COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Sistema basado en Inteligencia Artificial para el análisis de texturas musculares en pacientes con enfermedades neurodegenerativas (ELA) a partir de biomarcadores ecográficos

Autor: José Juan Cortina Galindo

Director: Constantino Malagón Luque

Madrid, Junio 2026

Declaración de originalidad

Declaro bajo mi responsabilidad que el Proyecto presentado con el título **Sistema basado en Inteligencia Artificial para el análisis de texturas musculares en pacientes con enfermedades neurodegenerativas (ELA) a partir de biomarcadores ecográficos** en la Escuela Técnica Superior de Ingeniería ICAI de la Universidad Pontificia Comillas en el curso académico 2025/2026 es de mi autoría y no ha sido presentado anteriormente para otros fines. El Proyecto no ha sido plagiado de ningún otro, ni total ni parcialmente, y la información que ha sido tomada de otros documentos está debidamente referenciada.

Uso de Inteligencia Artificial[1]

Declaro bajo mi responsabilidad (indicar la opción correcta):

☐ No he utilizado Inteligencia Artificial en la elaboración de este documento.

☒ He utilizado Inteligencia Artificial en la elaboración de este documento y/o del Anexo B bajo las condiciones permitidas por la Universidad Pontificia Comillas, es decir, aplicando el Nivel 2 de la Escala de Evaluación de Perkins et al. (2024): *"La IA puede utilizarse para actividades previas a la tarea, como lluvia de ideas, descripción e investigación inicial. Este nivel se centra en el uso de la IA para planificar, sintetizar y generar ideas, pero las evaluaciones deben enfatizar la capacidad de desarrollar y perfeccionar estas ideas de forma independiente"*. En concreto, la Inteligencia Artificial se ha utilizado para:

[1] Esta declaración se refiere al uso de Inteligencia Artificial generativa para la elaboración de los documentos del Proyecto (Anexo B y Memoria). No se aplica a Proyectos en los que, por su naturaleza, deba utilizarse inteligencia artificial como parte de los mismos (aplicación de técnicas de aprendizaje automático, redes neuronales, análisis de datos...).

Se ha utilizado IA generativa (Nivel 2) como apoyo para comprender conceptos metodológicos y estadísticos, mejorar la redacción y la coherencia del texto, y verificar referencias. El diseño, la implementación, los experimentos y las decisiones del trabajo son obra propia del autor.

(firmar aquí)

Firma: José Juan Cortina Galindo

Fecha: 13/06/2026

Autorización para la entrega del Proyecto

Director del TFG	Codirector del TFG, en su caso
MALAGON LUQUE CONSTANTINO - 50855150B (firmar aquí) Firmado digitalmente por MALAGON LUQUE CONSTANTINO - 50855150B Fecha: 2026.06.15 13:25:57 +02'00'	(firmar aquí)
Firma: Constantino Malagón Luque	Firma:
Fecha: 15 / 06 /2026	Fecha: / /2026

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Sistema basado en Inteligencia Artificial para el análisis de texturas musculares en pacientes con enfermedades neurodegenerativas (ELA) a partir de biomarcadores ecográficos

Autor: José Juan Cortina Galindo

UNIVERSIDAD PONTIFICIA COMILLAS
Escuela Técnica Superior de Ingeniería (ICAI)
Grado en Ingeniería Matemática e Inteligencia Artificial

Director: Constantino Malagón Luque

Madrid, Junio 2026

UNIVERSIDAD PONTIFICIA COMILLAS
Escuela Técnica Superior de Ingeniería (ICAI)
Grado en Ingeniería Matemática e Inteligencia Artificial

SISTEMA BASADO EN INTELIGENCIA ARTIFICIAL PARA EL ANÁLISIS DE TEXTURAS MUSCULARES EN PACIENTES CON ENFERMEDADES NEURODEGENERATIVAS (ELA) A PARTIR DE BIOMARCADORES ECOGRÁFICOS

Autor: José Juan Cortina Galindo
Director: Constantino Malagón Luque

Entidad colaboradora: Fundación San Juan de Dios; Escuela de Enfermería y Fisioterapia San Juan de Dios (Universidad Pontificia Comillas); Hospital Universitario y Politécnico La Fe (Valencia)

Resumen

La Esclerosis Lateral Amiotrófica (ELA) carece de un biomarcador diagnóstico precoz accesible. Este trabajo desarrolla un sistema de aprendizaje profundo que detecta ELA a partir de ecografía muscular, integrando cuatro grupos musculares en una decisión a nivel paciente. Sobre 52 pacientes, el sistema alcanza un AUC del 98,67 %, sensibilidad del 100 % y especificidad del 92,31 %, a la altura de la línea base clínica basada en análisis textural clásico.

Palabras clave: Esclerosis Lateral Amiotrófica; ecografía muscular; aprendizaje profundo; redes neuronales convolucionales; fusión a nivel paciente; validación cruzada por sujeto.

Resumen ejecutivo

1 Introducción

La Esclerosis Lateral Amiotrófica (ELA) es una enfermedad neurodegenerativa de pronóstico fatal en la que el diagnóstico precoz sigue siendo un reto clínico: el retraso medio entre los primeros síntomas y el diagnóstico definitivo se sitúa en torno a los doce meses [1]. La ecografía muscular cuantitativa se ha propuesto como biomarcador no invasivo y de bajo coste [2], pero su interpretación requiere experiencia y los métodos clásicos, basados en características texturales calculadas manualmente, presentan limitaciones de generalización [3]. Este Trabajo Fin de Grado aborda la automatización del análisis de texturas musculares mediante aprendizaje profundo [4], integrando los cuatro grupos musculares clásicamente explorados (bíceps, antebrazo, cuádriceps y tibial) en una única decisión diagnóstica formulada a nivel de paciente.

2 Objetivos

El objetivo principal es diseñar y validar un sistema de soporte al diagnóstico de ELA basado en redes neuronales convolucionales sobre ecografía muscular, operando a nivel de paciente. Como objetivos específicos: entrenar y comparar cinco arquitecturas pre-entrenadas; garantizar una validación metodológicamente honesta mediante validación cruzada estratificada y agrupada por sujeto; cuantificar la incertidumbre mediante análisis estadístico inferencial; diseñar una estrategia de fusión multi-músculo a nivel paciente; interpretar las decisiones del modelo mediante mapas de explicabilidad; y comparar el rendimiento global frente a la línea base clínica.

3 Descripción del modelo/sistema/herramienta

El sistema parte de cinco arquitecturas convolucionales pre-entrenadas en ImageNet (ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B0 y ConvNeXt-Tiny), ajustadas por *transfer learning* sobre cada uno de los cuatro músculos. El entrenamiento emplea validación cruzada de cinco particiones con `StratifiedGroupKFold`, que preserva el balance de clases y garantiza que ningún paciente comparta imágenes entre entrenamiento y validación. A partir de las predicciones *out-of-fold*, el sistema agrega las imágenes a nivel de paciente y músculo y fusiona los cuatro músculos en una probabilidad única de ELA. El umbral de decisión se recalibra mediante el índice de Youden.

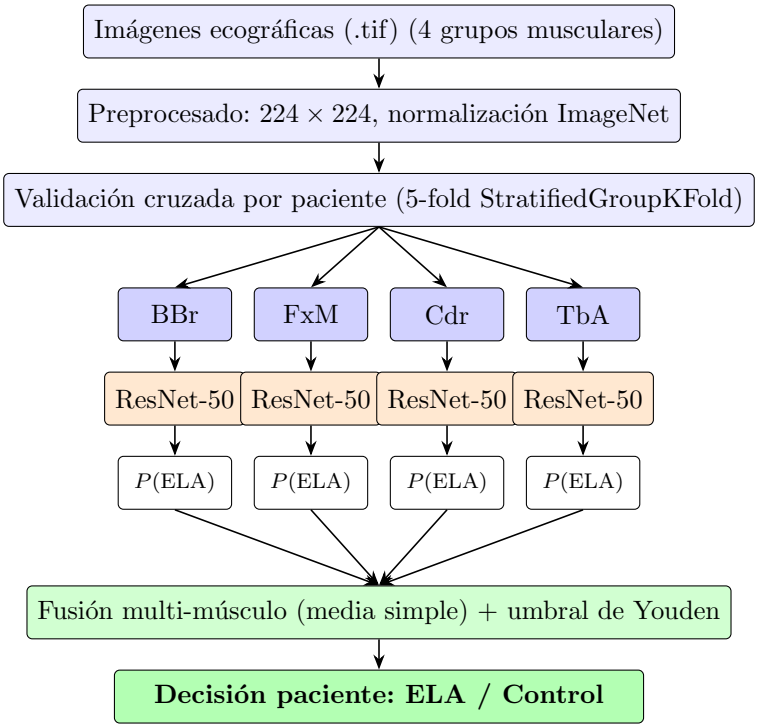


Figura 1: Esquema general del sistema: del preprocesado de las imágenes ecográficas a la decisión diagnóstica fusionada a nivel de paciente. BBr = bíceps braquial, FxM = flexores del antebrazo, Cdr = cuádriceps femoral, TbA = tibial anterior.

4 Resultados

La validación se realiza sobre 52 pacientes (26 con ELA, 26 controles) mediante predicciones *out-of-fold*, complementadas con intervalos de confianza al 95 % obtenidos por *bootstrap*. El sistema final, basado en ResNet-50 con fusión por media simple, alcanza un AUC del 98,67 % (IC95 % 95,4–100,0), una sensibilidad del 100 % y una especificidad del 92,31 %, sin ningún falso negativo y con dos falsos positivos. La aportación diferencial del trabajo es la integración multi-músculo a nivel paciente, ausente en la línea base clínica de Martínez-Payá.

Figura 2: Matriz de confusión del sistema final (ResNet-50 con fusión multi-músculo a nivel paciente) sobre los 52 pacientes evaluados *out-of-fold*: 24 verdaderos negativos, 26 verdaderos positivos, ningún falso negativo y 2 falsos positivos. Junto a la matriz se resumen las métricas globales (AUC 98,67 %, exactitud 96,15 %, sensibilidad 100 % y especificidad 92,31 %).

5 Conclusiones

La aportación principal del trabajo es la integración multi-músculo a nivel de paciente, validada con metodología estadística rigurosa e interpretada mediante mapas de activación. El sistema propuesto, reproducible y de código abierto, aporta evidencia de que los modelos profundos sobre ecografía muscular pueden alcanzar un rendimiento clínicamente operativo. Como líneas de trabajo futuro se plantean la validación externa con cohortes independientes, la calibración probabilística de las predicciones y la extensión a la monitorización longitudinal de la progresión de la enfermedad.

6 Referencias

[1] O. Hardiman et al., «Amyotrophic lateral sclerosis», *Nature Reviews Disease Primers*, vol. 3, n.º 17071, 2017.

[2] S. Pillen, I. M. P. Arts y M. J. Zwarts, «Muscle ultrasound in neuromuscular disorders», *Muscle & Nerve*, vol. 37, n.º 6, págs. 679-693, 2008.

[3] J. J. Martínez-Payá, J. Ríos-Díaz, M. E. del Baño-Aledo, J. I. Tembl-Ferrairó, J. F. Vázquez-Costa y F. Medina-Mirapeix, «Quantitative muscle ultrasonography using textural analysis in amyotrophic lateral sclerosis», *Ultrasonic Imaging*, vol. 39, n.º 6, págs. 357-368, 2017.

[4] G. Litjens et al., «A survey on deep learning in medical image analysis», *Medical Image Analysis*, vol. 42, págs. 60-88, 2017.

UNIVERSIDAD PONTIFICIA COMILLAS
Escuela Técnica Superior de Ingeniería (ICAI)
Grado en Ingeniería Matemática e Inteligencia Artificial

ARTIFICIAL INTELLIGENCE-BASED SYSTEM FOR MUSCLE TEXTURE ANALYSIS IN PATIENTS WITH NEURODEGENERATIVE DISEASES (ALS) FROM ULTRASOUND BIOMARKERS

Author: José Juan Cortina Galindo
Director: Constantino Malagón Luque

Collaborating entity: San Juan de Dios Foundation; San Juan de Dios School of Nursing and Physical Therapy (Comillas Pontifical University); Hospital Universitario y Politécnico La Fe (Valencia)

Abstract

Amyotrophic Lateral Sclerosis (ALS) lacks an accessible early diagnostic biomarker. This work develops a deep-learning system that detects ALS from muscle ultrasound, integrating four muscle groups into a patient-level decision. On 52 patients, the system reaches an AUC of 98.67 %, a sensitivity of 100 % and a specificity of 92.31 %, on par with the clinical baseline based on classical textural analysis.

Keywords: Amyotrophic Lateral Sclerosis; muscle ultrasound; deep learning; convolutional neural networks; patient-level fusion; subject-wise cross-validation.

Executive Summary

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease in which early diagnosis remains a clinical challenge: the average delay between the first symptoms and the definitive diagnosis is around twelve months [1]. Quantitative muscle ultrasonography has been proposed as a non-invasive, low-cost biomarker [2], but its interpretation requires expertise and classical methods, based on hand-crafted textural features, suffer from limited generalisation [3]. This Bachelor's Thesis addresses the automation of muscle texture analysis through deep learning [4], integrating the four classically explored muscle groups (biceps, forearm, quadriceps and tibial) into a single diagnostic decision formulated at the patient level.

2 Objectives

The main goal is to design and validate an ALS diagnostic support system based on convolutional neural networks applied to muscle ultrasound, operating at the patient level. The specific objectives are: to train and compare five pre-trained architectures; to ensure a methodologically honest validation through subject-wise stratified cross-validation; to quantify uncertainty by means of inferential statistical analysis; to design a multi-muscle fusion strategy at the patient level; to interpret the model's decisions through explainability maps; and to benchmark the overall performance against the clinical baseline.

3 Description of the Model/System/Tool

The system starts from five convolutional architectures pre-trained on ImageNet (ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B0 and ConvNeXt-Tiny), fine-tuned via transfer learning on each of the four muscles. Training uses five-fold cross-validation with `StratifiedGroupKFold`, which preserves class balance and guarantees that no patient shares images between training and validation. From the out-of-fold predictions, the system aggregates images at the patient-and-muscle level and fuses the four muscles into a single ALS probability. The decision threshold is recalibrated using Youden's J index.

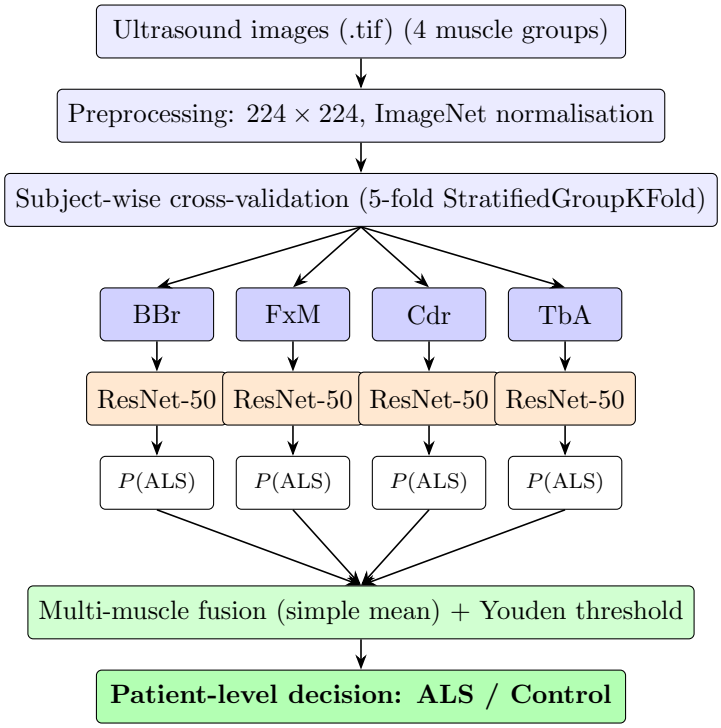


Figure 1: Overview of the system: from preprocessing of the ultrasound images to the fused patient-level diagnostic decision. BBr = biceps brachii, FxM = forearm flexors, Cdr = quadriceps femoris, TbA = tibialis anterior.

4 Results

Validation is performed on 52 patients (26 with ALS, 26 controls) using out-of-fold predictions, complemented with 95 % confidence intervals obtained by bootstrap. The final system, based on ResNet-50 with simple-mean fusion, reaches an AUC of 98.67 % (95 % CI 95.4–100.0), a sensitivity of 100 % and a specificity of 92.31 %, with no false negatives and two false positives. The differential contribution of this work is the multi-muscle integration at the patient level, absent in the clinical baseline of Martínez-Payá.

Figure 2: Confusion matrix of the final system (ResNet-50 with patient-level multi-muscle fusion) over the 52 patients evaluated out-of-fold: 24 true negatives, 26 true positives, no false negatives and 2 false positives. The global metrics are summarised next to the matrix (AUC 98.67 %, accuracy 96.15 %, sensitivity 100 % and specificity 92.31 %).

5 Conclusions

The main contribution of this work is the multi-muscle integration at the patient level, validated with a rigorous statistical methodology and interpreted through activation maps. The proposed system, reproducible and open-source, provides evidence that deep-learning models on muscle ultrasound can reach clinically operational performance. Future work includes external validation on independent cohorts, probabilistic calibration of the predictions and an extension to longitudinal monitoring of disease progression.

6 References

[1] O. Hardiman et al., «Amyotrophic lateral sclerosis», *Nature Reviews Disease Primers*, vol. 3, no. 17071, 2017.

[2] S. Pillen, I. M. P. Arts, and M. J. Zwarts, «Muscle ultrasound in neuromuscular disorders», *Muscle & Nerve*, vol. 37, no. 6, pp. 679–693, 2008.

[3] J. J. Martínez-Payá, J. Ríos-Díaz, M. E. del Baño-Aledo, J. I. Tembl-Ferrairó, J. F. Vázquez-Costa, and F. Medina-Mirapeix, «Quantitative muscle ultrasonography using textural analysis in amyotrophic lateral sclerosis», *Ultrasonic Imaging*, vol. 39, no. 6, pp. 357–368, 2017.

[4] G. Litjens et al., «A survey on deep learning in medical image analysis», *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

Índice

Índice de figuras

Índice de tablas

Capítulo 1 Introducción

La Esclerosis Lateral Amiotrófica (ELA) es una enfermedad neurodegenerativa caracterizada por la degeneración progresiva de las motoneuronas superior e inferior, que conduce a debilidad muscular, atrofia y, finalmente, fallo respiratorio [1], [2]. Se trata de una patología de pronóstico fatal, con una supervivencia mediana de entre tres y cinco años desde la aparición de los primeros síntomas. En Europa, su incidencia anual se estima en torno a 2 casos por cada 100.000 habitantes [3].

En ausencia de un biomarcador diagnóstico específico, el diagnóstico de la ELA se apoya en los criterios de El Escorial [4], que combinan la exploración clínica con la electromiografía. Este procedimiento conlleva un retraso medio cercano a los doce meses entre los primeros síntomas y el diagnóstico definitivo, una ventana temporal crítica en la que la intervención terapéutica podría tener un mayor impacto. En este contexto, la ecografía muscular cuantitativa ha emergido como una técnica de imagen no invasiva, accesible y de bajo coste, capaz de detectar los cambios estructurales que la enfermedad provoca en el tejido muscular [5], [6].

Este Trabajo Fin de Grado propone un sistema basado en aprendizaje profundo que automatiza el análisis de las texturas musculares a partir de imágenes ecográficas, integrando los cuatro grupos musculares clásicamente explorados en una única decisión diagnóstica formulada a nivel de paciente.

1.1. Motivación

El análisis de la ecografía muscular en la ELA se ha abordado tradicionalmente mediante características texturales calculadas de forma manual (parámetros de la matriz de co-ocurrencia de niveles de gris, ecogenicidad o ecovariación [6], [7]). Aunque estos métodos han demostrado capacidad discriminativa, presentan dos limitaciones relevantes. En primer lugar, dependen de descriptores diseñados a mano que pueden no capturar la totalidad de la información presente en la imagen. En segundo lugar, los trabajos previos reportan resultados a nivel de imagen y por músculo de forma aislada, sin proponer una estrategia que integre la exploración completa del paciente en una única decisión, que es la unidad de interés clínico.

El aprendizaje profundo ha transformado el análisis de imagen médica al permitir que los descriptores se aprendan directamente de los datos [8]. Aplicar este paradigma a la ecografía muscular en la ELA, y hacerlo de forma metodológicamente rigurosa y a nivel de paciente, constituye la motivación central de este trabajo.

1.2. Objetivos

El objetivo principal de este Trabajo Fin de Grado es diseñar, implementar y validar un sistema de soporte al diagnóstico de la ELA, basado en redes neuronales convolucionales aplicadas a la ecografía muscular, que opere a nivel de paciente. Para alcanzarlo se definen los siguientes objetivos específicos:

- Entrenar y comparar cinco arquitecturas convolucionales pre-entrenadas sobre cada uno de los cuatro músculos analizados.
- Garantizar la integridad metodológica de la validación mediante validación cruzada estratificada y agrupada por sujeto, evitando la fuga de información entre los conjuntos de entrenamiento y validación.
- Cuantificar la incertidumbre de las métricas obtenidas mediante análisis estadístico inferencial: intervalos de confianza, recalibración de umbral y contrastes de hipótesis pareados.
- Diseñar y validar una estrategia de fusión que integre los cuatro músculos en una única probabilidad de ELA a nivel de paciente.
- Generar mapas de explicabilidad que permitan verificar que las decisiones del modelo se sustentan en regiones clínicamente relevantes.
- Comparar el rendimiento global del sistema con la línea base clínica vigente.

1.3. Alineación con los ODS

Este trabajo se alinea con dos de los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 de Naciones Unidas. El proyecto contribuye principalmente al **ODS 3 (Salud y Bienestar)**: el sistema desarrollado constituye una herramienta de apoyo diagnóstico no invasiva y de bajo coste, en línea con la meta 3.4, orientada a reducir la mortalidad prematura por enfermedades no transmisibles, y con la meta 3.b, de apoyo a la investigación y el desarrollo de tecnologías sanitarias. De forma secundaria, el trabajo contribuye al **ODS 9 (Industria, Innovación e Infraestructura)**, al materializar la transferencia de técnicas avanzadas de inteligencia artificial a un dominio clínico tradicionalmente dependiente de la inspección cualitativa.

1.4. Estructura del trabajo

El resto de la memoria se organiza como sigue. El capítulo 2 revisa el estado del arte, tanto clínico como técnico, y sitúa el trabajo respecto a la literatura previa. El capítulo 3 describe

en detalle el sistema desarrollado: el planteamiento del problema, el diseño de la solución y su implementación. El capítulo 4 presenta los resultados experimentales y su análisis crítico. Finalmente, el capítulo 5 recoge las conclusiones del trabajo y las líneas de investigación futuras.

Capítulo 2 Estado del arte

Este capítulo revisa el conocimiento previo sobre el que se construye el proyecto. Se parte de los conceptos clínicos de la enfermedad y de la técnica de imagen empleada (secciones 2.1 y 2.2), se examina la línea base clínica que ha aplicado análisis cuantitativo de imagen al diagnóstico de la Esclerosis Lateral Amiotrófica (sección 2.3), se describen las técnicas de aprendizaje profundo relevantes para el problema (sección 2.4) y se discuten los retos metodológicos asociados a la validación con conjuntos de datos clínicos reducidos (sección 2.5). El capítulo concluye identificando el vacío en la literatura que justifica el presente trabajo (sección 2.6).

2.1. La Esclerosis Lateral Amiotrófica

La Esclerosis Lateral Amiotrófica (ELA) es una enfermedad neurodegenerativa caracterizada por la pérdida progresiva de las motoneuronas superiores e inferiores, lo que provoca debilidad muscular, atrofia y, en última instancia, parálisis y fallo respiratorio [1], [2]. Su incidencia anual en las poblaciones europeas se estima en torno a 2 casos por cada 100.000 habitantes, y la supervivencia mediana desde el inicio de los síntomas es de tres a cinco años [2], [3].

El diagnóstico de la ELA es eminentemente clínico y se apoya en los criterios de El Escorial y sus revisiones posteriores, que combinan la exploración neurológica con estudios electrofisiológicos para confirmar la afectación de la motoneurona en distintas regiones corporales [4]. Estos criterios fueron concebidos con fines de investigación y exigen una afectación ya extendida, por lo que su sensibilidad en las fases iniciales es limitada. En la práctica clínica esto se traduce en un retraso diagnóstico medio de entre diez y dieciséis meses desde la aparición de los primeros síntomas [2]. Dado que el deterioro neuronal es irreversible y que las terapias disponibles son más eficaces cuanto antes se instauran, reducir ese retraso constituye una necesidad clínica de primer orden y motiva la búsqueda de marcadores objetivos, accesibles y no invasivos que apoyen la sospecha diagnóstica.

2.2. Ecografía muscular cuantitativa

La ecografía muscular ha emergido como una herramienta complementaria en la evaluación de las enfermedades neuromusculares. Frente a la electromiografía, que es dolorosa e invasiva, y frente a la resonancia magnética, costosa y de disponibilidad limitada, la ecografía es indolora, de bajo coste, portátil y permite explorar varios músculos en una misma sesión [5]. Cuando una enfermedad neuromuscular sustituye tejido muscular sano por tejido fibroso y graso, la cantidad de ultrasonidos reflejados aumenta y el músculo aparece más brillante

en la imagen. Esta propiedad se cuantifica mediante la *ecointensidad*, definida como el nivel medio de gris del músculo en una región de interés [5], [9].

El uso clínico de la ecointensidad presenta, sin embargo, limitaciones conocidas. Su valor depende del equipo, de los ajustes del ecógrafo y del operador, lo que obliga a disponer de valores de referencia normalizados por edad y por músculo [9]. Además, la ecointensidad media resume en un único número toda la información de la imagen y descarta la organización espacial de los píxeles, es decir, la *textura* del músculo. Esta observación ha impulsado el desarrollo de métodos de análisis textural cuantitativo, que extraen descriptores estadísticos de los patrones de gris de la imagen con el objetivo de capturar alteraciones que la ecointensidad media no refleja.

2.3. Línea base clínica: análisis textural en la ELA

El grupo de Martínez-Payá y colaboradores constituye la referencia más directa para este trabajo. En un primer estudio, los autores demostraron que descriptores de textura extraídos de ecografías de músculos de pacientes con ELA permitían diferenciarlos de sujetos sanos, estableciendo que la imagen muscular contiene información discriminante más allá de la ecointensidad media [6]. Trabajos posteriores del mismo grupo confirmaron que estos descriptores también reflejan la progresión de la enfermedad a lo largo del tiempo [7].

Estos trabajos comparten un mismo esquema metodológico: las características de textura se calculan mediante fórmulas estadísticas predefinidas (matrices de coocurrencia, dimensión fractal, descriptores de la transformada wavelet, entre otros) y se seleccionan e introducen después en clasificadores estadísticos clásicos. Este enfoque, que puede denominarse de *ingeniería manual de características*, tiene la ventaja de la interpretabilidad, pero presenta una limitación de fondo: el conjunto de descriptores se fija de antemano y, por tanto, el sistema sólo puede detectar aquellos patrones que el experto ha decidido cuantificar. Cualquier patrón discriminante no contemplado en la batería de fórmulas queda fuera del alcance del modelo. Esta limitación es precisamente la que motiva explorar técnicas capaces de aprender las características relevantes directamente de los datos.

2.4. Aprendizaje profundo en imagen médica

Las redes neuronales convolucionales (CNN) invierten el paradigma del análisis textural manual: en lugar de calcular descriptores fijados de antemano, aprenden de los propios datos qué patrones de la imagen son relevantes para la tarea [10]. Desde su irrupción, las CNN se han convertido en el estándar para tareas de clasificación de imagen y han sido adoptadas de forma generalizada en el análisis de imagen médica [8], [11].

El principal obstáculo para aplicar aprendizaje profundo en el ámbito clínico es la escasez de datos: entrenar una CNN desde cero requiere decenas de miles de imágenes etiquetadas,

una cifra inalcanzable en la mayoría de los estudios médicos. La solución habitual es el *aprendizaje por transferencia* (*transfer learning*), que consiste en partir de una red preentrenada sobre un gran corpus de imágenes naturales (típicamente ImageNet [12]) y reajustarla sobre el conjunto de datos específico. Las primeras capas de una CNN aprenden detectores de bordes, texturas y formas que son en buena medida genéricos y reutilizables entre dominios [13], lo que hace de esta estrategia la opción de referencia cuando el conjunto de datos es pequeño, como han confirmado numerosos estudios en imagen médica [8], [14].

A lo largo de la última década se han propuesto sucesivas arquitecturas de CNN que han mejorado progresivamente el rendimiento y la eficiencia. ResNet introdujo las conexiones residuales, que permiten entrenar redes muy profundas evitando la degradación del gradiente [15]. DenseNet llevó esta idea más lejos conectando cada capa con todas las posteriores, lo que favorece la reutilización de características [16]. EfficientNet propuso un escalado equilibrado de la profundidad, la anchura y la resolución de la red para maximizar la precisión por parámetro [17]. Por último, ConvNeXt reexaminó el diseño convolucional incorporando elementos de los *transformers* de visión, alcanzando un rendimiento competitivo con un esquema puramente convolucional [18]. La existencia de familias de arquitecturas con compromisos distintos entre capacidad, eficiencia y riesgo de sobreajuste justifica que un estudio sobre un dominio nuevo compare varias de ellas en lugar de comprometerse con una sola a priori. La tabla 1 resume las características principales de las cinco arquitecturas comparadas en este trabajo, y los párrafos siguientes desarrollan brevemente las ideas centrales de cada familia.

Tabla 1: Resumen comparativo de las cinco arquitecturas convolucionales empleadas en este trabajo. Los parámetros se reportan en millones para la variante específica utilizada.

Arquitectura	**Año**	**Parámetros**	**Idea principal**
ResNet-18	2016	∼11 M	Conexiones residuales (*skip connections*)
ResNet-50	2016	∼26 M	ResNet con bloques *bottleneck* más profundos
DenseNet-121	2017	∼8 M	Concatenación densa entre capas
EfficientNet-B0	2019	∼5 M	Escalado balanceado profundidad/anchura/resolución
ConvNeXt-Tiny	2022	∼28 M	CNN moderna inspirada en *vision transformers*

ResNet [15] (*residual networks*) resolvió uno de los problemas centrales del entrenamiento de redes profundas: a partir de cierta profundidad, añadir capas dejaba de mejorar el rendimiento y, paradójicamente, lo empeoraba. La causa no era el sobreajuste, sino una dificultad de optimización: los gradientes se atenuaban antes de llegar a las capas iniciales. La solución propuesta consiste en añadir conexiones de salto (*skip connections*) que cortocircuitan dos o tres capas, de modo que cada bloque aprende un residuo $\Delta x = F(x)$ en lugar de la representación completa $H(x) = x + F(x)$. Esta reformulación facilita el aprendizaje

de la identidad como punto de partida y permite entrenar redes de cientos de capas. Las variantes ResNet-18 y ResNet-50 se diferencian en el número de bloques y, sobre todo, en su tipo: la primera utiliza bloques sencillos con dos convoluciones 3×3, mientras que la segunda usa bloques *bottleneck* con tres convoluciones (1×1, 3×3, 1×1) que reducen y restauran la dimensionalidad de los canales, ahorrando cómputo. ResNet-50 es históricamente la línea base de referencia para experimentos de *transfer learning* sobre ImageNet, y por eso se ha incluido en este trabajo.

DenseNet [16] llevó la idea de las conexiones de salto a un extremo: en lugar de saltar dos o tres capas, cada capa recibe como entrada la concatenación de las salidas de *todas* las capas anteriores dentro del mismo *dense block*. Esta conectividad densa cumple tres funciones simultáneas: refuerza la propagación de las características, mitiga el problema de los gradientes y reduce drásticamente el número de parámetros necesarios, porque las capas profundas no necesitan re-aprender las características ya extraídas por capas previas. DenseNet-121 logra rendimientos comparables a ResNet-50 con unos ocho millones de parámetros, una característica deseable en escenarios de pocos datos donde el riesgo de sobreajuste es elevado.

EfficientNet [17] introduce una idea distinta: en lugar de proponer una arquitectura concreta, los autores formulan un *compound scaling* que escala simultáneamente la profundidad, la anchura y la resolución de entrada de la red con un único coeficiente. La intuición es que escalar una sola de las tres dimensiones, como había sido habitual, no aprovecha de forma óptima el presupuesto computacional. EfficientNet-B0 es la red base del escalado, optimizada mediante *neural architecture search*. Con cinco millones de parámetros, ofrece una eficiencia notable: alcanza precisiones equivalentes a redes mucho más grandes con una fracción del cómputo, lo que la convierte en una alternativa interesante cuando el conjunto de datos es modesto.

ConvNeXt [18] responde a una pregunta especialmente relevante tras la irrupción de los *vision transformers* (ViT): ¿es la naturaleza convolucional la que limita el rendimiento, o lo son únicamente sus elecciones de diseño históricas? Los autores construyen progresivamente una CNN moderna incorporando, una a una, las decisiones de diseño que distinguen a los ViT: *patchify* en lugar de convolución agresiva inicial, normalización por capa (LayerNorm) en lugar de *batch normalization*, activación GELU en lugar de ReLU, mayor proporción entre la dimensión interna y la externa del bloque, etc. ConvNeXt-Tiny es el resultado de esa modernización a escala pequeña; con 28 millones de parámetros, iguala o supera a ViT-Base manteniendo el sesgo inductivo convolucional, especialmente útil cuando el conjunto de datos es limitado y no permite que un *transformer* aprenda desde cero las invarianzas espaciales que una CNN incorpora por construcción.

Selección de las arquitecturas para este trabajo. La elección de las cinco arquitecturas anteriores busca cubrir el espacio de soluciones modernas con compromisos distintos. ResNet-18 y ResNet-50 representan la referencia residual a dos escalas. DenseNet-121 aporta una opción densa de pocos parámetros. EfficientNet-B0 explora la eficiencia escalada.

ConvNeXt-Tiny incorpora las lecciones del paradigma *transformer* dentro de un marco convolucional. Se descartaron explícitamente dos candidatas habituales en comparativas anteriores: VGG-16, por su excesivo número de parámetros (∼138 millones) que la hace particularmente propensa al sobreajuste sobre 104 imágenes por músculo; y MobileNet-V3 Small, por solapar funcionalmente con EfficientNet-B0 en el rol de arquitectura moderna y compacta sin aportar una perspectiva diferenciada para los objetivos del trabajo.

Una objeción recurrente al uso de aprendizaje profundo en medicina es su naturaleza de «caja negra». Para mitigarla se han desarrollado técnicas de explicabilidad que permiten visualizar qué regiones de la imagen sustentan la predicción del modelo, como la retropropagación guiada [19], los mapas de saliencia [20], la visualización de activaciones [21] y, en particular, Grad-CAM [22], que genera mapas de calor superpuestos a la imagen original. Estas herramientas resultan imprescindibles para que un modelo sea verificable y aceptable en un contexto clínico.

2.5. Retos de validación en conjuntos de datos clínicos reducidos

Aplicar aprendizaje profundo a un conjunto de datos clínico pequeño no sólo plantea un reto de entrenamiento, sino sobre todo un reto de validación. Con pocas muestras, una única partición en entrenamiento y test produce estimaciones de rendimiento muy inestables, por lo que se recurre a la validación cruzada, que promedia el rendimiento sobre varias particiones y aprovecha todos los datos disponibles [23].

El riesgo más grave y a la vez más sutil en estos escenarios es la *fuga de información* (*data leakage*). En estudios de imagen médica es habitual disponer de varias imágenes por paciente; si imágenes de un mismo paciente quedan repartidas entre los conjuntos de entrenamiento y de evaluación, el modelo puede aprender a reconocer al individuo en lugar de la enfermedad, y el rendimiento medido resulta optimista y no generalizable [24]. La literatura ha documentado de forma reiterada este y otros errores metodológicos (ausencia de agrupación por paciente, selección de características antes de particionar, recalibración de umbrales sobre el conjunto de test) como causas frecuentes de resultados irreproducibles en aprendizaje automático aplicado a la medicina [25]. La prevención de la fuga exige que la partición se realice a nivel de paciente, garantizando que todas las imágenes de un individuo permanezcan en el mismo subconjunto.

La comparación rigurosa de modelos sobre conjuntos pequeños requiere asimismo un instrumental estadístico adecuado. El área bajo la curva ROC (AUC) es la métrica de referencia para cuantificar la capacidad discriminante de un clasificador [26]; la prueba de DeLong permite comparar dos AUC calculadas sobre la misma muestra teniendo en cuenta su correlación [27]; la prueba de los rangos con signo de Wilcoxon ofrece una alternativa no paramétrica para comparar rendimientos emparejados [28]; el índice J de Youden proporciona un criterio objetivo para fijar el umbral de decisión [29]; y los métodos de remuestreo *bootstrap* permiten

estimar intervalos de confianza sin asumir una distribución concreta [30], [31]. El empleo conjunto de estas herramientas es lo que distingue una comparación de modelos metodológicamente sólida de una mera ordenación por rendimiento.

2.6. Limitaciones del estado del arte y justificación del proyecto

De la revisión anterior se desprenden dos observaciones que, en conjunto, definen el vacío que aborda este trabajo. Por un lado, el análisis cuantitativo de la ecografía muscular en la ELA se ha apoyado hasta la fecha en descriptores de textura diseñados manualmente [6], [7], un enfoque limitado por definición a los patrones que el experto decide cuantificar. Por otro lado, el aprendizaje profundo (capaz de aprender esas características directamente de los datos y consolidado ya en numerosas modalidades de imagen médica [8], [11]) no se ha aplicado, hasta donde alcanza esta revisión, a la detección de la ELA a partir de ecografía muscular.

Este trabajo se sitúa precisamente en esa intersección. Su aportación es doble. En primer lugar, aplica y compara de forma sistemática varias arquitecturas modernas de CNN sobre ecografías musculares de pacientes con ELA, sustituyendo la ingeniería manual de características por características aprendidas. En segundo lugar, lo hace bajo un protocolo de validación diseñado expresamente para el escenario de datos clínicos reducidos: partición agrupada por paciente para evitar la fuga de información, validación cruzada estratificada y comparación estadística rigurosa frente a la línea base clínica de Martínez-Payá. El capítulo siguiente describe en detalle la metodología con la que se materializa esta propuesta.

Capítulo 3 Metodología

Este capítulo describe en detalle el sistema desarrollado. La sección 3.1 formaliza el problema de clasificación, fija los requisitos y define las métricas de evaluación. La sección 3.2 presenta el diseño de la solución: el enfoque metodológico, el conjunto de datos, las arquitecturas comparadas, el protocolo de validación cruzada, el esquema de fusión a nivel de paciente y el análisis estadístico. La sección 3.3 detalla la implementación: las tecnologías empleadas, la estructura del código, el flujo de procesamiento y las medidas adoptadas para garantizar la reproducibilidad.

3.1. Planteamiento del problema

3.1.1. Definición formal

El problema se aborda en dos niveles encadenados. En el **nivel de imagen**, se trata de una tarea de clasificación binaria supervisada: dada una imagen de ecografía muscular x, el sistema debe estimar la probabilidad $P(\text{ELA} \mid x)$ de que el músculo represente a un paciente con Esclerosis Lateral Amiotrófica, y asignar una etiqueta $y \in \{\text{Control}, \text{ELA}\}$. Formalmente, se busca una función $f_\theta : \mathcal{X} \to [0, 1]$, parametrizada por los pesos θ de una red neuronal convolucional, que aproxime la probabilidad a posteriori de la clase ELA.

En el **nivel de paciente**, que es el objetivo clínico último, el sistema dispone de varias imágenes del mismo individuo correspondientes a distintos músculos y debe emitir un único veredicto. Si $\mathcal{I}_p$ denota el conjunto de imágenes del paciente p, se busca una función de agregación g tal que $\hat{y}_p = g\big(\{f_\theta(x) : x \in \mathcal{I}_p\}\big)$ produzca una decisión robusta a partir de la evidencia parcial de cada músculo. La distinción entre ambos niveles es esencial: el modelo se entrena y se valida sobre imágenes, pero la utilidad clínica se mide sobre pacientes.

3.1.2. Requisitos

El sistema debe satisfacer un conjunto de requisitos que condicionan todas las decisiones de diseño posteriores. Como *requisitos funcionales*, el sistema debe aceptar como entrada imágenes de ecografía muscular en escala de gris, producir una estimación de probabilidad de ELA tanto por imagen como por paciente, y ofrecer un mecanismo de interpretación visual que permita verificar en qué regiones de la imagen se apoya la predicción. Como *requisitos no funcionales*, la evaluación debe estar libre de fuga de información a nivel de paciente; el protocolo experimental debe ser íntegramente reproducible; los resultados deben ser comparables con la línea base clínica de Martínez-Payá [6]; y el sistema debe operar con un conjunto de datos reducido, lo que descarta el entrenamiento de redes desde cero y obliga a recurrir al aprendizaje por transferencia.

3.1.3. Métricas de evaluación

Dado que el conjunto de datos está balanceado entre clases, la métrica principal es el **área bajo la curva ROC** (AUC), que cuantifica la capacidad discriminante del modelo con independencia del umbral de decisión [26]. El AUC admite una interpretación probabilística directa: es la probabilidad de que, ante un paciente con ELA y un control elegidos al azar, el modelo asigne mayor puntuación al primero. Como métricas complementarias, orientadas a la interpretación clínica, se reportan la **sensibilidad** (proporción de pacientes con ELA correctamente identificados), la **especificidad** (proporción de controles correctamente identificados) y la **exactitud** (*accuracy*). A partir de la matriz de confusión, con verdaderos positivos TP, verdaderos negativos TN, falsos positivos FP y falsos negativos FN, estas métricas se definen como

$$\text{Sens} = \frac{TP}{TP+FN}, \qquad \text{Spec} = \frac{TN}{TN+FP}, \qquad \text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}. \tag{3.1}$$

La sensibilidad y la especificidad dependen del umbral que convierte la probabilidad continua en una decisión binaria. Como ese umbral no tiene por qué ser 0,5, se recalibra mediante el **índice J de Youden** [29], definido como $J = \text{Sens} + \text{Spec} - 1$, eligiendo el umbral que maximiza J. El AUC, en cambio, es independiente del umbral, lo que lo hace idóneo como criterio principal para comparar arquitecturas.

3.2. Diseño de la solución

3.2.1. Enfoque propuesto

El enfoque propuesto sustituye la ingeniería manual de características de la línea base clínica por características aprendidas directamente de los datos mediante redes neuronales convolucionales. Dado el tamaño reducido del conjunto de datos, no se entrena ninguna red desde cero: todas las arquitecturas parten de pesos preentrenados sobre ImageNet [12] y se reajustan al problema concreto (aprendizaje por transferencia). Como en este dominio no existe una arquitectura de referencia, no se compromete el trabajo con una sola red, sino que se comparan cinco arquitecturas representativas bajo idéntico protocolo. La evaluación se realiza mediante validación cruzada agrupada por paciente, lo que evita la fuga de información, y los modelos por músculo se combinan después en un único veredicto por paciente mediante reglas de fusión. Finalmente, todas las comparaciones (entre arquitecturas y frente a la línea base) se someten a contraste estadístico, y las predicciones se acompañan de mapas de interpretación visual.

El sistema se organiza en cinco etapas encadenadas: (1) preparación y preprocesado de los datos, (2) entrenamiento de cada arquitectura sobre cada músculo con validación cruzada,

(3) análisis estadístico de los resultados por músculo, (4) fusión de los cuatro músculos en una decisión a nivel de paciente y (5) generación de explicaciones visuales sobre el modelo final.

3.2.2. Conjunto de datos

El conjunto de datos procede de un estudio de ecografía musculoesquelética y está formado por imágenes de cuatro músculos: el bíceps braquial (BBr), los flexores del antebrazo (FxM), el cuádriceps femoral (Cdr) y el tibial anterior (TbA). Para cada músculo se dispone de regiones de interés (ROI) ya segmentadas, exportadas en formato TIFF, correspondientes a 52 imágenes de sujetos control y 52 de pacientes con ELA, lo que suma 104 imágenes por músculo y 416 imágenes en total. El conjunto está, por tanto, perfectamente balanceado entre clases. A nivel de paciente, el estudio agrupa 52 individuos (26 con ELA y 26 controles), cada uno con imágenes de los cuatro músculos y de ambos lados del cuerpo (dos imágenes por músculo, lado derecho e izquierdo).

Estas imágenes proceden de los estudios de análisis textural muscular en la ELA desarrollados por el grupo de Martínez-Payá y colaboradores [6], [7], fruto de una colaboración entre la Universidad Pontificia Comillas (Escuela de Enfermería y Fisioterapia San Juan de Dios), la Fundación San Juan de Dios y el Hospital Universitario y Politécnico La Fe (Valencia). Las regiones de interés fueron segmentadas manualmente por personal clínico especializado, y el estudio de origen contó con la aprobación del comité de ética correspondiente y con el consentimiento informado de los participantes.

Un aspecto crítico del conjunto de datos es la identificación del paciente. Los nombres de fichero codifican el identificador del sujeto, que se extrae mediante una expresión regular para poder agrupar todas las imágenes de un mismo individuo. Esta identificación es la que permite realizar particiones a nivel de paciente y, con ello, prevenir la fuga de información descrita en el capítulo anterior.

3.2.3. Arquitecturas comparadas

Se comparan cinco arquitecturas de red neuronal convolucional, todas ellas preentrenadas sobre ImageNet y adaptadas al problema sustituyendo su capa de clasificación final por una capa lineal de dos salidas. La selección busca cubrir distintos compromisos entre capacidad, eficiencia y riesgo de sobreajuste. **ResNet-18** y **ResNet-50** [15] representan, respectivamente, una variante ligera y una de capacidad media de la familia residual. **DenseNet-121** [16] maximiza la reutilización de características mediante conexiones densas entre capas. **EfficientNet-B0** [17] aporta un diseño optimizado para obtener la máxima precisión por parámetro. **ConvNeXt-Tiny** [18] incorpora elementos de diseño de los *transformers* de visión manteniendo la naturaleza convolucional. Se descartaron deliberadamente VGG-16, por su elevado número de parámetros (∼138 millones) y el consiguiente riesgo de sobreajuste

sobre un conjunto pequeño, y MobileNet-V3, por resultar redundante con EfficientNet-B0 en el papel de arquitectura moderna y eficiente.

3.2.4. Protocolo de validación cruzada

Con un conjunto de datos tan reducido, una única partición entrenamiento/test produce estimaciones de rendimiento muy inestables. Por ello se emplea **validación cruzada de 5 particiones** (*5-fold cross-validation*) [23], que promedia el rendimiento sobre cinco repeticiones y aprovecha todos los datos disponibles. La partición concreta utilizada es `StratifiedGroupKFold`, que combina dos restricciones simultáneas: la *estratificación*, que conserva la proporción de clases Control/ELA en cada partición, y la *agrupación por paciente*, que garantiza que todas las imágenes de un mismo individuo permanezcan en la misma partición. Esta segunda restricción es la salvaguarda frente a la fuga de información: impide que el modelo vea en entrenamiento una imagen de un paciente y sea evaluado después sobre otra imagen del mismo paciente. El proceso se repite de forma independiente para cada una de las veinte combinaciones de arquitectura y músculo, y para cada combinación se obtienen cinco valores de AUC (uno por partición) sobre los que se construyen los intervalos de confianza y los contrastes estadísticos.

La figura 1 ilustra esquemáticamente cómo se distribuyen los 52 pacientes en los cinco *folds*. La estratificación reparte las clases (Control y ELA) en proporciones cercanas a la global (50 % control y 50 % ELA) dentro de cada partición; la agrupación por paciente impide que sus dos lateralidades (d e i) queden en lados opuestos del corte. En cada uno de los cinco entrenamientos, cuatro *folds* se dedican al entrenamiento y el restante a la validación; tras la rotación completa, todos los pacientes han actuado exactamente una vez como conjunto de validación. Esta propiedad es la que habilita la inferencia *out-of-fold* utilizada después en la fusión a nivel paciente.

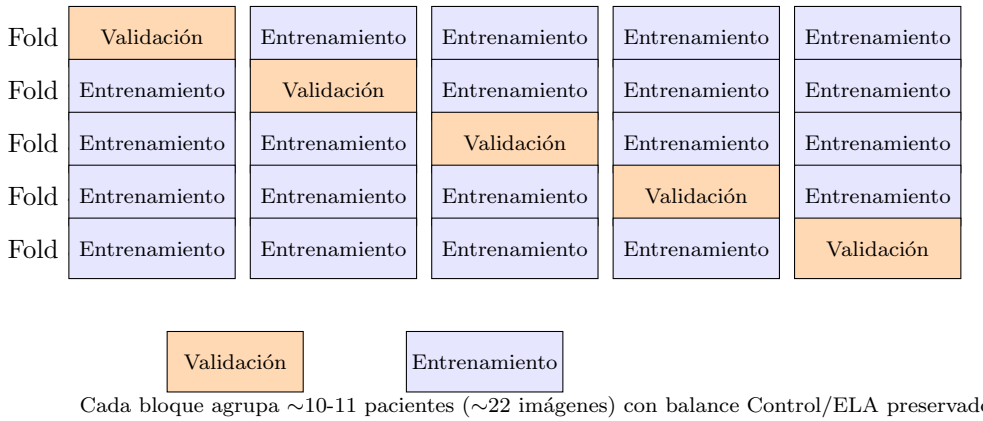


Figura 1: Distribución de los 52 pacientes en los cinco *folds* de la validación cruzada `StratifiedGroupKFold`. En cada iteración, cuatro *folds* (azul) se usan para entrenamiento y el restante (naranja) para validación. Tras las cinco iteraciones, cada paciente ha pasado exactamente una vez por validación, lo que permite construir las predicciones *out-of-fold* a nivel paciente.

Justificación de los hiperparámetros. Los hiperparámetros del entrenamiento se eligen siguiendo criterios estandarizados para *transfer learning* sobre conjuntos clínicos pequeños. La **tasa de aprendizaje inicial** 10^{-4} es el valor habitualmente recomendado para Adam cuando se parte de pesos preentrenados; valores más altos (típicos en entrenamientos desde cero) tienden a destruir las características aprendidas en ImageNet, mientras que valores significativamente más bajos prolongan el entrenamiento sin ganancia en rendimiento. El **tamaño de lote 16** es un compromiso entre estabilidad del gradiente y la memoria limitada disponible para entrenamientos en MPS. El **tope de 50 épocas** actúa como cota superior conservadora: en los 100 entrenamientos realizados, la mejor AUC en validación se alcanza con mediana en la época 8 y un 83 % de los *checkpoint* óptimos por debajo de la época 20, lo que confirma que el tope no constituye un cuello de botella sino un margen amplio para la activación del planificador. La **paciencia de 5 épocas** del `ReduceLROnPlateau` permite que el planificador reaccione a estancamientos cortos sin sobrerreaccionar a oscilaciones puntuales del AUC en validación, y el **factor 0,5** para la reducción de la tasa es el estándar de `scikit-learn` y `PyTorch` para este tipo de planificador. La **selección del *checkpoint* por mayor AUC en validación**, en lugar de por exactitud o por la última época, es preferible en conjuntos pequeños y balanceados porque el AUC se sustenta sobre las 22 predicciones probabilísticas completas y es menos sensible a la decisión binaria que puede oscilar por una

sola imagen mal clasificada. Finalmente, el **aumento de datos** (con sus rangos concretos de rotación ($\pm 10°$), brillo/contraste ($\pm 0{,}2$) y traslación afín (± 5 %)) se ha calibrado para reflejar la variabilidad operativa típica de la sonda y del protocolo de adquisición sin alterar el contenido textural relevante para el diagnóstico.

3.2.5. Fusión a nivel de paciente

Los modelos por músculo producen predicciones a nivel de imagen, pero la decisión clínica se toma sobre el paciente. La fusión combina la evidencia de los cuatro músculos en dos pasos. En el primero, para cada paciente y cada músculo se promedian las probabilidades $P(\text{ELA})$ de todas sus imágenes en ese músculo. Es esencial que estas probabilidades sean *out-of-fold*: se obtienen siempre del modelo de la partición en la que el paciente quedó en validación, de modo que ninguna predicción procede de un modelo que haya visto a ese paciente durante el entrenamiento. En el segundo paso se fusionan las hasta cuatro probabilidades musculares de cada paciente mediante tres reglas. La **media simple** promedia las cuatro probabilidades con igual peso. La **media ponderada** las promedia ponderando cada músculo por el AUC alcanzado por su modelo, de manera que los músculos más discriminantes pesan más:

$$P_{\text{fus}} = \frac{\sum_m w_m \, P_m}{\sum_m w_m}, \qquad w_m = \text{AUC}_m. \tag{3.2}$$

La tercera regla, el **voto mayoritario**, binariza primero la probabilidad de cada músculo con umbral 0,5 y declara ELA si al menos tres de los cuatro músculos votan ELA. Sobre la probabilidad fusionada de las dos primeras reglas se aplica el umbral de Youden para obtener la decisión final, y la incertidumbre se cuantifica mediante un intervalo de confianza al 95 % obtenido por *bootstrap* a nivel de paciente. Es importante destacar que la fusión no entrena ningún modelo nuevo: se limita a agregar predicciones ya existentes, por lo que no introduce sobreajuste adicional. La figura 2 resume el flujo completo de agregación.

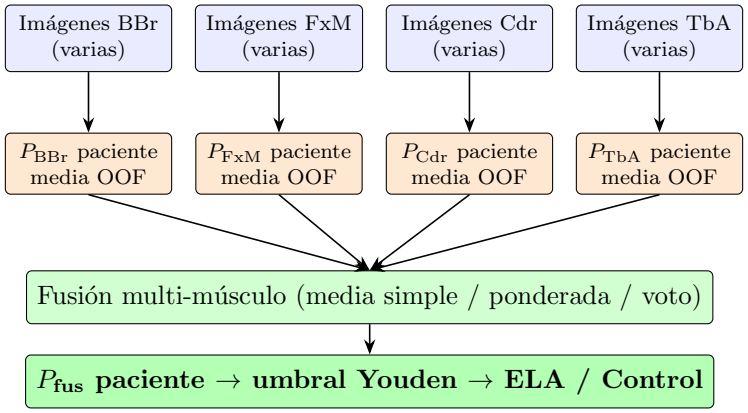


Figura 2: Esquema del proceso de fusión a nivel paciente. Para cada paciente y cada uno de los cuatro músculos, las probabilidades *out-of-fold* de sus imágenes se promedian para obtener una probabilidad muscular. Las cuatro probabilidades musculares se combinan mediante una de las tres reglas (media simple, media ponderada por AUC o voto mayoritario), y sobre la probabilidad fusionada se aplica el umbral de Youden para emitir la decisión final.

3.2.6. Análisis estadístico

Para que la comparación entre modelos sea metodológicamente sólida, y no una mera ordenación por rendimiento, se aplica un conjunto de herramientas estadísticas. A partir de los cinco valores de AUC por partición se calcula un **intervalo de confianza al 95 %** mediante la distribución t de Student, $\bar{x} \pm t_{n-1,\,\alpha/2} \cdot s/\sqrt{n}$, complementado con un intervalo *bootstrap* no paramétrico sobre esos mismos cinco valores [30]. La comparación frente a la línea base de Martínez-Payá se realiza comprobando si el AUC publicado por esos autores cae dentro o fuera del intervalo de confianza del modelo. La comparación entre arquitecturas dentro de un mismo músculo se aborda con tres pruebas: el test de los rangos con signo de **Wilcoxon** [28] sobre los AUC emparejados por partición, el test de **DeLong** [27] sobre las curvas ROC agregadas, y el test de **McNemar** [32] sobre los aciertos y fallos binarios. El empleo de tres pruebas complementarias, sensibles a aspectos distintos de la comparación, refuerza la robustez de las conclusiones.

3.2.7. Explicabilidad

Para mitigar la naturaleza de «caja negra» de las redes convolucionales y dotar al sistema de verificabilidad clínica, las predicciones del modelo final se acompañan de mapas de interpretación visual. Se emplean cuatro técnicas complementarias: **Grad-CAM** [22], que genera un mapa de calor de las regiones más influyentes a partir de los gradientes de la última capa convolucional; la **retropropagación guiada** [19]; los **mapas de saliencia** [20]; y el

análisis por **oclusión** [21], que mide la caída de la probabilidad al ocultar sistemáticamente distintas zonas de la imagen. Estas técnicas permiten comprobar que el modelo se apoya en el parénquima muscular y no en los bordes de la región de interés.

3.3. Implementación

3.3.1. Tecnologías y recursos empleados

El sistema se ha implementado íntegramente en Python. El entrenamiento y la inferencia de las redes se apoyan en **PyTorch** [33] y en **torchvision** [34], que proporciona tanto las arquitecturas como sus pesos preentrenados sobre ImageNet. El preprocesado de imágenes utiliza la biblioteca OpenCV; las particiones de validación cruzada y las métricas de clasificación, **scikit-learn** [35]; el análisis estadístico, **SciPy** [36] y statsmodels; el cálculo numérico, **NumPy** [37]; y la generación de gráficas, **Matplotlib** [38]. El cómputo se ha realizado en un equipo con procesador Apple Silicon M4, aprovechando la aceleración por GPU a través del *backend* Metal Performance Shaders (MPS); el código detecta automáticamente el dispositivo disponible y recurre a CUDA o a la CPU como alternativas.

3.3.2. Estructura del código

El código se organiza de forma modular, separando configuración, datos, modelos, entrenamiento y análisis. El módulo `config.py` centraliza todas las rutas, los hiperparámetros y la semilla de aleatoriedad. `preprocessing.py` convierte los segmentos TIFF originales en imágenes normalizadas. `dataset.py` construye los conjuntos de datos, extrae el identificador de paciente y genera las particiones de validación cruzada agrupadas. `models.py` actúa como factoría de arquitecturas, devolviendo cualquiera de las cinco redes preentrenadas con su capa de clasificación adaptada. `train_kfold.py` ejecuta el entrenamiento con validación cruzada de todas las combinaciones de arquitectura y músculo. `statistical_tests.py` realiza el análisis estadístico sobre los resultados. `patient_level_fusion.py` implementa la inferencia *out-of-fold* y la fusión a nivel de paciente. Por último, `explainability.py` genera los mapas de interpretación visual. El repositorio incluye además `organise_data.py` (organización de las imágenes por músculo durante la preparación del conjunto) y `create_figures.py` (regeneración de las figuras de resultados de esta memoria a partir de los ficheros de salida). Procedentes de una fase exploratoria previa con partición simple 80/20, se conservan también `train.py` y `evaluate_saved.py`, que no forman parte del experimento principal y se mantienen por trazabilidad.

3.3.3. Flujo de procesamiento

El *preprocesado* parte de las ROI segmentadas en formato TIFF y las redimensiona a una resolución fija de 224×224 píxeles, el tamaño de entrada estándar de las arquitecturas preentrenadas, guardándolas como imágenes JPEG de alta calidad. Durante el entrenamiento se aplica *aumento de datos* sobre el conjunto de entrenamiento (volteos horizontal y vertical, rotaciones de hasta diez grados, ligeras variaciones de brillo y contraste y pequeñas traslaciones) con el fin de simular la variabilidad de la sonda y del operador entre adquisiciones, mientras que el conjunto de validación solo se redimensiona y normaliza. En todos los casos las imágenes se normalizan con la media y la desviación típica de ImageNet, condición necesaria para el aprendizaje por transferencia.

El *entrenamiento* de cada combinación de arquitectura y músculo se realiza durante un máximo de 50 épocas, con tamaño de lote 16, función de pérdida de entropía cruzada y optimizador Adam con tasa de aprendizaje inicial 10^{-4}. Un planificador `ReduceLROnPlateau` reduce la tasa de aprendizaje a la mitad cuando el AUC de validación se estanca. De cada partición se conserva el punto de control correspondiente a la época de mayor AUC en validación, criterio más estable que la exactitud cuando el conjunto de validación es pequeño.

La *evaluación* se efectúa a dos niveles. A nivel de imagen, cada partición produce las métricas de AUC, sensibilidad, especificidad y exactitud, que se agregan en media y desviación típica sobre las cinco particiones. A nivel de paciente, se reconstruyen las particiones con la misma semilla, se generan las predicciones *out-of-fold* y se aplican las reglas de fusión descritas en la sección 3.2. Todos los resultados intermedios (predicciones por partición, predicciones *out-of-fold*, tablas de fusión e informes estadísticos) se persisten en disco, lo que permite reejecutar el análisis sin necesidad de reentrenar.

3.3.4. Reproducibilidad

La reproducibilidad se ha tratado como un requisito de primer orden. Se fija una única semilla de aleatoriedad (42) que se propaga a los generadores de Python, NumPy y PyTorch, así como a la construcción de las particiones de validación cruzada. Gracias a ello, las particiones son idénticas entre la fase de entrenamiento y la fase de fusión, condición indispensable para que las predicciones *out-of-fold* sean válidas. Los hiperparámetros se declaran en un único módulo de configuración, los puntos de control de cada partición se guardan en disco y los resultados intermedios se almacenan en ficheros estructurados. En conjunto, estas medidas permiten regenerar la totalidad de los resultados del trabajo a partir del código y de los datos originales. El código completo del proyecto está disponible públicamente en el repositorio `https://github.com/jjcortinaa/TFG`; los pesos entrenados (aproximadamente 5,5 GB), no versionados por su tamaño, se encuentran en `https://drive.google.com/drive/folders/1hDHp9OzT3U6jzT9ogTiLIab_jd5rGU96?usp=drive_link`. El conjunto de imágenes, propiedad del centro médico colaborador, no se publica por motivos de confidencialidad.

Capítulo 4 Resultados

Este capítulo presenta los resultados experimentales del sistema descrito en el capítulo 3. La exposición se organiza siguiendo el flujo de validación de lo más granular a lo más integrado: la sección 4.1 detalla la configuración experimental exacta; la sección 4.2 reporta el rendimiento de cada arquitectura sobre cada músculo de forma aislada, con sus intervalos de confianza; la sección 4.3 contrasta estadísticamente la igualdad de las arquitecturas dentro de cada músculo; la sección 4.4 mide el impacto de la recalibración del umbral mediante el índice de Youden; la sección 4.5 integra las predicciones de los cuatro músculos en un único veredicto por paciente y compara las tres reglas de fusión propuestas; la sección 4.6 confronta el sistema final con la línea base clínica de Martínez-Payá; y la sección 4.7 cierra con un análisis cualitativo basado en mapas de explicabilidad. La sección 4.8 sintetiza los hallazgos.

4.1. Configuración experimental

Todos los experimentos se ejecutan sobre el conjunto de datos descrito en el capítulo 3: 416 imágenes (104 por músculo, 52 control y 52 ELA) de 52 pacientes (26 control y 26 ELA), redimensionadas a 224×224 píxeles y normalizadas con la media y desviación típica de ImageNet. Las particiones se generan con `StratifiedGroupKFold` ($n = 5$, `shuffle=True`), que conserva el balance entre clases y garantiza la separación a nivel de paciente entre entrenamiento y validación. Una aserción explícita en `dataset.py` verifica en cada *fold* que la intersección entre sujetos de entrenamiento y validación es vacía.

Cada combinación (arquitectura, músculo, partición) se entrena durante un máximo de 50 épocas con tamaño de lote 16, optimizador Adam con tasa de aprendizaje inicial 10^{-4}, función de pérdida de entropía cruzada y planificador `ReduceLROnPlateau` (`mode=max`, factor 0,5 y `patience` $= 5$) sobre el AUC de validación. De cada *fold* se conserva el *checkpoint* de mayor AUC en validación. El cómputo se realiza en un MacBook con procesador Apple Silicon M4 aprovechando la aceleración por GPU mediante el *backend* MPS. La semilla de aleatoriedad se fija en 42 y se propaga a los generadores de Python, NumPy, PyTorch y scikit-learn, lo que garantiza que las particiones sean idénticas entre la fase de entrenamiento y la fase de fusión a nivel paciente.

El aumento de datos descrito en el capítulo 3 se aplica únicamente al conjunto de entrenamiento; el conjunto de validación se limita a redimensionar y normalizar. Las predicciones por imagen, los pesos del mejor *epoch* de cada *fold* y los informes estadísticos se persisten en disco, lo que permite reproducir el análisis completo sin necesidad de reentrenar.

4.2. Rendimiento por arquitectura y músculo

La tabla 2 reúne el AUC medio y la desviación típica de cada combinación de arquitectura y músculo, calculados sobre los cinco *folds* de la validación cruzada. La tabla 3 amplía esa información con dos intervalos de confianza al 95 % complementarios: el paramétrico, obtenido por aproximación de la t de Student sobre los cinco valores por *fold*, y el no paramétrico, obtenido por *bootstrap* ($N = 1000$) sobre esos mismos valores [30]. Los dos procedimientos arrojan estimaciones consistentes entre sí.

Tabla 2: AUC (en %) en validación cruzada de 5 *folds* para cada combinación de arquitectura y músculo. Se reporta media ± desviación típica. En negrita, el mejor valor por columna.

Arquitectura	**Bíceps**	**Antebrazo**	**Cuádriceps**	**Tibial**
ResNet-18	86,25 ± 5,80	91,38 ± 3,35	**99,33** ± 1,33	**93,96** ± 5,53
ResNet-50	86,35 ± 6,08	89,76 ± 3,71	**99,33** ± 1,33	92,37 ± 4,28
DenseNet-121	**90,30** ± 4,38	88,73 ± 3,90	**99,33** ± 1,33	93,17 ± 4,13
EfficientNet-B0	85,86 ± 5,54	**91,71** ± 5,09	99,13 ± 1,29	90,98 ± 5,72
ConvNeXt-Tiny	86,28 ± 5,62	91,03 ± 5,99	**99,33** ± 1,33	92,38 ± 5,24

Tres observaciones se desprenden de las tablas anteriores. En primer lugar, las diferencias entre arquitecturas dentro de un mismo músculo son pequeñas y caen casi siempre dentro de los intervalos de confianza de las demás, lo que sugiere que, una vez fijado el paradigma de aprendizaje por transferencia desde ImageNet, la elección concreta de arquitectura tiene un impacto secundario sobre el rendimiento por músculo. En segundo lugar, las diferencias entre músculos son mucho mayores: el cuádriceps alcanza un AUC en torno al 99 % con cualquier arquitectura, el bíceps se mueve en el rango 85–90 %, y antebrazo y tibial se sitúan a medio camino. Este gradiente refleja la heterogeneidad clínica de la afectación muscular en la ELA y es coherente con la mayor sensibilidad del cuádriceps a los cambios texturales descrita en la literatura previa [5], [6]. En tercer lugar, el cuádriceps muestra una saturación informativa: cuatro de las cinco arquitecturas alcanzan exactamente el mismo AUC (99,33 %) con la misma desviación típica, lo que indica que la información discriminante de ese músculo es prácticamente recuperable en su totalidad por cualquier red moderna y que el techo de rendimiento por músculo aislado se ha alcanzado.

La figura 3 visualiza la distribución de AUC entre los cinco *folds* para las veinte combinaciones de arquitectura y músculo. Estos diagramas de caja confirman gráficamente las dos observaciones anteriores: el cuádriceps presenta cajas muy estrechas y elevadas, indicando consistencia entre arquitecturas; el bíceps muestra cajas más anchas, reflejando mayor variabilidad entre *folds*; y antebrazo y tibial muestran patrones intermedios.

Tabla 3: Intervalos de confianza al 95 % para el AUC (en %) por arquitectura y músculo. Se reportan el IC obtenido por aproximación de la t de Student sobre los 5 valores por *fold* y el IC por *bootstrap* no paramétrico ($N = 1000$).

Músculo	**Arquitectura**	**IC95 % t-Student**	**IC95 % bootstrap**
Bíceps	ResNet-18	[78,20; 94,30]	[81,41; 91,18]
	ResNet-50	[77,92; 94,78]	[80,42; 91,24]
	DenseNet-121	[84,23; 96,37]	[86,47; 94,13]
	EfficientNet-B0	[78,17; 93,54]	[81,27; 91,03]
	ConvNeXt-Tiny	[78,48; 94,08]	[80,70; 90,32]
Antebrazo	ResNet-18	[86,73; 96,04]	[88,60; 94,55]
	ResNet-50	[84,61; 94,91]	[87,10; 93,53]
	DenseNet-121	[83,32; 94,15]	[85,63; 92,63]
	EfficientNet-B0	[84,65; 98,77]	[86,53; 95,23]
	ConvNeXt-Tiny	[82,72; 99,33]	[85,93; 96,16]
Cuádriceps	ResNet-18	[97,48; 100,00]	[98,00; 100,00]
	ResNet-50	[97,48; 100,00]	[98,00; 100,00]
	DenseNet-121	[97,48; 100,00]	[98,00; 100,00]
	EfficientNet-B0	[97,34; 100,00]	[97,80; 100,00]
	ConvNeXt-Tiny	[97,48; 100,00]	[98,00; 100,00]
Tibial	ResNet-18	[86,29; 100,00]	[88,96; 98,58]
	ResNet-50	[86,43; 98,31]	[88,60; 96,13]
	DenseNet-121	[87,43; 98,90]	[89,83; 97,33]
	EfficientNet-B0	[83,05; 98,92]	[86,15; 96,25]
	ConvNeXt-Tiny	[85,11; 99,64]	[88,00; 97,21]

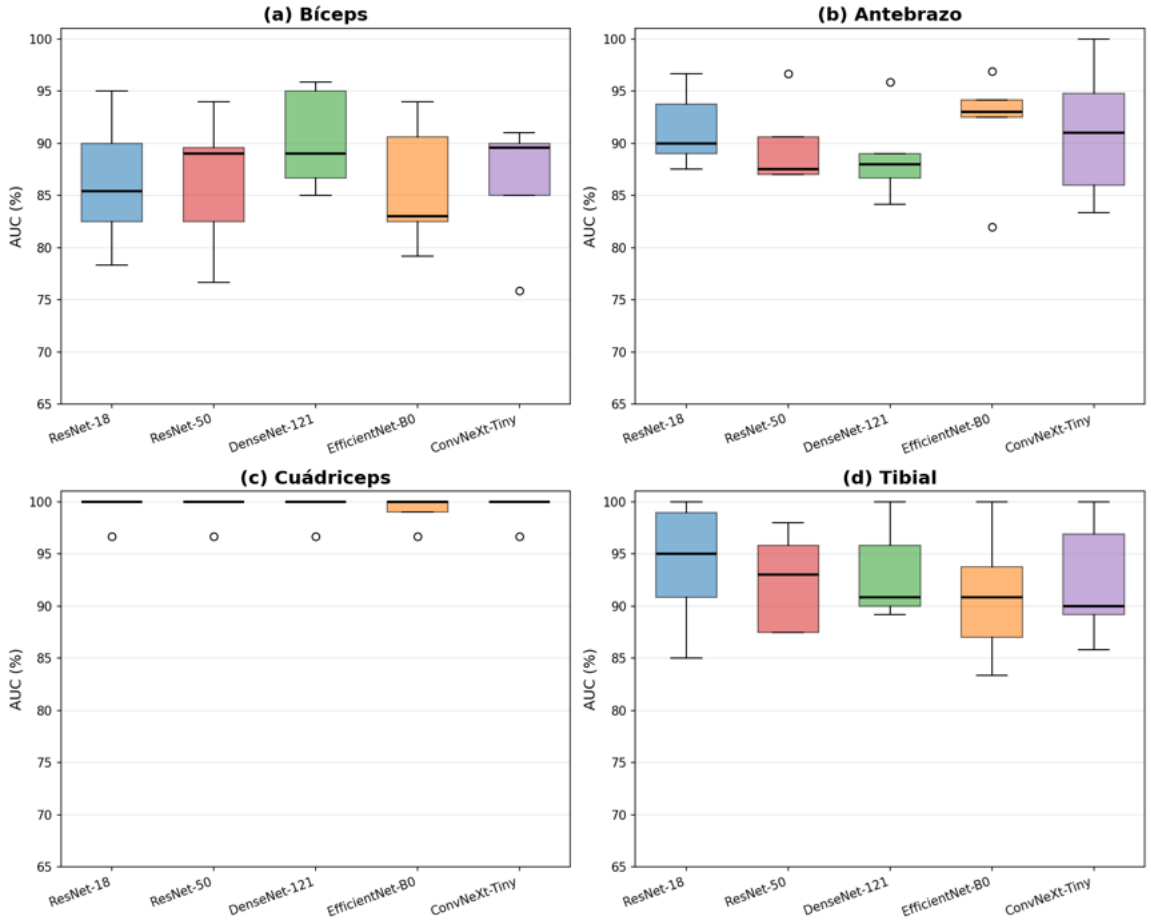
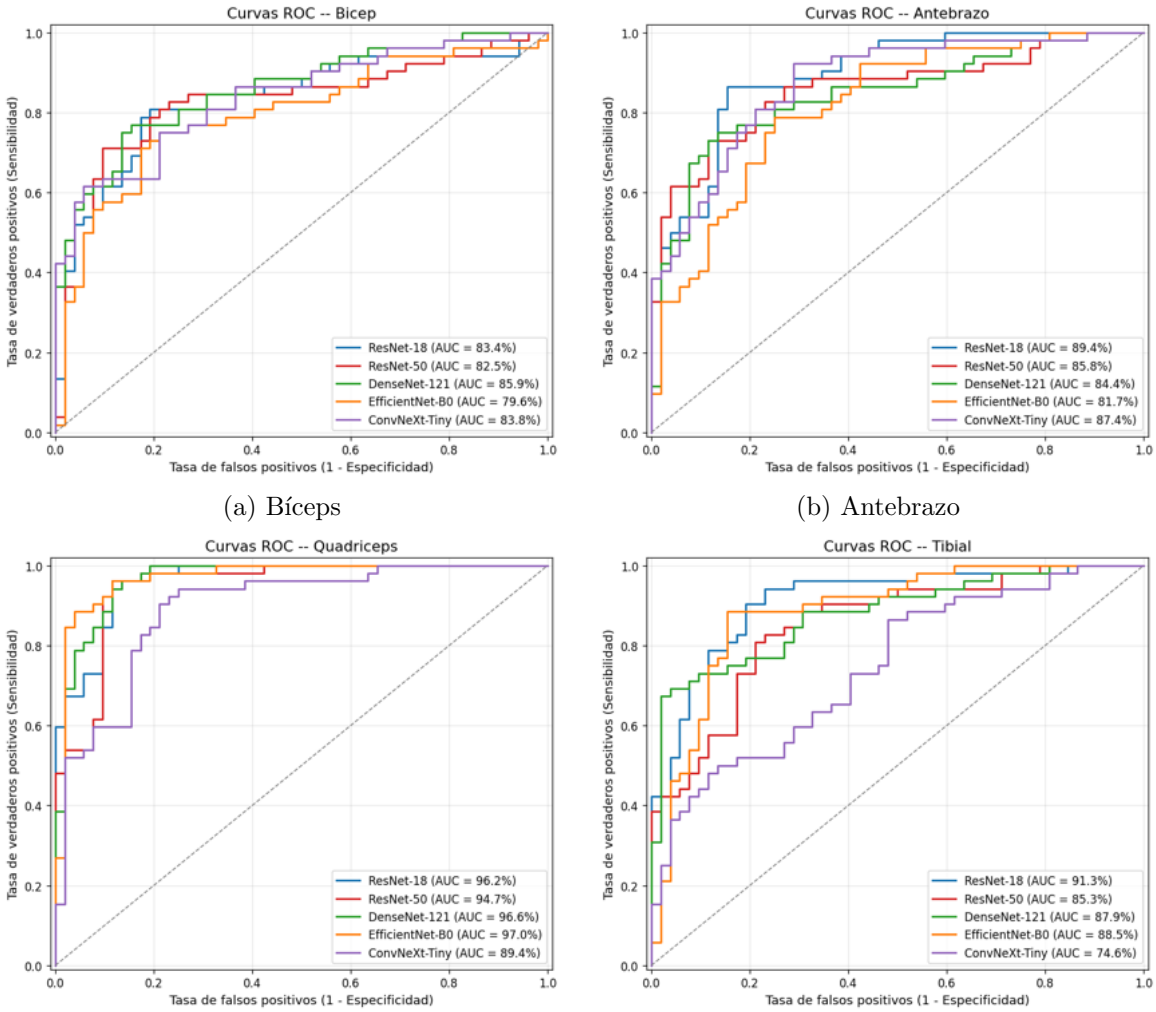


Figura 3: Distribución del AUC por *fold* (5 valores) para las cinco arquitecturas en cada uno de los cuatro músculos. La altura y dispersión de cada caja resumen la consistencia del modelo: cajas más estrechas indican mayor estabilidad entre particiones. El cuádriceps presenta el patrón más consistente y elevado; el bíceps, el más disperso.

Para visualizar el comportamiento de las cinco arquitecturas a lo largo de todo el rango de umbrales (no solo en la métrica resumen del AUC), la figura 4 muestra las curvas ROC obtenidas a partir de las predicciones *out-of-fold* concatenadas de los cinco *folds*. Cada curva agrupa las 104 predicciones (52 control + 52 ELA) de cada combinación de arquitectura y músculo.



(a) Bíceps (b) Antebrazo

(c) Cuádriceps (d) Tibial

Figura 4: Curvas ROC *out-of-fold* para las cinco arquitecturas en cada uno de los cuatro músculos. Las curvas se construyen sobre las 104 predicciones concatenadas de los cinco *folds*. La proximidad de las cinco curvas en cada panel ilustra gráficamente la equivalencia entre arquitecturas, y la diferencia visible entre paneles refleja la heterogeneidad entre músculos: el cuádriceps satura cerca del codo superior izquierdo, mientras que el bíceps muestra un compromiso menos favorable entre sensibilidad y especificidad.

Las curvas confirman gráficamente las dos observaciones cuantitativas anteriores. En cuádriceps las cinco curvas se solapan casi por completo y se aproximan a la esquina superior izquierda, dejando un AUC en torno al 99 % y un compromiso sensibilidad-especificidad prácticamente óptimo en cualquier umbral. En bíceps, el más exigente de los cuatro, las

curvas se separan más entre sí y muestran codos más suaves, lo que se traduce en una mayor dependencia del umbral elegido y justifica el peso clínico de la recalibración de Youden discutida en la sección 4.4.

4.3. Comparación estadística entre arquitecturas

Para contrastar formalmente si las diferencias de AUC observadas son estadísticamente significativas, se aplican tres pruebas complementarias a cada par de arquitecturas dentro de cada músculo: el contraste de los rangos con signo de Wilcoxon sobre los cinco valores de AUC por *fold* [28], el contraste de DeLong sobre las curvas ROC agregadas [27] y el contraste de McNemar sobre los aciertos y fallos binarios [32]. Las tres pruebas son sensibles a aspectos distintos de la comparación: Wilcoxon a la dispersión de los AUC entre *folds*, DeLong a la forma global de la curva ROC y McNemar a la concordancia de las predicciones binarias.

La tabla 4 sintetiza, para cada músculo, el resultado del contraste entre la arquitectura con mayor AUC medio y la segunda mejor. Con un nivel de significación $\alpha = 0{,}05$, ninguna comparación entre la mejor y la segunda mejor arquitectura resulta significativa en bíceps, antebrazo, cuádriceps ni tibial, lo que confirma la lectura inicial: las arquitecturas son, en la práctica, intercambiables a nivel de músculo individual.

Tabla 4: Contraste estadístico entre la arquitectura de mayor AUC y la segunda mejor en cada músculo. Se reportan los AUC medios (en %) y los p-valores de las tres pruebas. Ningún p-valor es inferior a 0,05.

Músculo	**Mejor**	**2.ª mejor**	**AUC$_1$**	**AUC$_2$**	**p Wilc.**	**p DeLong**	**p McNem.**
Bíceps	DenseNet-121	ResNet-50	90,30	86,35	0,125	0,348	0,286
Antebrazo	EfficientNet-B0	ResNet-18	91,71	91,38	0,812	0,071	0,711
Cuádriceps	ResNet-18*	EfficientNet-B0	99,33	99,13	1,000	0,746	0,424
Tibial	ResNet-18	DenseNet-121	93,96	93,17	0,750	0,350	0,481

*En cuádriceps, ResNet-18, ResNet-50, DenseNet-121 y ConvNeXt-Tiny empatan en AUC 99,33 %.

El caso del cuádriceps merece un comentario adicional. Cuatro arquitecturas comparten exactamente el mismo AUC medio y la misma desviación típica entre *folds*, por lo que el contraste de Wilcoxon, basado en las diferencias por *fold*, no puede calcular un p-valor para algunas de esas parejas: la igualdad es exacta. El contraste de DeLong, que opera sobre las curvas ROC agregadas, sí detecta diferencias significativas en alguna pareja minoritaria (por ejemplo, ConvNeXt-Tiny frente a EfficientNet-B0, $p_{\text{DeLong}} = 0{,}011$). Esto refleja que los modelos producen *rankings* ligeramente distintos de la misma colección de imágenes, aun cuando la métrica resumen sea idéntica. La conclusión práctica es la misma: en cuádriceps el techo está tan alto que la elección de arquitectura es indiferente desde un punto de vista clínico.

El tibial es el músculo donde las pruebas pareadas detectan más diferencias entre arquitecturas. En particular, ConvNeXt-Tiny pierde frente a ResNet-18 con $p_{\text{DeLong}} = 0{,}0012$ y $p_{\text{McNemar}} = 0{,}0076$, y frente a DenseNet-121 con $p_{\text{DeLong}} = 0{,}0004$ y $p_{\text{McNemar}} = 0{,}0294$. Aunque la diferencia en AUC medio es modesta (92,38 % frente a 93,96 %), las pruebas sugieren que las arquitecturas residuales y de conexiones densas se adaptan mejor a este músculo que ConvNeXt-Tiny.

4.4. Recalibración del umbral con el índice de Youden

Las arquitecturas devuelven una probabilidad continua de pertenencia a la clase ELA, $P(\text{ELA} \mid x) \in [0, 1]$. La conversión a una decisión binaria requiere fijar un umbral, y la elección por defecto ($t = 0{,}5$) no es necesariamente óptima en un escenario con clases balanceadas pero con un coste asimétrico entre falsos positivos y falsos negativos. Para cada combinación (arquitectura, músculo) se recalibra el umbral mediante el índice de Youden [29], que escoge el valor que maximiza $J = \text{Sens} + \text{Spec} - 1$ sobre la curva ROC. La tabla 5 resume el efecto de esta recalibración.

El impacto de la recalibración es notable, especialmente en bíceps y cuádriceps. En bíceps, la sensibilidad de DenseNet-121 pasa del 49,7 % con umbral 0,5 al 87,7 % con el umbral de Youden, manteniendo una especificidad por encima del 86 %; un cambio cualitativo, no marginal. En cuádriceps, varias arquitecturas alcanzan simultáneamente sensibilidades del 98–100 % y especificidades del 96–100 % tras la recalibración, lo que confirma que la limitación a umbral 0,5 no era un límite de capacidad del modelo, sino una mera ineficiencia del punto de operación. Los umbrales óptimos resultantes se sitúan en su mayoría por debajo de 0,5 (mediana cercana a 0,45), lo que indica que las arquitecturas tienden a ser conservadoras al asignar probabilidades altas a la clase ELA en este conjunto de datos.

4.5. Fusión a nivel paciente

El objetivo clínico último del sistema es emitir una decisión a nivel de paciente, no a nivel de imagen. Para ello se construye, para cada paciente, una probabilidad fusionada que integra la evidencia de los cuatro músculos. El procedimiento se realiza en dos pasos. En el primero, para cada paciente y cada músculo se promedian las probabilidades *out-of-fold* producidas por el modelo del *fold* en el que ese paciente quedó en validación; este detalle es lo que garantiza que ninguna predicción provenga de un modelo que haya visto al paciente durante el entrenamiento. En el segundo, las cuatro probabilidades resultantes se combinan mediante las tres reglas descritas en la sección 3.2: media simple, media ponderada por el AUC del modelo en ese músculo y voto mayoritario.

Una precisión sobre el conjunto de datos: en el directorio de imágenes los sujetos control aparecen identificados como `Cnnn` en bíceps, antebrazo y cuádriceps y como `RCnnn` en tibial.

Tabla 5: Efecto de la recalibración del umbral con el índice de Youden. Se reportan sensibilidad y especificidad (en %) con umbral fijo $t = 0{,}5$ y con el umbral óptimo de Youden, así como el umbral medio resultante con su desviación típica entre *folds*.

Músculo	**Arquitectura**	**Sens**$_{0,5}$	**Spec**$_{0,5}$	**Sens**$_Y$	**Spec**$_Y$	t_Y
Bíceps	DenseNet-121	49,7	96,0	87,7	86,7	0,26 ± 0,18
	ResNet-50	71,8	86,3	81,8	88,3	0,53 ± 0,10
	ResNet-18	76,2	83,0	83,5	85,3	0,49 ± 0,23
	ConvNeXt-Tiny	61,0	90,0	78,5	88,3	0,36 ± 0,15
	EfficientNet-B0	63,3	82,7	85,3	81,0	0,43 ± 0,09
Antebrazo	EfficientNet-B0	69,7	78,7	89,3	88,3	0,52 ± 0,15
	ConvNeXt-Tiny	59,5	88,3	86,5	90,3	0,38 ± 0,14
	ResNet-50	78,5	78,7	83,3	92,7	0,56 ± 0,09
	ResNet-18	65,8	86,3	84,5	88,3	0,44 ± 0,18
	DenseNet-121	67,3	92,3	80,3	92,3	0,55 ± 0,16
Cuádriceps	ResNet-18	78,0	91,3	98,3	100,0	0,56 ± 0,22
	ConvNeXt-Tiny	69,3	84,0	100,0	98,0	0,34 ± 0,22
	ResNet-50	68,7	90,7	98,3	98,0	0,51 ± 0,11
	DenseNet-121	74,2	96,0	98,3	98,0	0,46 ± 0,23
	EfficientNet-B0	98,0	76,3	100,0	96,0	0,57 ± 0,10
Tibial	EfficientNet-B0	77,0	84,3	91,8	88,3	0,48 ± 0,04
	ResNet-18	78,8	86,3	88,3	92,3	0,48 ± 0,28
	ConvNeXt-Tiny	51,7	83,7	98,3	80,7	0,38 ± 0,14
	DenseNet-121	72,3	84,7	86,3	92,7	0,41 ± 0,26
	ResNet-50	74,0	83,0	91,3	84,7	0,48 ± 0,21

Ambas convenciones se refieren al mismo individuo, lo que se confirmó con el responsable del dataset; la función `patient_uid` en `patient_level_fusion.py` normaliza el identificador antes de la fusión, garantizando que las cuatro probabilidades musculares se atribuyan al mismo paciente físico.

La tabla 6 contiene los resultados de fusión a nivel paciente para cada una de las cinco arquitecturas y las tres reglas. La métrica principal es el AUC, acompañado de su intervalo de confianza al 95 % obtenido por *bootstrap* no paramétrico ($N = 1000$) sobre los 52 pacientes. La sensibilidad, especificidad y exactitud se reportan con el umbral de Youden aplicado a la probabilidad fusionada; el voto mayoritario, al binarizar antes de la fusión, no admite un AUC fusionado y se reporta solo con sus valores de operación.

El sistema final del trabajo es **ResNet-50 con fusión por media simple**, que alcanza un AUC del 98,67 % (IC95 % [95,38; 100,00]), una exactitud del 96,15 % (IC95 % [90,38; 100,00]), una sensibilidad del 100 % (IC95 % [100,00; 100,00]) y una especificidad del 92,31 % (IC95 % [79,31; 100,00]) sobre los 52 pacientes del estudio. Las cinco arquitecturas

Tabla 6: Resultados de fusión a nivel paciente (52 pacientes; 26 ELA, 26 control) con las tres reglas estudiadas. Se reportan AUC, exactitud, sensibilidad y especificidad (en %) y la matriz de confusión TP/TN/FP/FN. En negrita, el sistema final del trabajo; las últimas filas (Campeones) corresponden a la fusión de la mejor arquitectura por músculo.

Arquitectura	Regla	AUC	Acc	Sens	Spec	TP/TN/FP/FN
ResNet-18	media simple	99,26	96,15	96,15	96,15	25/25/1/1
	ponderada	99,26	96,15	96,15	96,15	25/25/1/1
	voto mayoritario	–	90,38	80,77	100,00	21/26/0/5
ResNet-50	**media simple**	**98,67**	**96,15**	**100,00**	**92,31**	**26/24/2/0**
	ponderada	98,82	96,15	100,00	92,31	26/24/2/0
	voto mayoritario	–	84,62	69,23	100,00	18/26/0/8
DenseNet-121	media simple	97,93	96,15	96,15	96,15	25/25/1/1
	ponderada	97,93	96,15	96,15	96,15	25/25/1/1
	voto mayoritario	–	78,85	61,54	96,15	16/25/1/10
EfficientNet-B0	media simple	98,22	96,15	96,15	96,15	25/25/1/1
	ponderada	98,37	96,15	96,15	96,15	25/25/1/1
	voto mayoritario	–	88,46	80,77	96,15	21/25/1/5
ConvNeXt-Tiny	media simple	98,52	96,15	100,00	92,31	26/24/2/0
	ponderada	98,52	96,15	100,00	92,31	26/24/2/0
	voto mayoritario	–	73,08	46,15	100,00	12/26/0/14
Campeones[†]	media simple	99,11	94,23	100,00	88,46	26/23/3/0
	ponderada	99,11	94,23	100,00	88,46	26/23/3/0
	voto mayoritario	–	84,62	69,23	100,00	18/26/0/8

[†] Campeones: *ensemble* heterogéneo que combina la mejor arquitectura por músculo (DenseNet-121 en bíceps, EfficientNet-B0 en antebrazo y ResNet-18 en cuádriceps y tibial).

resultan estadísticamente indistinguibles también a nivel de fusión (todos los AUC se sitúan entre 97,9 y 99,3 % con intervalos de confianza solapados); entre opciones equivalentes se adopta ResNet-50 como sistema final por ser la arquitectura de referencia para *transfer learning* y por ofrecer el perfil de error clínicamente más conservador, con **sensibilidad del 100 %** (ninguna ELA sin detectar). La matriz de confusión a nivel paciente revela dos errores, ambos falsos positivos (controles clasificados como ELA); no hay ningún falso negativo.

Varios hallazgos del proceso de fusión merecen subrayarse. En primer lugar, la media ponderada apenas mejora a la media simple: las dos reglas producen exactamente las mismas predicciones binarias en las cinco arquitecturas, y solo alteran ligeramente el AUC (a lo sumo 0,15 puntos, en ResNet-50 y EfficientNet-B0). Esto se debe a que los pesos (AUC por músculo) están comprimidos en un rango estrecho (entre 86 % y 99 %) y, una vez normalizados, no priorizan ningún músculo de forma decisiva. La regla simple, al no requerir hiperparámetro

alguno, es preferible. En segundo lugar, el voto mayoritario funciona muy por debajo de las dos reglas continuas: su exactitud no supera el 90,4 % en ninguna arquitectura y su sensibilidad se deteriora hasta el 46,2 % en ConvNeXt-Tiny. La explicación es metodológica: binarizar a $t = 0{,}5$ por músculo antes de fusionar descarta toda la información de confianza de las probabilidades y exige el acuerdo simultáneo de al menos tres de los cuatro músculos, un criterio demasiado estricto que penaliza a los pacientes cuya señal de ELA es nítida solo en algunos músculos. La conclusión es que fusionar probabilidades es estrictamente mejor que fusionar votos.

Una estrategia alternativa que combinaría los *campeones* por músculo (el modelo con mayor AUC en cada uno: DenseNet-121 en bíceps, EfficientNet-B0 en antebrazo, ResNet-18 en cuádriceps y ResNet-18 en tibial) alcanza un AUC fusionado del 99,11 % con la regla de media simple (IC95 % [97,00; 100,00]; últimas filas de la tabla 6), un valor estadísticamente indistinguible del de las arquitecturas únicas (cuyos AUC fusionados oscilan entre 97,9 y 99,3 %, con intervalos de confianza solapados). Mezclar arquitecturas no aporta, por tanto, una ventaja apreciable sobre emplear una sola, mientras que sí introduce complejidad operativa (cuatro arquitecturas y cuatro procedimientos de entrenamiento distintos) y, además, exhibe un perfil de error ligeramente peor (especificidad 88,5 % frente al 92,3 % del sistema ResNet-50, por tres falsos positivos frente a dos). Este resultado refuerza la elección de diseño adoptada: usar una única arquitectura entrenada por separado sobre cada músculo es preferible, por simplicidad y mantenibilidad, a un *ensemble* heterogéneo de los mejores modelos por músculo.

4.6. Comparación con la línea base clínica

La línea base de referencia para este trabajo es el sistema de análisis textural manual de Martínez-Payá y colaboradores [6]. La tabla 7 resume el AUC publicado por esos autores para cada músculo, las técnicas que emplean y la comparación frente al mejor modelo de este trabajo en validación cruzada de 5 *folds*.

Tabla 7: Comparación con la línea base clínica [6]. La columna "Mejor arquitectura" indica el modelo de mayor AUC en cada músculo, con su IC95 % *bootstrap*. La columna "Veredicto" resume si el AUC publicado por la línea base cae dentro (equivalente) o fuera (mejora) del IC95 %.

Músculo	**Mejor arquitectura**	**Nuestro AUC (IC95 %)**	**Baseline**	**Veredicto**
Bíceps	DenseNet-121	90,30 [86,5; 94,1]	92,6 (EV+MTh)	equivalente
Antebrazo	EfficientNet-B0	91,71 [86,5; 95,2]	90,5 (GLCM+MTh)	equivalente
Cuádriceps	ResNet-18	99,33 [98,0; 100,0]	98,3 (GLCM+MTh)	equivalente
Tibial	ResNet-18	93,96 [89,0; 98,6]	95,3 (EV+MTh)	equivalente

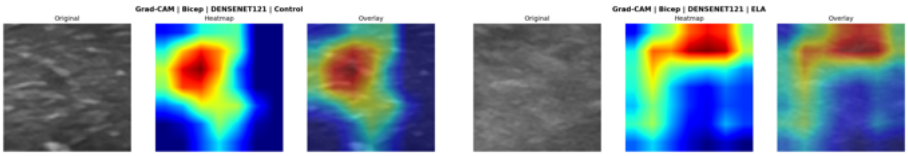
En los cuatro músculos, el AUC publicado por la línea base cae dentro del intervalo de confianza al 95 % del mejor modelo de este trabajo. Por tanto, a nivel de músculo individual, ninguna arquitectura mejora de forma estadísticamente concluyente al análisis textural clásico. Lejos de ser un resultado negativo, esta observación es coherente con la naturaleza del problema y con la propuesta del trabajo: la aportación no es ganar al análisis textural cuando se compite músculo a músculo, sino integrar los cuatro músculos en una única decisión clínica.

La comparación adecuada del sistema completo es con la mejor combinación que la línea base puede producir a nivel paciente. El trabajo original de Martínez-Payá no propone una estrategia explícita de fusión multi-músculo a nivel paciente; reporta resultados por músculo de manera aislada. Por consiguiente, la comparación directa entre el sistema fusionado ($\text{AUC}_{\text{fusión}} = 98{,}67$ %) y el AUC máximo por músculo del *baseline* (98,3 % en cuádriceps) no es estrictamente equivalente, pero sí ilustra la magnitud de la contribución: el sistema final mantiene un AUC cercano al 99 % en una clasificación a nivel paciente, mientras que el *baseline* alcanza ese rango únicamente cuando se examina el músculo individualmente más informativo. La aportación fundamental del trabajo es, por tanto, la integración multi-músculo a nivel paciente con un protocolo de validación libre de fuga de información, no la mejora marginal por músculo.

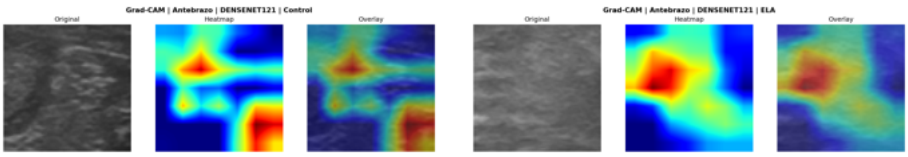
4.7. Explicabilidad: análisis cualitativo

Para mitigar la naturaleza de caja negra del modelo y verificar la plausibilidad clínica de sus decisiones, se generan cuatro tipos de mapas de explicabilidad sobre el conjunto de validación: Grad-CAM [22], Grad-CAM guiado, mapas de saliencia [20] y análisis por oclusión [21], complementado con retropropagación guiada [19]. La figura 5 muestra una selección representativa de mapas Grad-CAM sobre los cuatro músculos, con un sujeto control y un paciente ELA para cada uno.

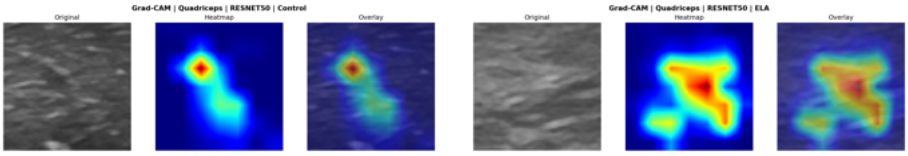
El análisis cualitativo aporta tres conclusiones principales. En primer lugar, las activaciones de mayor magnitud se localizan sobre el tejido muscular y no sobre los bordes o márgenes de la región de interés; este hallazgo es necesario para considerar al modelo confiable desde un punto de vista clínico. En segundo lugar, no se aprecia una diferencia visual evidente en la extensión o la heterogeneidad de las activaciones entre los pacientes con ELA y los controles. Esto es esperable por dos motivos. Por un lado, los cambios texturales asociados a la enfermedad son sutiles (de hecho, esa es la motivación para emplear una CNN en lugar de la inspección visual directa). Por otro, las regiones de interés analizadas son recortes seleccionados de antemano por personal clínico como las zonas más relevantes de cada ecografía; al tratarse de entradas ya acotadas al tejido informativo, el mapa no necesita descartar grandes áreas irrelevantes, como sí ocurriría si se hubiera empleado la ecografía completa, donde el modelo podría mostrar de forma más llamativa que ignora el fondo, el hueso o las fascias.



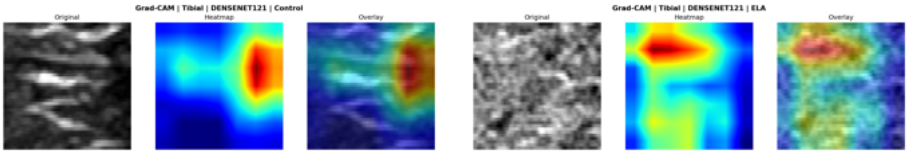
(a) Bíceps — control (izq.) vs ELA (der.)



(b) Antebrazo — control (izq.) vs ELA (der.)



(c) Cuádriceps — control (izq.) vs ELA (der.)



(d) Tibial — control (izq.) vs ELA (der.)

Figura 5: Mapas Grad-CAM sobre los cuatro músculos analizados. Para cada músculo se muestra un sujeto control (izquierda) y un paciente ELA (derecha). Las activaciones se concentran sobre el parénquima muscular y no sobre los bordes de la región de interés, lo que indica que el modelo se apoya en regiones clínicamente plausibles para emitir su predicción.

La explicabilidad confirma, por tanto, la plausibilidad de la localización de la predicción más que una diferencia cualitativa entre clases. En tercer lugar, los mapas de las cuatro técnicas (Grad-CAM, Grad-CAM guiado, saliencia y oclusión) coinciden cualitativamente entre sí en la localización de las regiones discriminantes, lo que refuerza la robustez de la interpretación: no se trata de un artefacto particular de un método de explicabilidad concreto, sino de una propiedad reproducible del modelo.

Es importante puntualizar dos limitaciones intrínsecas de este tipo de análisis para el problema concreto que se aborda. La primera es de **resolución**: la última capa convolucional de ResNet-50 sobre una entrada de 224×224 píxeles tiene una resolución espacial de 7×7, que es la rejilla nativa sobre la que Grad-CAM produce sus mapas antes de reescalarlos por in-

terpolación bilineal al tamaño original. Esta resolución intrínseca limita el detalle anatómico que pueden mostrar los focos de calor y explica el aspecto «suave» o difuso característico de las activaciones que se observan en la figura 5. La segunda limitación es de **naturaleza del problema**: Grad-CAM, y en general las técnicas de explicabilidad basadas en localización, responden a la pregunta de *dónde* mira el modelo, no a la de *qué* ve en la región atendida. En un problema de clasificación textural como el aquí planteado, en el que la información discriminante reside en el patrón fino del parénquima muscular (granulado, fibrosis, distribución de la ecointensidad) y no en la posición espacial del músculo en la imagen, es esperable que las regiones atendidas para ELA y control compartan ubicaciones similares (el centro del músculo) aunque la decisión final difiera. Precisamente por esta limitación se emplean cuatro técnicas complementarias (Grad-CAM, Grad-CAM guiado, mapas de saliencia y análisis por oclusión) que sondean el modelo desde mecanismos distintos y, cuando coinciden en señalar las mismas regiones, refuerzan la robustez de la interpretación más allá de lo que cualquiera de ellas podría sostener por separado.

Los mapas de explicabilidad son, en última instancia, una herramienta de verificación cualitativa, no una prueba formal de causalidad. Demuestran que el modelo se apoya en regiones clínicamente plausibles, no que sus decisiones individuales sean clínicamente correctas. Una validación clínica completa requeriría el contraste con un radiólogo experto sobre cada caso, un trabajo que excede el alcance de este Trabajo Fin de Grado pero que se identifica como línea de investigación futura en el capítulo 5.

4.8. Síntesis del capítulo

A nivel de músculo, las cinco arquitecturas estudiadas son estadísticamente equivalentes entre sí y, simultáneamente, equivalentes a la línea base clínica de Martínez-Payá: ningún contraste pareado entre la mejor y la segunda mejor arquitectura alcanza significación en ninguno de los cuatro músculos, y todos los AUC publicados por la línea base caen dentro de los intervalos de confianza de los modelos de este trabajo. La recalibración del umbral mediante el índice de Youden mejora la sensibilidad de forma sustancial sin sacrificar especificidad, hasta el punto de que en cuádriceps varias arquitecturas alcanzan simultáneamente sensibilidad y especificidad próximas al 100 %.

El aporte central del trabajo aparece al integrar los cuatro músculos en una única decisión por paciente. La fusión por media simple de las probabilidades *out-of-fold* de ResNet-50 produce el sistema final: AUC del 98,67 %, exactitud del 96,15 %, sensibilidad del 100 % y especificidad del 92,31 % sobre los 52 pacientes, con dos errores de clasificación, ambos falsos positivos y ninguna ELA sin detectar. La regla ponderada por AUC no aporta mejora apreciable sobre la media simple, y el voto mayoritario se descarta por su deterioro sistemático de la sensibilidad. Una estrategia alternativa que mezcle arquitecturas distintas (campeones por músculo) obtiene un rendimiento equivalente al de la arquitectura única, por lo que esta

última se prefiere por su mayor simplicidad operativa. Los mapas de explicabilidad confirman que la predicción se sustenta sobre el parénquima muscular, en regiones clínicamente plausibles. El siguiente capítulo discute las implicaciones de estos resultados y las líneas de trabajo futuro.

Capítulo 5 Conclusiones y Trabajo Futuro

Este capítulo cierra la memoria evaluando el grado de cumplimiento de los objetivos planteados en el capítulo 1, resumiendo las aportaciones técnicas y metodológicas del sistema, exponiendo las limitaciones del trabajo y trazando las líneas de investigación que de forma natural lo prolongan.

5.1. Cumplimiento de los objetivos

El objetivo principal del trabajo (diseñar, implementar y validar un sistema de soporte al diagnóstico de la ELA basado en redes neuronales convolucionales aplicadas a ecografía muscular y operando a nivel de paciente) se considera **alcanzado**. El sistema final, ResNet-50 entrenado de forma independiente sobre cada uno de los cuatro músculos y fusionado mediante la media simple de las probabilidades *out-of-fold*, alcanza sobre los 52 pacientes del estudio un AUC del 98,67 %, una exactitud del 96,15 %, una sensibilidad del 100 % y una especificidad del 92,31 %, con dos errores de clasificación (ambos falsos positivos). Cada objetivo específico planteado en el capítulo 1 se ha cubierto del siguiente modo:

- Se han entrenado y comparado las cinco arquitecturas previstas (ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B0 y ConvNeXt-Tiny) sobre los cuatro músculos del estudio, con un total de 100 modelos entrenados ($5 \times 4 \times 5$ *folds*), reportados en la sección 4.2.

- Se ha garantizado la integridad metodológica de la validación mediante `StratifiedGroupKFold` con agrupación por sujeto. Una aserción explícita en `dataset.py` verifica en cada *fold* la ausencia de solapamiento entre los conjuntos de entrenamiento y validación, salvaguarda frente a la fuga de información descrita por la literatura [24], [25].

- La incertidumbre se ha cuantificado con tres herramientas complementarias: intervalos de confianza al 95 % por aproximación de la t de Student y por *bootstrap* sobre los AUC por *fold* (tabla 3), recalibración del umbral mediante el índice de Youden [29] (tabla 5) y contrastes de hipótesis pareados entre arquitecturas (Wilcoxon, DeLong y McNemar, tabla 4).

- Se ha diseñado y validado una estrategia de fusión a nivel paciente con tres reglas (media simple, media ponderada por AUC y voto mayoritario) reportada en la tabla 6. La media simple resultó la mejor opción por su equilibrio entre sensibilidad y especificidad, y la fusión con una única arquitectura superó a la mezcla de campeones por músculo, una conclusión metodológica relevante.

- Se han generado mapas de explicabilidad (Grad-CAM, Grad-CAM guiado, mapas de saliencia y análisis por oclusión) sobre el sistema final (sección 4.7), que confirman que las activaciones del modelo se concentran sobre el parénquima muscular y no sobre los bordes de la región de interés.
- La comparación con la línea base clínica de Martínez-Payá [6] se ha realizado a nivel músculo mediante el criterio del intervalo de confianza (tabla 7), arrojando un veredicto de equivalencia en los cuatro músculos, y se ha discutido cualitativamente la contribución diferencial de la fusión multi-músculo, que el trabajo base no contempla.

5.2. Conclusiones principales

Del trabajo se desprenden cuatro conclusiones de fondo.

En primer lugar, a nivel de músculo individual, **las arquitecturas convolucionales modernas son intercambiables entre sí y estadísticamente equivalentes a la línea base de análisis textural clásico**. Ninguna comparación pareada entre la mejor y la segunda mejor arquitectura alcanza significación en bíceps, antebrazo, cuádriceps ni tibial, y todos los AUC publicados por Martínez-Payá caen dentro de los intervalos de confianza al 95 % de los modelos de este trabajo. Lejos de invalidar el enfoque, este resultado lo enmarca con precisión: el aprendizaje profundo no aporta una mejora marginal por músculo, sino que sustituye un descriptor diseñado a mano por uno aprendido y libera la metodología para incorporar etapas posteriores (en este caso, la fusión) que el análisis textural clásico no soporta de forma natural.

En segundo lugar, **la integración multi-músculo a nivel paciente es la aportación fundamental del trabajo**. La fusión por media simple de las probabilidades de los cuatro músculos eleva el rendimiento desde un rango del 86–99 % por músculo (con varianza apreciable) a un AUC del 98,67 % a nivel paciente, con solo dos errores sobre 52 sujetos (ambos falsos positivos; ninguna ELA sin detectar). Este salto no proviene de mejorar el clasificador por imagen, sino de utilizar de forma estadísticamente honesta toda la evidencia disponible para cada paciente: una observación coherente con la naturaleza clínica del problema, en el que el especialista nunca emite un diagnóstico a partir de un solo músculo.

En tercer lugar, **el diseño con una única arquitectura entrenada de forma independiente por músculo es preferible, por simplicidad operativa, a un *ensemble* heterogéneo de los mejores modelos por músculo**. La fusión de los campeones alcanza un AUC del 99,11 %, estadísticamente indistinguible del de las arquitecturas únicas (cuyos AUC fusionados oscilan entre 97,9 y 99,3 % con intervalos de confianza solapados). Mezclar arquitecturas no aporta, por tanto, una ventaja apreciable de rendimiento y sí añade complejidad. Esta observación tiene implicaciones prácticas: el sistema final es operativamente más simple (un único tipo de red, un único procedimiento de entrenamiento) y, por tanto,

más mantenible.

En cuarto lugar, **la transparencia metodológica es indistinguible del rendimiento medido**. La validación cruzada agrupada por paciente, la recalibración del umbral mediante el índice de Youden, los intervalos de confianza por *bootstrap* y los contrastes pareados han permitido reportar resultados que son simultáneamente altos y honestos. Esta exigencia es la salvaguarda frente al fenómeno, recurrentemente documentado en la literatura [25], de modelos con métricas espectaculares sobre el conjunto de validación que no generalizan más allá.

5.3. Limitaciones del trabajo

Una evaluación responsable de los resultados anteriores debe ponerlos en contexto frente a las limitaciones del trabajo, que son las propias de un estudio piloto con un conjunto clínico reducido.

La limitación más relevante es el **tamaño y la procedencia única del conjunto de datos**. Los 52 pacientes proceden de un mismo centro y han sido adquiridos con un protocolo y un equipo homogéneos. Aunque la validación cruzada agrupada por sujeto previene la fuga de información dentro de este conjunto, no aporta evidencia sobre la generalización a otras instituciones, otros ecógrafos u otras poblaciones. Un AUC del 98,67 % sobre 52 pacientes es una estimación con un intervalo de confianza ancho (límite inferior del IC95 % *bootstrap* en 95,38 %) y debe interpretarse como tal.

La segunda limitación es la **ausencia de validación externa con una cohorte independiente**. El conjunto disponible se ha agotado en la propia validación cruzada, sin reservar una porción para evaluación final ciega; esa decisión de diseño se justifica por el tamaño muestral, pero implica que el sistema no ha sido contrastado nunca contra datos que no hayan participado en la elección de hiperparámetros y arquitectura. Conviene precisar, además, dos aspectos en los que la estimación por validación cruzada es ligeramente optimista. Por un lado, el *epoch* de cada modelo se selecciona por su AUC sobre el propio *fold* de validación; por otro, el umbral de Youden se calcula sobre las mismas predicciones *out-of-fold* sobre las que después se reportan la sensibilidad y la especificidad. El AUC, al ser independiente del umbral, no se ve afectado por este segundo punto, pero las métricas de operación (sensibilidad, especificidad y exactitud) deben interpretarse como una cota superior del rendimiento esperable sobre datos verdaderamente nuevos. La separación estricta entre la fase de selección y la de evaluación (ya sea mediante validación cruzada anidada o fijando el umbral únicamente sobre los datos de entrenamiento) se materializaría de forma natural en la validación externa planteada como trabajo futuro.

La tercera limitación es la **ausencia de contraste con un radiólogo experto**. Los mapas de explicabilidad demuestran que el modelo se apoya en regiones plausibles, pero esta es una verificación cualitativa: no se ha llevado a cabo una concordancia paciente a paciente

entre las decisiones del sistema y el juicio independiente de un especialista, contraste que sería necesario antes de cualquier integración clínica.

Finalmente, el sistema produce una probabilidad y una decisión binaria, pero **no proporciona una estimación de incertidumbre por paciente individual**. Para un uso clínico responsable sería deseable acompañar cada predicción de un intervalo de confianza propio del paciente, que distinga las decisiones rotundas de las situadas cerca del umbral.

5.4. Líneas de trabajo futuro

Las limitaciones anteriores sugieren las líneas naturales de continuación del proyecto.

Validación externa con una cohorte independiente. La primera prioridad metodológica es contrastar el sistema con un conjunto de pacientes adquiridos en otra institución y con otro equipo de ecografía. Durante el desarrollo de este trabajo se ha tenido conocimiento de una cohorte adicional, adquirida en otra institución y compatible con la metodología propuesta, cuyos datos no se han llegado a incorporar; aplicar el modelo ya entrenado a esa cohorte (sin reentrenar) constituiría la prueba de generalización más exigente y se plantea como el siguiente paso inmediato del proyecto.

Calibración y cuantificación de la incertidumbre. Las probabilidades emitidas por el modelo se han recalibrado mediante el índice de Youden a nivel de umbral, pero no se han calibrado a nivel de distribución; técnicas como la regresión isotónica o el escalado de Platt producirían probabilidades cuya magnitud reflejara la confianza real del modelo, requisito necesario para integrarlo con otras fuentes de información clínica. De forma complementaria, métodos como *deep ensembles* o *Monte Carlo dropout* permitirían acompañar cada decisión de una estimación de incertidumbre por paciente, distinguir los casos rotundos de los marginales y, en último término, recomendar la derivación de los casos dudosos a una segunda lectura humana.

Extensión a monitorización longitudinal. Trabajos previos del grupo de Martínez-Payá han documentado que los descriptores texturales reflejan la progresión de la enfermedad a lo largo del tiempo [7]. Adaptar la arquitectura propuesta a una entrada longitudinal (varias adquisiciones por paciente a lo largo de meses) permitiría pasar del diagnóstico al seguimiento, una aplicación clínica de gran valor en una enfermedad como la ELA.

Integración clínica. Más allá del modelo, la incorporación efectiva del sistema a un flujo asistencial requiere un protocolo de operación definido (umbrales aprobados, criterios de decisión asistida y no automática), una interfaz integrada con el ecógrafo o el sistema PACS y, sobre todo, un estudio de concordancia con radiólogos sobre una cohorte suficientemente amplia.

5.5. Reflexión final

El trabajo demuestra que es posible aplicar aprendizaje profundo a la detección de ELA a partir de ecografía muscular alcanzando un rendimiento clínicamente operativo y, simultáneamente, sostenido por un protocolo de validación riguroso. La aportación no reside únicamente en los números (que son altos, pero acotados por las limitaciones de un estudio piloto), sino en la combinación de tres elementos: la sustitución del análisis textural manual por descriptores aprendidos, la integración multi-músculo a nivel paciente y el control explícito de la fuga de información por sujeto. Cualquiera de los tres sin los otros dos produciría un sistema más débil. Tomados conjuntamente, definen una metodología reproducible y extensible que constituye el legado técnico de este Trabajo Fin de Grado y la base sobre la que se asientan las líneas de trabajo futuro identificadas.

Capítulo 6 Bibliografía

[1] R. H. Brown y A. Al-Chalabi, «Amyotrophic Lateral Sclerosis», *New England Journal of Medicine*, vol. 377, n.º 2, págs. 162-172, 2017.

[2] O. Hardiman et al., «Amyotrophic lateral sclerosis», *Nature Reviews Disease Primers*, vol. 3, n.º 17071, 2017.

[3] G. Logroscino et al., «Incidence of amyotrophic lateral sclerosis in Europe», *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 81, n.º 4, págs. 385-390, 2010.

[4] B. R. Brooks, R. G. Miller, M. Swash y T. L. Munsat, «El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis», *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, n.º 5, págs. 293-299, 2000.

[5] S. Pillen, I. M. P. Arts y M. J. Zwarts, «Muscle ultrasound in neuromuscular disorders», *Muscle & Nerve*, vol. 37, n.º 6, págs. 679-693, 2008.

[6] J. J. Martínez-Payá, J. Ríos-Díaz, M. E. del Baño-Aledo, J. I. Tembl-Ferrairó, J. F. Vázquez-Costa y F. Medina-Mirapeix, «Quantitative muscle ultrasonography using textural analysis in amyotrophic lateral sclerosis», *Ultrasonic Imaging*, vol. 39, n.º 6, págs. 357-368, 2017.

[7] J. J. Martínez-Payá, J. Ríos-Díaz, F. Medina-Mirapeix, J. F. Vázquez-Costa y M. E. del Baño-Aledo, «Monitoring progression of amyotrophic lateral sclerosis using ultrasound morpho-textural muscle biomarkers: a pilot study», *Ultrasound in Medicine & Biology*, vol. 44, n.º 1, págs. 102-109, 2018.

[8] G. Litjens et al., «A survey on deep learning in medical image analysis», *Medical Image Analysis*, vol. 42, págs. 60-88, 2017.

[9] I. M. P. Arts, S. Pillen, H. J. Schelhaas, S. Overeem y M. J. Zwarts, «Normal values for quantitative muscle ultrasonography in adults», *Muscle & Nerve*, vol. 41, n.º 1, págs. 32-41, 2010.

[10] A. Krizhevsky, I. Sutskever y G. E. Hinton, «ImageNet classification with deep convolutional neural networks», en *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012.

[11] D. Shen, G. Wu y H.-I. Suk, «Deep learning in medical image analysis», *Annual Review of Biomedical Engineering*, vol. 19, págs. 221-248, 2017.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li y L. Fei-Fei, «ImageNet: A large-scale hierarchical image database», en *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, págs. 248-255.

[13] J. Yosinski, J. Clune, Y. Bengio y H. Lipson, «How transferable are features in deep neural networks?», en *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[14] N. Tajbakhsh et al., «Convolutional neural networks for medical image analysis: full training or fine tuning?», *IEEE Transactions on Medical Imaging*, vol. 35, n.º 5, págs. 1299-1312, 2016.

[15] K. He, X. Zhang, S. Ren y J. Sun, «Deep residual learning for image recognition», en *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, págs. 770-778.

[16] G. Huang, Z. Liu, L. van der Maaten y K. Q. Weinberger, «Densely connected convolutional networks», en *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, págs. 4700-4708.

[17] M. Tan y Q. V. Le, «EfficientNet: rethinking model scaling for convolutional neural networks», en *Proc. Int. Conf. Machine Learning (ICML)*, 2019, págs. 6105-6114.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell y S. Xie, «A ConvNet for the 2020s», en *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, págs. 11 976-11 986.

[19] J. T. Springenberg, A. Dosovitskiy, T. Brox y M. Riedmiller, «Striving for simplicity: the all convolutional net», en *Proc. Int. Conf. Learning Representations (ICLR) Workshop*, 2015.

[20] K. Simonyan, A. Vedaldi y A. Zisserman, «Deep inside convolutional networks: visualising image classification models and saliency maps», en *Proc. Int. Conf. Learning Representations (ICLR) Workshop*, 2014.

[21] M. D. Zeiler y R. Fergus, «Visualizing and understanding convolutional networks», en *Proc. European Conf. Computer Vision (ECCV)*, 2014, págs. 818-833.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh y D. Batra, «Grad-CAM: visual explanations from deep networks via gradient-based localization», en *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, págs. 618-626.

[23] R. Kohavi, «A study of cross-validation and bootstrap for accuracy estimation and model selection», en *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, vol. 14, 1995, págs. 1137-1145.

[24] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr y K. P. Kording, «The need to approximate the use-case in clinical machine learning», *GigaScience*, vol. 6, n.º 5, págs. 1-9, 2017.

[25] M. Roberts et al., «Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans», *Nature Machine Intelligence*, vol. 3, n.º 3, págs. 199-217, 2021.

[26] J. A. Hanley y B. J. McNeil, «The meaning and use of the area under a receiver operating characteristic (ROC) curve», *Radiology*, vol. 143, n.º 1, págs. 29-36, 1982.

[27] E. R. DeLong, D. M. DeLong y D. L. Clarke-Pearson, «Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach», *Biometrics*, vol. 44, n.º 3, págs. 837-845, 1988.

[28] F. Wilcoxon, «Individual comparisons by ranking methods», *Biometrics Bulletin*, vol. 1, n.º 6, págs. 80-83, 1945.

[29] W. J. Youden, «Index for rating diagnostic tests», *Cancer*, vol. 3, n.º 1, págs. 32-35, 1950.

[30] B. Efron, «Bootstrap methods: another look at the jackknife», *The Annals of Statistics*, vol. 7, n.º 1, págs. 1-26, 1979.

[31] B. Efron y R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: Chapman & Hall/CRC, 1993.

[32] Q. McNemar, «Note on the sampling error of the difference between correlated proportions or percentages», *Psychometrika*, vol. 12, n.º 2, págs. 153-157, 1947.

[33] A. Paszke et al., «PyTorch: an imperative style, high-performance deep learning library», en *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, págs. 8024-8035.

[34] S. Marcel e Y. Rodriguez, «Torchvision: the machine-vision package of Torch», en *Proc. ACM Int. Conf. Multimedia*, 2010, págs. 1485-1488.

[35] F. Pedregosa et al., «Scikit-learn: machine learning in Python», *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[36] P. Virtanen, R. Gommers, T. E. Oliphant et al., «SciPy 1.0: fundamental algorithms for scientific computing in Python», *Nature Methods*, vol. 17, n.º 3, págs. 261-272, 2020.

[37] C. R. Harris, K. J. Millman, S. J. van der Walt et al., «Array programming with NumPy», *Nature*, vol. 585, n.º 7825, págs. 357-362, 2020.

[38] J. D. Hunter, «Matplotlib: a 2D graphics environment», *Computing in Science & Engineering*, vol. 9, n.º 3, págs. 90-95, 2007.

Capítulo A Material complementario

Este anexo recoge información complementaria al cuerpo principal de la memoria: la configuración exacta del entrenamiento, la tabla completa de las comparaciones estadísticas pareadas entre arquitecturas y los mapas de explicabilidad generados sobre el sistema final.

A.1. Configuración completa del entrenamiento

La tabla 8 reúne todos los hiperparámetros, semillas y rutas relevantes del experimento principal, tal y como aparecen en `src/config.py` y en los módulos de entrenamiento (`train_kfold.py`) y fusión (`patient_level_fusion.py`). Su propósito es permitir la reproducción exacta del experimento.

Tabla 8: Hiperparámetros y configuración del experimento principal del trabajo.

Parámetro	**Valor**
Semilla de aleatoriedad	42 (propagada a Python, NumPy, PyTorch y scikit-learn)
Tamaño de imagen	224×224 píxeles, 3 canales (RGB)
Normalización	Media [0,485, 0,456, 0,406], std [0,229, 0,224, 0,225] (ImageNet)
Aumento entren.	Flip h. $p = 0{,}5$, flip v. $p = 0{,}2$, rotación $\pm 10°$, color jitter $\pm 0{,}2$, traslación ± 5 %
Aumento valid.	Solo redimensionar y normalizar
Particionado	`StratifiedGroupKFold`, $n = 5$, `shuffle=True`
Tamaño de lote	16
Función de pérdida	`CrossEntropyLoss`
Optimizador	Adam, $\eta_0 = 10^{-4}$
Scheduler	`ReduceLROnPlateau` (mode=max, factor=0,5, patience=5) sobre AUC val
Número máx. de épocas	50
Selección de *checkpoint*	Mejor AUC en validación por *fold*
Bootstrap (fusión)	$N = 1000$ réplicas sobre los 52 pacientes
Umbral de decisión	Recalibrado por *fold* con índice de Youden $J = \text{Sens} + \text{Spec} - 1$
Dispositivo	Apple Silicon M4 vía MPS (con *fallback* a CUDA o CPU)
Versiones (mínimas)	`torch`$\geq$2.3, `torchvision`$\geq$0.18, `scikit-learn`$\geq$1.4, `scipy`$\geq$1.12, `statsmodels`$\geq$0.14

Las cinco arquitecturas comparadas son las siguientes, todas inicializadas con los pesos preentrenados sobre ImageNet de `torchvision`: ResNet-18 ($\sim$11 M parámetros), ResNet-50 ($\sim$26 M), DenseNet-121 ($\sim$8 M), EfficientNet-B0 ($\sim$5 M) y ConvNeXt-Tiny ($\sim$28 M). En todas ellas se reemplaza la última capa lineal por una nueva con dos salidas. Se descartan deliberadamente VGG-16 (por su número excesivo de parámetros, $\sim$138 M, propenso al sobreajuste sobre el conjunto disponible) y MobileNet-V3 Small (por resultar redundante con EfficientNet-B0 en el rol de arquitectura moderna y eficiente).

A.2. Comparaciones estadísticas pareadas completas

La tabla 9 contiene las 40 comparaciones pareadas entre arquitecturas dentro de cada músculo (10 pares por músculo $\times$ 4 músculos). Para cada par se reportan los AUC medios, el p-valor de la prueba de Wilcoxon sobre los AUC por *fold*, el p-valor del test de DeLong sobre las curvas ROC agregadas y el p-valor del test de McNemar sobre los aciertos binarios. Los p-valores marcados con $<$ no alcanzan la significación al nivel $\alpha = 0{,}05$ habitual; los marcados con asterisco sí.

Tabla 9: Comparaciones estadísticas pareadas entre las cinco arquitecturas, por músculo. AUC en %. "–" indica p-valor no calculable por igualdad exacta de los AUC por *fold*.

Músculo	Modelo A	Modelo B	AUC_A	AUC_B	p Wilc.	p DeLong	p McNem.
Antebrazo	ResNet-18	ResNet-50	91,38	89,76	0,250	0,345	0,701
Antebrazo	DenseNet-121	EfficientNet-B0	88,73	91,71	0,312	0,519	0,286
Antebrazo	DenseNet-121	ResNet-18	88,73	91,38	0,312	0,144	0,629
Antebrazo	EfficientNet-B0	ResNet-50	91,71	89,76	0,375	0,344	0,362
Antebrazo	ConvNeXt-Tiny	ResNet-50	91,03	89,76	0,688	0,653	0,442
Antebrazo	ConvNeXt-Tiny	EfficientNet-B0	91,03	91,71	0,812	0,149	1,000
Antebrazo	ConvNeXt-Tiny	ResNet-18	91,03	91,38	0,812	0,370	0,774
Antebrazo	EfficientNet-B0	ResNet-18	91,71	91,38	0,812	0,071	0,711
Antebrazo	ConvNeXt-Tiny	DenseNet-121	91,03	88,73	0,875	0,403	0,267
Antebrazo	DenseNet-121	ResNet-50	88,73	89,76	1,000	0,634	1,000
Bíceps	DenseNet-121	EfficientNet-B0	90,30	85,86	0,062	0,130	1,000
Bíceps	DenseNet-121	ResNet-50	90,30	86,35	0,125	0,348	0,286
Bíceps	ConvNeXt-Tiny	DenseNet-121	86,28	90,30	0,250	0,447	0,581
Bíceps	DenseNet-121	ResNet-18	90,30	86,25	0,312	0,407	0,230
Bíceps	ResNet-18	ResNet-50	86,25	86,35	0,812	0,798	1,000
Bíceps	EfficientNet-B0	ResNet-18	85,86	86,25	0,875	0,264	0,189
Bíceps	EfficientNet-B0	ResNet-50	85,86	86,35	0,875	0,434	0,345
Bíceps	ConvNeXt-Tiny	EfficientNet-B0	86,28	85,86	1,000	0,325	0,648
Bíceps	ConvNeXt-Tiny	ResNet-18	86,28	86,25	1,000	0,895	0,523
Bíceps	ConvNeXt-Tiny	ResNet-50	86,28	86,35	1,000	0,712	0,648
Cuádriceps	ConvNeXt-Tiny	EfficientNet-B0	99,33	99,13	1,000	0,011*	0,089
Cuádriceps	DenseNet-121	EfficientNet-B0	99,33	99,13	1,000	0,767	0,839
Cuádriceps	EfficientNet-B0	ResNet-18	99,13	99,33	1,000	0,746	0,424
Cuádriceps	EfficientNet-B0	ResNet-50	99,13	99,33	1,000	0,302	0,163
Cuádriceps	ConvNeXt-Tiny	DenseNet-121	99,33	99,33	–	0,022*	0,150
Cuádriceps	ConvNeXt-Tiny	ResNet-18	99,33	99,33	–	0,049*	0,392
Cuádriceps	ConvNeXt-Tiny	ResNet-50	99,33	99,33	–	0,048*	0,839
Cuádriceps	DenseNet-121	ResNet-18	99,33	99,33	–	0,858	0,701
Cuádriceps	DenseNet-121	ResNet-50	99,33	99,33	–	0,413	0,296
Cuádriceps	ResNet-18	ResNet-50	99,33	99,33	–	0,543	0,597
Tibial	ConvNeXt-Tiny	ResNet-18	92,38	93,96	0,250	0,001*	0,008*
Tibial	DenseNet-121	EfficientNet-B0	93,17	90,98	0,250	0,866	0,804
Tibial	ResNet-18	ResNet-50	93,96	92,37	0,312	0,139	0,362
Tibial	EfficientNet-B0	ResNet-18	90,98	93,96	0,375	0,373	0,832
Tibial	EfficientNet-B0	ResNet-50	90,98	92,37	0,438	0,399	0,481
Tibial	ConvNeXt-Tiny	DenseNet-121	92,38	93,17	0,500	0,000*	0,029*
Tibial	ConvNeXt-Tiny	EfficientNet-B0	92,38	90,98	0,625	0,003*	0,011*
Tibial	DenseNet-121	ResNet-50	93,17	92,37	0,625	0,489	0,856
Tibial	DenseNet-121	ResNet-18	93,17	93,96	0,750	0,350	0,481
Tibial	ConvNeXt-Tiny	ResNet-50	92,38	92,37	1,000	0,027*	0,071

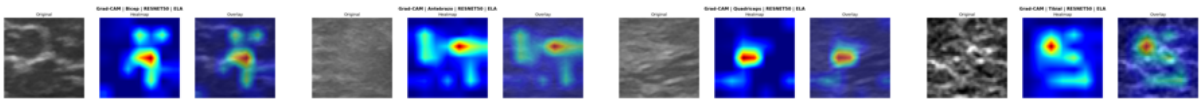
* marca $p < 0{,}05$. "–" en Wilcoxon indica que las series de AUC por *fold* son exactamente iguales, por lo que no hay diferencias rangadas que evaluar.

La lectura global de la tabla confirma lo expuesto en la sección 4.3: en bíceps y antebrazo ninguna comparación alcanza significación; en cuádriceps las diferencias significativas (por DeLong) involucran exclusivamente a la única arquitectura que se separa del techo (99,13 % de EfficientNet-B0 frente al 99,33 % del resto, o ConvNeXt frente a los demás); en tibial,

ConvNeXt-Tiny aparece sistemáticamente en el lado perdedor de las comparaciones significativas, una observación coherente con la lectura general del músculo y con la elección final de ResNet-50 como arquitectura del sistema.

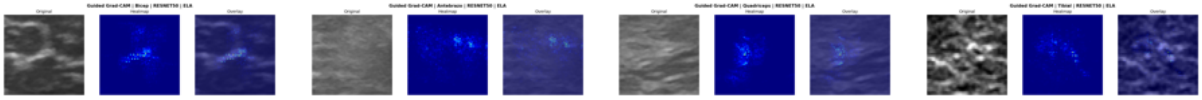
A.3. Mapas de explicabilidad completos

Esta sección recoge los mapas de explicabilidad generados sobre ResNet-50 (el sistema final) para los cuatro músculos, con las cuatro técnicas empleadas en el trabajo: Grad-CAM [1], Grad-CAM guiado [2], mapas de saliencia [3] y análisis por oclusión [4]. Para cada técnica se muestra un ejemplo representativo de un paciente con ELA por cada músculo, tomado del conjunto de validación *out-of-fold*.



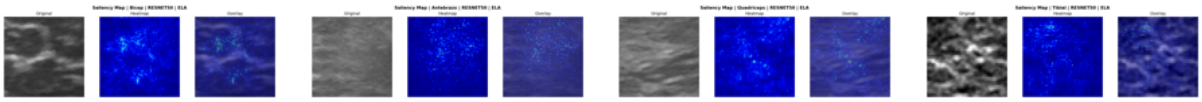
Bíceps Antebrazo Cuádriceps Tibial

Figura 6: Grad-CAM sobre un paciente con ELA en cada uno de los cuatro músculos.



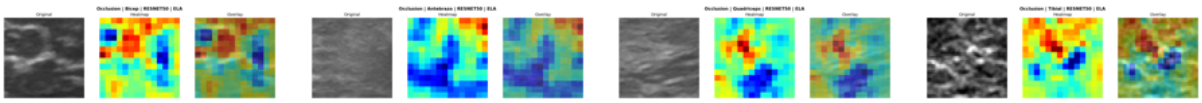
Bíceps Antebrazo Cuádriceps Tibial

Figura 7: Grad-CAM guiado sobre los mismos pacientes con ELA. La técnica combina la localización de Grad-CAM con la retropropagación guiada para refinar la atribución a píxeles concretos.



Bíceps Antebrazo Cuádriceps Tibial

Figura 8: Mapas de saliencia sobre los mismos pacientes. Visualizan el gradiente de la probabilidad de ELA respecto a cada píxel de entrada.



Bíceps Antebrazo Cuádriceps Tibial

Figura 9: Análisis por oclusión sobre los mismos pacientes. Cuantifica la caída de la probabilidad de ELA al ocultar parches de la imagen, identificando las regiones cuya información es más crítica para la decisión del modelo.

Las cuatro técnicas coinciden cualitativamente: las regiones de mayor influencia se concentran sobre el tejido muscular, no sobre los bordes de la región de interés. La convergencia de cuatro métodos de explicabilidad mecánicamente distintos sobre los mismos focos refuerza la interpretación: las decisiones del modelo se apoyan en patrones reproducibles y plausibles desde el punto de vista clínico.