluyeron columnas no relevantes para el procesamiento de los datos de este trabajo. Además, se categorizaron los espectros reales según su grupo etario. Para orientar el estudio, se tomaron en cuenta los estándares de los grupos etarios en México, conforme a los datos reportados por el INEGI [2], [43], [88]. Estos grupos se clasificaron en rangos de edad de la siguiente manera (en años): 19-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, y mayores de 85.

El siguiente paso fue la normalización de los conjuntos de datos, un proceso fundamental para garantizar que los diferentes rangos de potencia de los espectros se ajustaran a un mismo intervalo, evitando aquellas potencias predominen en el análisis. Para ello, se utilizó el método de *MinMaxScaler* de la librería *scikitearn*, que permite reescalar los valores de las potenciases de los espectros en un rango definido [89]. En este trabajo, se definió un rango de normalización de 0 a 5. Esto implica que, para cada espectro en el

conjunto de datos, el valor mínimo de la potencia se encuentra en un rango de 0 a el valor máximo 5, mientras que los intermedios se ajustan proporcionalmente dentro de este intervalo. Esta elección, en lugar del rango más común de 0 a 1, se fundamenta en la necesidad de representar las potencias de los espectros en un intervalo más amplio que nos permitió identificar patrones que podrían pasar desapercibidos en un intervalo más estrecho [19].

Lo anterior facilitó la aplicación de un PCA que permitió obtener una representación visual estadística de los espectros derivados de ERG basal dentro del contexto de cada grupo etario, mediante la reducción de dimensionalidad. Este análisis nos permitió identificar patrones ocultos que no eran evidentes en el espacio de características original, reduciendo los datos a dimensiones más significativas [90].

Para determinar cuántos componentes principales eran necesarios para capturar la mayor cantidad de información sin comprometer la interpretabilidad ni introducir redundancias, se utilizó el gráfico de sedimentación. En este gráfico, el eje Y muestra la varianza explicada por cada componente principal, mientras que el eje X ordena los componentes principales (PC1, PC2, PC3, etc.) según su importancia o la cantidad de varianza explicada. Al trazar la varianza explicada en función del número de componentes, el gráfico generalmente presenta una disminución rápida de la varianza para los primeros componentes, seguida de una pendiente más suave. El "codo" o punto de inflexión de la curva indica el punto en el que añadir más componentes no aporta una ganancia significativa en la explicación de la varianza [91].

Finalmente, se generó un gráfico para cada clase subdividida en grupos etarios, lo que permitió visualizar la disposición de los datos dentro del espacio seleccionado a partir de las componentes principales.

4.1.2 Detección de valores atípicos

Para detectar valores atípicos, se implementó un algoritmo que permitió identificar aquellos espectros que se desvían significativamente de la distribución normal, es decir, los *outliers*. Esta detección no solo buscó identificar anomalías, sino también permitió, a través de la desviación típica Sigma (σ), asignar un nivel de importancia relativo a las áreas donde no se concentra la mayor parte de los espectros. Cabe precisar que, aunque casos presentan propiedades diferentes a la distribución normal, se incluyeron en el trabajo, ya que forman parte del grupo etario correspondiente y representan la variabilidad humana.

El funcionamiento de este algoritmo se basó en la selección de un grupo etario específico, a partir del cual se generaron dos subconjuntos de datos: *data_split*, que incluye los datos del grupo etario en cuestión y *data_residuo*, que comprende el resto de los datos. Esto para cada clase: *control* y *enfermos*.

El primer paso consistió en introducir el grupo etario normalizado anteriormente realizado para facilitar la aplicación del PCA, seleccionando los mismos componentes principales encontrados en el proceso previo. Posteriormente, se calculó el centroide del conjunto *data_split*, representando el punto medio de la distribución de datos en el espacio multidimensional generado por el PCA. Mediante el uso de la distancia euclidiana y un umbral predefinido por la desviación estándar (σ), se determinaron los índices de aquellos casos cuyos datos se encuentran más alejados del centroide, es decir, los más distantes de la concentración principal de puntos. Este enfoque permitió identificar de manera efectiva a los *outliers*, otorgando un marco cuantitativo para evaluar la dispersión de los espectros en relación con la norma del grupo etario específico. El pseudocódigo del algoritmo (**Figura 4.2**) detalla paso a paso este proceso.

Algorithm 1 Procesamiento de datos y detección de outliers

Require: *data_espectros* (dataframe), *condicion* (rango grupo etario)
Ensure: Indices y datos con ouliers

```
function DETECCIÓN OUTLIERS(data_espectros, condicion)
    pca ← Inicializar PCA con N componentes
    Aplicar PCA data_espectros

    Separar datos transformados en subconjuntos (data_split, data_residuo)

    Calcular el centroide de data_split
    Calcular la distancia euclidiana de cada punto al centroide
    Determinar un umbral a partir de σ

    Identificar outliers basados en el umbral
    data_outliers ← Índices con los sujetos outliers

    Return Indices outliers y data_outliers
end function
```

Figura 4.2: *Pseudocódigo del Algoritmo 1. (Procesamiento de los datos y detección)*

4.2 Generación de espectros sintéticos con EspectraGAN

Posteriormente, se procedió a la generación de espectros sintéticos a partir del modelo *espectraGAN*, la cual se llevó a cabo en cinco fases principales.

Primero, la implementación del modelo GAN, que detalla las arquitecturas de las redes neuronales *generador* y *discriminador* que lo componen, así como la metodología utilizada para entrenar con los espectros orgánicos.

La segunda fase se enfocó en las consideraciones necesarias para entrenar el modelo *espectraGAN*, incluyendo los ajustes realizados en la tasa de aprendizaje durante el entrenamiento y aspectos clave para asegurar que cada lote de datos incluye espectros con una mayor capacidad de generalización.

La tercera fase trata del análisis de las funciones de pérdida mediante ecuaciones que se actualizan automáticamente para optimizar el desempeño del modelo.

La cuarta fase consiste en la generación de los espectros sintéticos a partir de los modelos entrenados exitosamente.

Finalmente, en la 5ta fase, de filtrado, se aplicaron métricas estadísticas, como el RMSE y la prueba de Mann-Whitney U, para seleccionar los espectros sintéticos más similares a sus contrapartes orgánicas.

4.2.1 Implementación del modelo EspectraGAN

Las GANs constan de dos redes neuronales nombradas *generador* y *discriminador* [24].Su entrenamiento se basa en un juego 'minmax' entre el generador y el discriminador, donde el generador intenta minimizar el valor de V(D,G) (**Ecuación 1**) mientras que el discriminador intenta maximizarla [60], [86]:

$$V\ (D,G) = E_{x\sim p_{dt}(x)}[log\ log\ D\ (x)] + E_{z\sim p_z(z)}\left[log\ log\ \left(1 - D\big(G(z)\big)\right)\right] \quad (1)$$

La señal de entrada del *generador* corresponde a un vector de ruido aleatorio z, el cul se extrae de una distribución de datos P_z. Este vector suele tener baja dimensionalidad y sus valores típicamente de una distribución normal con media cero y desviación estándar uno , es decir, $N(\mu = 0, \sigma = 1)$ [24]. Por su parte, $D(x)$ represente la función del discriminador, la cual estima la probabilidad de que una muestra X provenga de los datos reales. Aquí, E_x denota el valor esperado sobre la distribución de los datos reales P_{dt} mientras que $G(z)$ corresponde a la salida generada por el generador.

El termino $D(G(z))$ indica la probabilidad, evaluada por el discriminador, de que un espectro generado sea auténtico o sintético. El valor esperado sobre $P_{z(z)}$ se representa como E_z. Tanto el *generador* como el *discriminador* son entrenados de forma adversarial, con el objetivo de mejorar mutuamente su desempeño[60], [86].

Durante el proceso de entrenamiento, el *discriminador* D recibe tanto los espectros generados $D(G(z))$ como los espectros orgánicos D(x), y debe aprender a distinguir entre ellos. A su vez, el *generador* G ajusta sus parámetros con base en la retroalimentación, intentando producir espectros cada vez más parecidos a los reales. La función de pérdida utilizada para entrenar al discriminador puede formularse como la **Ecuación 2** [60], [86]:

$$V\ (D,G) = E_{x\sim p_{dt}(x)}[log\ log\ D\ (x)] + E_{x\sim p_z(z)}\left[log\ \left(1 - D\big(G(z)\big)\right)\right] \quad (2)$$

De forma simultánea, el *generador* G se ajusta para minimizar el termino $log\ \left(1 - D\left(G(z)\right)\right)$ con el fin de que las muestras producidas por este se asemejen lo más posible a los datos orgánicos. La función de pérdida del generador G puede expresarse con la **Ecuación 3** [60], [86]:

$$V\ (D, G) = E_{x \sim p_{dt}(x)}[log\ log\ D\ (x)] \qquad (3)$$

A partir del concepto general de las redes GAN, se propuso el modelo *espectraGAN*. Para describir la organización de los datos de entrada, se definió el conjunto $x = x_1, \ldots, x_n$, que representa una secuencia temporal de cada componente espectral. En este conjunto n, corresponde al número total de muestras, y cada $x^i \in R^{1xd}$, donde d indica la dimensión de cada vector de entrada [86].

La arquitectura propuesta fue desarrollada utilizando *TensorFlow* en *Python 3.8*, una elección estratégica debido a varias razones fundamentales. *TensorFlow* es una plataforma de código abierto líder en el aprendizaje automático, que permite la construcción y despliegue eficiente de modelos de ML en diversos entornos y plataformas [92]. Por otro lado, la utilización de *Python 3.8* ofrece una interfaz amigable y bien integrada con *TensorFlow*, permitiendo a los desarrolladores utilizar diversidad de bibliotecas para manipular datos de manera eficiente antes de convertirlos a tensores para el entrenamiento de modelos [93]. Además, la capacidad de *TensorFlow* para el funcionar tanto con CPU como en GPU optimiza significativamente el proceso de entrenamiento, como es el caso del uso de una NVIDIA GeForce RTX 2060 SUPER, que permite acelerar las operaciones de cálculo en comparación con el uso exclusivo de la CPU [94]. Esta configuración tanto de software como hardware no solo facilitó la investigación y desarrollo en los modelos de aprendizaje profundo, sino que también permitió asegurar la transición en despliegue de modelos en servidores de alto rendimiento.

El primer paso de la metodología consistió en introducir un vector de ruido en la entrada de la red *generador* y escalar este vector a una matriz que pudiera contener las dimensiones de cada espectro. La arquitectura de la red *generador* fue diseñada especialmente para generar un espectro a partir de una representación latente. La primera capa, una *ConvTranspose1D,* fue esencial para convertir el ruido en espectros con longitud

de acuerdo a su frecuencia, preservando las relaciones temporales [86]. Este tipo de capa, también conocida como convolución transpuesta o deconvolución, permite expandir las dimensiones de los datos mientras mantiene las características esenciales, lo que es fundamental para la construcción precisa de los espectros desde un espacio latente más comprimido [95].

Se utilizaron cuatro capas *ConvTranspose1D* con un tamaño de kernel de 5, un stride de 1 y padding del tiempo de tipo *same.* Este diseño permitió una expansión gradual y controlada de las dimensiones de los datos a lo largo de la red, asegurando que el modelo pueda reconstruir detalladamente los espectros desde el espacio latente. La inclusión de capas *BatchNormalization* entre las capas *ConvTranspose1D* fue crucial para acelerar el entrenamiento y mejorar la estabilidad y robustez de la red, mientras que la función de activación *Swish* ayudó a adaptar la red más eficientemente a la distribución de los datos y suavizar los gradientes, evitando problemas comunes en el entrenamiento como la desaparición de los gradientes [96].

Finalmente, se añadió una capa de regularización *Dropout* para prevenir el sobreajuste y garantizar que el modelo generalice las características de nuevos espectros que no fueron utilizados durante el entrenamiento.

El número de mapas de características de cada capa *ConvTranspose1D* fue de 142, 253, 253 y 142, respectivamente, culminando en un tamaño que corresponde a la longitud del espectro en términos de frecuencia. Es decir, los espectros de la base de datos representan la potencia en función de la frecuencia. Para formar el espectro de un paciente, se utilizaron 71 frecuencias que abarcan un rango de 0.1 a 40 Hz [22].Por lo tanto, la salida del modelo *generador* consistió en 71 neuronas simulando ese rango de frecuencias, cada una generando la potencia relacionada a la frecuencia de los espectros con los que fueron entrenados.

La red *discriminador*, inspirada en la arquitectura *inception*, es eficaz para la clasificación de series de tiempo debido a su capacidad para manejar múltiples escalas de tiempo y capturar características relevantes de forma automática [97]. Se utilizó el módulo *inception* porque permitió aplicar simultáneamente filtros de diferentes longitudes, lo que mejora la capacidad de la red para extraer características tanto de series de tiempo largas

como cortas [97]. Esta flexibilidad es fundamental en el procesamiento de series de tiempo, donde se necesita capturar patrones tanto locales como globales [98]. La primera capa del módulo, conocida como *bottleneck*, reduce la dimensionalidad de las entradas, lo que no solo disminuye el coste computacional sino también el número de parámetros necesarios. Este enfoque es efectivo para pre-procesar los datos antes de aplicar operaciones más complejas. Luego, un conjunto de capas convolucionales paralelas actúa sobre el mismo mapa de características, permitiendo que la red procese diferentes aspectos de las series de tiempo simultáneamente. La tercera capa, *MaxPooling1D*, ayuda a hacer el modelo más robusto a pequeñas perturbaciones y variantes en los datos. La última capa es una concatenación de la salida de todas las capas anteriores, integrando todas las características extraídas en una representación unificada. Finalmente, la arquitectura completa del *discriminador* en el modelo llamado *espectraGAN* incluye dos módulos *inception* con capas *Conv1D,* optimizadas para trabajar con series de tiempo y mantener la integridad temporal mientras se extraen características detalladas para la clasificación [98]. Esto se concatena a una capa densa con función de activación *Sigmoid* obteniendo así la función de pérdida del discriminador (**Ecuación 2**). En la **Figura 4.3**, se muestra la arquitectura final de *espectraGAN* para cada una de las redes *generador* y *discriminador*.

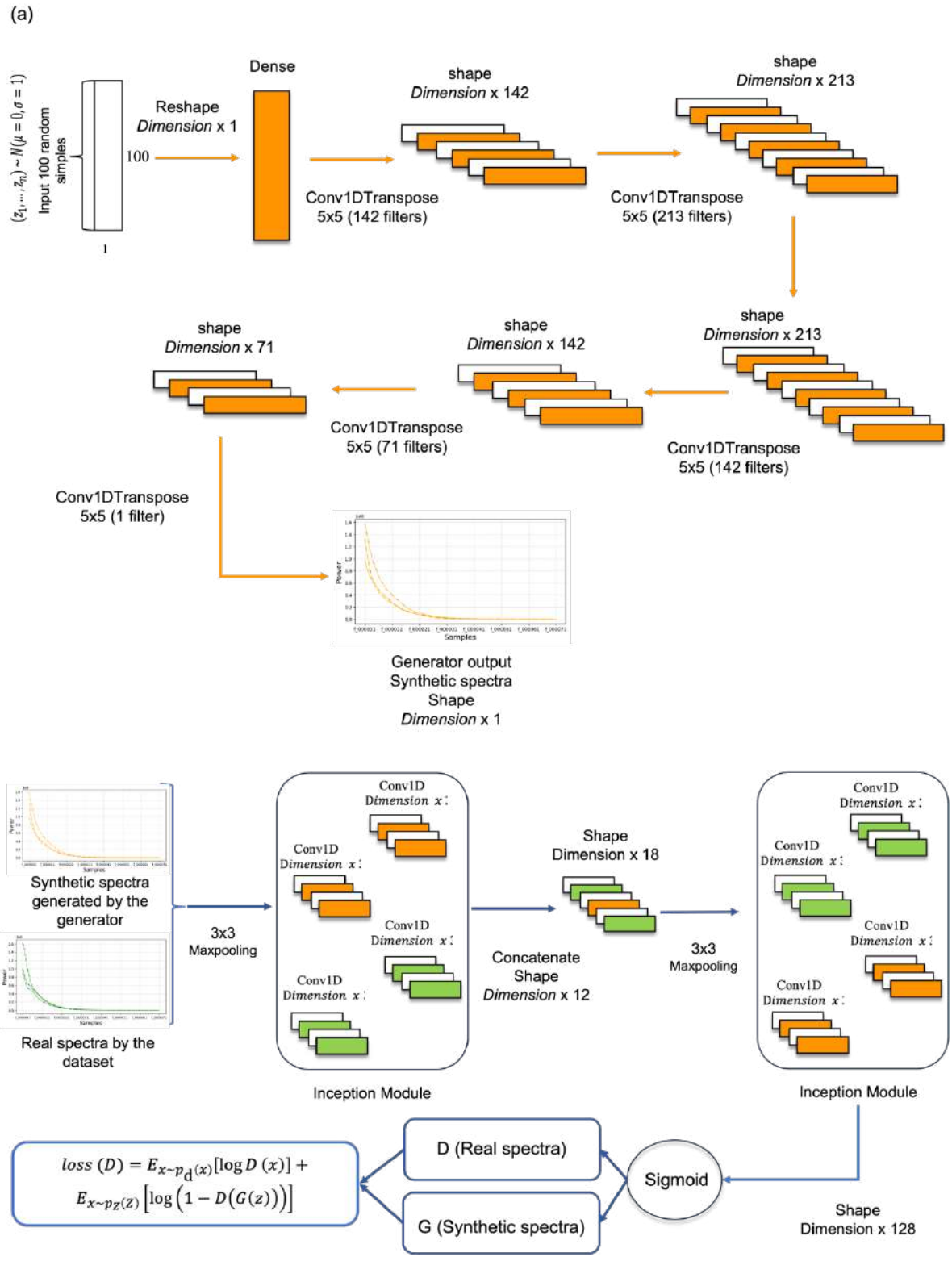


Figura 4.3: *Arquitectura espectraGAN. a) Modelo generador. b) modelo discriminador.*

La **Figura 4.4** muestra el Algoritmo 2 o pseudocódigo empleado por *espectraGAN* para crear espectros sintéticos para los distintos grupos etarios. Este pseudocódigo detalla cómo el algoritmo recibe como entradas los datos específicos del grupo etario, el conjunto de datos *outliers* identificados previamente (**Algoritmo 1**), y los hiper parámetros definidos en el entrenamiento del modelo.

El procedimiento inicia transformando los espectros de los *dataframes* a arreglos de NumPy. Esta conversión se debe a las ventajas significativas que ofrecen los arreglos de NumPy en términos de rendimiento y eficiencia en cálculos numéricos, además de una indexación y segmentación versátiles [98]. Creamos un conjunto de datos denominado *spectrums* que contenía el arreglo con la estructura (x,y,z), donde X representa el número de espectros correspondientes a los ERG basales de cada paciente, Y la longitud del espectro en términos de frecuencia y Z es la profundidad del arreglo **(líneas 1-2)**.

Algorithm 2 EpectraGAN

Require: *data_split* (dataframe), *data_outliers* (dataframe), *Epocas* (Número de épocas), *steps* (número de iteraciones), *batch* (Tamaño del batch)

Ensure: Entrenamiento de modelo generador y discriminador

Convertir *data_split* a un arreglo de numpy

Expandir la dimensión del arreglo y crear *spectrums*

function GENERAR N MUESTRAS REALES(*data_outliers*, *spectrums*, *n*)

  Encontrar el mínimo y máximo espectro en *spectrums*

  $min,\ max \leftarrow$ Índices con mayor y menor amplitud

  Recuperar muestras de ERG con min y max valores

  $Muestras\_aleatorias \leftarrow$ de *spectrums* tomar N muestras aleatorias

  $X \leftarrow stack$(Muestras_outliers, min, max,Muestras_aleatorias)

  $Y \leftarrow$ generar la clase 1 (Real)

  return X (Muestras reales), Y (Clase real)

end function

function GENERAR MUESTRAS FALSAS(*Generador*, *num_datos_sinteticos*)

  Generar ruido aleatorio $N(\mu = 0, \sigma = 1)$

  Usar *Generador* para crear muestras sintéticas

  $X \leftarrow$ muestras sintéticas

  $Y \leftarrow$ generar la clase 0 (Falsa)

  return X (Muestras falsas), Y (Clase falsa)

end function

function ARQUITECTURA GENERADOR

  Definir el modelo generador "Decoder"

  return Modelo Generador

end function

function ARQUITECTURA DISCRIMINADOR

  Definir la arquitectura usando modulos "Inception"

  return Modelo discriminador

end function

Definir funciones de decaimiento exponencial para las tasas de aprendizaje

Definir optimizadores de las tasas de aprendizaje

for *epoch* $\in$ range(*epoch*) **do**

  for *steps* $\in$ range(*steps*) **do**

    $X\_real, Y\_real \leftarrow$ **Funcion** Generar muestras reales

    $X\_false, Y\_false \leftarrow$ **Funcion** Generar muestras falsas

    $X\_train \leftarrow$ Stack (X_real, X_false)

    $Y\_train \leftarrow$ Stack (Y_real, Y_false)

    Actualizar el discriminador del gradiente descentente

$$\theta_d \leftarrow \theta_d + \text{ADAM}\left(\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\left[D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right]\right)$$

    Actualizar el generador del gradiente ascendente

$$\theta_g \leftarrow \theta_g - \text{ADAM}\left(\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right)$$

  end for

end for

Figura 4.4: *Pseudocódigo del Algoritmo 2 (Entrenamiento espectraGAN).*

En el desarrollo del pseudocódigo, se incorporaron diversas funciones clave para optimizar el proceso de entrenamiento. La función inicial, denominada *Generar N muestras reales*, implicó la selección de N muestras del conjunto de datos real para ser clasificados como auténticas, asignándoles la etiqueta 1. Para maximizar la generalización de los datos, se tomaron en consideración tres aspectos: los espectros con valores mínimos y máximos, los espectros más alejados y más cercanos a la concentración de los datos, y espectros

aleatorios seleccionadas de entre la cantidad N seleccionadas. El resultado de esta función es un conjunto de datos nombrados X, que incorporaron los aspectos mencionados, junto con su correspondiente conjunto de etiquetas Y, marcando estas muestras como auténticas **(línea 3 - 11)**. Por otro lado, la función *Generar N muestras falsas* se encargó de producir un conjunto de muestras sintéticas, iniciando con la generación de un vector de ruido aleatorio, o espacio latente, basado en N muestras y siguiendo $N(\mu = 0, \sigma = 1)$. La arquitectura de la red *generador* se emplea para transformar este ruido en muestras sintéticas, las cuales se agrupan en un conjunto de datos X, asignándoles la etiqueta 0 para indicar que son muestras falsas, y su correspondiente conjunto de etiquetas Y **(línea 12 - 18)**.

La red *generador* se inspiró en el modelo *Decoder*, que se emplea para convertir representaciones latentes en secuencias temporales [98]. Por su parte, la red *discriminador* contiene módulos *inception* para evaluar si las muestras son reales o sintéticas, facilitando así la distinción entre los dos tipos de datos generados durante el entrenamiento **(línea 19 - 26)**.

4.2.2 Consideraciones para el entrenamiento

La estabilidad durante el entrenamiento de GANs representó un reto significativo, debido a la naturaleza dinámica de entrenar dos redes en competencia [24]. Para abordar esto, se implementó una técnica conocida como "decaimiento exponencial" [99] para ajustar la tasa de aprendizaje a medida que avanzaba el entrenamiento. Esta técnica permite un ajuste más refinado de la tasa de aprendizaje en función del número de iteraciones. El decaimiento exponencial reduce gradualmente la tasa de aprendizaje, lo que puede mitigar los problemas de sobreajuste y contribuir a que el modelo alcance los mínimos de la función de costo de manera más efectiva [99]. La reducción de la tasa de aprendizaje se rige por la fórmula (**Ecuación 4**):

$$lr(t) = lr_{inicio} \times decay\_rate^{\left(\frac{t}{decay\_steps}\right)} \qquad (4)$$

En esta fórmula, $lr(t)$ denota la tasa de aprendizaje en el tiempo t; lr_{inicio} es la tasa de aprendizaje inicial; $decay\_rate$ es el factor de decaimento; y $decay\_steps$ es el número de pasos tras los cuales se aplica el decaimiento (línea 27 de la **Figura 4.4**).

Una vez establecida la tasa de aprendizaje y el factor de decaimiento, se seleccionaron optimizadores específicos para ajustar los parámetros del modelo con el objetivo de minimizar (o maximizar) la función de pérdida (**Ecuación 1)**. Es crucial señalar que, aunque los optimizadores utilizados para el *discriminador* y el *generador* en el modelo *espectraGAN* son los mismos para ambas redes, funcionan de manera independiente para cada una de ellas. Esto refleja la necesidad de entrenar ambas por separado y con objetivos diferentes: el *generador* debe crear espectros lo más reales posibles y el *discriminador* debe identificar estos espectros sintéticos. Este enfoque garantiza que tanto el *generador* como el *discriminador* mejoren de forma equilibrada y coherente durante el proceso de entrenamiento.

En el entrenamiento del modelo *espectraGAN* se consideraron tres aspectos importantes para asegurar que cada lote contuviera datos con mayor generalización de las características de los espectros o, en otras palabras, que los espectros generados sintéticamente se asemejara lo máximo posible a la base de datos original.

En primer lugar, con el objetivo de definir los límites para la generación de datos sintéticos en el modelo, se estableció un rango de potencia basado en la detección de los valores mínimos y máximos para cada grupo etario. Los espectros se consideraron como series temporales, por lo que la identificación de las potencias extremas se llevó a cabo localizando los picos más altos y bajos de cada espectro.

Para implementar este procedimiento, se diseñó el algoritmo mostrado en la **Figura 4.5.** Este algoritmo recibe como entrada los datos correspondientes a un grupo etario específico y tienen como propósito identificar y almacenar las potencias mínimas y máximas de cada espectro, así como los índices de los espectros que corresponden a dichos valores. El algoritmo se inicia estableciendo dos variables auxiliares (*min_value* y *max_value*), las cuales se inicializan con valores infinitos opuestos para asegurar que cualquier potencia observada en el primer recorrido las sustituya. Simultáneamente, se crean dos listas vacías (*min_spec_index* y *mas_spec_index*) donde se guardan los índices de los espectros con las potencias mínimas y máximas detectadas, respectivamente.
Luego, el algoritmo recorre cada espectro dentro del grupo etario y calcula la potencia mínima (*current_min*) y máxima (*current_max*) del espectro en cuestión. Si *current_min* es

menor que *min_value*, *min_value* se actualiza con este nuevo valor, y el índice del espectro se almacena en la lista *min_spec_index*. Asimismo, si *current_max* supera a *max_value*, *max_value* se actualiza y se guarda el índice en *max_spec_index*. Este proceso se repite para cada espectro hasta que se identifican las potencias extremas del grupo etario y se registran los correspondientes índices. Finalmente, el algoritmo retorna como salida los valores de las potencias mínima y máxima, junto con las listas de índices que indican los espectros asociados a dichos valores.

Algorithm 3 Min Max Espectros

Require: *data* (lista de los espectros)
Ensure: *min_value*, *max_value*, *min_spec_index* list, *max_spec_index* list

```
    function MinMaxSpectrum(data)
        min_value ← ∞
3:      max_value ← −∞
        min_spec_index ← Inicializar una lista vacía
        max_spec_index ← Inicializar una lista vacía
6:      for spec_index, spec ∈ enumerate(data) do
            current_min ← min(espectra)
            current_max ← max(espectra)
9:          if current_min ≤ min_value then
                min_value ← current_min
                Append spec_index to min_spec_index
12:         end if
            if current_max ≥ max_value then
                max_value ← current_max
15:             Append spec_index to max_spec_index
            end if
        end for
18:     return min_value, max_value, min_spec_index, max_spec_index
    end function
```

Figura 4.5: *Pseudocódigo del algoritmo 3 para detectar los espectros con la mínima y máxima potencia por cada grupo etario.*

En segundo lugar, se consideraron tanto los espectros más alejados como los más cercanos a la concentración de los datos. Para localizar estos grupos de espectros más alejados (outliers), se utilizó la implementación del **algoritmo 2**. Aplicando la transformación de los espectros mediante la reducción de dimensionalidad con PCA antes introducida, se obtuvo una representación tridimensional donde cada punto corresponde a la reducción del espectro de cada paciente. Con esta distribución de los espectros, se identificaron los outliers utilizando el parámetro σ. Se calculó la distancia euclidiana de cada espectro al centroide (media de los datos). El umbral para los outliers se definió como

cualquier punto a una distancia superior a σ. Por otra parte, para encontrar la concentración de los espectros más cercanos a la media de los datos, se consideraron todas las distancias por debajo del parámetro σ .

En tercer lugar, se consideró solo el 10 % del porcentaje restante de los espectros de manera aleatoria durante cada época de entrenamiento para garantizar una mayor diversidad de los datos de entrada.

En términos generales, la distribución de porcentajes para la selección de espectros durante el entrenamiento del modelo *espectraGAN* (35 %, 35 % y 10 %) fue arbitraria, pero razonada. Se fundamentó en la necesidad de exponer al modelo a una muestra representativa y equilibrada de los datos, asegurando tanto la captura de patrones dominantes como la capacidad de generalización frente a casos diversos. El 35 % de los espectros se compone de outliers validados, con el objetivo de garantizar que el modelo aprenda a manejar variaciones extremas que pueden presentarse en datos reales con alta variabilidad en las señales [100]. Consideramos que elegir un porcentaje menor hubiera limitado la exposición del modelo a estos casos extremos, reduciendo su capacidad para adaptarse a situaciones de alta incertidumbre.

Por otro lado, el otro 35 % de los espectros está conformado por muestras cercanas a la media. Este grupo representa patrones más frecuentes y estables de la señal, esenciales para que el modelo capte características predominantes y no se sesgue hacia los casos extremos. De acuerdo con [100], la inclusión de una cantidad robusta de espectros promedio permite al modelo entender las características más comunes de los datos y evita el sobreajuste a situaciones específicas. Usar un porcentaje mayor, como un 40 % podría sesgar el entrenamiento hacia los patrones más comunes, dificultando que el modelo reconozca casos menos frecuentes pero relevantes.

Finalmente, el 10 % de los datos se seleccionó de manera aleatoria del 30 % restante, lo que permitió introducir heterogeneidad en el conjunto de entrenamiento. Este enfoque ayuda a mejorar la robustez del modelo al incluir espectros diversos que no caen necesariamente en las categorías de outliers o promedio. La aleatorización con un conjunto

de datos que varía entre épocas, fomenta un aprendizaje más generalizado y flexible [24], [101].

Continuando con el algoritmo 2 Pseudocódigo de *espectraGAN*, estos espectros mencionados anteriormente, se obtienen utilizando la función *Generar N muestras reales* conteniendo un conjunto de datos *(X_real, Y_real),* asignándoles la etiqueta 1 para indicar que son espectros auténticos (línea 32, **Figura 4.4**). De manera similar (línea 33, **Figura 4.4**), la función *Generar N muestras falsas* permitió generar espectros falsos *(X_false, Y_false)* en una cantidad equivalente al lote de entrenamiento original, asignándoles la etiqueta 0 para diferenciarlos de los espectros reales. Es decir, una vez obtenido el lote de datos reales, se requiere un lote de datos sintéticos que contengan la misma cantidad de espectros que el conjunto de datos originales. De esta manera, el lote final utilizado por *espectraGAN* durante el entrenamiento contiene el doble de espectros que el lote real correspondiente a cada grupo etario.

La decisión de utilizar un lote combinado con un 50 % de espectros reales y un 50 % de espectros sintéticos responde a la necesidad de un balance entre ambos tipos de datos durante el entrenamiento del modelo *espectraGAN*, para que este tenga una exposición equitativa tanto a ejemplos orgánicos como sintéticos [101]. Este enfoque balanceado permite optimizar tanto la precisión como la generalización del modelo, garantizando que EspectraGAN no se sobreentrene en los datos reales ni se sobreentrene en los datos sintéticos.

4.2.3 Análisis de las funciones de pérdida

El entrenamiento comenzó con la actualización del *discriminador* (D), cuyo objetivo es maximizar la probabilidad de asignar la etiqueta de ambos espectros, tanto los reales como los generados por el *generador* (G). Esta etapa se realizó mediante un proceso de descenso del gradiente, ajustando los parámetros del *discriminador* (D), para minimizar su función de pérdida. Para este propósito, se empleó el optimizador ADAM (Adaptive Moment Estimation, línea 35) que ajusta dinámicamente los gradientes de los parámetros a lo largo del proceso de entrenamiento [102], y es reconocido por combinar las ventajas de los algoritmos AdaGrad y RMSProp, logrando un balance entre alta eficiencia computacional y un bajo consumo de memoria. La función de pérdida del *discriminador* es el promedio de la

suma de dos términos: el logaritmo de la probabilidad de que D asigne los espectros reales y el logaritmo de uno menos la probabilidad de que D asigne a los espectros generados por G de ser reales [24].

Paralelamente, se llevó a cabo el entrenamiento del *generador* con un proceso de ascenso del gradiente. El *generador* (G) se actualiza con la meta de confundir al discriminador para que considere los espectros sintéticos como auténticos. Para ello, se modifican los parámetros del *generador* (G) para maximizar la probabilidad de que el *discriminador* cometa errores al clasificar estos espectros generados como falsos. Esta fase se realizó también con el optimizador ADAM (línea 36, **Figura 4.4**). La función de pérdida del *generador* se basa en el logaritmo de uno menos la probabilidad de que el *discriminador* (D) identifique los espectros generados por *generador* (G) como falsos. El propósito es aumentar este valor lo máximo posible, lo cual sucede cuando el *discriminador* (D) clasifica incorrectamente los espectros generados como verdaderos.

Cabe enfatizar que la función de pérdida del *generador* se enfoca exclusivamente en los espectros sintéticos, dado que el propósito del *generador* es fabricar espectros falsos que sean clasificados como reales por el *discriminador*, mientras que la función de pérdida del *discriminador* evalúa tanto los espectros reales como los falsos, ya que su función es la correcta clasificación de ambos.

4.2.4 Generación de espectros sintéticos

Con base en el análisis exploratorio, se definieron los grupos etarios correspondiente a las clases *control* y *enfermos.* A cada uno de estos grupos se les realizó el análisis PCA para examinar su estructura morfológica, revelando estructuras distintivas en algunos grupos específicos. En particular, estas diferencias estructurales fueron más evidentes en el grupo etario 55-64 años de la clase *control*, y en el grupo etario 45-54, 55-64, 65-74 y más de 85 años de la clase *enfermos*. Para incluir esta variabilidad en la generación de los datos sintéticos, estos datos mostraron una estructura dividida en dos grupos, siendo uno de ellos un subgrupo claramente alejado del centroide del conjunto principal. Al calcular la distancia euclidiana entre estos datos atípicos y el centroide, se observó que dicha distancia superaba la desviación estándar, representando aproximadamente el 30 % de los datos. Debido a esto,

dicho subgrupo fue entrenado utilizando un modelo generador independiente. En la **Figura 4.6** presenta un diagrama de flujo que detalla cómo se realizó la generación específica de estos grupos etarios particulares. La principal diferencia respecto a los demás grupos etarios es el uso de un único modelo generador en lugar de dos, lo cual está ilustrado mediante el cuadro y líneas violetas en dicha figura.

Este proceso comenzó con la generación de un vector de ruido aleatorio con una distribución normal $N(\mu = 0, \sigma = 1)$ para cada uno de los generadores, similar al utilizado en el entrenamiento del modelo *espectraGAN*. Estos vectores se introducen en los modelos *generador* del *espectraGAN,* el cual, utilizando los pesos entrenados almacenados en el archivo .h5, se aplicó la función de predicción. El formato .h5 ofrece varias ventajas significativas para este propósito. En primer lugar, es altamente compatible con diversas bibliotecas de aprendizaje automático, como TensorFlow y Keras, facilitando la carga y guardado de modelos sin pérdida de información [92]. Otra ventaja importante es la capacidad de almacenar no sólo los pesos de la red neuronal, sino también la estructura del modelo y cualquier otro metadato necesario, permitiendo una reutilización y transferencias de modelos más sencilla entre diferentes entornos de desarrollo [92]. Cada modelo generador transformó el vector de ruido aleatorio en espectros sintéticos, específicos para grupo etario.

Siguiendo el diagrama de flujo, una vez generados los espectros sintéticos, se implementó un filtro, el cual se explica más a detalle en el siguiente subcapítulo, para garantizar la fiabilidad del conjunto de espectros sintéticos del grupo etario en cuestión, con base al cálculo del RMSE se obtuvieron los espectros sintéticos con mayor similitud a los orgánicos. Posteriormente, se compararon las distribuciones de los datos mediante la prueba Mann-Whitney U, una técnica no paramétrica que permite determinar si dos muestras independientes provienen de la misma distribución [103]. Adicionalmente, se realizaron comparativas visuales directas 1:1 entre los espectros orgánicos y sintéticos para confirmar la calidad y fiabilidad de los datos generados.

Finalmente, se definió un factor multiplicador que indica la proporción de espectros sintéticos generados respecto a los orgánicos. Por ejemplo, un factor de uno implica la generación del mismo número de espectros sintéticos que orgánicos. que determina la cantidad de espectros sintéticos a generar en relación con los espectros reales. Por ejemplo,

un factor de 1 genera una cantidad igual de espectros sintéticos con respecto al grupo etario seleccionado, mientras que un factor de 10 genera diez veces más espectros sintéticos que los orgánicos. Este proceso se realizó para cada grupo etario obteniendo así una base de datos balanceada x1, x10 y x100 veces la cantidad de los datos orgánicos.

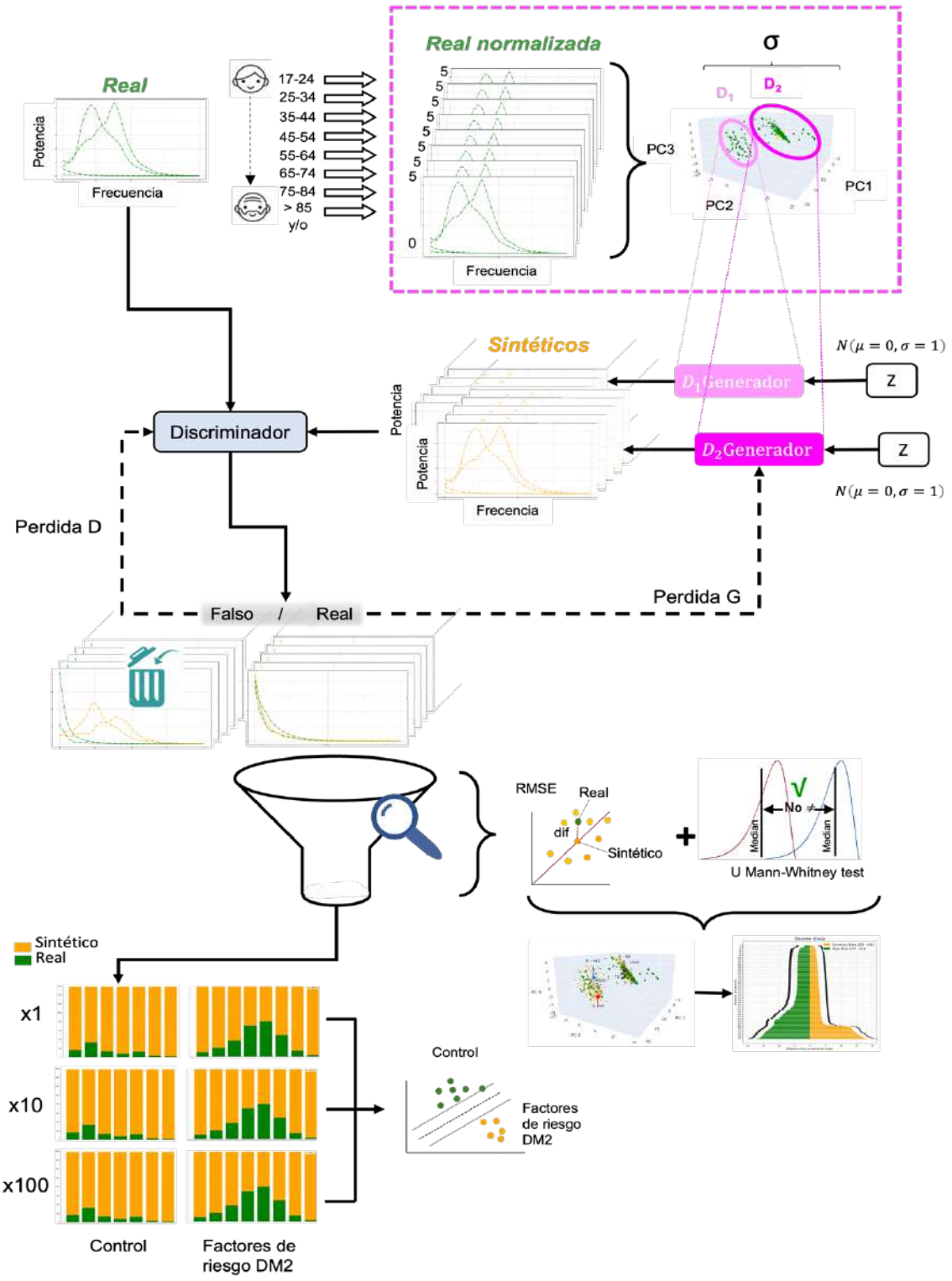


Figura 4.6: *Diagrama de flujo para la generación de espectros sintéticos mediante espectraGAN.*

4.2.5 Filtrado del conjunto de los espectros sintéticos

El RMSE mide la diferencia promedio entre los valores reales observados (R) y los valores predichos (S) por el modelo [104] . Específicamente, el RMSE se calcula tomando la diferencia entre cada valor real y su correspondiente valor predicho, elevando estas diferencias al cuadrado, promediando los resultados y finalmente tomando la raíz cuadrada del promedio. Esta métrica proporciona una medida cuantitativa del error promedio en las predicciones del modelo, donde valores más altos indican mayores discrepancias entre los espectros orgánicos y sintéticos, mientras más bajos reflejan un mejor ajuste entre ellos [104]. La **ecuación 4**, muestra la implementación de esta métrica:

$$RMSE = \sqrt{(\sum_{i=1}^{N}(R_i - S_i)^2) \,/\, N} \qquad (4)$$

Descripción de la Ecuación:

- R_i : valor real del espectro en el punto i
- S_i : valor sintético por el modelo en el punto i
- N : Número total de puntos en el espectro

Dado este contexto, y consideraciones de dos conjuntos de datos independientes para los datos reales $X = \{x_1, x_2, \ldots, x_n\}$ y los datos sintéticos $Y = \{y_1, y_2, \ldots, y_m\}$, la prueba evalúa la hipótesis nula:

$$H_o: F_{real}(x) = F_{sintético}(y)$$

Donde F_{real} y $F_{sintético}$ representan las distribuciones de las muestras X y Y, respectivamente. Es decir, se evalúa si ambas muestras provienen de la misma distribución.

El estadístico U se calcula a partir de los rangos de los datos combinados de las muestras. Para ello, primero se ordenan los datos de ambas muestras juntas y se les asignan rangos $R(x_i)$ y $R(y_j)$. Luego, el estadístico se calcula como (**ecuación 5 y 6**):

$$U_x = n \cdot m + \frac{n(n+1)}{2} - \sum_{i=1}^{n} R(x_i) \qquad (5)$$

$$U_y = n \cdot m + \frac{m(m+1)}{2} - \sum_{j=1}^{m} R(y_j) \quad (6)$$

Donde:

- U_x es el estadístico U calculado sobre la muestra X
- U_y es el estadístico U calculado sobre la muestra Y
- n es el tamaño de la muestra X
- m es el tamaño de la muestra Y

El menor de U_x y U_y se utiliza como estadístico de prueba, es decir:

$$U = min(U_x, U_u)$$

El p-valor asociado con el estadístico U se calcula utilizando distribuciones de referencia o aproximaciones basados en la distribución normal cuando los tamaños de las muestras son grandes. Dado un nivel de significancia α, se comparan el p-valor con α para decidir si se rechaza la hipótesis nula:

$$Si\ p\_valor \leq \alpha \rightarrow Rechazar\ H_p$$

$$Si\ p\_valor > \alpha \rightarrow No\ rechazar\ H_p$$

Si el p-valor obtenido para cada par de datos es mayor que α, no se rechaza la hipótesis nula, lo que significa que el modelo ha generado datos sintéticos que son estadísticamente similares a los datos reales.

El **algoritmo 4** de la **Figura 4.7** presenta la implementación del cálculo del RMSE. El resultado es una lista de valores RMSE (diccionario), cada uno asociado al índice de un paciente. Se identifican los casos cuyos espectros orgánicos tienen mayor similitud con los espectros sintéticos, tomando como referencia el valor mínimo, a saber, más cercano a 0 de RMSE, creando así parejas de mayor similitud a nivel cuantitativo.

Algorithm 4 Calcular RMSE

```
Require: data_real, data_synthetic
Ensure: metrics_rmse, averages_rmse, index_min
function CALCULAR_RMSE(data_real, data_synthetic)
    def RMSE(real_data, data_synthetic):
        MSE ← mean_squared_error(data_synthetic, real_data)
        return math.sqrt(MSE)
    def Calcular_RMSE_por_Fila_Sintética(row_syn, real_data):
        rmse_values ← real_data.apply(λ row_real: RMSE(row_real, row_syn), axis=1)
        average_rmse ← rmse_values.mean()
        return rmse_values.to_dict(), average_rmse
    metrics_rmse ← {}
    averages_rmse ← []
    index_min ← []
    for index_s, row_syn in tqdm(data_synthetic.iterrows()) do
        metrics_rmse[index_s], average_rmse ← Calcular_RMSE_por_Fila_Sintética(row_syn, data_real)
        averages_rmse.append(average_rmse)
    end for
    min_value ← float('inf')
    min_indices ← None
    for index_s, inner_dict in tqdm(metrics_rmse.items()) do
        for index_r, rmse_value in inner_dict.items() do
            if rmse_value ¡ min_value then
                index_min.append((index_r, index_s))
                min_value ← rmse_value
                min_indices ← (index_r, index_s)
            end if
        end for
    end for
    return metrics_rmse, averages_rmse, index_min
end function
```

Figura 4.7: *Algoritmo 4, implementación para la comparación de espectros sintéticos con los orgánicos mediante la métrica RMSE.*

A partir de los espectros sintéticos con mayor similitud a los espectros orgánicos (determinados por índices de similitud con el menor valor de RMSE) y la prueba Mann-Whitney U, se desarrolló un algoritmo denominado *"Filtrar dataframe"*. El algoritmo inicia creando una variable para almacenar los espectros sintéticos que superan la prueba de Mann-Whitney U. Posteriormente, el algoritmo compara cada par de espectros a partir del diccionario de los índices de similitud (**Algoritmo 4**), designado como **X** al espectro orgánico y como **Y** al espectro sintético. Se aplica la prueba de Mann-Whitney U para cada par. Si el valor obtenido supera el nivel de significancia α (0.05), no se rechaza la hipótesis nula, lo que indica que ambos espectros provienen de la misma distribución. Esto respalda que el modelo genera datos sintéticos estadísticamente similares a los orgánicos, con un margen de error aceptado del 5 % para un error tipo I (rechazar incorrectamente la hipótesis nula). La **Figura 4.8** muestra el diagrama de flujo de la implementación del algoritmo 5.

Algorithm 5 Filtrar Dataframe

Require: $rmse\_dict, data\_real, data\_synthetic, \alpha$ (default 0.05)
Ensure: $data\_min\_list$ dataframe

```
function FILTRAR_DATAFRAME(rmse_dict, data_real, data_synthetic, α)
    Importar PruebaUdeMann − Whitney
    data_min_list ← Inicializar una lista vacia
    for all idx_syn ∈ data_synthetic do
        idx_min ← Obtener el indice con el min RMSE de rmse_dict para cada idx_syn
        X ← Convertir data_real de idx_min a un arreglo
        Y ← convertir data_synthetic de idx_syn a un arreglo
        stat, p_value ← mannwhitneyu(X, Y)
        if p_value > α then
            Agregar data_synthetic de idx_syn a data_min_list
        end if
    end for
    data_min_list ← Crear un Dataframe de la lista
    return data_min_list
end function
```

Figura 4.8: *Algoritmo 5, implementación del algoritmo 5 para filtrar el conjunto de datos sintéticos a partir de la prueba de Mann-Whitney U.*

4.3 Evaluación estadística de los espectros

Se completó la evaluación estadística de los espectros sintéticos con su comparación con los orgánicos desde un enfoque visual, cualitativo y cuantitativo. Este análisis buscó verificar la coherencia entre los espectros generados y las características principales de los datos originales, validando su precisión y aplicabilidad en estudios posteriores. Además, se emplearon diversas herramientas y métricas estadísticas para proporcionar una evaluación objetiva de la calidad de los espectros.

4.3.1 Comparación visual de los espectros sintéticos frente a los orgánicos

La representación gráfica de los espectros permitió identificar de manera inmediata las similitudes y diferencias entre ambos conjuntos de datos. A través de la visualización de curvas y patrones, se evaluó el grado de aproximación de los espectros sintéticos respecto a los originales. Este análisis preliminar proporcionó una visión intuitiva sobre la validez del modelo y su capacidad para replicar las características generales de los espectros orgánicos.

Al obtener el conjunto de datos sintéticos correspondiente al grupo etario seleccionado, este se concatenó con el conjunto de datos orgánicos de la clase analizada (ya sea de *control* o de *enfermos*), incluyendo tanto el grupo etario elegido como los espectros

restantes. Se aplicó nuevamente el PCA como al principio del análisis de los espectros, pero ahora considerando el grupo etario sintético. Esto con la finalidad de realizar una comparación global entre ambos grupos (orgánico y sintético).

Después de aplicar el PCA a la base de datos y seleccionar el grupo etario sintético junto con su grupo etario orgánico correspondiente, se procedió a graficar utilizando los mismos componentes principales obtenidos en el análisis inicial.

Por otro lado, se graficaron espectros de casos individuales, seleccionados aleatoriamente dentro del espacio de los componentes principales previamente seleccionados, con el fin de verificar que los espectros obtenidos, acorde a la distancia euclidiana más cercana, presentaban mayor similitud sin ser réplicas exactas generadas por el modelo *espectraGAN*. Con este tipo de gráficos, se compararon los espectros de potencia uno a uno. Además, para realizar un análisis más detallado de las similitudes y los rangos de potencia, ambos grupos etarios se normalizaron en un rango de 0 a 5, lo que permitió observar las morfologías desde otra perspectiva [105].

Finalmente, para comprender mejor cómo se distribuyen las distancias entre los datos sintéticos generados por el modelo, se realizó un gráfico adicional que muestra la distribución de las distancias de los espectros generados en comparación con los datos orgánicos. Este gráfico se obtuvo a partir del análisis PCA con los componentes principales seleccionados, y se calculó la distancia euclidiana de cada punto (orgánico y sintético) con respecto al centroide orgánico. El objetivo fue identificar qué espectros están más cercanos al centroide y, por lo tanto, evaluar con mayor precisión su similitud y diferencias a nivel local y global [105].

4.3.2 Análisis Cualitativo

El análisis cualitativo profundizó las características visuales observadas, como la forma de las señales, los picos de amplitud y la continuidad de los espectros. Para validar la efectividad del modelo *espectraGAN*, se emplearon dos técnicas de reducción de dimensionalidad: PCA antes introducido y el algoritmo de aprendizaje no supervisado t-SNE. Estas técnicas permiten visualizar y explorar los datos en un espacio de menor dimensión, preservando al mismo tiempo las características esenciales de los datos originales [86]. Como

ya dicho antes, PCA busca preservar la varianza de los datos al identificar las direcciones de máxima variabilidad, proporcionando una representación de los datos en función de sus componentes principales [91]. El enfoque no lineal de t-SNE se centra en mantener las relaciones de similitud entre las instancias, lo cual es especialmente útil para identificar agrupaciones complejas y patrones no lineales en los datos [86], [106].

4.3.3 Análisis Cuantitativo

Se calcularon métricas y estadísticas de error, similitud y variabilidad para determinar objetivamente en qué medida los espectros generados se asemejan a los espectros orgánicos. Así, el análisis cuantitativo no solo complementa el análisis cualitativo, sino que también proporciona una validación objetiva y reproducible de la eficacia del modelo [24]. Las métricas utilizadas fueron el MAE, el FID y PRD.

El MAE se calcula tomando el promedio de las diferencias absolutas entre los valores del espectro original y espectro sintético. Esta métrica evalúa la precisión promedio de las señales generadas, midiendo qué tan cercanas son las amplitudes de ambas señales en cada punto. Un valor bajo de MAE refleja una mayor similitud entre los espectros comparados [104]. La **ecuación 8**, muestra la implementación de esta métrica:

$$MAE = \frac{1}{n}\sum_{i=1}^{N}|R_i - S_i| \qquad (8)$$

Descripción de la Ecuación:

R_i : valor real del espectro en el punto i

S_i : valor sintético por el modelo en el punto i

N : Número total de puntos en el espectro

El PRD se calcula tomando la raíz cuadrada del cociente entre la suma de las diferencias cuadradas de los espectros (original y sintético) y la suma de los valores cuadrados del espectro original. Para expresarlo en porcentaje, este valor se multiplica por 100. Esta métrica evalúa las diferencias entre dos señales, el espectro real y el espectro sintético, de forma puntual. Esta medida es especialmente útil porque cuantifica el grado de

similitud o discrepancia entre ambas señales [107]. Su relevancia radica en que permite comparar tanto la forma general como la proporción relativa de amplitud de las señales, garantizando que la estructura del espectro conserve una coherencia con la señal original. La **ecuación 7**, muestra la implementación de esta métrica:

$$PRD = \sqrt{100 \frac{\sum_{i=1}^{N}(R_i - S_i)^2}{\sum_{i=1}^{N}(R_i)^2}} \qquad (7)$$

Descripción de la Ecuación:

R_i : valor real del espectro en el punto i

S_i : valor sintético por el modelo en el punto i

N : Número total de puntos en el espectro

El FID se basa en la distancia entre las medias y las covarianzas de las distribuciones de características de los espectros orgánicos (μ_r, Σ_r) y los sintéticos (μ_s, Σ_s). Es una métrica utilizada para comparar las distribuciones de los espectros reales y sintéticos. Se basa en la evaluación de espectros globales como la similitud en la estructura de los picos, así como la dispersión en términos de amplitud y frecuencia [108]. Un valor bajo de FID indica que los espectros generados presentan una distribución de características similar a la de los espectros reales. La **ecuación 9**, muestra la implementación de la métrica:

$$FID = || \mu_r - \mu_s ||^2 + T_r(\Sigma_r + \Sigma_s - 2\,(\Sigma_r \Sigma_s)^{1/2}) \qquad (9)$$

Donde:

μ_r y μ_s son las medias de las distribuciones de características de los espectros reales y sintéticos, respectivamente

Σ_r y Σ_s son las matrices de covarianza de las distribuciones de los espectros reales y sintéticos.

$(\Sigma_r \Sigma_s)^{1/2}$ representa la raíz cuadrada matricial del producto de las covarianzas.

Estas métricas ayudan a identificar áreas de mejora en el proceso de generación, como el ajuste de parámetros, y proporcionaron un marco confiable asegurando que los espectros generados puedan usarse con precisión en aplicaciones posteriores.

4.3.4 Balance de clases

El balance de clases en una base de datos es un aspecto crucial a procurar durante el entrenamiento de modelos de aprendizaje automático [39], [109], [110]. La base de datos de espectros presentaba un desbalance tanto a nivel de clases principales (*control* y *enfermos*) como en las subclases basadas en rango de edades. En este contexto, una de las decisiones clave en el proyecto fue entrenar el modelo *espectraGAN* de forma separada para cada grupo etario dentro de cada clase (*control* y *enfermos*) dividida en grupos etarios para capturar las particularidades de cada uno de estos subconjuntos y, posteriormente, generar un conjunto de datos equilibrado. Para hacer esto, primero duplicamos la cantidad de datos en el grupo que tenía la mayor cantidad de datos (por lo que este grupo tiene 50 % de datos orgánicos y 50 % de datos sintéticos), lo que nos dio un número de datos a alcanzar para equilibrar los otros grupos.

Capítulo 5

Resultados

5.1 Análisis exploratorio de los espectros orgánicos correspondientes al ERG basal de humanos

En la **Figura 5.1** se muestra la distribución de los grupos etarios para cada clase. Se encontró que el grupo predominante dentro de la clase *control* son los casos de 25-34 años, con 328 observaciones, lo que representa el 38.4 % del total de la clase, mientras que el grupo de 75-84 años tiene el menor porcentaje con solo un 1.64 % (14 observaciones). Es importante mencionar la ausencia de datos de ERG basal para casos mayores a 85 años y que en esta clase, el grupo etario con mayor representación es el de 55-64 años, con 790 observaciones (28.40 %). Mientras tanto, el grupo de más de 85 años tiene el menor porcentaje, con solo un 1.29 % (36 observaciones) en la clase de casos *enfermos*.

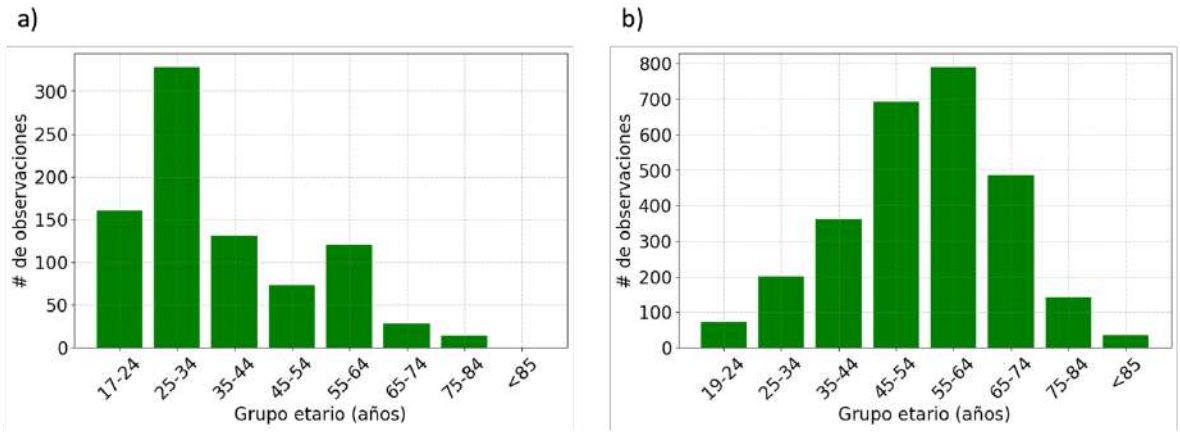


Figura 5.1: *Distribución de grupos etarios para la base de datos de ERG basal. a) Gráfica para la clase ¨Control¨. b) Gráfica para la clase ¨Enfermos¨.*

En relación con el análisis de componentes principales, se generaron gráficas de sedimentación aplicada a la base de datos de ERG basal para cada clase, donde se visualiza

la varianza explicada por cada componente principal (Línea verde), así como la razón de varianza explicada acumulada (Línea naranja) (**Figura 5.2**).

Se observó una disminución pronunciada de la varianza explicada después de los tres primeros componentes. Es decir, a partir del componente 4, la contribución es mínima: en el caso de la clase *control,* solo aporta el 3.21 % y, en el caso de la clase *enfermos,* el 2.27 %. Además, el punto de inflexión se muestra para ambas clases en el componente 3, explicando los tres primeros componentes el 92.58 % y 92.08 % de la varianza total del conjunto de datos para la clase *control* y *enfermos*, respectivamente.

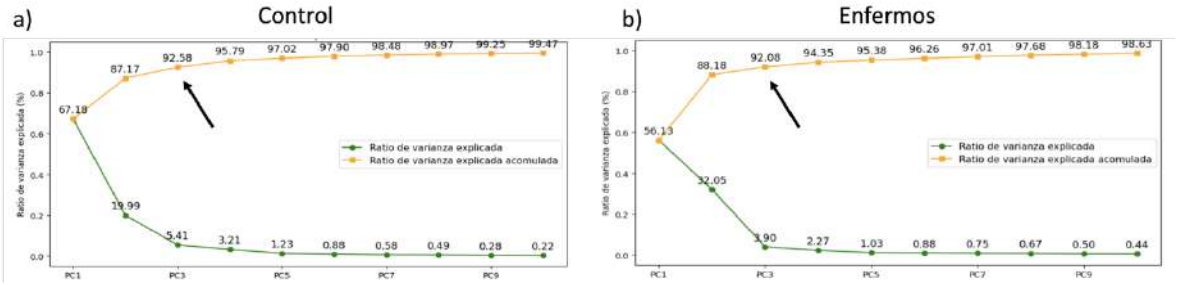


Figura 5.2: *Gráfica de sedimentación de la base de datos de ERG basales. a) Clase 'control'. b) Clase 'Enfermos'. Con flechas negras se muestra el punto de inflexión para cada gráfica de sedimentación. En el eje X del gráfico se representan los componentes principales ordenados de mayor a menor varianza explicada. En el eje Y se representan los valores asociados a cada componente, lo cual refleja la cantidad de varianza total del conjunto de datos.*

En la **Figura 5.3** se presentan gráficos tridimensionales de la dispersión de datos según los tres primeros componentes principales en el conjunto de datos correspondiente a las clases *control* (**Figura 5.3(a)**) y *enfermos* (**Figura 5.3(b)**). Cada punto en estos espacios tridimensionales representa a un paciente, y estos están codificados por colores para diferenciar los grupos etarios, que abarcan desde jóvenes de 17-24 años hasta adultos mayores de más de 85 años. Se pueden apreciar zonas de mayor y menor concentración de datos. Las primeras, indican que los casos dentro de estas zonas comparten patrones o características típicas asociadas a su grupo etario, mientras que las segundas sugieren que los espectros de potencia de los ERG basales de ciertos casos presentan patrones distintos a los predominantes.

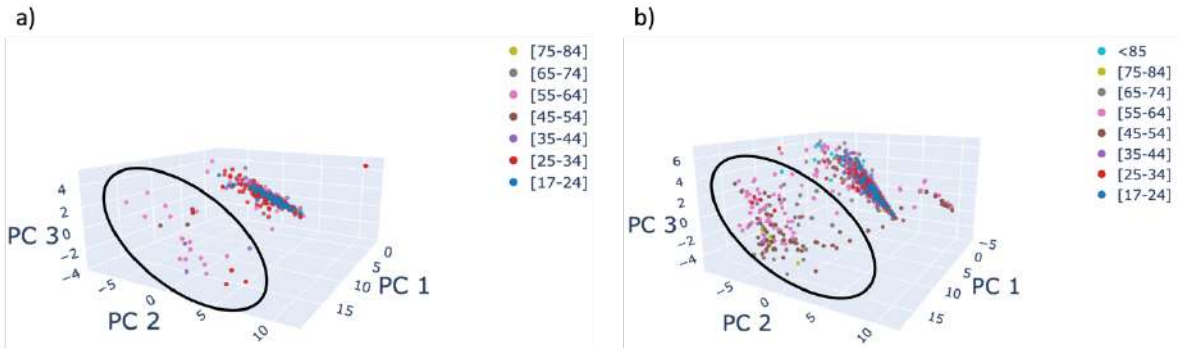


Figura 5.3: *Gráficos en 3 dimensiones (3D) de PCA de los espectros de potencia obtenidos del ERG basal organizados por grupos etarios tal como se describen, correspondiente a la clase a) control y b) enfermos. Las elipses negras resaltan los grupos de puntos que se alejan de la mayor concentración de datos.*

En más detalle, al analizar el espacio tridimensional generado por las mismas componentes, se observó que en la mayoría de los grupos etarios de la clase *control* (**Figura 5.4**) una proporción considerable de muestras se agrupó cerca del centroide. Sin embargo, también se identificaron casos que se alejan significativamente del centroide, distribuyéndose de forma polar o dispersa, con excepción de grupo etario de 55-64, el cual mostró una mayor aglomeración muy alejada al centroide. En la clase *enfermos* (**Figura 5.5**), el comportamiento general fue similar; no obstante, se destacaron como excepciones los grupos etarios 45-54, 55-64, 65-74, los cuales exhibieron una mayor dispersión alejada del centroide principal.

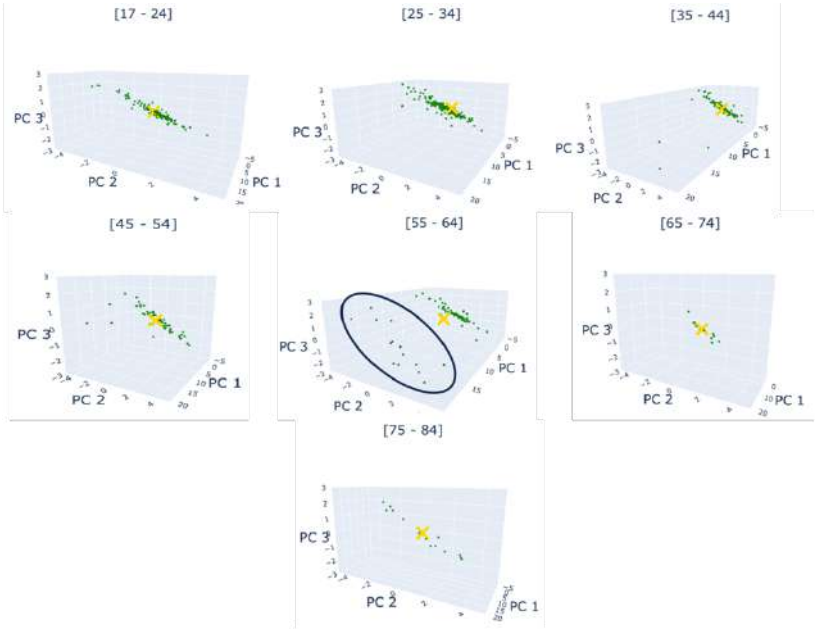


Figura 5.4: *Gráficos en 3D de PCA de los espectros de potencia obtenidos del ERG basal para la clase control organizados por grupos etarios: a) 17 – 24 años, b) 25 – 34 años, c) 35 – 44 años, d) 45 – 54 años, e) 55 – 64 años, f) 65 –*

74 años, g) 75 – 84 años. El centroide está marcado con el símbolo X amarillo , representando la media de los datos. Las elipses negras encierran los casos que se encuentran más alejados con respecto al centroide.

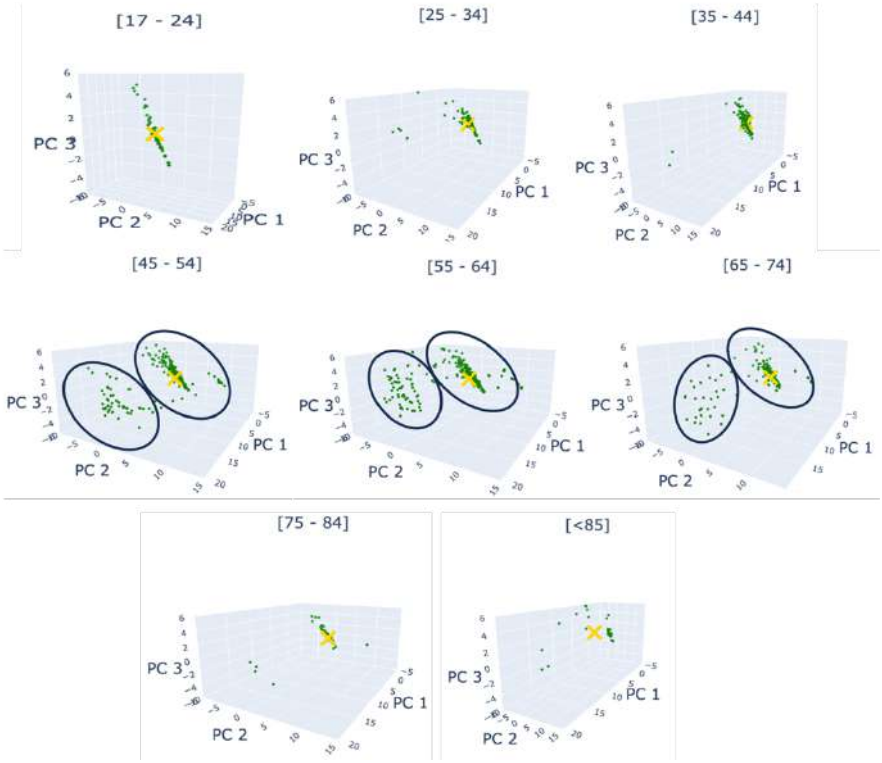


Figura 5.5: *Gráficos en 3D de PCA de los espectros de potencia obtenidos del ERG basal para la clase enfermos organizados por grupos etarios: a) 17 – 24 años, b) 25 – 34 años, c) 35 – 44 años, d) 45 – 54 años, e) 55 – 64 años, f) 65 – 74 años, g) 75 – 84 años. El centroide está marcado con el símbolo X amarillo, representando la media de los datos. Las elipses negras encierran los casos que se encuentran más alejados con respecto al centroide.*

A continuación, se optó por mostrar principalmente los resultados de las implementaciones de cada uno de los algoritmos desarrollados, así como la generación de datos sintéticos en el grupo etario de 17-24 de la clase *control*. Esto debido a que en los otros grupos etarios se aplicaron con los mismos algoritmos y técnicas, y sus resultados no presentan diferencias significativas en comparación con el grupo seleccionado. De cualquier forma, los resultados de los análisis en los demás grupos se encuentran disponibles en el material complementario al final del documento.

Primero, se definieron los límites para la generación de datos sintéticos aplicando el **algoritmo 3** que permite detectar las potencias mínimas y máximas de los espectros presentes en el grupo etario elegido (**Figura 5.6**). Se observa el espectro orgánico correspondiente con el pico de potencia máxima y, por otra parte, en línea roja, el espectro con el pico de potencia

mínimo. Estos espectros se tomaron como límites de rango en el cual se generaron los datos sintéticos.

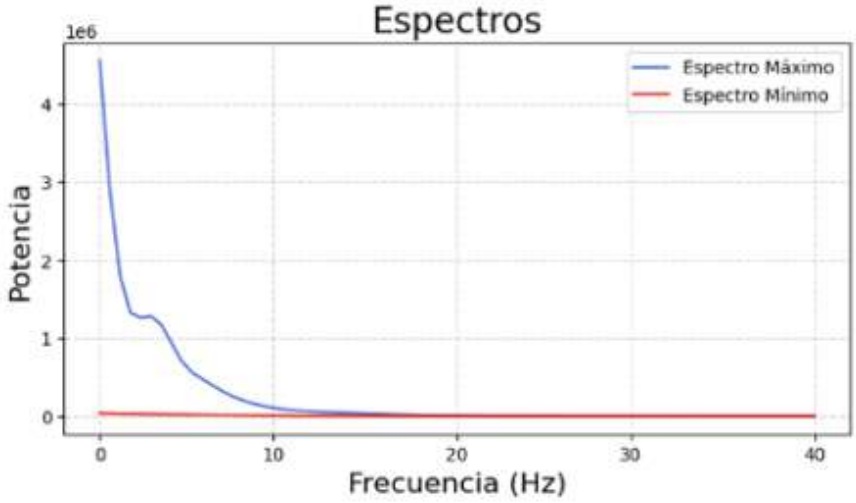


Figura 5.6: *Gráfica de los espectros obtenidos del algoritmo 3 para el grupo etario 17-24 de la clase ¨control¨. La línea azul corresponde al espectro orgánico con el pico de potencia máxima, mientras que la línea roja representa el espectro con el pico de potencia mínima.*

Segundo, se localizaron tanto los valores atípicos como los más cercanos a la media de los datos utilizando el **algoritmo 1**. La **Figura 5.7(a)** muestra la representación tridimensional del análisis PCA combinado con el cálculo del centroide o media de los espectros. Para establecer un umbral que definiera a los valores atípicos, se consideró cualquier punto cuya distancia al centroide supere un valor específico de σ. Por otro lado, los puntos más cercanos al centroide se definieron como aquellos con una distancia inferior σ.

En nuestro grupo ilustrativo, para obtener los valores considerando las proporciones específicas de los datos (35 % outliers, 35 % cercanos, 10 % aleatorio), se calculó el valor de σ= 0.8, identificando 56 observaciones más alejadas del centroide que, por ende, se validaron como valores atípicos (**Figura 5.7(a),** color púrpura). Asimismo, para identificar los espectros más cercanos a la media, correspondientes al otro 35 % de los datos, se tomaron en cuenta todas las observaciones con distancias inferiores a σ= 0.33 (**Figura 5.7(a),** color negro).

Este análisis permitió observar dos comportamientos importantes: por una parte, los casos cuyos espectros mantienen una morfología muy similar se concentran en el centro del

espacio tridimensional. Por otra parte, a medida que los puntos se alejan de esta concentración central, la morfología de los espectros va cambiando.

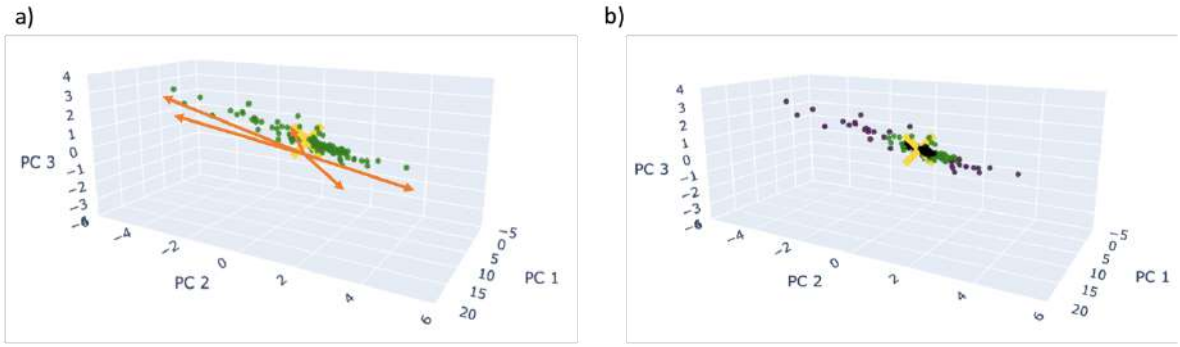


Figura 5.7: *Gráfica en 3 dimensiones del PCA aplicado a los espectros de potencia normalizados del grupo etario 17-24 años de la clase control. Cada punto corresponde a la reducción del espectro de cada paciente en el grupo antes mencionado. La media de los datos en el espacio tridimensional se presenta con una cruz dorada. a) Se presenta la distancia euclidiana de la media con respecto a cada uno de los espectros orgánicos (puntos verdes) con flechas naranjas. b) se representa los espectros orgánicos considerados como outliers (color purpura) y espectros orgánicos cercanos al centroide (color negro) en base a* σ.

Tercero, el 10 % de los datos aleatorios se tomaron a partir de los espectros residuales (**Figura 5.7(b),** color verde). Durante la primera época del entrenamiento de *espectraGAN,* el 70 % de los datos (compuestos por los valores atípicos y los espectros cercanos al centroide) se mantuvo en el lote de entrenamiento. Del 30 % restante disponible, se seleccionó aleatoriamente un 10 %. En la siguiente época, este proceso se repitió, tomando nuevamente un 10 % aleatorio de los datos restantes disponibles. El 20 % de los datos restantes se utilizó en cada época como datos de validación.

En el ejemplo ilustrativo, el lote de entrenamiento de espectros originales constó de un total de 160 observaciones. De estas, 56 correspondieron a valores atípicos, 56 fueron espectros cercanos al centroide, y 16 fueron seleccionados aleatoriamente de los 48 espectros restantes. La **Tabla 4**, detalla los valores de parámetro σ utilizados para cada grupo etario, así como el total de espectros incluidos en el lote de entrenamiento.

	Grupo etario (años)	Observaciones	σ (Lejanos)	Outliers	σ (Cercanos)	Outliers	Aleatorio	Total
Control								
	17-24	160	0.8	56	0.33	56	16	128
	25-34	328	0.48	114	0.33	110	38	262
	35-44	131	0.53	46	0.35	46	12	104
	45-54	73	0.65	25	0.42	28	5	58
	55-64*	90	0.87	31	0.51	31	9	72
		30	1.4	10	0.8	10	4	24
	65-74	28	1.3	10	0.5	9	3	22
	75-84	14	2.7	5	1.3	5	1	11
Enfermos								
	19-24	101	1.1	34	0.45	33	13	80
	25-34	201	0.47	71	0.3	68	21	160
	35-44	362	0.44	126	0.25	126	37	289
	45-54*	506	0.6	175	0.28	171	59	405
		187	5.93	64	5.47	57	28	149
	55-64*	571	0.73	205	0.47	196	56	446
		219	5.96	73	5.13	78	24	175
	65-74*	335	0.59	117	0.28	118	33	268
		152	3.6	53	3.15	50	18	121
	75-84	141	0.6	52	0.47	47	13	112
	<85	36	2.1	12	1.8	11	5	28

Tabla 4: *Total de datos considerados en el entrenamiento del modelo espectraGAN.*

5.2 Análisis del entrenamiento del modelo EspectraGAN

Posterior al análisis exploratorio de la distribución y características de los espectros orgánicos, se procedió a entrenar el modelo *espectraGAN* para generar los espectros sintéticos. La tasa de aprendizaje, que controla la magnitud de las actualizaciones de los pesos, fue ajustada de manera diferenciada para el *generador* y el *discriminador* a fin de equilibrar el proceso de entrenamiento [24]. Específicamente, se emplearon tasas de 0.0002 para el *generador* y 0.0004 para el *discriminador*, junto con un esquema de decaimiento exponencial. Este enfoque permitió una convergencia más eficiente, minimizando oscilaciones cerca de los mínimos locales [111]. Para lograr ajustes más precisos en las etapas finales del entrenamiento y mitigar el riesgo de sobreajuste, se establecieron tasas de decaimiento diferenciadas: 0.1 cada 50 épocas para el generador y 0.45 cada 70 épocas para el discriminador [111].

La inicialización de los pesos se realizó utilizando una distribución Gaussiana con media 0 y desviación estándar de 0.02, lo que favoreció una convergencia estable del modelo durante el entrenamiento [111]. Por otro lado, el espacio latente empleado para alimentar el *generador* fue una distribución normal estándar $N(\mu = 0, \sigma = 1)$ consistente con un enfoque

aleatorio controlado. Finalmente, el tamaño del lote de entrenamiento se determinó en función de los grupos etarios utilizados, como está detallado en la **Tabla 4**.

El modelo fue entrenado durante un total de 256 épocas, con 128 iteraciones por cada una de ellas. Para la red *discriminadora*, se empleó la función de pérdida de entropía cruzada, la cual es ampliamente utilizada en tareas de clasificación por su capacidad para manejar problemas binarios de manera eficiente [24]. En cuanto a la salida binaria del discriminador, se implementó la función de activación Sigmoid, que proporciona probabilidades normalizadas y facilita la interpretación de las decisiones del modelo [112].

Por otro lado, en la red *generadora*, se utilizó la función de activación Swish, seleccionada por su capacidad de mejorar la precisión y estabilidad del modelo durante el proceso de generación [113]. Los valores utilizados para los parámetros del entrenamiento del modelo *espectraGAN* se presentan de manera detallada en la **Tabla 5**.

Parámetro	Valor
Optimizer	Adam
Learning rate (G and D)	$2 \times 10^{-4}, 4 \times 10^{-4}$
β_1, β_2	0.5, 0.99
Decay rate and step (G)	0.1, 50
Decay rate and step (D)	0.45, 70
Weight's initializer	N(0, 0.02)
Noise distribution	N(0,1)
Batch size	Age Group
No steps	128
No epochs	256
Cost Function	Cross-Entropy loss
Activation Function (G)	Swish
Activation Function (D)	Sigmoid

Tabla 5: *Hiperparametros para el entrenamiento del modelo espectraGAN*

Como explicado en el capítulo anterior, el proceso de entrenamiento involucró tanto los espectros reales como los sintéticos. Se le asignó una etiqueta a cada espectro, 1 a los reales y 0 a los sintéticos. Este conjunto de datos combinados se ingresó al modelo *discriminador*, el cual realizó una predicción de la etiqueta para cada espectro, ya sea real o sintético. Las predicciones se compararon con las etiquetas y se calculó la función de pérdida mediante la función de entropía cruzada binaria (*Binary crossentropy*). Con base a esta pérdida, se actualizaron los pesos del discriminador aplicando los gradientes relacionados

con la técnica de optimización de tipo retropropagación del error [24]. En este proceso, el gradiente de la función de pérdida con respecto a cada peso del modelo se calculó y usó para ajustar los pesos en la dirección que minimiza la pérdida.

Una vez actualizado los pesos del discriminador, el proceso de entrenamiento continuó con la generación de nuevos espectros sintéticos por parte del *generador,* a partir del espacio latente N(0,1). A estos nuevos espectros sintéticos, se les asignó la etiqueta 1 para que el *discriminador* los considerara como provenientes del conjunto de datos original. Este conjunto sintético recién generado se ingresó al *discriminador* con los pesos actualizados, para obtener una nueva función de pérdida. Estos espectros sintéticos etiquetados como reales, se utilizaron para calcular la pérdida del generador. El objetivo siendo que el *discriminador*, con sus pesos recién actualizados, clasifique estos nuevos espectros sintéticos ya sea reales o sintéticos. A partir de esta iteración, se actualizaron los pesos del generador.

Es decir, este proceso fue iterativo, ajustando gradualmente los pesos de ambas redes. Lo que permite que el *generador* aprenda a crear espectros cada vez más convincentes, mientras que el discriminador mejora su capacidad para diferenciar entre lo real y lo sintético.

La selección de los hiper parámetros se realizó mediante un enfoque que garantiza una adaptación eficiente del modelo. En primer lugar, se emplearon los valores más comunes reportados en el estado del arte para el entrenamiento de modelos GANs. Posteriormente, estos valores fueron ajustados de acuerdo con las gráficas de pérdida observadas durante el entrenamiento del modelo *espectraGAN.* La **Figura 5.8** ilustra diferentes tipos de entrenamientos con comportamientos no deseados.

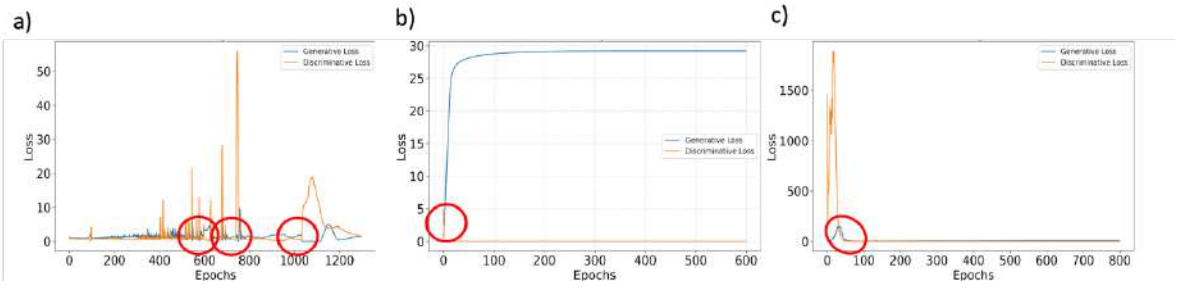


Figura 5.8: *Gráfica de entrenamiento de las pérdidas de espectraGAN. a) b) c) Gráfica de las pérdidas de experimentos fallidos durante la selección de hiper parámetros. Línea naranja pérdida de la red discriminador. Línea azul pérdida de la red generador. En círculo rojo marcan las fluctuaciones y las inestabilidades observadas al finalizar el entrenamiento.*

En la **Figura 5.8 (a)**, las grandes fluctuaciones en la pérdida del *discriminador* (marcadas en rojo) indicaron inestabilidad en el entrenamiento, lo que fue problemático para la convergencia del modelo. Asimismo, los picos abruptos en la pérdida del *discriminador* indicaron momentos en los que el *generador* creó muestras que el discriminador no pudo clasificar correctamente [59], [111].

En la **Figura 5.8 (b)**, la pérdida del generador permaneció alta durante todo el entrenamiento, con un valor por encima de 25, lo que indicó que el *generador* dominó al *discriminador*, llevando a un sobreajuste y a la generación de espectros sintéticos que no eran suficientemente variados. La falta de fluctuaciones en la pérdida del *discriminador* es una señal de que el *discriminador* no aprendió adecuadamente y se ha estancado en su capacidad para diferenciar los espectros reales de los generados, ya que su valor se mantiene en 0 desde prácticamente el inicio del entrenamiento [59], [111].

Por último, en la **Figura 5.8 (c)**, la alta pérdida inicial, especialmente en el *discriminador* (línea naranja), es un signo de que el modelo comenzó con un desajuste significativo. Aunque se aprecia que las pérdidas se estabilizan, la magnitud inicial de la pérdida indicó que el modelo necesitaba ajustes adicionales en los hiper parámetros o en la arquitectura para mejorar la eficiencia del entrenamiento [59], [111].

En contraste, en la **Figura 5.9**, se muestra la gráfica del entrenamiento exitoso para el grupo etario en cuestión. El modelo final *espectraGAN* se entrenó con 256 épocas, observándose la convergencia después de las 100 primeras épocas. Analizando las pérdidas de las redes, se observó que el modelo *generador* converge una vez que su pérdida se mantiene en su valor máximo de 4.44 (línea azul). Mientras tanto, el discriminador muestra una pérdida constante de 1.67.

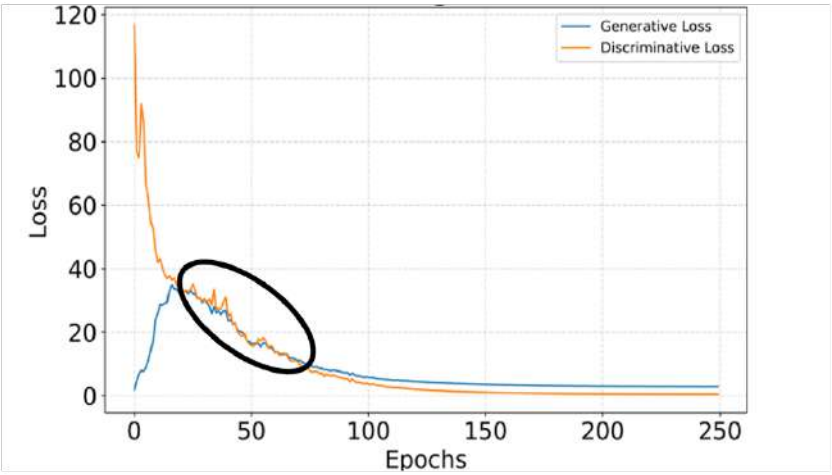


Figura 5.9: *Gráfica de entrenamiento exitoso de las pérdidas de espectraGAN para el grupo etario 17-24 años de la clase ¨Control¨. En elipse negra se muestra ligeras fluctuaciones de las redes, pero al finalizar se observa la convergencia del modelo.*

De acuerdo con la literatura sobre el uso de modelos generativos, un equilibrio entre las pérdidas del *generador* y *discriminador* se indica cuando la función de pérdida del *generador* se mantiene por arriba que la del discriminador [59], [111]. En la elipse negra marcada en la **Figura 5.9,** se muestra que las funciones de pérdida de ambas redes se cruzan, lo cual indica ese juego adversarial en el que, en ciertas épocas, el discriminador es mejor que el generador y viceversa [59], [111]. Cabe mencionar que la estabilidad de este modelo para el grupo etario seleccionado, así como para el resto de los grupos en la base de datos, se evidenció con el decaimiento exponencial, el cual aseguró que el modelo no se estanca, no tiene algún sobreajuste y alcanza el comportamiento deseado.

5.3 Análisis comparativo entre espectros sintéticos y orgánicos

A continuación, se presentan los resultados de la implementación del **Algoritmo 4**, utilizado en la fase de filtrado una vez que el modelo fue entrenado. Para ilustrar su funcionamiento, se incluyen ejemplos de espectros orgánicos y sus respectivas parejas de similitud sintética, seleccionados en función del valor mínimo de RMSE.

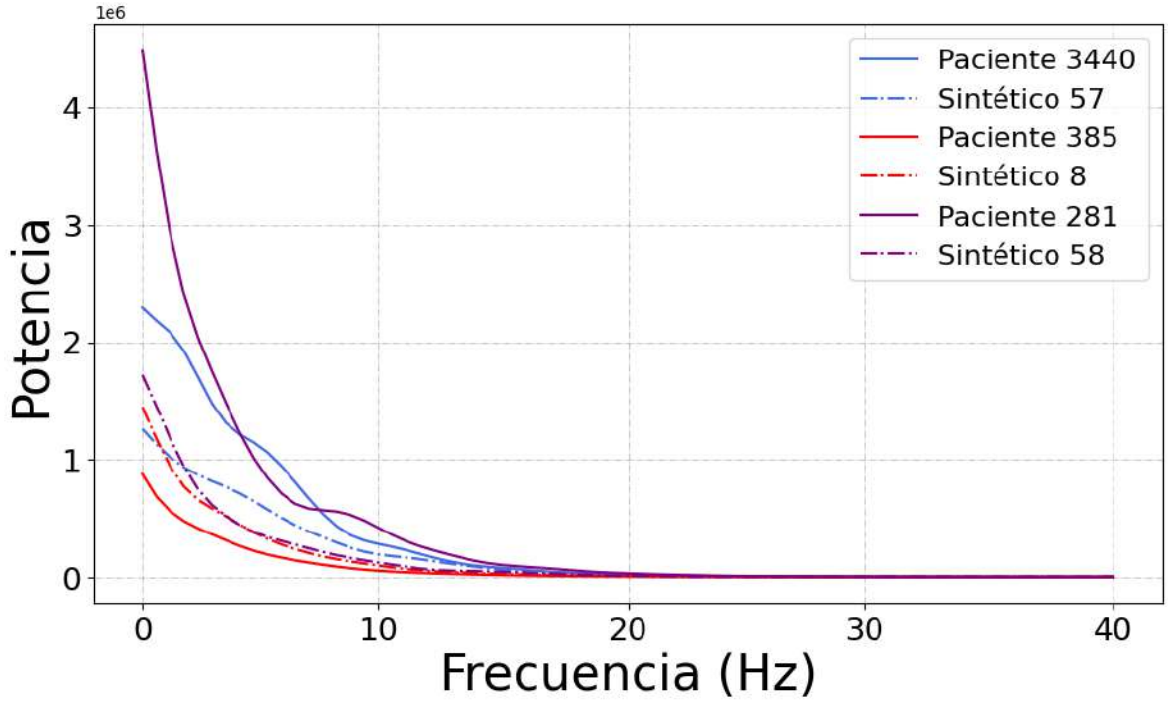


Figura 5.10: *Gráfica comparativa de los espectros orgánicos contra los sintéticos para 3 sujetos del grupo etario 17-24 clase ¨Control¨. Únicamente implementando el algoritmo 4. Líneas continuas se muestran los espectros orgánicos. La líneas punteadas muestran su correspondiente espectro sintético a partir del valor mínimo RMSE.*

La inspección visual de estos resultados (**Figura 5.10**) permitió destacar los siguientes puntos:

- Similitud de patrones: Las curvas sintéticas siguen un patrón similar al de los espectros orgánicos para cada paciente.
- Concordancia en los rangos de potencia: Los espectros de potencia para los sintéticos y orgánicos se encuentran dentro del mismo rango, lo que indicó que el modelo *espectraGAN* capturó correctamente la distribución de potencia de los espectros orgánicos.

En el ejemplo de la **Figura 5.10** se muestran únicamente un ejemplo con 3 espectros, sin embargo, se obtuvo con toda el grupo etario y su correspondiente pareja de similitud sintética para posteriormente ser aplicada la prueba Mann-Whitney U (**Algoritmo 5**).

En la **Figura 5.11** se presentan los resultados de la comparación del grupo etario 17-24, tanto en los datos orgánicos como en los sintéticos. Después de aplicar el PCA a toda la base de datos y seleccionar el grupo etario sintético generado junto con su grupo etario

orgánico correspondiente, se procedió a graficar ambos grupos de datos dentro del mismo espacio tridimensional, utilizando los mismos componentes principales obtenidos en el análisis inicial, con el objetivo de identificar posibles diferencias significativas entre ambos grupos.

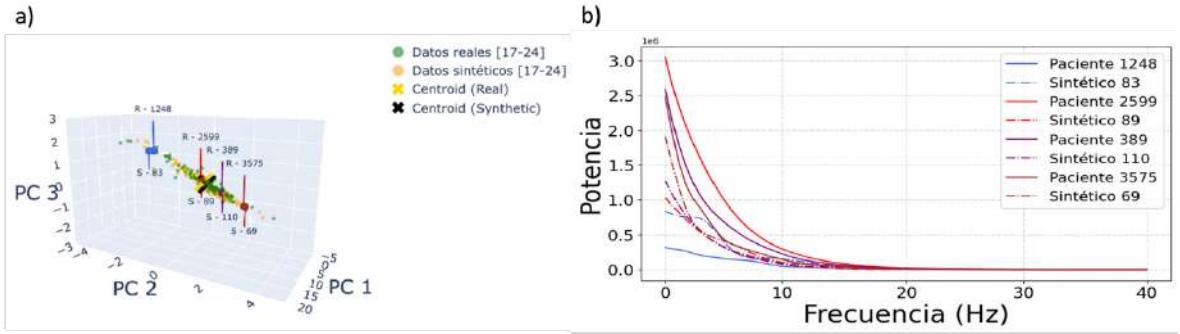


Figura 5.11: *Análisis de los espectros sintéticos en comparación con los espectros orgánicos. a) Gráfico de PCA 3D por grupo etario (control) para el grupo etario 17-24 años, los puntos verdes: datos orgánicos, los puntos naranjas: datos sintéticos. Los centroides de cada clase están marcados mediante cruces: cruz amarilla para los datos orgánicos y una cruz negra para los sintéticos. Las líneas en verticales señalan los espectros seleccionados aleatoriamente para la comparativa orgánica contra sintética cada color representa un paciente diferente. b) Espectros* de potencia *de los casos seleccionados en a). Las líneas continuas muestran los casos orgánicos. Las líneas punteadas de mismo color muestran su correspondiente espectro sintético seleccionado en base al valor mínimo RMSE.*

El gráfico de PCA 3D (**Figura 5.11 a**) muestra una distribución de los datos sintéticos notablemente alineada con los datos reales. Ambos conjuntos concentran la mayor parte de las combinaciones hacia el centro del espacio tridimensional, mientras que, en regiones más alejadas, existe baja densidad de espectros. [122], [123][123]

Además, se seleccionaron cuatro casos sintéticos y cuatro orgánicos dentro del espacio tridimensional (**Figura 5.11 a**) para comprobar si los espectros sintéticos no son copias directas de los orgánicos, sino que presentan una gran similitud.

- La **Figura 5.11 (b)** incluye espectros seleccionados aleatoriamente de distintas zonas del espacio tridimensional generado por el PCA (**Figura 5.11 a**) que son los espectros de casos reales etiquetados como R-1248, R-2599, R-389 y R-3575, cuyos espectros sintéticos más cercanos fueron seleccionados según la distancia euclidiana en el espacio. Las curvas generadas por el modelo muestran una alta consistencia con las tendencias de las curvas orgánicas, especialmente en el rango de frecuencias bajas (0 a 10 Hz), donde la potencia es mayor (**Figura 5.11 b**). Por ejemplo, en el caso del paciente R-

2599 (línea continua roja) y su espectro sintético de mayor similitud S-89 (línea discontinua roja), el modelo genera una curva que sigue un comportamiento similar, replicando la tendencia logarítmica de disminución de potencia (**Figura 5.11 b**). Sin embargo, las potencias no coinciden exactamente en valores absolutos. Esto indica que el modelo no se limita a copiar los espectros orgánicos, sino que aprende sus características principales para generar nuevos que, si bien no se sobrelapan perfectamente en el espacio tridimensional, son cercanos a los reales.

Lo anterior valida que el modelo *espectraGAN* puede generar nuevos datos manteniendo la variabilidad natural de los datos orgánicos.

Por otro lado, el análisis de los espectros seleccionados en diferentes áreas revela que el comportamiento de los datos orgánicos varía conforme se alejan del centroide del grupo (**Figura 5.11 b**). Por ejemplo, al comparar los espectros del casos R-1248, se observa una morfología similar en el rango de 0 a 10 Hz, donde la potencia disminuye de manera logarítmica. No obstante, al analizar otros casos, como R-3575, comienzan a evidenciarse variaciones en la potencia para ciertas frecuencias. Este patrón indica que los espectros más alejados del centroide representan una mayor diversidad estructural, reflejando la complejidad de los datos originales.

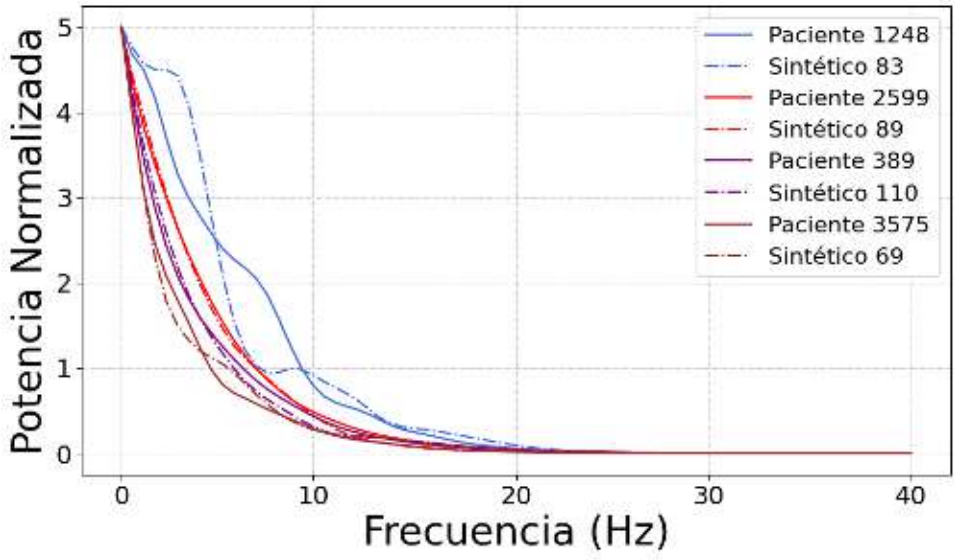


Figura 5.12: *Gráfica de comparación entre espectros orgánicos grupo etario 17-24 y los correspondientes sintéticos normalizados. Líneas continuas se muestran los casos orgánicos. Las líneas punteadas muestran su correspondiente espectro sintético a partir del valor mínimo RMSE.*

Para realizar un análisis más detallado de la similitud entre los espectros orgánicos y los sintéticos, se normalizaron ambos grupos etarios. En la **Figura 5.12**, se muestra esta normalización, limitada al rango de potencias de 0 a 5, con el objetivo de observar las morfologías desde otra perspectiva. Observamos que el caso R-2599 y su espectro sintético S-89 tienen formas de onda muy similares. Por otro lado, en el caso R-1248, se nota una mayor variación en la potencia dentro del rango de frecuencias. Aunque la potencia sigue disminuyendo, ya no lo hace de forma logarítmica, si no que presenta variaciones en diferentes frecuencias. A pesar de que el espectro sintético de mayor similitud S-83, según la distancia euclidiana, no es completamente idéntico al espectro orgánico del R-1248, ambos comparten características generales en sus formas de onda. Estos resultados refuerzan que nuestro modelo *espectraGAN* genera espectros basados en las características aprendidas durante el entrenamiento, logrando espectros con una estructura coherente para su posterior procesamiento.

Posteriormente se analizó la distribución de las distancias de los datos sintéticos generados por el modelo. La **Figura 5.13** permite contrastar las distancias de los datos sintéticos con las de los datos orgánicos reales, tomando como referencia el centroide calculado a partir de los datos orgánicos.

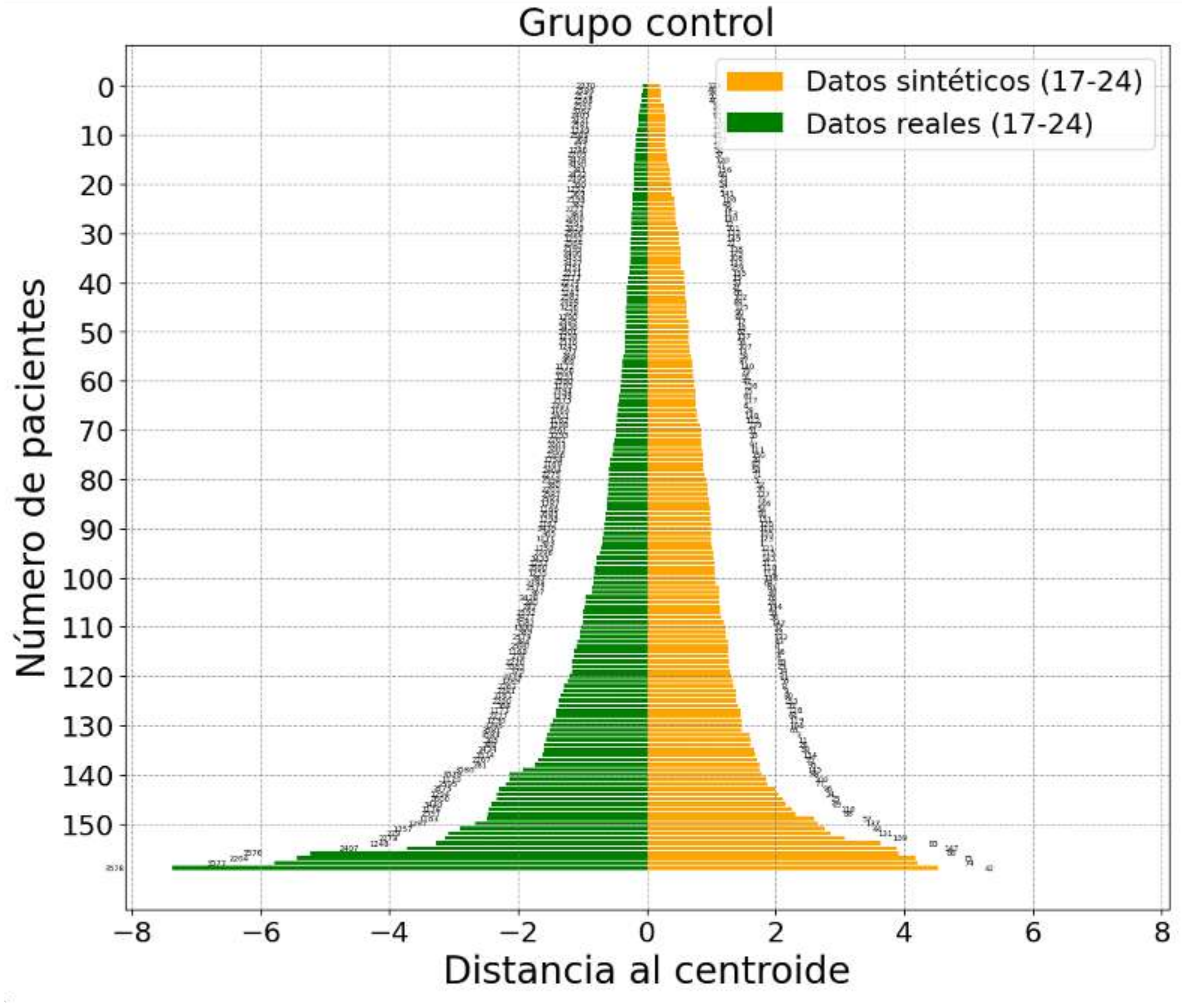


Figura 5.13: *Distribución de los espectros orgánicos (color verde) y sintéticos (color naranja) grupo etario 17-24 según las distancias euclidianas en el espacio tridimensional definido por el PCA, donde el 0 corresponde al centroide de los datos orgánicos. El eje Y indica la cantidad de observaciones del grupo etario y los números que bordean los datos son el número de paciente.*

Lo anterior permite examinar las siguientes características clave:

- Distribución de los datos:

  Se observa que tanto los espectros orgánicos como los espectros sintéticos presentan una distribución altamente concentrada alrededor del centroide del grupo etario (distancia 0). Esto indica que, en términos de las principales características identificadas por el PCA, los espectros generados por el modelo *espectraGAN* son muy similares a los espectros orgánicos. [43], [145]

- Discrepancia en las colas de la distribución:

Si bien las distribuciones muestran una concentración significativa en torno al centroide, se identifican ligeras diferencias en las colas, que representan las distancias más alejadas del centroide. Específicamente, hay una mayor cantidad de datos orgánicos (color verde) en las colas en comparación con los datos sintéticos (color naranja) [43].

- Similitud entre grupos sintéticos y orgánicos:
  Ambos conjuntos de datos exhiben picos en las mismas posiciones, lo que refuerza la demostración de la capacidad del modelo *espectraGAN* para replicar eficazmente las principales características de los datos orgánicos [114]. Este resultado es especialmente significativo, ya que demuestra que el modelo logra generar datos sintéticos con una estructura estadística y morfológica consistente con la distribución de los datos reales, al menos en las zonas más representativas del grupo etario analizado. Esto pone en evidencia la habilidad del modelo para aprender las características subyacentes más importantes de los datos originales [24].

- Coincidencia global y pequeñas diferencias locales:
  En términos generales, las formas de ambas distribuciones son muy similares, lo que confirma la efectividad del modelo *espectraGAN* en la generación de espectros sintéticos. Sin embargo, se detectan pequeñas diferencias en ciertas distancias, particularmente en las regiones más alejadas del centroide. Estas discrepancias podrían representar una oportunidad para futuras mejoras en el modelo, con el objetivo de capturar de manera más precisa las variaciones extremas o valores atípicos que están presentes en los datos orgánicos [24].

Tras culminar el entrenamiento de los grupos etarios, se identificaron algunos conjuntos que presentaban estructuras distintivas en comparación con el resto. En particular, estas diferencias estructurales fueron más evidentes en el grupo etario 55-64 años del grupo control, y en el grupo etario 45-54, 55-64, 65-74 y más de 85 años del grupo enfermos. En la **Figura 5.14** se ilustra un ejemplo representativo de estas estructuras distintivas, utilizando el grupo etario de 45-54 años en el grupo de casos enfermos.

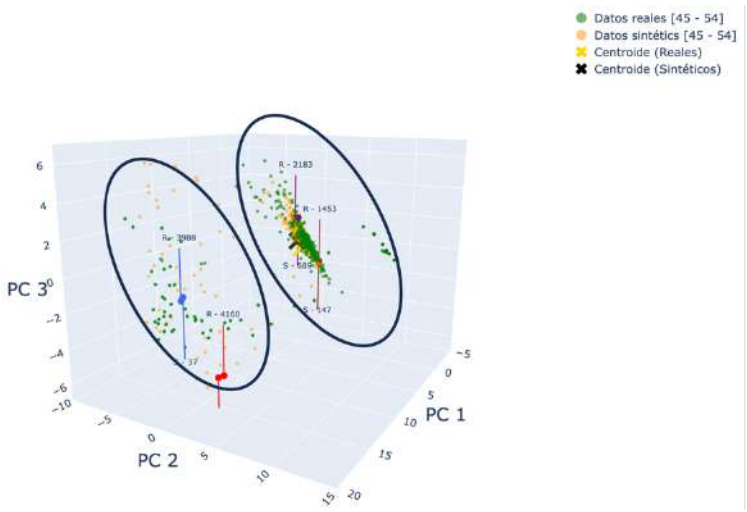


Figura 5.14: *Gráfica de PCA 3D de los casos enfermos reales y sintéticos para el grupo etario 45-54 años. Las elipses negras enmarcan las aglomeraciones que se encuentran en el espacio 3D.*

Algunas observaciones claves son:

- Aglomeración en subgrupos: Los datos orgánicos (en verde) están distribuidos en dos aglomeraciones principales, delimitadas por elipses negras. Estas aglomeraciones reflejan una segmentación natural dentro del grupo etario analizado, que podría estar asociada a subpoblaciones con características comunes.

- Limitaciones del modelo inicial: Durante el entrenamiento se observó que el modelo *espectraGAN* tiende a aprender predominantemente las características de una sola aglomeración dentro del grupo etario, dejando las otras sin representación adecuada.

- Solución propuesta: Para abordar esta problemática, se decidió dividir el entrenamiento de los grupos etarios que presentan esta divergencia en dos etapas. En la primera etapa, se entrenó el modelo con las observaciones pertenecientes a una elipse, mientras que la segunda etapa se entrenó el modelo independiente con las observaciones de la otra elipse. Finalizado el entrenamiento con ambas elipses, se utilizaron ambos modelos entrenados con

el formato .h5 para producir los espectros sintéticos. Para garantizar una representación equilibrada, el factor de generación de cada modelo fue ajustado a la misma cantidad de observaciones presentes en las elipses, de modo que al combinar los resultados se obtuvieron una cantidad equitativa de espectros sintéticos representando ambas aglomeraciones.

Para evaluar la morfología de los espectros, se seleccionaron aleatoriamente muestras representativas de cada una de las aglomeraciones previamente identificadas. Los resultados se presentan en la **Figura 5.15**. En el panel a) se observa que los espectros sintéticos correspondientes a las diferentes aglomeraciones tienen características morfológicas diversas. Por ejemplo, el espectro del paciente R-4160 (línea roja) presenta picos de potencia significativamente más altos en ciertas frecuencias, así como el espectro del paciente R-3988 y su correspondiente sintético S-37. Estos resultados confirman que, a medida que los espectros se encuentran más alejados del centroide o de la principal aglomeración, su morfología tiende a ser más variable y diferenciada. Por el contrario, espectros de grupos más densamente aglomerados presentan similitudes muy parecidas al comportamiento logarítmico.

El panel b) muestra los mismos espectros, pero después de ser normalizados en términos de potencia. Esta normalización permite analizar con mayor detalle la similitud estructural entre los espectros sintéticos generados por el modelo *espectraGAN* y sus contrapartes orgánicas. A pesar de las diferencias observadas en el panel a), la normalización revela que los espectros sintéticos conservan características generales importantes del espectro original.

Por ejemplo, en el caso del paciente R-3988, aunque los picos de potencia en el espectro original son más marcados y diferentes, el espectro sintético S-37 generado logra replicar características clave del perfil morfológico. Esto sugiere que, aunque el modelo *espectraGAN* puede no capturar todas las variaciones específicas en potencias extremas, sí es capaz de reproducir patrones generales de los espectros. Un patrón similar se observa en el caso del paciente R-4160, cuyo espectro muestra una morfología que se asemeja más a las aglomeraciones de las elipses más alejadas que a las más densas. Por el contrario, los casos R-2183 y R-1453, pertenecientes a aglomeraciones más cercanas al centroide y con mayor

densidad de datos, presentan espectros con morfologías menos variables en cuestión de potencia tanto en el espectro original como en el sintético.

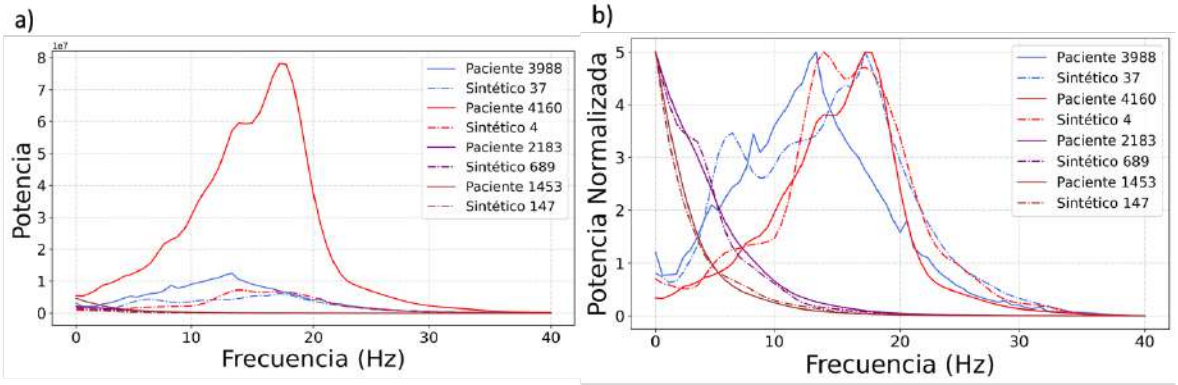


Figura 5.15: *Gráfica de comparación entre espectros orgánicos grupo etario 45-54 y los correspondientes sintéticos. a) espectros generados directamente del modelo espectraGAN. b) espectros normalizados en términos de potencia.*

Este análisis confirma que la morfología de los espectros depende significativamente de su posición dentro de las aglomeraciones identificadas en el PCA. Espectros pertenecientes a aglomeraciones más alejadas tienden a tener mayor variabilidad y diferencias en sus picos de potencia, mientras que aquellos más cercanos a la media presentan una morfología más uniforme.

Finalmente, la **Figura 5.16** presenta la distribución de las distancias de los espectros orgánicos (color verde) y sintéticos (color naranja) con respecto al centroide orgánico del grupo etario 45-54 de la clase enfermos.

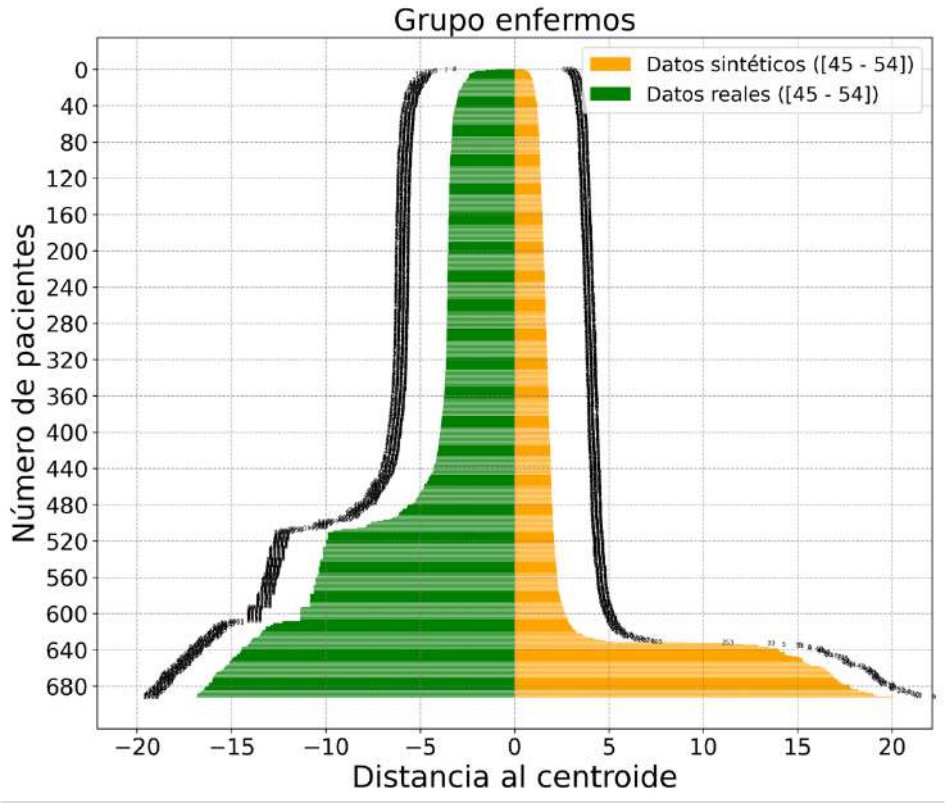


Figura 5.16: *Distribución de las distancias de los espectros orgánicos (color verde) y espectros sintéticos (color naranja) con respecto al centroide orgánico del grupo etario 45-54 de la clase ¨Enfermos¨.*

De la misma forma que para la clase *sanos*:

- Distribución de los datos: Tanto la distribución de los datos orgánicos como la de los sintéticos está altamente centrada en torno al centroide del grupo (distancia 0). A medida que las distancias se alejan del centroide, se observa una disminución de la densidad de ambos tipos de datos, lo que refleja un comportamiento esperado para los datos generados, a saber, que sigan la distribución original.

- Discrepancia en las colas de la distribución: En las regiones más alejadas del centroide, se observa pequeñas discrepancias entre los datos. Este comportamiento, similar al descrito anteriormente, muestra que el modelo tiene dificultades para generar espectros que reflejen las variaciones más extremas o atípicas de los datos reales, posiblemente debido a una menor representación de estas características en el conjunto de entrenamiento.

- Similitud entre grupos sintéticos y orgánicos: La mayor simetría entre los espectros orgánicos y los sintéticos se observa en las distancias cercanas al centroide (entre -5 y +5). Este resultado es clave, ya que representa las zonas de mayor densidad de datos en el grupo etario, donde las características estructurales principales se encuentran más definidas. La coincidencia en esta región evidencia la efectividad del modelo *espectraGAN* para reproducir las características más representativas de los datos reales.

- Coincidencia global y pequeñas diferencias locales: La distribución global de los datos sintéticos abarca prácticamente el mismo rango que los datos orgánicos, lo que indica que el modelo logra capturar de manera global la variabilidad presente en el grupo etario. Sin embargo, las diferencias en las densidades de las regiones extremas destacan una oportunidad para refinar el modelo en futuras iteraciones.

5.4 Análisis Cualitativo

Una vez concluido el entrenamiento para los distintos grupos etarios, se generó una base de datos compuestos por espectros sintéticos, seguido de un análisis integral comparando estos resultados sintéticos con los datos orgánicos. La **Figura 5.17** presenta gráficos tridimensionales (3D) de los espectros de potencia obtenidos mediante ERG basal, junto con los datos sintéticos generados por el modelo *espectraGAN*.

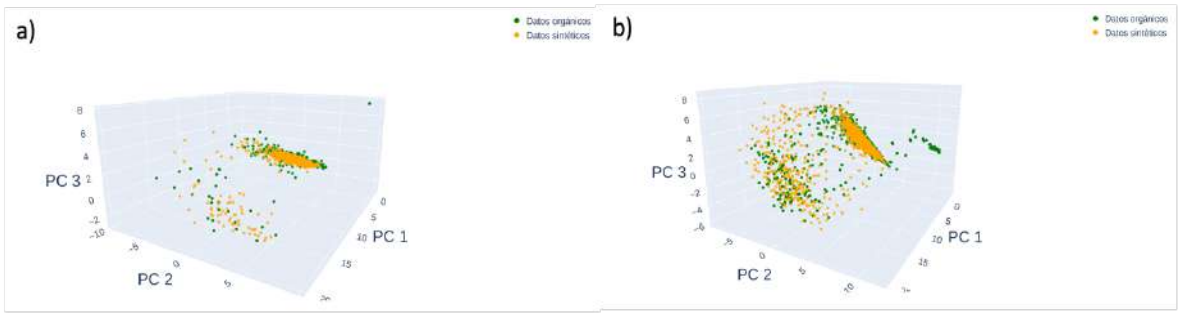


Figura 5.17: *Gráficos en 3 dimensiones (3D) de PCA de los espectros de potencia obtenidos del ERG basal y sintéticos generados por el modelo correspondiente a la clase a) control y b) enfermos. Espectros orgánicos de color verde y espectros sintéticos de color naranja.*

En ambos gráficos (a y b) puede apreciarse que los datos sintéticos (color naranja) mantienen la distribución global observada en la datos orgánicos (color verde), indicando que el modelo espectraGAN ha capturado adecuadamente las características globales de los espectros orgánicos. En el gráfico a (clase *control*) muestra claramente cómo se superponen los datos sintéticos y los datos orgánicos. En el gráfico b (clase *enfermos*), se observa en general una correspondencia satisfactoria entre ambas distribuciones, los datos sintéticos exhiben una dispersión más amplia respecto a los datos orgánicos. Este comportamiento podría reflejar una mayor variabilidad inherente en los espectros para la clase enfermos.

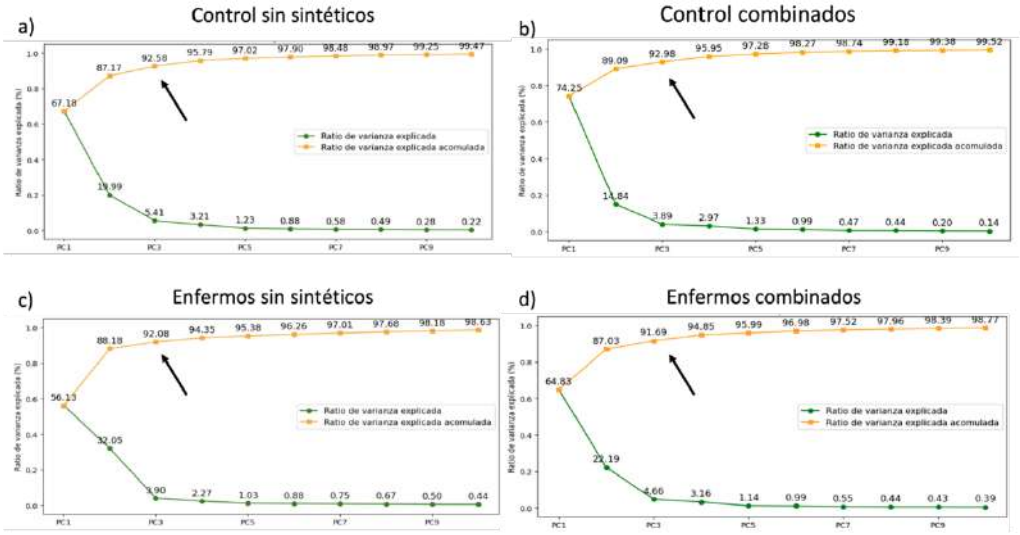


Figura 5.18: *Análisis de PCA aplicado a los espectros orgánicos a) clase control y c) clase enfermos. Análisis de PCA para la base de datos combinados tanto espectros orgánicos como sintéticos b) clase control y d) clase enfermos.*

En la **Figura 5.18** se muestra el análisis de PCA aplicado a los espectros reales y su combinación con los espectros sintéticos. En el gráfico a), correspondiente a la clase *control*

sin sintéticos, la mayor proporción de la varianza explicada se concentra en los dos primeros componentes principales (PC1 y PC2). El primer componente explica el 67.18 % de la varianza, mientras que el segundo añade un 19.99 %, alcanzando un 87.17 % de varianza acumulada. En el gráfico c), correspondiente a la clase *enfermos sin sintéticos*, se observa un patrón similar, aunque con algunas diferencias clave. El primer componente explica un 56.13 % de la varianza, mientras que el segundo añade un 32.05 %, acumulando un total de 88.18 % en los dos primeros componentes. Esto indica que las muestras de la clase *enfermos* presentan características que se concentran más fuertemente en los primeros componentes principales, en comparación con la clase control.

La inclusión de datos sintéticos modifica significativamente la distribución de la varianza explicada en ambas clases, como se observa en los gráficos b) y d). En el gráfico b) correspondiente a la clase *control combinado*, la varianza explicada por el primer componente aumenta a 74.25 %, mientras que el segundo componente contribuye con un 14.84 %, alcanzando una varianza acumulada del 89.09 % en los dos primeros componentes. Este cambio muestra que los datos sintéticos introducen una mayor dispersión en la estructura de los datos, distribuyendo la varianza de manera más uniforme entre los componentes adicionales.

Por otro lado, en el gráfico d) de la clase *enfermos combinados*, el impacto de los datos sintéticos es aún más evidente. La varianza explicada por el primer componente aumenta a 64.83 %, mientras que el segundo componente disminuye a 22.19 %, acumulando un total de 87.03 % en los dos primeros componentes. Esto apunta a que los datos sintéticos en la clase *enfermos* aportan una nueva variabilidad que amplifica la estructura principal de los datos.

La reducción en la proporción de varianza explicada por los primeros componentes y el aumento en la acumulación de varianza tras la adición de datos sintéticos reflejan la capacidad de estos datos para introducir nueva variabilidad. Esta nueva variabilidad puede estar relacionada con aspectos de la variabilidad de las características de los espectros reales que no estaban completamente capturados en el análisis inicial. En este sentido, la combinación de datos reales y sintéticos podría mejorar la representatividad del espacio de

características, especialmente en el caso de la clase *enfermos*, donde la complejidad de los datos parece ser mayor.

En la **Figura 5.19**a de la clase control, el PC1 explica el 74.24 % de la varianza total, el gráfico muestra un pico prominente alrededor de la región espectral cercana a la componente espectral f000021, seguido de un descenso gradual hacia la región f000041. Este rango de componentes son las más cruciales para explicar el aparte significativo del PC1 [115]. Para el caso del PC2, con un 14.84 % de la varianza, exhibe oscilaciones que señalan contribuciones relevantes de diferentes zonas espectrales (principalmente en el rango f000031-f00041 y cerca de f000046). Aunque estas características no son dominantes, resultan relevantes para la variabilidad adicional de los datos [115]. Por último, la PC3 que aporta un 3.89 % de la varianza, revela detalles más específicos representados por múltiples picos y valles que denotan la variabilidad localizada en ciertas regiones espectrales específicas (componentes espectrales f000011, f000038, f000045).

En la **Figura 5.19**b, se presentan las parcelas de puntuación junto con sus histogramas correspondientes (diagonal de la **Figura 5.19**b), donde se comparan las distribuciones de los datos. En esta clase *control*, las mayores diferencias entre ambos conjuntos se observan principalmente en PC1. Las gráficas de dispersión (PC1 vs. PC2 y PC1 vs. PC3) reflejan una dispersión particular en los datos orgánicos que el modelo *espectraGAN* no logró capturar correctamente. A pesar de esta limitación, existe notable correspondencia general entre ambos conjuntos de datos, aspecto reforzado por la alta similitud observada en los histogramas.

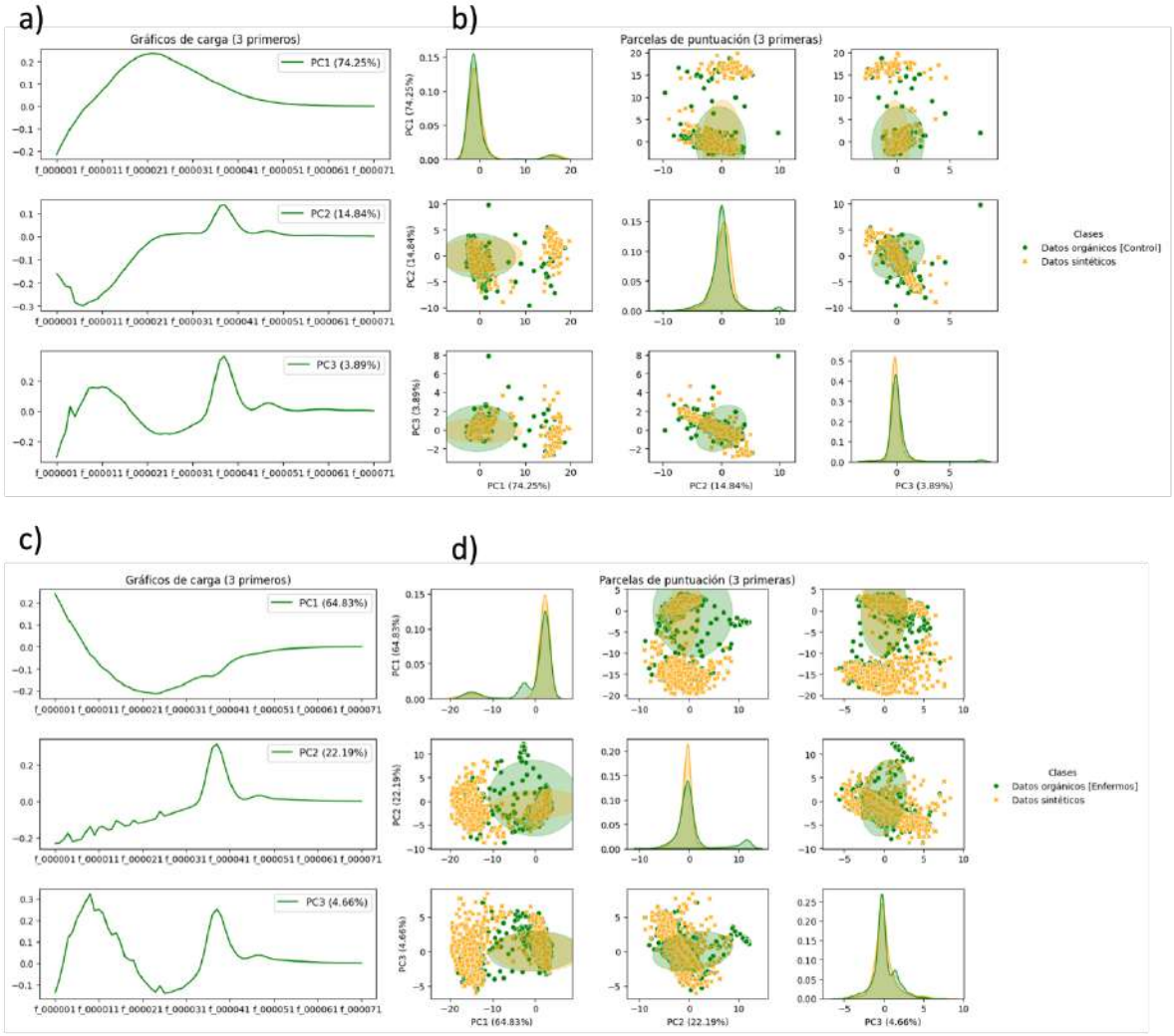


Figura 5.19: *a) y c) se muestran los gráficos de carga para la clase control y enfermos: representan los pesos de las variables originales en los tres primeros componentes, el eje x corresponde a las variables espectrales (longitud de onda), mientras que el Y indica la potencia y dirección de la contribución de cada una de las variable. b) y d) se muestran parcelas de puntuación (scores) para la clase control: representan las distribuciones y las relaciones entre los datos en el espacio reducido de las tres primeras componentes, mostrado tanto los datos orgánicos (verde) como los sintéticos (naranja) y distribuciones univariadas: histogramas y curvas de densidad para los escores de los tres primero componentes.*

En la **Figura 5.19**c y d, corresponde al análisis de la clase enfermos. El PC1 explica un 64.28 % de la varianza, seguida por PC2 con un 22.19 % y PC3 con un 4.66 %. Las curvas de carga presentan patrones más complejos, con múltiples picos y oscilaciones que sugieren características espectrales más variadas con respecto a la clase control. Destaca particularmente que, en PC1, el rango espectral más relevante es f000001-f000006, mientras que en PC2, el rango predominante es f000033-f000041, siendo considerablemente más influyente en esta clase que en la clase control. Finalmente, la PC3 presenta contribuciones

importantes en los rangos espectrales (f000005-f000015 y f000031-f000041), resaltando diferencias morfológicas significativas con respecto a la clase control [114], [115].

En la **Figura 5.19**d, en las parcelas de puntuación de la clase *enfermos*, la superposición entre los datos orgánicos y sintéticos evidente, especialmente en la gráfica PC2 vs. PC3, indicando que el modelo *espectraGAN* logró capturar satisfactoriamente las características de los espectros para esas regiones. Sin embargo, en la gráfica PC1 vs. PC2, se identificaron regiones específicas de datos orgánicos sin correspondencia de datos sintéticos. Lo cual reflejó la dificultad del modelo en producir espectros alejados de los centroides característicos de los grupos etarios. Respecto a los histogramas, ambos conjunto de datos muestran distribuciones semejantes con ligeras diferencias en amplitud, posición de los picos principales y colas, lo que confirmó que, aunque el modelo capturó adecuadamente la estructura global, presentó limitaciones para producir características espectrales más específicas.

En la **Figura 5.20** se presenta la proyección tridimensional de los datos orgánicos y sintéticos utilizando la técnica de reducción de dimensionalidad t-SNE para la clase *control*. Esta técnica permite visualizar la estructura subyacente de los datos sintéticos en un espacio reducido, preservando las relaciones locales y globales del conjunto orgánico [116]. El análisis mediante t-SNE con diferentes combinaciones de perplexity y learning rate facilitó la identificación de la configuración más adecuada para revelar patrones relevantes y destacar diferencias o similitudes entre los conjuntos de datos. Modificar el perplexity permitió controlar cómo t-SNE equilibra las relaciones locales y globales en la representación; valores bajos de perplexity enfatizan las relaciones locales, mientras que valores altos dan mayor peso a las relaciones globales, mientras que ajustar el learning rate influye en la estabilidad y precisión de la convergencia del algoritmo [117]. Esto optimizó así la interpretación visual de estructuras complejas en los datos analizados. Para la selección de los parámetros óptimos que lograron capturar una mejor representación visual, se llevó a cabo una búsqueda de hiper parámetros, evaluando diferentes combinaciones de perplexity y learning rate. Los resultados más representativos se muestran en la figura nombrarla con el objetivo de realizar una comparación visual entre configuraciones.

En la configuración con perplexity=60 y learning rate=100, se observó que el clúster principal tiene una estructura compacta y bien definida. Aunque hay una notable superposición entre datos orgánicos y sintéticos, existen pequeños subgrupos o estructuras internas que podrían asociarse con las características específicas presentes en los datos orgánicos [115]. Esta configuración de parámetros favoreció las relaciones locales entre los puntos, resultando en agrupaciones claras y coherentes, aunque podría descuidar relaciones globales más amplias.

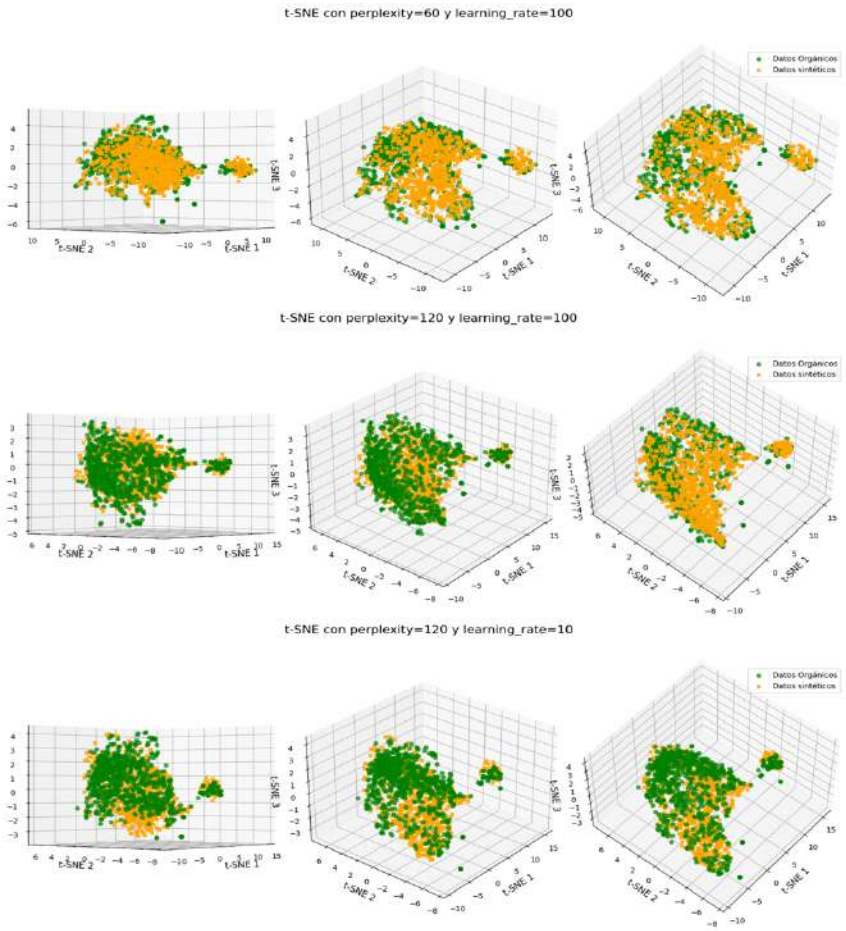


Figura 5.20: *Análisis de componentes t-SNE (t-Distributed Stochastic Neighbor Embeddings) para la clase control. Se utilizan diferentes parámetros de hiperconfiguración de t-SNE, como perplexity y el learning rate con la finalidad de explorar cómo afectan la distribución y separación entre los datos. Color verde para datos orgánicos y color naranja para los datos sintéticos.*

Al incrementar la perplexity a 120 (manteniendo el learning rate en 100), la estructura del cluster principal se volvió más amplia y definida en comparación la configuración previa. Esta modificación permitió a t-SNE capturar relaciones a mayor escala, ofreciendo una representación más integral y balanceada de las características globales y locales del

conjunto. Se logró apreciar que persiste la superposición entre los datos orgánicos y sintéticos, indicando que la distribución general mejora al considerar relaciones más globales [116].

Al mantener una perplexity de 120 pero reduciendo el learning rate a valores más bajos, se observó una la compactación del clúster principal. Persistió cierta superposición entre los datos orgánicos y sintéticos, la agrupación siguió siendo homogénea. Con todas las configuraciones se reflejaron claramente las estructuras internas locales y globales, sugiriendo que el modelo espectraGAN capturó adecuadamente las características fundamentales.

En la **Figura 5.21**, se presentan los resultados de la proyección tridimensional obtenida mediante t-SNE para la clase enfermos. Con una configuración de perplexity=60, los datos orgánicos y sintéticos exhibieron una distribución compacta y una superposición significativa en el espacio tridimensional. Esta configuración prioriza fuertemente las relaciones locales, resultando en una agrupación densa que reflejó alta similitud ente los conjuntos de datos [116].

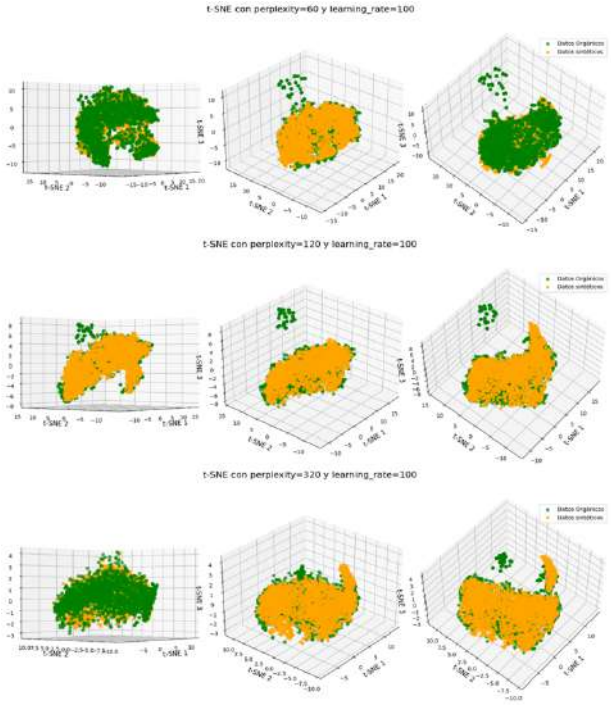


Figura 5.21: *Análisis de componentes t-SNE (t-Distributed Stochastic Neighbor Embeddings) para la clase enfermos. Se utilizan diferentes parámetros de hiperconfiguración de t-SNE, como perplexity y el learning rate con la*

finalidad de explorar cómo afectan la distribución y separación entre los datos. Color verde para datos orgánicos y color naranja para los datos sintéticos.

Al incrementar la perplexity a 120, el clúster principal muestra una estructura más amplia y menos compacta. Anqué persistió la superposición, se aprecian diferencias sutiles en la distribución, particularmente hacia los bordes del clúster, donde los datos sintéticos abarcaron ligeramente más espacio. Por último, con una perplexity de 320, se obtuvo una representación aún más dispersa y global. Los datos sintéticos orgánicos y sintéticos se mostraron distribuidos de manera uniforme, disminuyendo notablemente la compactación observada en configuraciones precias. En cada una de las tres representación, se observó claramente áreas específicas de datos orgánicos donde el modelo *espectraGAN* tuvo dificultades para replicar adecuadamente las características originales. Esta observación indica la existencia de subgrupos específicos en los datos orgánicos que presentan características únicas y más alejadas del comportamiento general del grupo global [117].

La **Figura 5.22** y **Figura 5.23** presentan muestras tomadas entre los datos orgánicos y sintéticos en términos de sus espectros de frecuencia, separados por las clases *control* y *enfermos*. En ambas figuras se representa la relación entre la frecuencia (Hz) y la potencia de los espectros, facilitando la evaluación visual de la capacidad del modelo *espectraGAN* para replicar las características espectrales orgánicas.

En la **Figura 5.22**, correspondiente a la clase *control*, se observó que los datos orgánicos muestran una disminución suave y continua en la potencia conforme aumenta la frecuencia, con picos definidos en ciertas zonas de baja frecuencia, indicando características específicas propias de estas señales. Por su lado, los datos sintéticos generados replicaron eficazmente esta estructura global de los espectros orgánicos y los patrones principales. Aunque existen ligeras variaciones en la intensidad de algunos picos, estas diferencias no comprometieron significativamente la representación global. Consideramos que más bien, reflejaron la capacidad del modelo para incorporar variabilidad natural dentro del rango espectral típico observado en los datos orgánicos, destacando así una fortaleza intrínseca del modelo *espectraGAN*.

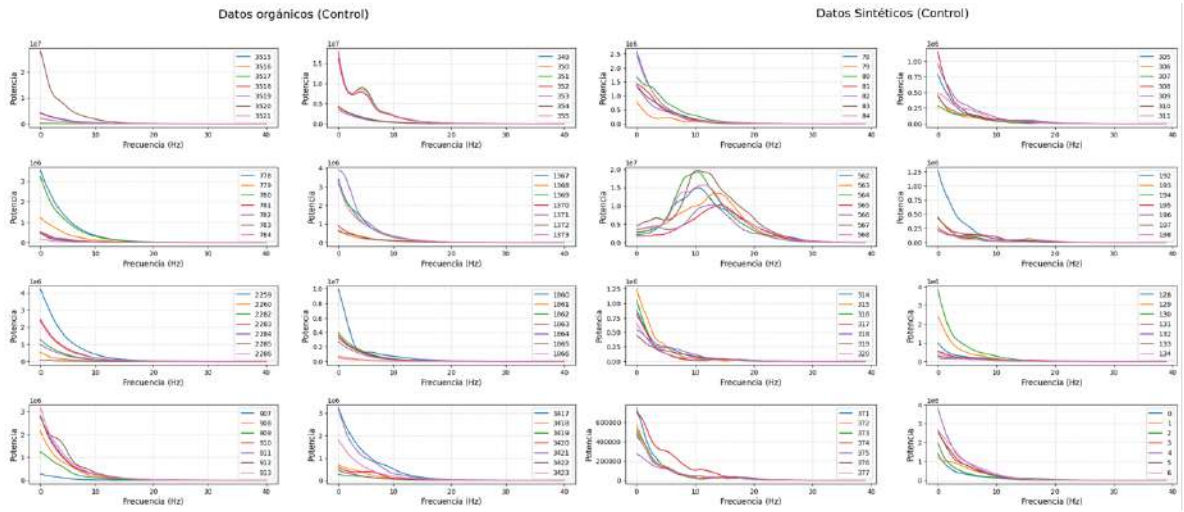


Figura 5.22: *Gráfica de componentes espectrales reales y sintéticos provenientes del ERG basal de la clase control. Estos espectros son tomados aleatoriamente de la base de datos de la clase correspondiente.*

Por otro lado, la **Figura 5.23** muestra los espectros correspondientes a la clase *enfermos*, los cuales presentan una variabilidad mayor tanto de los datos orgánicos como en los sintéticos. Los espectros orgánicos en esta clase exhiben un patrón general similar al de la clase *control*, caracterizado por caídas suaves en la potencia al incrementar frecuencia. No obstante, muestran mayor número de picos y oscilaciones, lo que reflejó heterogeneidad y complejidad espectral superior. Los datos sintéticos generados para esta clase también fueron reproducidos adecuadamente la forma general de los espectros, incluyendo la caída gradual de la potencia; sin embargo, presentó algunas discrepancias en regiones específicas.

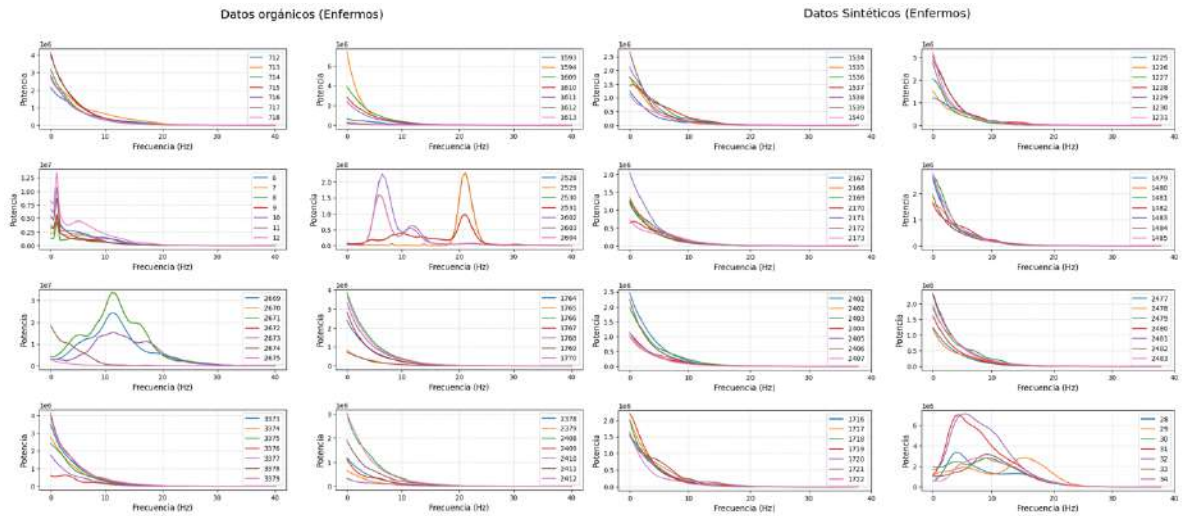


Figura 5.23: *Gráfica de componentes espectrales reales y sintéticos provenientes del ERG basal de la clase enfermos. Estos espectros son tomados aleatoriamente de la base de datos de la clase correspondiente.*

5.5 Análisis Cuantitativo

Con el objetivo de evaluar la calidad de los espectros generados por el modelo *espectraGAN*, se utilizaron como referencias las métricas empleadas en el estudio de Hazra y col. [85], quienes desarrollaron el modelo SynSigGAN para la generación de señales biomédicas artificiales. Dicho modelo trabaja con cuatro tipos de señalesfisiológicas: ECG, EEG, EMG y PPG [86].

	RMSE	PRD	MAE	FD
EspectraGAN				
Control	0.173	6.52	**0.014**	**0.113**
Disorder	0.22	7.26	**0.01**	**0.51**
SynSinGAN				
ECG	0.126	6.343	0.218	0.926
EEG	**0.0314**	5.985	0.047	0.982
EMG	0.0529	**2.971**	0.053	0.921
PPG	0.0596	5.167	0.063	0.783

Tabla 6: *Tabla comparativa de las métricas cuantitativas obtenidas por el modelo espectraGAN y el modelo SynSigGAN [85] los valores marcados en negritas se consideran como el mejor valor.*

Los valores de RMSE reportados por Hazra y col. [85] (<0.126) son notablemente menores a los de *espectraGAN*, indicando un menor valor de la discrepancia entre las potencias de los espectros sintéticos y los orgánicos. Esta diferencia podría atribuirse a que, la investigación de *SynSinGAN*, las señales biomédicas fueron normalizadas previamente en un rango entre 0 y 1, lo que condiciona la salida del modelo generador dentro de este mismo rango, mientras que nuestro modelo *espectraGAN* no impone restricciones de rango de potencias en su salida, generando espectros con potencias más variadas, incrementando la discrepancia observadas para le métrica.

En cuanto al PRD, esta métrica evalúa el nivel de discrepancia entre los conjuntos de datos. Un valor de 6.52 indica que, para la clase control, los espectros sintéticos por *EspectraGAN* presentan una desviación promedio del 6.52 % con respecto a las señales organicas. Para la clase enfermos, el PRD asciende a 7.26 %, lo que refleja una mayor diferencia.

El valor de MAE, cuantifica la diferencia promedio entre los conjuntos de datos [45]. En el contexto del modelo *espectraGAN* se calculó un valor de 0.01, valor que refleja un

nivel bajo de error medio, demostrando que, en promedio, los espectros sintéticos presentaron mínimas diferencias respecto a los orgánicos, lo que respalda la presión de nuestro modelo generativo.

Finalmente, la Fréchet Distance (FD) mide la similitud entre las distribuciones de los datos orgánicos y sintéticos en un espacio específico, donde valores más bajos indican mayor similitud y, por lo tanto, menor discrepancia entre ambas distribuciones [46]. Los valores obtenidos de 0.113 para la clase *control* y 0.51 para la clase *enfermos*, indicaron que la similitud entre los espectros sintéticos y orgánico es alta, particularmente en la clase *control.* El incremento observado en la clase *enfermos* (0.51) señaló una mayor discrepancia, confirmando nuevamente la mayor complejidad de replicar adecuadamente las características espectrales específicas presentes en este grupo.

5.6 Balance de clases

Las conclusiones antes enunciadas nos llevaron al objetivo principal de este trabajo, que es generar una base de datos sintética balanceada por grupo etario. Para ello, se seleccionó como referencia el grupo etario con mayor número de observaciones dentro de la base de datos orgánica. En este caso, dicho grupo etario correspondió a los de 55-64 años de la clase *enfermos*, con un total de 790 observaciones. Para establecer un criterio uniforme de generación de datos sintéticos, se decidió duplicar este valor, considerando así un máximo de datos sintéticos para las demás subclases.

En la **Tabla 7** se presentan en detalle los factores de generación aplicados a cada subclase, así como el total de datos sintéticos generados que conforman la base de datos sintética final. Por otro lado, la Figura 5.24 ilustra la distribución global de los datos orgánicos y los sintéticos.

	Grupo etario (años)	Observaciones	Factor de generación	Sintéticos	Total
Control					
	17-24	160	8.87	1420	1580
	25-34	328	3.82	1252	1580
	35-44	131	11.06	1449	1580
	45-54	73	20.64	1507	1580
	55-64	120	12.17	1460	1580
	65-74	28	55.43	1552	1580
	75-84	14	111.86	1566	1580
Total Control		**854**		**10206**	**11060**
Disorder					
	19-24	101	14.65	1479	1580
	25-34	201	6.86	1379	1580
	35-44	362	3.36	1218	1580
	45-54	693	1.28	887	1580
	55-64	790	1.00	790	1580
	65-74	487	2.24	1093	1580
	75-84	141	10.21	1439	1580
	85	36	42.89	1544	1580
Total Disorder		**2811**		**10829**	**12640**

Tabla 7: *Balance de clases para la base de datos ERG basal.*

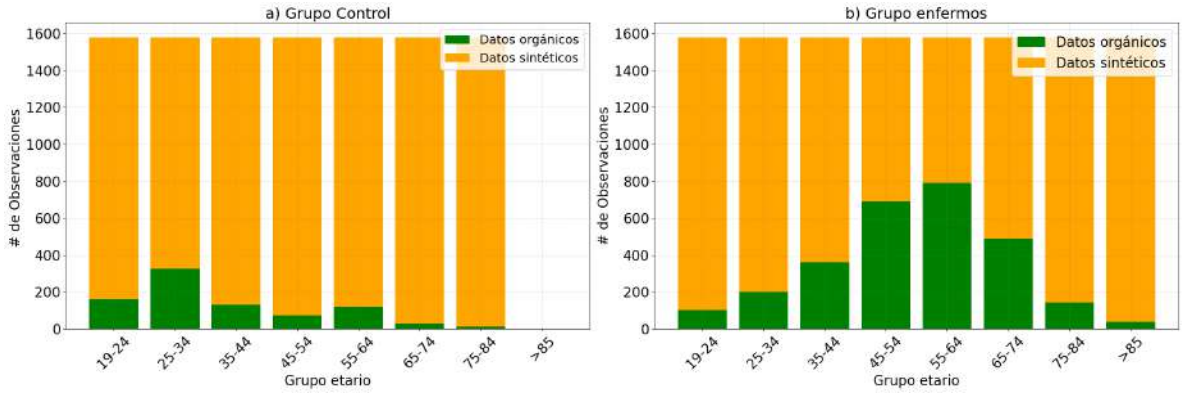


Figura 5.24: *Distribución de la base de datos de los espectros tanto orgánicos como sintéticos. En color verde se muestran los datos orgánicos y en naranja los datos sintéticos generados para obtener una base de datos balanceada.*

Capítulo 6

Discusión y conclusión

El objetivo principal de este estudio fue la generación de una base de datos de muestras sintéticas provenientes de oscilaciones espontáneas del ERG basal mediante la implementación de modelos generativos adversarios (GANs), con el fin de mejorar el rendimiento de un modelo de detección no invasiva y temprana de factores de riesgo modificables asociados a la DM tipo 2. A partir de esta investigación, se destacan varios hallazgos relevantes. En primer lugar, se estableció un marco metodológico para la generación de espectros sintéticos utilizando GANs, demostrando su capacidad para aprender la estructura y dinámica de estas series temporales, y evaluar la fidelidad de las características principales de los datos orgánicos tanto en nuestros grupos control y enfermos. En segundo lugar, se propuso e implementó un filtro estadístico basado en RMSE y la prueba de Mann-Whitney U, que permitió seleccionar únicamente aquellas muestras sintéticas con alta similitud a los datos orgánicos, sin que estas fueran réplicas exactas de los datos reales. En tercer lugar, se diseñó una metodología robusta para la evaluación cualitativa y cuantitativa de las series de tiempo sintéticas, lo cual representa una aportación metodológica al estado del arte. Finalmente, se abordó el problema del desbalance de clases mediante la ampliación de la base de datos orgánicas existente, demostrando que esta estrategia puede mejorar sustancialmente la calidad de las predicciones en modelos de DL al mitigar los sesgos asociados a la edad presentes en los conjuntos de datos orgánicos.

Investigaciones recientes han demostrado que el desempeños de modelos predictivos puede verse comprometido por base de datos pequeñas o desbalanceadas, especialmente cuando existen sesgos asociados con la edad [22], [39]. Este fenómeno particularmente relevante en estudios sobre DM tipo 2, donde la prevalencia de la enfermedad varia significativamente según los grupos etarios, ya que, en la mayoría de los estudios, los conjuntos de datos son limitados o no reflejan la diversidad de la población general, lo que

puede afectar la capacidad de generalización de los modelos [24], [39]. Para mitigar este sesgo, se implementó *espectraGAN* para generar señales sintéticas provenientes de ERG basal, equilibrando así la distribución de los grupos etarios.

Los análisis gráficos confirmaron que nuestro modelo fue capaz de generar espectros sintéticos con alto grado de fidelidad respecto a los datos orgánicos, preservando las estructuras globales y manteniendo proximidad entre los centroides de ambos conjuntos, lo cual respalda la coherencia fisiológica de las muestras generadas [91]. Por otra parte, se identificaron limitaciones al replicar características atípicas o complejas, este comportamiento se notó en la clase enfermos, que presente una heterogeneidad y dispersión significativamente mayores comparadas con la clase control [59]. A pesar de ello, el normalizar los espectros con respecto a la potencia demostró que, a pesar de las diferencias de potencias absolutas, las características morfológicas dominantes son reproducidas efectivamente [118]. Este patrón fue consistente a lo largo de varios grupos etarios, lo que reafirmó la capacidad de *espectraGAN* para mantener la integridad estructural de los espectros. Sin embargo, aunque el modelo logró reproducir satisfactoriamente las características generales de ambas clases, los retos persisten en particularmente en la representación más precisa de características específicas de los espectros sintéticos, como se mostró en la clase *control* 45-54 y la clase *enfermos* de los grupos etarios 45-54, 55-64, 65-74. Este hallazgo subraya la importancia de considerar ajustes adicionales en nuestro modelo o explorar técnicas complementarias para superar estas limitaciones y mejorar aún más la capacidad de *espectraGAN* [60], [95].

Un aspecto clave de esta investigación fue la decisión metodológica de emplear espectros de ERG transformados mediante wavelets como series temporales en lugar de imágenes (escalogramas), que suelen ser más comunes en el uso de GANs [24]. Esta selección de fundamente debido a que las series de tiempo preservan mejor la estructura temporal de la señal original [119]. Las señales de ERG basal presentan una estructura inherentemente temporal, donde la información clínica más valiosa está relacionada con patrones específicos de potencia y frecuencia, los cuales pueden perderse o distorsionarse al convertirlos en representaciones de imágenes [120]. Aunque las GAN basadas en imágenes son frecuentes en la literatura, no garantizan preservar adecuadamente las dependencias y

correlaciones temporales fundamentales de estas señales [121]. Además, utilizar imágenes como entrada puede implicar ciertas limitaciones, entre ellas la alta densidad de información en cada pixel, lo que incrementa significativamente la dimensionalidad de los datos, así como los requerimientos computacionales para su almacenamiento. También se requiere el uso de métodos especializados para el procesamiento, lo que se complica aún más por el carácter complejo, costoso e incluso invasivo para ciertas adquisición de estas[86], [122]. Por último, los modelos entrenados con imágenes suelen depender fuertemente del dominio de origen, dificultando su capacidad para generalizar a otros conjuntos y añadiendo una capa adicional de complejidad que restriegue el uso de métodos basados en imágenes [120], [121].

Por el contrario, la implementación de GAN orientadas a series temporales permitió mantener intactas estas características esenciales, asegurando que los espectros sintéticos generados tuvieran la calidad y coherencia fisiológica necesarias para ser utilizados eficazmente en modelos predictivos relacionados con factores de riesgo modificables de la DM tipo 2 [59], [84]. Las series de tiempo suelen ser unidimensional o multidimensional, pero con mucha menos dimensionalidad que una imagen, esto se traduce a una reducción de recursos para su almacenamiento [59].

Para validar los datos sintéticos generados, seleccionamos aquellos espectros más parecidos a los orgánicos utilizando la implementación de un filtro, a partir del RMSE y la prueba de U-Man. La similitud espectral fue clave para asegurar la coherencia fisiológica y evitar sesgos artificiales en los modelos predictivos. Además, dado el efecto documentado en esta tesis de la edad sobre las señales ERG, analizamos los espectros por grupo etario, garantizando así que las variaciones asociadas al envejecimiento fueran adecuadamente producidas.

Con la finalidad de complementar el estudio anterior, nuestro equipo de trabajo evaluó el desempeño de diferentes modelos de DL, específicamente, SVM, RF, CNN, RNN y regresión logística, utilizando la transformación mediante la onda Wavelet Real Morlet. Los cinco mejores modelos se clasificaron según su ROC-AUC y los valores de exactitud, precisión, sensibilidad, especificidad y puntuación F1 (**Figura 6.1 (a))**.

El modelo que mostro los resultados más destacados fue el SVM, alcanzando un valor de ROC-AUC de 0.60. Es modelo se distinguió especialmente en la métrica de especificidad, obteniendo el valor más alto entre todos los modelos evaluados (0.90). Sin embargo, presento una baja sensibilidad, con un valor de 0.26. Esta disparidad indica que, aunque SVM es eficiente identificando correctamente casos negativos (especificidad alta), su capacidad para detectar correctamente casos negativos (sensibilidad baja) [123]. Por otro lado, los modelos RF, CNN, RNN y Regresión Logística mostraron valores similares de ROC-AUC, variando entre 0.59 y 0.60, mostrando desempeños comparables en términos generales. En particular, las RNN destacaron al obtener la precisión más alta (0.70), indicando un buen desempeño en cuanto a la proporción de predicciones correctas sobre todas las predicciones positivas realizadas. No obstante, este modelo también presento una sensibilidad relativamente baja (0.34), lo que implica que, aunque es preciso, su capacidad para identificar casos positivos reales es muy moderada [123].

Es relevante mencionar que las RNN obtuvieron un mejor equilibrio general, representando en un valor de sensibilidad más alto (0.47) respecto a los otros modelos, así como una puntuación F1 más equilibrada (0.51).

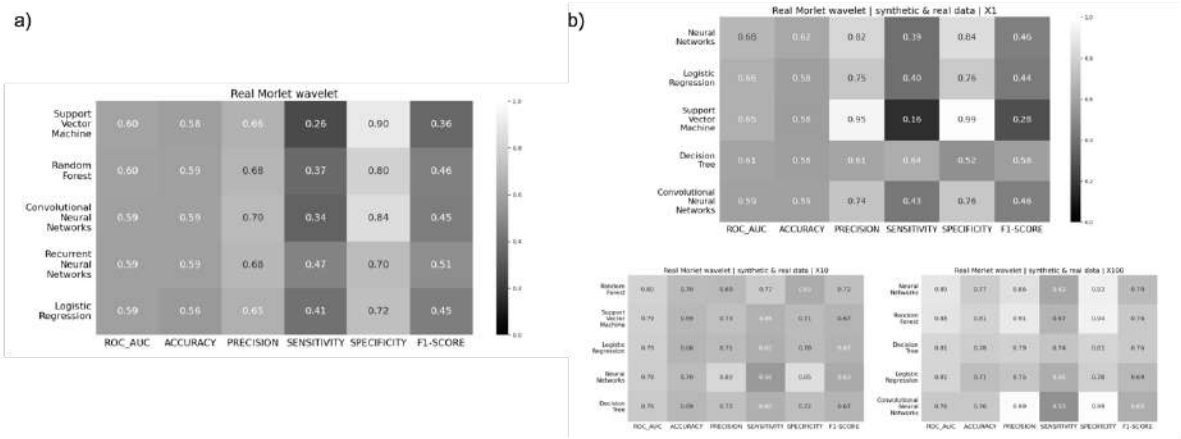


Figura 6.1: *Comparativa de los cinco mejores modelos para predecir factores de riesgo asociados con la DM tipo 2 usando ERG basal entrenados con real Morlet Wavelet. a) Modelos entrenados únicamente con los datos orgánicos. b) Modelos entrenados con balance de clases generando espectros sintéticos mediante espectraGAN, generando x1, x10 y x100 la cantidad de espectros sintéticos.*

Adicionalmente como se muestra en la **Figura 6.1 (b)**, se realizó una evaluación comparativa incorporando los datos sintéticos. Ese análisis tuvo como objetivo determinar el efecto que tiene el balancear las clases de la base de datos orgánica con datos sintéticos. Se

compararon escenarios en los cuales se incrementó la cantidad de datos combinados de 1 (únicamente balance de clase), 10 y 100 veces respecto al tamaño de la base de datos balanceada.

En el caso del conjunto de datos balanceado (x1), nuevamente se observó que el modelo SVM destacó significativamente en términos de precisión (0.95) y especificidad (0.99). No obstante, se evidenció una notable limitación en la sensibilidad (0.16), reiterando las deficiencias del modelo para detectar casos positivos reales cuando no existe suficiente cantidad de estos [123]. Otros modelos como las RNN y la regresión logística mostraron un equilibrio más aceptable, con especificidades altas (0.84 y 0.77 respectivamente), aunque sus sensibilidades (0.39 y 0.40) aún fueron limitadas. Al analizar los resultados incrementar 10 veces (x10) la cantidad de datos combinados, se observó una mejora considerable generalizada en casi todas las métricas. El modelo RF obtuvo un desempeño más equilibrado con un ROC-AUC de 0.80, sensibilidad de 0.77 y puntuación F1 de 0.72, indicando una eficiencia importante al balancear precisión y sensibilidad [123]. Otros modelos como SVM y regresión logística también mostraron mejoras notables en sensibilidad y puntuación F1 respecto al primer experimento (**Figura 6.1 (a))**, lo que reflejó que el aumento en el volumen de datos mejoró la detección de casos positivos. Finalmente, al evaluar el desempeño tras incrementar 100 veces los datos sintéticos, se obtuvo el mayor rendimiento en todas las métricas evaluadas. Las RNN alcanzaron valores particularmente altos, obteniendo un ROC-AUC de 0.89, una precisión de 0.86 y una sensibilidad de 0.62. De forma similar, RF mostró un desempeño sobresaliente con un ROC-AUC de 0.88, precisión 0.91, y especificidad particularmente alta de 0.94. Este escenario demuestra cómo un incremento sustancial en la cantidad de datos puede mitigar eficientemente problemas de desequilibrio de clase, incrementando la capacidad predictiva global de los modelos analizados.

En esta investigación se logró diseñar satisfactoriamente un catálogo de muestras sintéticas provenientes de señales ERG basal mediante el uso un de modelos GAN, llamado *espectraGAN*. La calidad de estos espectros sintéticos fue evaluada tanto cuantitativa como cualitativa, demostrando resultados robustos con un alto grado de similitud espectral y coherencia fisiológica respecto a los datos orgánicos. La implementación de datos sintéticos mostró ser especialmente útil para balancear las clases y reducir significativamente el sesgo

por desbalance en los datos orgánicos, lo que permitió una notable aumento en el desempeño del modelo predictivo existente. Esta metodología contribuye a solventar problemas de sesgo desbalance en la base de datos original, ofreciendo una herramienta potente y accesible para la detección temprana y no invasiva a factores de riesgo asociados con la DM tipo 2.

[1] “International Diabetes Federation (IDF).” Accessed: Jul. 18, 2023. [Online]. Available: https://idf.org/

[2] INEGI, “Estadísticas a propósito del Día Mundial de la Diabetes (14 de noviembre).” Accessed: Jan. 02, 2025. [Online]. Available: https://www.inegi.org.mx/app/saladeprensa/noticia/7746

[3] A. Basto-Abreu *et al.*, “[06] Prevalence of prediabetes and diabetes in Mexico: Ensanut 2022.,” *Salud Publica Mex*, vol. 65, 2023.

[4] I. De and L. A. Ocde, “[09] Panorama de la Salud 2021,” 2022, Accessed: Dec. 02, 2023. [Online]. Available: https://doi.org/10.1787/ae3016b9-en.

[5] R. Baralt-Zamudio *et al.*, “[10] ¿Por qué es necesario actualizar la información epidemiológica en México?,” *Revista mexicana de oftalmología*, vol. 96, no. 4, pp. 188–189, Jul. 2022, Accessed: Dec. 02, 2023. [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2604-12272022000400188&lng=es&nrm=iso&tlng=es

[6] L. Daniel Tavares *et al.*, “[70] Prediction of metabolic syndrome: A machine learning approach to help primary prevention,” *Diabetes Res Clin Pract*, Sep. 2022.

[7] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[8] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 706–716. doi: 10.1016/j.procs.2020.03.336.

[9] O. T. Kee *et al.*, “Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review,” Dec. 01, 2023, *BioMed Central Ltd*. doi: 10.1186/s12933-023-01741-7.

[10] H. M. Deberneh and Kim Intaek, “Prediction of Type 2 Diabetes Based on Machine Learning Algorithm | Enhanced Reader,” 2021.

[11] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, “[15] Artificial intelligence in retina,” 2018, *Elsevier* .

[12] American Society of Retina Specialists, “[12] Guía de imágenes diagnosticas de la retina,” 2022.

[13] G. Ghirlanda *et al.*, “[40] Detection of inner retina dysfunction by steady-state focal electroretinogram pattern and flicker in early IDDM,” *Diabetes*, vol. 40, no. 9, pp. 1122–1127, 1991, Accessed: Nov. 28, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1936619/

[14] R. An, J. Shen, and Y. Xiao, “[69] Applications of Artificial Intelligence to Obesity Research: Scoping Review of Methodologies,” *J Med Internet Res 2022;24(12):e40589 https://www.jmir.org/2022/12/e40589*, vol. 24, no. 12, p. e40589, Dec. 2022, Accessed: Dec. 06, 2023. [Online]. Available: https://www.jmir.org/2022/12/e40589

[15] M. W. Stewart, “[52] Diabetic Retinopathy Current Pharmacologic Treatment and Emerging Strategies,” *Springer*, 2017.

[16] K. Kato, M. Kondo, M. Sugimoto, K. Ikesugi, and H. Matsubara, “[75] Effect of Pupil Size on Flicker ERGs Recorded With RETeval System: New Mydriasis-Free Full-Field ERG System,” *Invest Ophthalmol Vis Sci*, Jun. 2015, Accessed: Dec. 06, 2023. [Online]. Available: http://www.lkc.com/

[17] İ. S. Yapici, O. Erkaymaz, and R. U. Arslan, “A hybrid intelligent classifier to estimate obesity levels based on ERG signals,” *Phys Lett A*, vol. 399, p. 127281, May 2021, doi: 10.1016/J.PHYSLETA.2021.127281.

[18] R. Noguez Imm *et al.*, “Preventable risk factors for type 2 diabetes can be detected using noninvasive spontaneous electroretinogram signals,” 2023, doi: 10.1371/journal.pone.0278388.

[19] J. M. Barcala Riveira, J. L. Fernandez Marron, J. Alberdi Primicia, J. J. Navarrete Marin, and J. C. Oller Gonzalez, “Application of Wavelets and Quaternions to NIR Spectra Classification; Aplicacion de las Wavelests y los Cuaterniones a la Clasificaciond e Espectros NIR,” 2003.

[20] C. A. Rodríguez-Arzate *et al.*, “Potential contributions of the intrinsic retinal oscillations recording using non-invasive electroretinogram to bioelectronics,” *Front Cell Neurosci*, vol. 17, p. 1224558, Jan. 2023, doi: 10.3389/FNCEL.2023.1224558/BIBTEX.

[21] R. Noguez Imm *et al.*, “Preventable risk factors for type 2 diabetes can be detected using noninvasive spontaneous electroretinogram signals,” 2023, doi: 10.1371/journal.pone.0278388.

[22] R. N. Imm *et al.*, “Preventable risk factors for type 2 diabetes can be detected using noninvasive spontaneous electroretinogram signals,” *PLoS One*, vol. 18, no. 1 January, Jan. 2023, doi: 10.1371/JOURNAL.PONE.0278388.

[23] N. Japkowicz and S. Stephen, “[43] The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[24] I. J. Goodfellow *et al.*, “[44] Generative Adversarial Nets,” 2014.

[25] S. Wan, Y. Liang, and Y. Zhang, “[45] Deep convolutional neural networks for diabetic retinopathy detection by image classification,” *Computers and Electrical Engineering*, vol. 72, pp. 274–282, 2018.

[26] D. Nankani and R. Dutta Baruah, “[46] Improved Diagnostic Performance of Arrhythmia Classification Using Conditional GAN Augmented Heartbeats,” *Intelligent Systems Reference Library*, vol. 217, pp. 275–304, 2022, doi: 10.1007/978-3-030-91390-8_12/COVER.

[27] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, “[47] EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals,” *ArXiv*, 2018.

[28] Y. Xia, Y. Xu, P. Chen, J. Zhang, and Y. Zhang, "[48] Generative adversarial network with transformer generator for boosting ECG classification," *Biomed Signal Process Control*, vol. 80, 2023, doi: 10.1016/j.bspc.2022.104276.

[29] T. Golany, G. Lavee, S. T. Yarden, and K. Radinsky, "[49] Improving ECG Classification Using Generative Adversarial Networks," 2020.

[30] A. M. Delaney, E. Brophy, and T. E. Ward, "[50] Synthesis of Realistic ECG using Generative Adversarial Networks," *ArXiv*, 2019.

[31] K. F. Hossain *et al.*, "[51] ECG-Adv-GAN: Detecting ECG Adversarial Examples with Conditional Generative Adversarial Networks," in *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*, 2021, pp. 50–56. doi: 10.1109/ICMLA52953.2021.00016.

[32] "Desbalance total en la formación de médicos." Accessed: Apr. 07, 2025. [Online]. Available: https://www.eleconomista.com.mx/opinion/Desbalance-total-en-la-formacion-de-medicos-20180520-0030.html

[33] "Déficit de médicos especialistas asciende a 154 mil 786 profesionales." Accessed: Apr. 07, 2025. [Online]. Available: https://contralinea.com.mx/interno/semana/deficit-de-medicos-especialistas-asciende-a-154-mil-786-profesionales/

[34] "México tiene falta de control oportuno en pacientes con diabetes - Federación Mexicana de Diabetes, A.C." Accessed: Apr. 07, 2025. [Online]. Available: https://fmdiabetes.org/mexico-falta-control-oportuno-pacientes-diabetes/

[35] M. F. Faruque, Asaduzzaman, S. M. M. Hossain, M. H. Furhad, and I. H. Sarker, "Predicting diabetes mellitus and analysing risk-factors correlation," *EAI Endorsed Trans Pervasive Health Technol*, vol. 5, no. 20, 2020, doi: 10.4108/eai.13-7-2018.164173.

[36] I. Lemieux, "Reversing Type 2 Diabetes: The Time for Lifestyle Medicine Has Come!," *Nutrients 2020, Vol. 12, Page 1974*, vol. 12, no. 7, p. 1974, Jul. 2020, doi: 10.3390/NU12071974.

[37] B. Gómez Marín *et al.*, *[59] Manual de Riesgo Cardiovascular*. 2021.

[38] “How ERG Helps Boost Clinical & Financial Success - Review of Optometric Business.” Accessed: May 10, 2025. [Online]. Available: https://reviewob.com/how-erg-helps-boost-clinical-financial-success/?utm_source=chatgpt.com

[39] Z. Chen, J. Duan, L. Kang, and G. Qiu, “Class-Imbalanced Deep Learning via a Class-Balanced Ensemble,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 10, pp. 5626–5640, Oct. 2022, doi: 10.1109/TNNLS.2021.3071122.

[40] F. Torres, A. Rojas, F. Torres, and A. Rojas, “[66] Obesidad y salud pública en México: transformación del patrón hegemónico de oferta-demanda de alimentos,” *Probl Desarro*, vol. 49, no. 193, pp. 145–169, Apr. 2018, Accessed: Nov. 28, 2023. [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0301-70362018000200145&lng=es&nrm=iso&tlng=es

[41] D. Lanzagorta-Ortega, D. L. Carrillo-Pérez, R. Carrillo-Esper, D. Lanzagorta-Ortega, D. L. Carrillo-Pérez, and R. Carrillo-Esper, “[63] Inteligencia artificial en medicina: presente y futuro,” *Gac Med Mex*, vol. 158, pp. 17–21, Dec. 2022, Accessed: Dec. 03, 2023. [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0016-38132022001100017&lng=es&nrm=iso&tlng=es

[42] J. F. Ávila-Tomás, M. A. Mayer-Pujadas, and V. J. Quesada-Varela, “[64] Artificial intelligence and its applications in medicine II: Current importance and practical applications,” *Aten Primaria*, 2021.

[43] INEGI, “Economía y Sectores Productivos.” Accessed: Jan. 03, 2025. [Online]. Available: https://www.inegi.org.mx/temas/saludsat/

[44] A. Bener, M. T. Yousafzai, S. Darwish, A. O. A. A. Al-Hamaq, E. A. Nasralla, and M. Abdul-Ghani, “[68] Obesity index that better predict metabolic syndrome: Body mass index, waist circumference, waist hip ratio, or waist height ratio,” *J Obes*, vol. 2013, 2013.

[45] G. Peltz, M. T. Aguirre, M. Sanderson, and M. K. Fadden, “[67] The role of fat mass index in determining obesity,” *Am J Hum Biol*, vol. 22, no. 5, pp. 639–647, Sep. 2010, Accessed: Nov. 30, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20737611/

[46] P. Lachapelle, “[23] The human suprathreshold photopic oscillatory potentials: method of analysis and clinical application,” *Doc Ophthalmol*, vol. 88, no. 1, pp. 1–25, 1994, Accessed: Nov. 28, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/7743909/

[47] H. Chen, M. Zhang, S. Huang, and D. Wu, “[36] The photopic negative response of flash ERG in nonproliferative diabetic retinopathy,” *Doc Ophthalmol*, vol. 117, no. 2, pp. 129–135, 2008, Accessed: Nov. 28, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18214565/

[48] D. Yonemura, K. Tsuzuki, and T. Aoki, “[32] Clinical importance of the oscillatory potential in the human ERG,” *Acta Ophthalmol Suppl*, vol. Suppl 70, no. 70 S, pp. 115–123, 1962, Accessed: Nov. 28, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/14040380/

[49] B. D. Kels, A. Grzybowski, and J. M. Grant-Kels, “[76] Human ocular anatomy,” *Clin Dermatol*, vol. 33, no. 2, pp. 140–146, 2015.

[50] Rojas Juárez Sergio and Saucedo Castillo Adriana, *[77] Retina y vítreo*, vol. 2. 2012.

[51] J. E. HALL, *[78] Guyton y Hall. Tratado de fisiología médica*, vol. 14. 2011.

[52] Fernando Arévalo *et al.*, *[05] Retina 2019*. 2019.

[53] Y. Kanagasingam, A. Bhuiyan, M. D. Abràmoff, R. T. Smith, L. Goldschmidt, and T. Y. Wong, “Progress on retinal image analysis for age related macular degeneration,” *Prog Retin Eye Res*, vol. 38, pp. 20–42, Jan. 2014, doi: 10.1016/J.PRETEYERES.2013.10.002.

[54] N. M. Bressler, “Photodynamic therapy of subfoveal choroidal neovascularization in age- related macular degeneration with verteporfin: One-year results of 2 randomized

clinical trials - TAP report 1," *Archives of Ophthalmology*, vol. 117, no. 10, pp. 1329–1345, 1999, doi: 10.1001/ARCHOPHT.117.10.1329.

[55] D. I. Conget, "Diagnóstico, clasificación y patogenia de la diabetes mellitus," *Rev Esp Cardiol*, vol. 55, no. 5, pp. 528–535, Jan. 2002, doi: 10.1016/S0300-8932(02)76646-3.

[56] J. G. WEBSTER, *[84] Medical instrumentation: application and design*, vol. 14. 2009.

[57] T. H. Yang, E. Y. C. Kang, P. H. Lin, P. L. Wu, J. A. Sachs, and N. K. Wang, "The Value of Electroretinography in Identifying Candidate Genes for Inherited Retinal Dystrophies: A Diagnostic Guide," Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/diagnostics13193041.

[58] J. A. Hughes-Cano *et al.*, "Improved predictive diagnosis of diabetic macular edema based on hybrid models: An observational study," *Comput Biol Med*, vol. 170, p. 107979, Mar. 2024, doi: 10.1016/J.COMPBIOMED.2024.107979.

[59] J. Brownlee, "[85] Generative Adversarial Networks with Python Deep Learning Generative Models for Image Synthesis and Image Translation," 2019.

[60] J. Langr and V. Bok, "Deep learning with Generative Adversarial Networks."

[61] A. Basto-Abreu *et al.*, "Prevalence of prediabetes and diabetes in Mexico: Ensanut 2022.," *Salud Publica Mex*, vol. 65, 2023, doi: 10.21149/14832.

[62] D. Magliano and E. Boyko, "IDF Diabetes Atlas 11th Edition," 2025.

[63] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *Int J Environ Res Public Health*, vol. 18, no. 14, Jul. 2021, doi: 10.3390/ijerph18147346.

[64] I. Gandin *et al.*, "Deep-learning-based prognostic modeling for incident heart failure in patients with diabetes using electronic health records: A retrospective cohort study," *PLoS One*, vol. 18, no. 2 February, Feb. 2023, doi: 10.1371/journal.pone.0281878.

[65] S. L. Cichosz, M. D. Johansen, S. T. Knudsen, T. K. Hansen, and O. Hejlesen, "A classification model for predicting eye disease in newly diagnosed people with type 2 diabetes," *Diabetes Res Clin Pract*, vol. 108, no. 2, pp. 210–215, May 2015, doi: 10.1016/j.diabres.2015.02.020.

[66] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[67] O. Metsker *et al.*, "Identification of risk factors for patients with diabetes: Diabetic polyneuropathy case study," *BMC Med Inform Decis Mak*, vol. 20, no. 1, Aug. 2020, doi: 10.1186/s12911-020-01215-w.

[68] N. Al-Sari *et al.*, "Precision Diagnostic Approach to Predict 5-Year Risk for Microvascular Complications in Type 1 Diabetes," Sep. 29, 2021. doi: 10.1101/2021.09.28.21264161.

[69] Y. Wang *et al.*, "Genetic Risk Score Increased Discriminant Efficiency of Predictive Models for Type 2 Diabetes Mellitus Using Machine Learning: Cohort Study | Enhanced Reader," 2021.

[70] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," Dec. 01, 2021, *BioMed Central Ltd*. doi: 10.1186/s13098-021-00767-9.

[71] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[72] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J Clin Epidemiol*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006, doi: 10.1016/J.JCLINEPI.2006.01.014.

[73] S. Jaiswal and P. Gupta, "GLSTM: A novel approach for prediction of real & synthetic PID diabetes data using GANs and LSTM classification model," *International Journal*

of Experimental Research and Review, vol. 30, pp. 32–45, 2023, doi: 10.52756/ijerr.2023.v30.004.

[74] D. Boughareb, H. Bensalah, and H. Seridi, "A Hybrid GAN-ANN-Based Model for Diabetes Prediction," *Research and Technology In Association with International Journal of Scientific Research in Science and Technology*, vol. 10, 2023, [Online]. Available: www.ijsrst.com

[75] D. Chushing-Muzo, H. Calero-Díaz, F. Lara-Abelenda, and Gómez-Martínez Vanesa, "Interpretable Data-Driven Approach Based on Feature Selection Methods and GAN-Based Models for Cardiovascular Risk Prediction in Diabetic Patients | Enhanced Reader," 2024.

[76] T. Golany, G. Lavee, S. T. Yarden, and K. Radinsky, "Improving ECG Classification Using Generative Adversarial Networks," 2020. [Online]. Available: https://bitbucket.org/

[77] K. Fariha-Hossain, S. Amit-Kamram, A. Tavakkoli, L. Pan, and X. Ma, "ECG-adv-GAN detecting ECG adversarial examples with conditional generative adversarial network," 2021.

[78] X. Li, A. Hiong Ngu, and V. Metsis, "TT-CCGAN a transformer times series conditional GAN for biodignal," 2022.

[79] S. Festag, J. Denzler, and C. Spreckelsen, "Generative adversarial networks for biomedical time series forecasting and imputation: A systematic review," May 01, 2022, *Academic Press Inc.* doi: 10.1016/j.jbi.2022.104058.

[80] S. Liao, H. Ni, M. Sabate-Vidales, L. Szpruch, M. Wiase, and B. Xiao, "Sig-Wasserstein GANs for conditional time series generation | Enhanced Reader," 2022.

[81] M. Kaisti, J. Laitala, D. Wong, and A. Airola, "Domain randomization using synthetic electrocardiograms for training neural networks," *Artif Intell Med*, vol. 143, Sep. 2023, doi: 10.1016/j.artmed.2023.102583.

[82] N. Audebert, B. Le Saux, and S. Lefèvre, “Generative adversarial networks for realistic synthesis of hyperspectral samples,” *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2018-July, pp. 4359–4362, Oct. 2018, doi: 10.1109/IGARSS.2018.8518321.

[83] K. E. Smith and A. O. Smith, “Conditional GAN for time series generation,” *arXiv preprint*, 2020.

[84] E. Pavlou and N. Kourkoumelis, “Deep adversarial data augmentation for biomedical spectroscopy: Application to modelling Raman spectra of bone,” *Chemometrics and Intelligent Laboratory Systems*, vol. 228, p. 104634, Sep. 2022, doi: 10.1016/J.CHEMOLAB.2022.104634.

[85] D. Hazra and Y. C. Byun, “SynSigGAN: Generative Adversarial Networks for Synthetic Biomedical Signal Generation,” *Biology 2020, Vol. 9, Page 441*, vol. 9, no. 12, p. 441, Dec. 2020, doi: 10.3390/BIOLOGY9120441.

[86] O. Hernandez, J. Muñoz, and S. Thebault, “Redes generativas adversarias para la generación de componentes espectrales sintéticos,” *Memorias XVII Coloquio de Posgrado FI UAQ*, vol. XVII, pp. 172–184, 2023.

[87] G. Perea and V. Benfenati, “Potential contributions of the intrinsic retinal oscillations recording using non-invasive electroretinogram to bioelectronics,” 2024, doi: 10.3389/fncel.2023.1224558.

[88] “Estadísticas a propósito del Día Mundial de la Diabetes (14 de noviembre),” INEGI. Accessed: Nov. 25, 2023. [Online]. Available: https://www.inegi.org.mx/app/saladeprensa/noticia.html?id=7746

[89] “scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation.” Accessed: Feb. 21, 2025. [Online]. Available: https://scikit-learn.org/stable/

[90] L. Ferré, “Selection of components in principal component analysis: A comparison of methods,” *Comput Stat Data Anal*, vol. 19, no. 6, pp. 669–682, 1995.

[91] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers 2022 2:1*, vol. 2, no. 1, pp. 1–21, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

[92] "TensorFlow." Accessed: Feb. 21, 2025. [Online]. Available: https://www.tensorflow.org/?hl=es-419

[93] "What's New In Python 3.8 — Python 3.8.20 documentation." Accessed: Feb. 21, 2025. [Online]. Available: https://docs.python.org/3.8/whatsnew/3.8.html

[94] "Usa una GPU | TensorFlow Core." Accessed: Feb. 21, 2025. [Online]. Available: https://www.tensorflow.org/guide/gpu?hl=es-419

[95] V. Dumoulin, F. Visin, and G. E. P. Box, "A guide to convolution arithmetic for deep learning," Mar. 2016, Accessed: Feb. 21, 2025. [Online]. Available: https://arxiv.org/abs/1603.07285v2

[96] M. A. Mercioni and S. Holban, "P-Swish: Activation Function with Learnable Parameters Based on Swish Activation Function in Deep Learning," *2020 14th International Symposium on Electronics and Telecommunications, ISETC 2020 - Conference Proceedings*, Nov. 2020, doi: 10.1109/ISETC50328.2020.9301059.

[97] P.-A. Muller, J. Laetitia, and J. Weber, "Deep learning for time series classification," Oct. 2020, Accessed: Feb. 21, 2025. [Online]. Available: https://arxiv.org/abs/2010.00567v1

[98] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Deep learning for time series classification: a review," *Data Min Knowl Discov*, vol. 33, no. 4, pp. 917–963, Jul. 2019, doi: 10.1007/S10618-019-00619-1/METRICS.

[99] A. Sokolowski, "Modeling with Exponential Decay Function," *Scientific Inquiry in Mathematics - Theory and Practice*, pp. 65–82, 2018, doi: 10.1007/978-3-319-89524-6_7.

[100] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep Anomaly Detection with Outlier Exposure," *7th International Conference on Learning Representations, ICLR 2019*,

Dec. 2018, Accessed: Feb. 21, 2025. [Online]. Available: https://arxiv.org/abs/1812.04606v3

[101] L. Rice, E. Wong, and J. Z. Kolter, “Overfitting in adversarially robust deep learning,” Nov. 21, 2020, *PMLR*. Accessed: Feb. 21, 2025. [Online]. Available: https://proceedings.mlr.press/v119/rice20a.html

[102] D. P. Kingma and J. Lei Ba, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION”.

[103] P. E. McKnight and J. Najab, “Mann-Whitney U Test,” *The Corsini Encyclopedia of Psychology*, pp. 1–1, Jan. 2010, doi: 10.1002/9780470479216.CORPSY0524.

[104] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geosci Model Dev*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/GMD-15-5481-2022.

[105] P. Maragos and R. W. Schafer, “Morphological Systems for Multidimensional Signal Processing,” *Proceedings of the IEEE*, vol. 78, no. 4, pp. 690–710, 1990, doi: 10.1109/5.54808.

[106] T. Wright, F. Cortese, J. Nilsson, and C. Westall, “[42] Analysis of multifocal electroretinograms from a population with type 1 diabetes using partial least squares reveals spatial and temporal distribution of changes to retinal function,” *Doc Ophthalmol*, vol. 125, no. 1, pp. 31–42, 2012, Accessed: Nov. 28, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22610144/

[107] C.-L. Tseng, C.-C. ¶ Hsiao, I.-C. Chou, C.-J. Hsu, Y.-J. Chang, and R.-G. Lee, “DESIGN AND IMPLEMENTATION OF ECG COMPRESSION ALGORITHM WITH CONTROLLABLE PERCENT ROOT-MEAN-SQUARE DIFFERENCE,” 2007. [Online]. Available: www.worldscientific.com

[108] A. Obukhov and M. Krasnyanskiy, “Quality Assessment Method for GAN Based on Modified Metrics Inception Score and Fréchet Inception Distance,” *Advances in Intelligent Systems and Computing*, vol. 1294, pp. 102–114, 2020, doi: 10.1007/978-3-030-63322-6_8.

[109] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach Learn*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/S10994-022-06268-8/FIGURES/27.

[110] Fanny and T. W. Cenggoro, "Deep Learning for Imbalance Data Classification using Class Expert Generative Adversarial Network," *Procedia Comput Sci*, vol. 135, pp. 60–67, Jan. 2018, doi: 10.1016/J.PROCS.2018.08.150.

[111] J. Brownlee, "[31] Generative Adversarial Networks with Python Deep Learning Generative Models for Image Synthesis and Image Translation," 2019.

[112] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," in *Proceedings of the VLDB Endowment*, 2018, pp. 1071–1083.

[113] S. Ramachandra, A. Hoelzemann, and K. Van Laerhoven, "Transformer Networks for Data Augmentation of Human Physical Activity Recognition," *ArXiv*, Sep. 2021.

[114] T. Kurita, "Principal Component Analysis (PCA)," *Computer Vision*, pp. 1–4, 2020, doi: 10.1007/978-3-030-03243-2_649-1.

[115] "Análisis factorial: Extracción - Documentación de IBM." Accessed: Aug. 24, 2024. [Online]. Available: https://www.ibm.com/docs/es/spss-statistics/saas?topic=analysis-factor-extraction

[116] M. Wattenberg, F. Viégas, and I. Johnson, "How to Use t-SNE Effectively," *Distill*, vol. 1, no. 10, p. e2, 2016, doi: 10.23915/distill.00002.

[117] G. H. de Rosa, J. R. F. Brega, and J. P. Papa, "How optimizing perplexity can affect the dimensionality reduction on word embeddings visualization?," *SN Appl Sci*, vol. 1, no. 12, pp. 1–17, Dec. 2019, doi: 10.1007/S42452-019-1689-4/FIGURES/14.

[118] V. V. Dvoeglazov, "On the importance of the normalization," *Czechoslovak Journal of Physics*, vol. 50, no. 2, pp. 225–237, Dec. 1997, doi: 10.1023/A:1022895923841.

[119] N. Wulan, W. Wang, P. Sun, K. Wang, Y. Xia, and H. Zhang, “Generating electrocardiogram signals by deep learning,” *Neurocomputing*, vol. 404, pp. 122–136, Sep. 2020, doi: 10.1016/J.NEUCOM.2020.04.076.

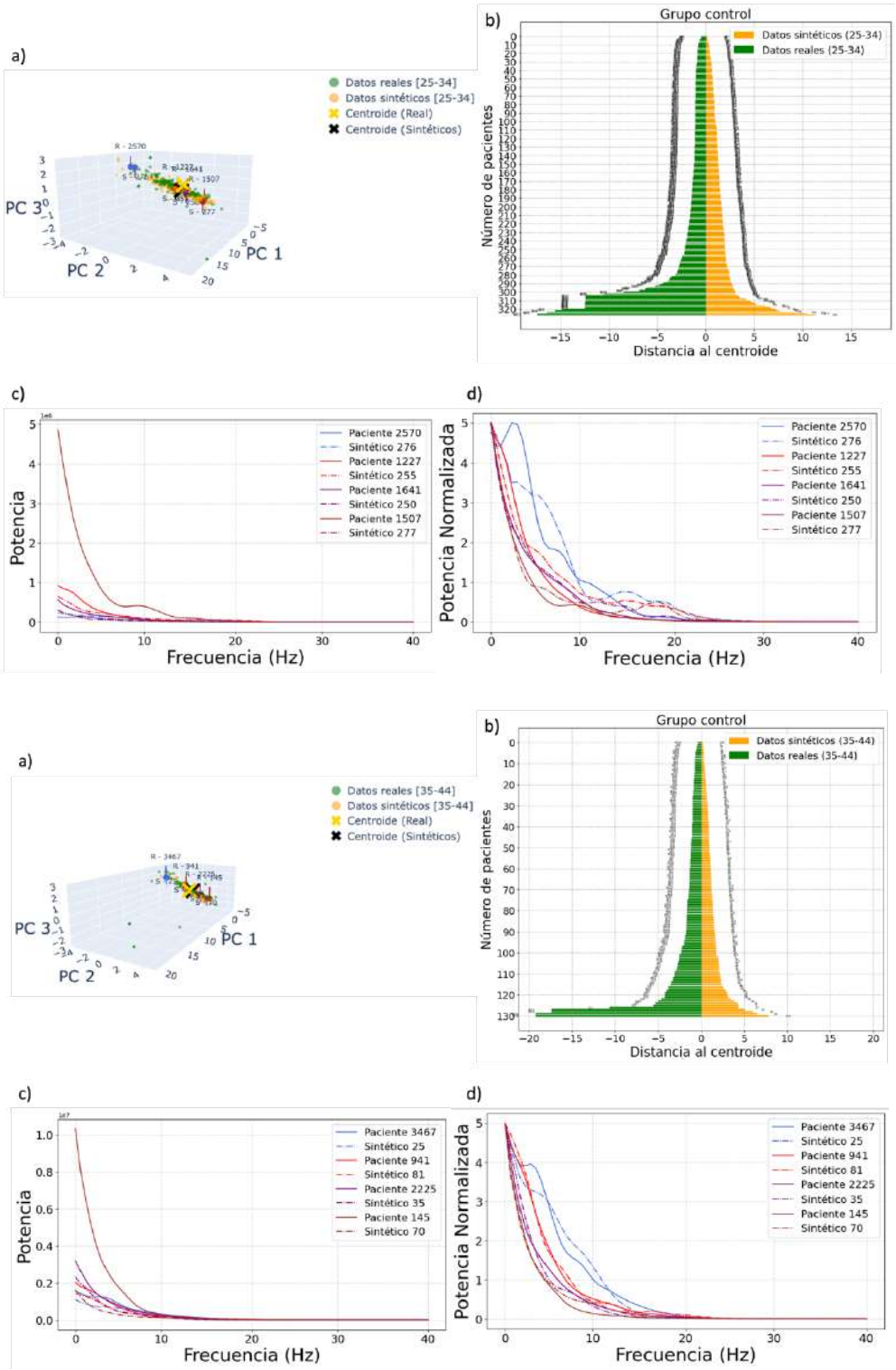
[120] K. H. Chon *et al.*, “Enhancing Electroretinogram Classification with Multi-Wavelet Analysis and Visual Transformer,” 2023, doi: 10.3390/s23218727.

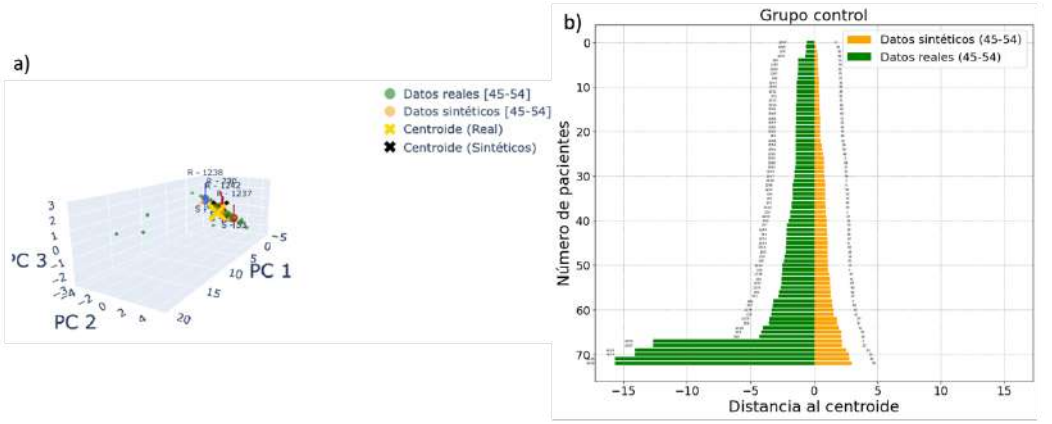
[121] C. Gupta, P. Kamath, and L. Wyse, “Signal Representations for Synthesizing Audio Textures with Generative Adversarial Networks,” *Proceedings of the Sound and Music Computing Conferences*, vol. 2021-June, pp. 159–166, Mar. 2021, doi: 10.5281/zenodo.5054145.

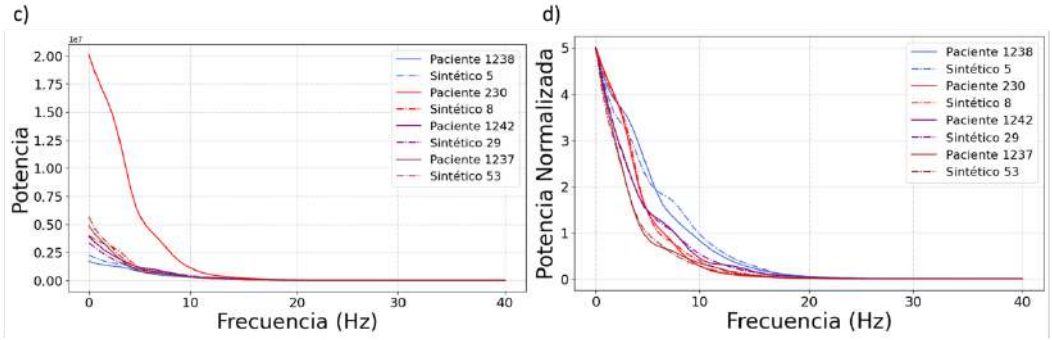
[122] Z. Yang and J. Hirschberg, “Predicting arousal and valence from waveforms and spectrograms using deep neural networks,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2018, pp. 3092–3096. doi: 10.21437/Interspeech.2018-2397.

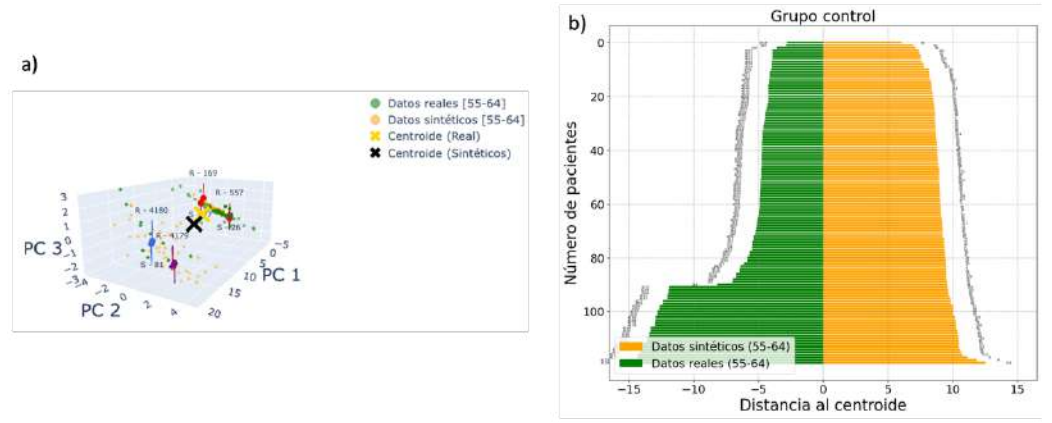
[123] A. Swift, R. Heale, and A. Twycross, “What are sensitivity and specificity?,” *Evid Based Nurs*, vol. 23, no. 1, pp. 2–4, Jan. 2020, doi: 10.1136/EBNURS-2019-103225.

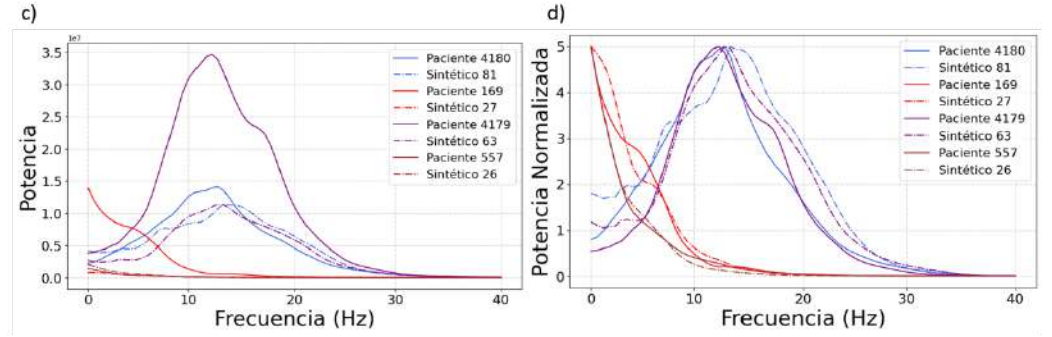
Material complementario (Control)

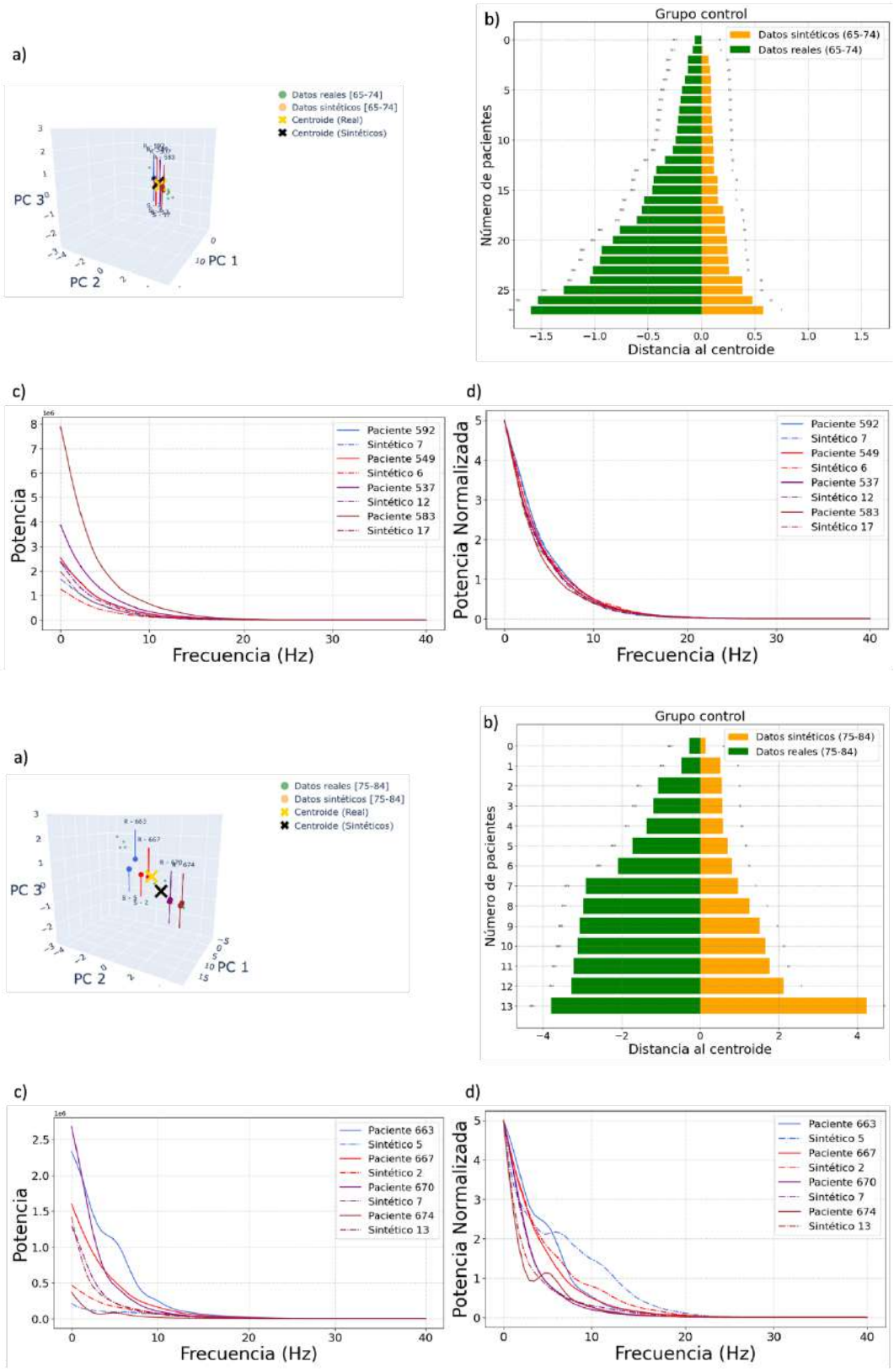



a)
Datos reales [45-54]
Datos sintéticos [45-54]
Centroide (Real)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo control
Datos sintéticos (45-54)
Datos reales (45-54)
Número de pacientes
Distancia al centroide

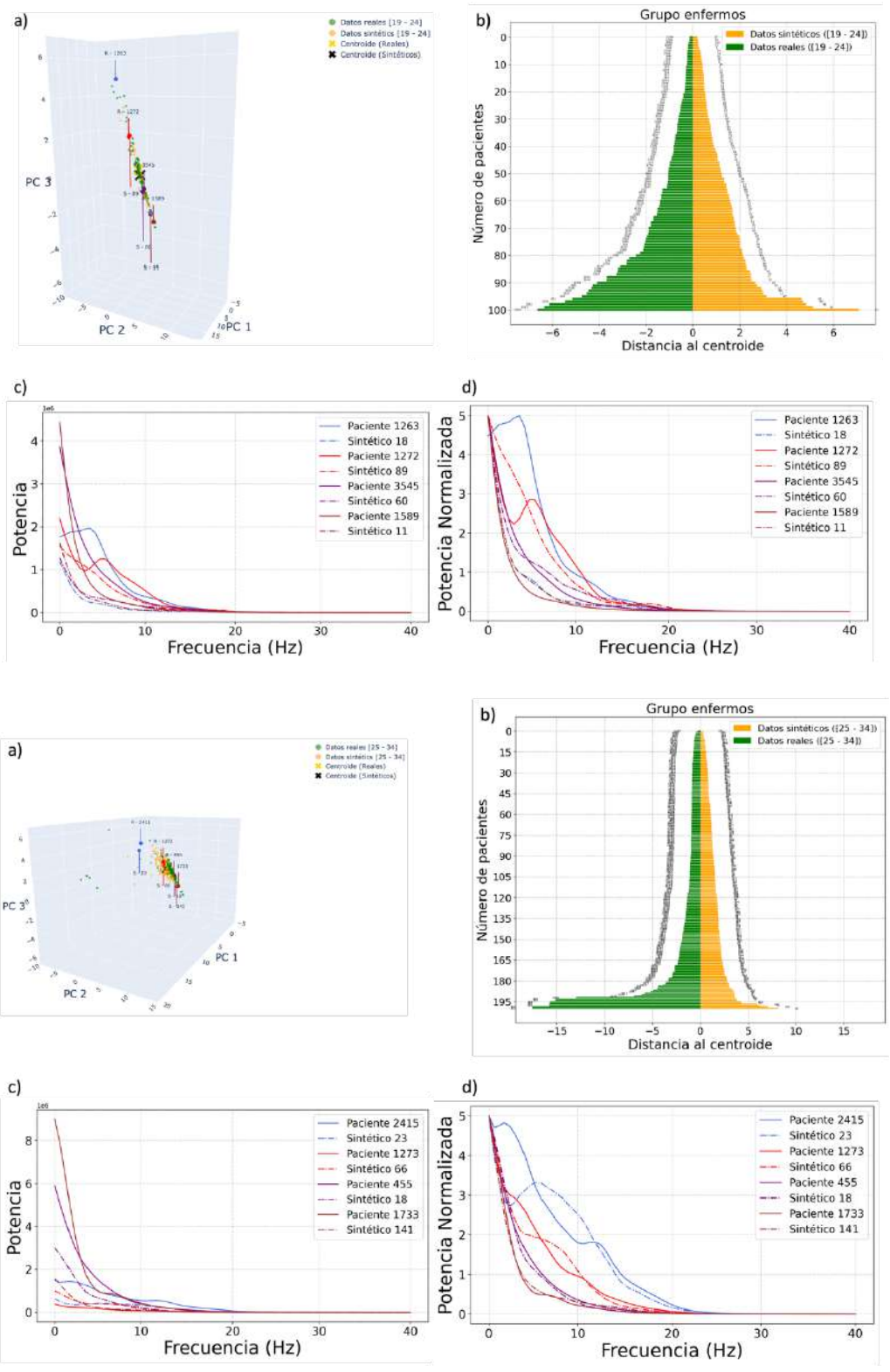

c)
Potencia
Frecuencia (Hz)
Paciente 1238
Sintético 5
Paciente 230
Sintético 8
Paciente 1242
Sintético 29
Paciente 1237
Sintético 53
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 1238
Sintético 5
Paciente 230
Sintético 8
Paciente 1242
Sintético 29
Paciente 1237
Sintético 53

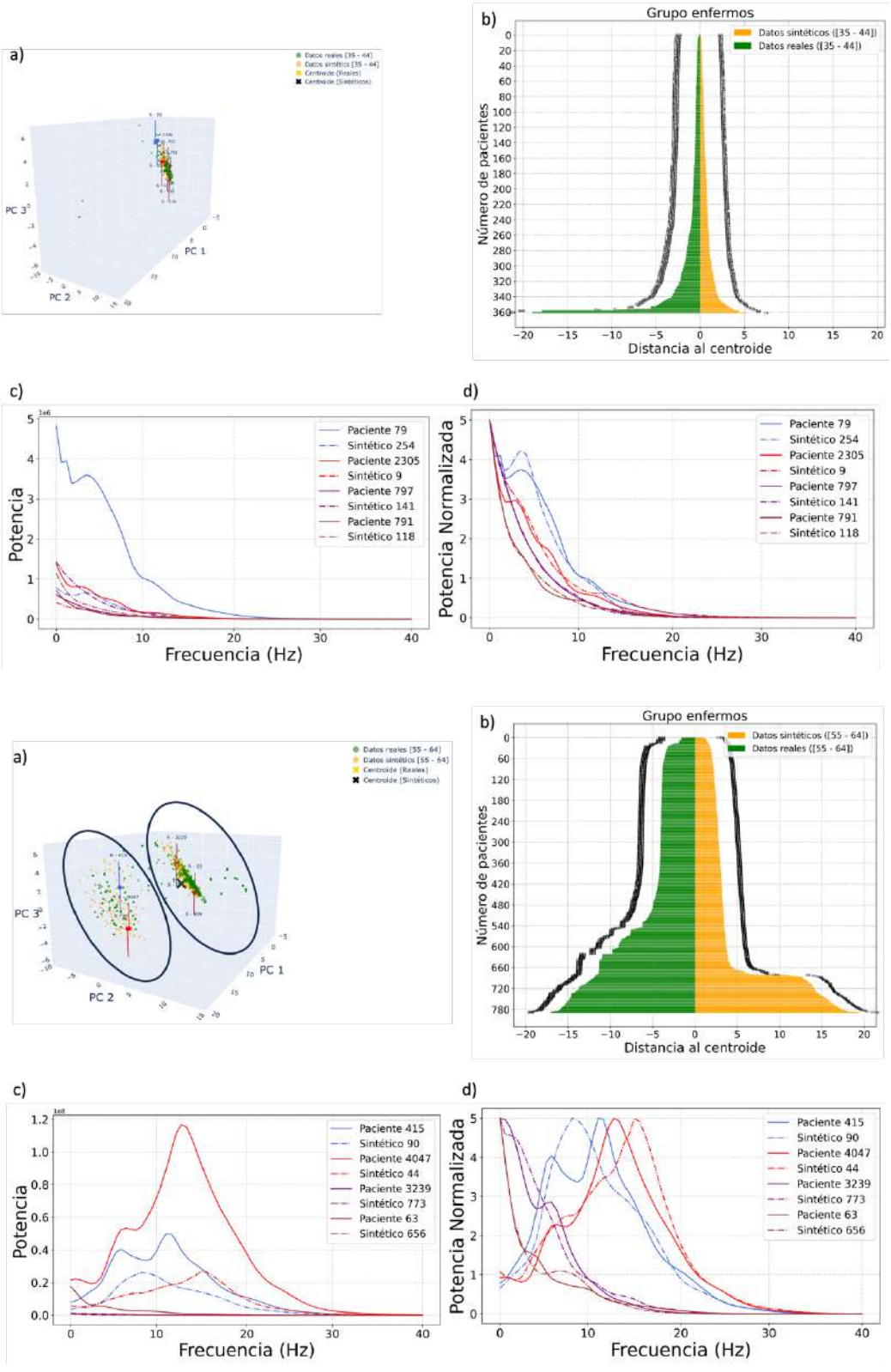

a)
Datos reales [55-64]
Datos sintéticos [55-64]
Centroide (Real)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo control
Datos sintéticos (55-64)
Datos reales (55-64)
Número de pacientes
Distancia al centroide

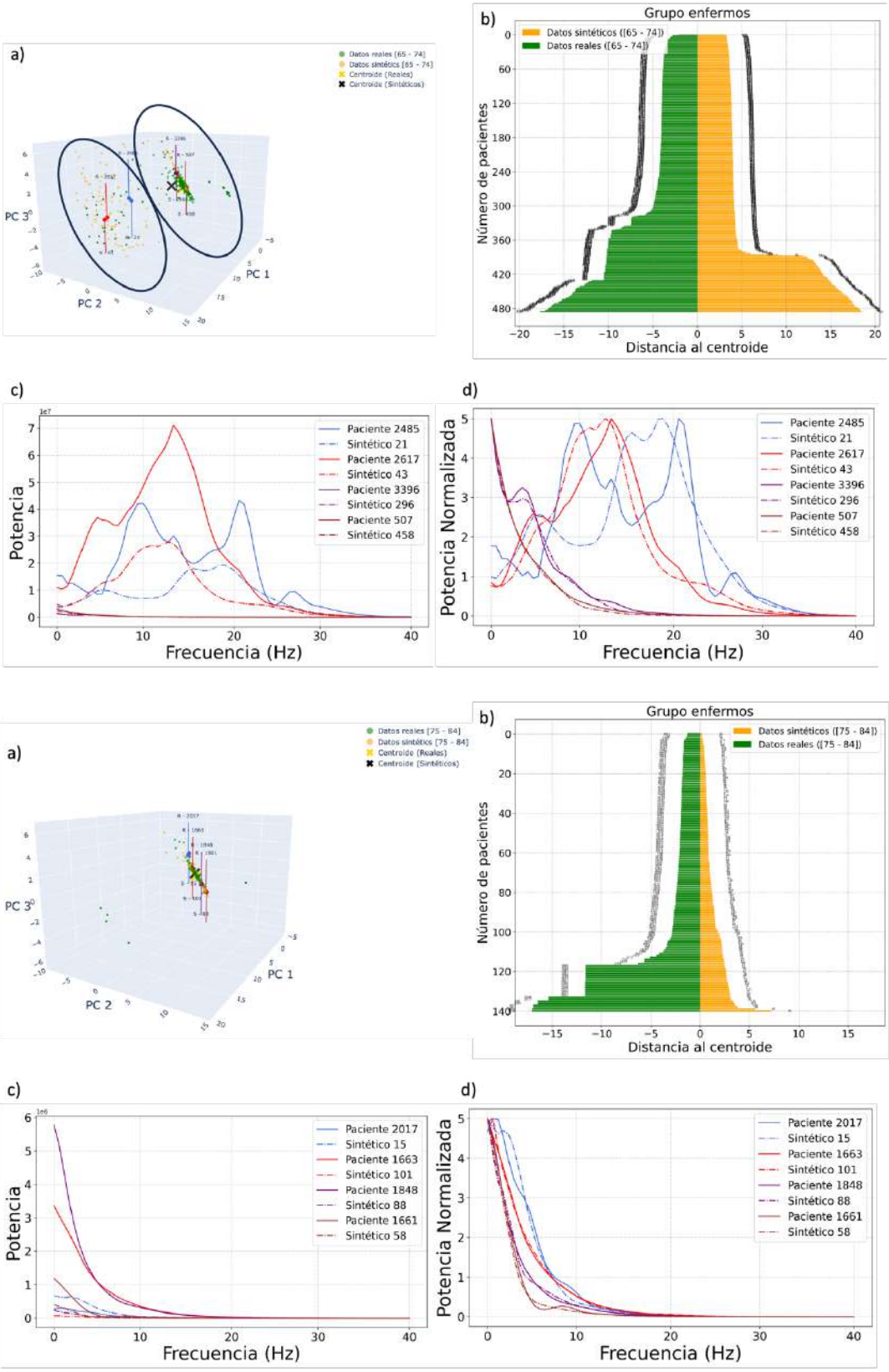

c)
Potencia
Frecuencia (Hz)
Paciente 4180
Sintético 81
Paciente 169
Sintético 27
Paciente 4179
Sintético 63
Paciente 557
Sintético 26
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 4180
Sintético 81
Paciente 169
Sintético 27
Paciente 4179
Sintético 63
Paciente 557
Sintético 26

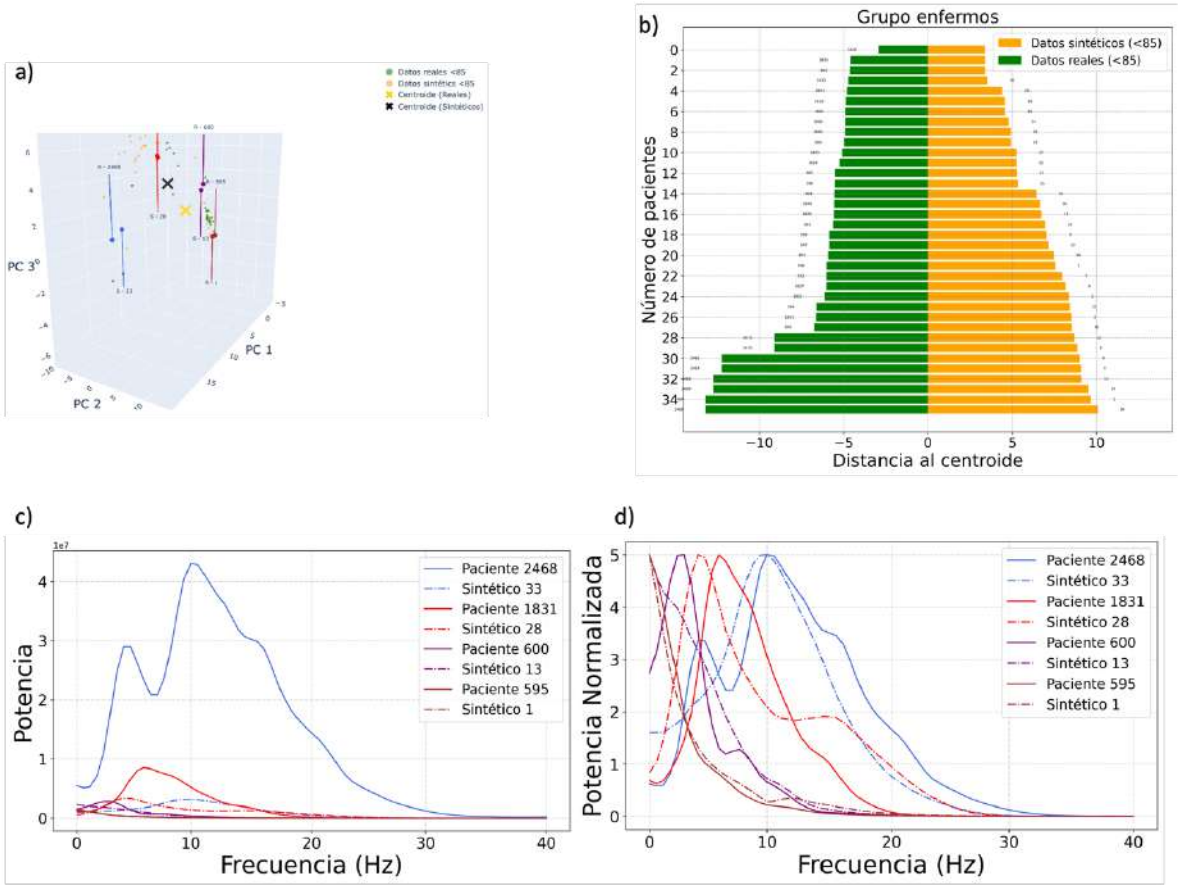

a)
Datos reales [65-74]
Datos sintéticos [65-74]
Centroide (Real)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo control
Datos sintéticos (65-74)
Datos reales (65-74)
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 592
Sintético 7
Paciente 549
Sintético 6
Paciente 537
Sintético 12
Paciente 583
Sintético 17
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 592
Sintético 7
Paciente 549
Sintético 6
Paciente 537
Sintético 12
Paciente 583
Sintético 17
a)
Datos reales [75-84]
Datos sintéticos [75-84]
Centroide (Real)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo control
Datos sintéticos (75-84)
Datos reales (75-84)
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 663
Sintético 5
Paciente 667
Sintético 2
Paciente 670
Sintético 7
Paciente 674
Sintético 13
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 663
Sintético 5
Paciente 667
Sintético 2
Paciente 670
Sintético 7
Paciente 674
Sintético 13

Material complementario (Enfermos)




a)
Datos reales ([35 - 44])
Datos sintéticos ([35 - 44])
Centroide (Reales)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo enfermos
Datos sintéticos ([35 - 44])
Datos reales ([35 - 44])
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 79
Sintético 254
Paciente 2305
Sintético 9
Paciente 797
Sintético 141
Paciente 791
Sintético 118
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 79
Sintético 254
Paciente 2305
Sintético 9
Paciente 797
Sintético 141
Paciente 791
Sintético 118
a)
Datos reales ([55 - 64])
Datos sintéticos ([55 - 64])
Centroide (Reales)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo enfermos
Datos sintéticos ([55 - 64])
Datos reales ([55 - 64])
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 415
Sintético 90
Paciente 4047
Sintético 44
Paciente 3239
Sintético 773
Paciente 63
Sintético 656
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 415
Sintético 90
Paciente 4047
Sintético 44
Paciente 3239
Sintético 773
Paciente 63
Sintético 656


a)
Datos reales ([65 - 74])
Datos sintéticos ([65 - 74])
Centroide (Reales)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo enfermos
Datos sintéticos ([65 - 74])
Datos reales ([65 - 74])
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 2485
Sintético 21
Paciente 2617
Sintético 43
Paciente 3396
Sintético 296
Paciente 507
Sintético 458
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 2485
Sintético 21
Paciente 2617
Sintético 43
Paciente 3396
Sintético 296
Paciente 507
Sintético 458
a)
Datos reales ([75 - 84])
Datos sintéticos ([75 - 84])
Centroide (Reales)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo enfermos
Datos sintéticos ([75 - 84])
Datos reales ([75 - 84])
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 2017
Sintético 15
Paciente 1663
Sintético 101
Paciente 1848
Sintético 88
Paciente 1661
Sintético 58
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 2017
Sintético 15
Paciente 1663
Sintético 101
Paciente 1848
Sintético 88
Paciente 1661
Sintético 58


a)
Datos reales <85
Datos sintéticos <85
Centroide (Reales)
Centroide (Sintéticos)
PC 1
PC 2
PC 3
b)
Grupo enfermos
Datos sintéticos (<85)
Datos reales (<85)
Número de pacientes
Distancia al centroide
c)
Potencia
Frecuencia (Hz)
Paciente 2468
Sintético 33
Paciente 1831
Sintético 28
Paciente 600
Sintético 13
Paciente 595
Sintético 1
d)
Potencia Normalizada
Frecuencia (Hz)
Paciente 2468
Sintético 33
Paciente 1831
Sintético 28
Paciente 600
Sintético 13
Paciente 595
Sintético 1