Clasificación de la enfermedad de Alzheimer y de la Demencia Frontotemporal

Un estudio observacional transversal utilizando medidas clínicas y cognitivas de rutina en muestras multicéntricas subrepresentadas.

Tesis presentada en el cumplimiento de los requisitos para obtener el grado de Magíster en Explotación de Datos y Gestión del Conocimiento.

Departamento de Ingeniería
Universidad Austral
Buenos Aires, Argentina
Marzo de 2024

Autor: Lic. Marcelo A. Maito
Co-Directores: Dr. Hernando Santamaría-García, Ing. Martín Volpacchio

Resumen

Las iniciativas globales de salud cerebral exigen mejorar los métodos para diagnosticar la enfermedad de Alzheimer (EA) y la demencia frontotemporal (DFT) en poblaciones subrepresentadas. Sin embargo, los procedimientos de diagnóstico en países de ingresos medianos altos (PIMA) y países de ingresos medianos bajos (PIMB), como los países de América Latina (PAL), enfrentan múltiples desafíos. Estos incluyen, la heterogeneidad en los métodos de diagnóstico, la falta de armonización clínica y el acceso limitado a los biomarcadores de los pacientes que padecen estas enfermedades.

Este estudio observacional transversal tuvo como objetivo identificar la mejor combinación de predictores para discriminar entre EA y DFT utilizando datos tomados en el contexto clínico, incluyendo información demográfica, datos clínicos y cognitivos entre 1792 participantes [904 diagnosticados con EA, 282 diagnosticados con DFT y 606 controles sanos (CN)] recopilados en 11 centros clínicos en cinco PAL (cohorte ReDLat -Multipartner Consortium to expand dementia research in Latin America-).

Se desarrolló una metodología de imputación y armonización de datos, y se aplicó un enfoque computacional que incluyó métodos estadísticos clásicos, técnicas de aprendizaje supervisado y procedimientos de selección de características secuenciales. Los resultados demostraron una clasificación precisa de los pacientes con EA, DFT y CN. Un modelo Random Forest produjo los mejores resultados para diferenciar la EA de los pacientes con DFT con un accuracy de 0.85 y un área bajo la curva ROC de 0.82 sobre datos no vistos en la fase de entrenamiento. Las características principales fueron el año de nacimiento, función ejecutiva (IFS), cognición (ACE-III normalizado siguiendo el método de Matías-Guiu), cognición social (Mini-SEA total) y síntomas neuropsiquiátricos (NPI).

Los resultados demuestran que las técnicas de imputación, armonización y aprendizaje supervisado aplicadas a conjuntos de datos clínicos de archivo podrían mejorar los procedimientos de diagnóstico en regiones con recursos limitados. Estos resultados también sugieren que las medidas cognitivas y conductuales pueden ayudar a diagnosticar la EA y la DFT en PAL. Además, nuestros resultados destacan una oportunidad para la armonización de herramientas clínicas para el diagnóstico de demencia en la región.

Abstract

Global brain health initiatives demand improvement in methods for diagnosing Alzheimer's disease (AD) and frontotemporal dementia (FTD) in underrepresented populations. However, diagnostic procedures in upper-middle-income countries (UMICs) and lower-middle-income countries (LMICs), such as countries in Latin America (LAC), face multiple challenges. These include heterogeneity in diagnostic methods, lack of clinical harmonization, and limited access to biomarkers for patients with these diseases.

This cross-sectional observational study aimed to identify the best combination of predictors to discriminate between AD and FTD using routine medical data including demographic, clinical, and cognitive data among 1792 participants [904 diagnosed with AD, 282 diagnosed with FTD, and 606 healthy controls (HC)] collected at 11 clinical centers in five LAC countries (ReDLat - Multipartner Consortium to expand dementia research in Latin America- cohort).

A data imputation and harmonization methodology was developed, applying a computational approach that included classical statistical methods, supervised learning techniques, and sequential feature selection procedures. The results demonstrated accurate classification of patients with AD, FTD, and HC. A Random Forest model produced the best results for distinguishing AD from FTD patients with an accuracy of 0.85 and an area under the ROC curve of 0.82 on unseen data. Key features included year of birth, executive function (IFS), cognition (ACE-III normalized following the Matías-Guiu method), social cognition (Mini-SEA) and neuropsychiatric symptoms (NPI).

The findings show that imputation, harmonization, and supervised learning techniques applied to archived clinical datasets could enhance diagnostic procedures in resource-limited regions. These results also suggest that cognitive and behavioral measures may aid in diagnosing AD and FTD in LAC. Furthermore, our results highlight an opportunity for harmonizing clinical tools for dementia diagnosis in the region.

ÍNDICE GENERAL

Índice de Tablas

Índice de figuras

Abreviaturas

EA - Enfermedad de Alzheimer

DFT - Demencia frontotemporal

PIA - Países de ingresos altos

PIMA - Países de ingresos medios-altos

PIMB - Países de ingresos medios-bajos

PAL - Países de América Latina

CN - Controles sanos

ReDLat - Multipartner Consortium to expand dementia research in Latin America

IFS - Evaluación Frontal INECO

ACE-III - Examen Cognitivo de Addenbrooke

NPI - Inventario Neuropsiquiátrico

MMSE - Examen de Estado Mental Mini-Mental

MoCA - Evaluación Cognitiva de Montreal

FAQ - Cuestionario de Actividad Funcional de Pfeffer

Mini-SEA - Evaluación de Cognición Social y Emocional (corta)

ToM - Teoría de la Mente

MICE - Multiple Imputation by Chained Equations

MAR - Missing at random

MCAR - Missing completely at random

MNAR - Missing not at random

RMSE - Raíz del error cuadrático medio

MSE - Error cuadrático medio

LR - Regresión logística

RF - Random Forest

SVM - Support Vector Machines

XGBoost - Extreme Gradient Boosting

1. Introducción

1.1 Definición del problema

El diagnóstico preciso de la enfermedad de Alzheimer (EA) y de la demencia frontotemporal (DFT) continúa siendo un desafío mundial para la salud cerebral[1–4]. Para el año 2050, la prevalencia de la demencia aumentará alrededor de un 75% en los países de ingresos altos (PIA) y alrededor de un 200% en los países de ingresos medianos-bajos (PIMB)[5,6]. Además, el escenario actual de prevalencia de la demencia muestra que dos tercios de las personas con demencia viven en PIMB[6,7]. La mejora de la precisión diagnóstica en una región con un crecimiento dramático y progresivo del número de casos de demencia es fundamental para brindar intervenciones personalizadas[8].

En los países de América Latina (PAL) existen numerosos desafíos que afectan la precisión diagnóstica de la EA y la DFT, incluidos (a) la diversidad y heterogeneidad de los instrumentos para evaluar el estado clínico y cognitivo; (b) la ausencia de procedimientos estandarizados para incorporar factores socio-demográficos en el diagnóstico; (c) una escasez de procedimientos de diagnóstico armonizados entre países; (d) baja concientización sobre la demencia entre los médicos generales y (e) formación poco desarrollada en el diagnóstico de la demencia[1,8,9]. Además, aunque los métodos que han sido efectivos en países de altos ingresos (imágenes PET de amiloide y tau, y evaluación de biomarcadores basados en fluidos) podrían ser soluciones efectivas en el futuro; actualmente no están ampliamente disponibles en los PAL por razones tanto financieras como logísticas, limitando su utilidad para informar decisiones clínicas[1,10] . Por lo tanto, desarrollar métodos para armonizar las evaluaciones clínicas, cognitivas y

funcionales es el enfoque escalable más prometedor disponible para diagnosticar la EA y la DFT en los PAL[9]. Para ello, utilizamos conjuntos de datos clínicos de archivo no armonizados de poblaciones heterogéneas de once centros de cinco PAL.

Aunque las medidas clínicas y cognitivas se usan ampliamente en PAL, existen limitaciones que restringen su uso para la caracterización multicéntrica de la demencia[1]. La falta de datos normativos regionales apropiados en la evaluación clínica dificulta las comparaciones directas entre países[8]. Las diferencias en los entornos socioculturales, la mezcla genética y la experiencia clínica en la región también aumentan la heterogeneidad clínica. La ausencia de procedimientos pre-armonizados entre centros y países impone barreras adicionales[7]. No tener en cuenta estas limitaciones puede dar lugar a resultados que no se pueden generalizar a otros entornos, la extrapolación injustificada de patrones locales a tendencias regionales y la incapacidad de determinar qué puntos de datos demuestran ser los impulsores más sólidos de los hallazgos. Por estas razones, muchos estudios en los PAL se basan en pequeños conjuntos de datos de cohortes de regiones geográficas restringidas que pueden conducir a resultados que no son generalizables. Así, se requieren nuevos procedimientos de armonización de datos de rutina clínica para permitir comparaciones multicéntricas entre los PAL.

1.2 Objetivos

Objetivo principal: Implementar y evaluar un modelo de clasificación binario para determinar si los datos clínicos de archivo resultan suficientes para obtener una buena clasificación entre Alzheimer y demencia frontotemporal.

Objetivo específico #1: Desarrollar una metodología de mitigación de la variabilidad y la heterogeneidad de los datos multicéntricos no armonizados para desarrollar un abordaje escalable a nivel regional.

Objetivo específico #2: Evaluar las diferencias de rendimiento que se obtienen mediante diferentes técnicas de tratamiento de valores perdidos para determinar cuál es la mejor estrategia de imputación para un conjunto de datos con éstas características.

Objetivo específico #3: Determinar cuáles son los mejores predictores para cada una de las clases para informar los procesos de armonización de protocolos de investigación actuales y futuros.

1.3 Contribución de este trabajo

Aquí presentamos un abordaje computacional para la clasificación de EA, DFT y controles sanos (CN) en muestras multicéntricas, heterogéneas de PAL utilizando evaluaciones clínicas, demográficas y cognitivas no armonizadas. Utilizamos conjuntos de datos clínicos archivados recopilados en poblaciones heterogéneas y protocolos de once centros en cinco PAL. Cada centro que contribuyó con datos es un miembro del *Multipartner consortium to expand dementia research in Latin America* (ReDLat)[9].

Nuestro enfoque combinó métodos estadísticos clásicos (imputación por medias, modelos de regresión logística y bootstrapping), y procedimientos de aprendizaje automático[11] (imputación mediante algoritmos, escalado, clasificación y selección secuencial de características) para identificar los mejores factores para discriminar entre EA, DFT y CN. Estudios previos que analizaron el diagnóstico de demencia utilizando bases de datos con alta multidimensionalidad han revelado mejores rendimientos con métodos de aprendizaje automático que con modelos estadísticos clásicos[12]. Por lo tanto, anticipamos una mayor precisión con modelos de aprendizaje automático que con modelos de regresión logística. Estudios en PIA han revelado una alta capacidad de las pruebas cognitivas para distinguir a los CN de los pacientes con demencia y discriminar entre sus diferentes tipos[13,14]. En PAL, encontramos más heterogeneidad en las baterías de evaluación cognitiva aplicadas[13,15]. Sin embargo, en la mayoría de los países, se utilizan baterías básicas para rastrear la cognición, los síntomas neuropsiquiátricos y el nivel de funcionalidad en las poblaciones envejecidas[1,8]. En este contexto, esperábamos lograr una discriminación elevada entre los subtipos de demencia y CN a pesar de la heterogeneidad de los datos. Además, esperábamos que las medidas

cognitivas generales ayudaran a discriminar a los pacientes con EA y DFT de los CN. Finalmente, considerando que los pacientes que viven con DFT se caracterizan por trastornos conductuales y ejecutivos frontales[16,17], anticipamos que los instrumentos que rastrean el funcionamiento frontal y los cambios conductuales (incluyendo IFS, NPI y Mini-SEA) serían los mejores para discriminar FTD de EA.

2. Antecedentes

Teniendo en cuenta un trabajo publicado con anterioridad[18], no existen otros estudios que hayan desarrollado un abordaje para distinguir entre pacientes con EA y DFT utilizando una muestra recolectada de archivos clínicos multicéntricos y heterogéneos, procedentes de entornos subrepresentados y de bajos recursos. El desarrollo de modelos capaces de realizar tareas de clasificación bajo estas condiciones requiere un esfuerzo adicional tendiente a la armonización de los datos, la minimización de la variabilidad y de las diferencias regionales y el tratamiento de los valores perdidos mediante diferentes estrategias. En este sentido, existen algunos trabajos que han obtenido buenos resultados lidiando con algunos de estos factores, pero no todos a la vez.

Específicamente, en la investigación de Moguilner et al. (2022) se desarrollaron modelos de clasificación XGBoost (ROC AUC > .9) sobre una muestra subrepresentada y emparejada de tres países de Latinoamérica (n= 282) de participantes sanos, con EA y DFT. Obtuvieron resultados robustos a pesar de la heterogeneidad multimodal, la variabilidad sociodemográfica y los valores perdidos de la muestra. Para el desarrollo de los modelos, utilizaron datos multimodales (neuropsicológicos, demográficos, y

marcadores de resonancias magnéticas, conectividad y electroencefalogramas) de participantes con EA, DFT y CN.

Por otra parte, en el estudio de Donnelly-Kehoe et al. (2019) se utilizó un abordaje similar con muy buenos resultados, para desarrollar modelos de clasificación para pacientes con DFT y CN sobre datos de resonancia magnética de 2 países latinoamericanos y Australia.

En el caso de Gupta y Kahali (2020) se entrenaron modelos de clasificación multiclase para pacientes con EA, deterioro cognitivo leve y CN sobre dos muestras desbalanceadas. La única medida tomada contra la variabilidad de los datos proveniente de múltiples centros fue el emparejamiento de los casos de una de las muestras. Sus modelos clasificaron correctamente las últimas dos clases, pero tuvieron problemas para clasificar correctamente EA.

Por otra parte, también existen trabajos con muestras homogéneas de distintos tamaños. En el caso de Kwak et al. (2022) se testearon modelos de machine-learning para predecir el deterioro funcional en pacientes con EA sobre una muestra homogénea de gran tamaño (n=2642) y con variables de pruebas funcionales y neuropsicológicas. Los resultados muestran que a medida que crece la muestra los modelos lineales pierden capacidad predictiva mientras que modelos más complejos y no lineales aumentan su capacidad predictiva, lo que se explica, según los autores, por la posibilidad de explotar los patrones complejos en los datos.

Por otra parte, en el estudio de Garcia-Gutierrez et al. (2021) se obtuvo una alta precisión en la clasificación de una muestra homogénea de un centro (n=329) de participantes con EA y DFT usando una combinación de métodos de machine-learning (imputación de valores perdidos, algoritmos evolutivos, selección de características, clasificadores multiclase y binarios) con datos de pruebas neuropsicológicas exclusivamente.

Finalmente, en el trabajo de Grassi et al. (2018) se entrenó un clasificador utilizando Support Vector Machines para predecir la aparición de la EA en participantes con deterioro cognitivo leve y pre deterioro cognitivo leve con una alta precisión sobre una muestra homogénea (n= 184), utilizando variables recolectadas mediante métodos no invasivos (pruebas neuropsicológicas y cognitivas, datos clínicos y demográficos).

En contraste con estos estudios, este trabajo analizará la productividad de los datos clínicos de archivo no armonizados, estructurados en una muestra relativamente grande (n=1792), heterogénea (once centros de cinco PAL) y con protocolos clínicos diferentes, en función de su rendimiento para la clasificación de EA, DFT y CN.

3. Materiales y Métodos

3.1 Tipo de estudio y muestra

Para este estudio observacional analítico de corte transversal se utilizó una muestra de conveniencia que representaba el total de casos de EA, DFT y CN reclutados por los centros incluidos en este estudio. Los casos fueron reclutados entre enero de 2015 y octubre de 2021 en los diferentes centros del consorcio ReDLat.

Tabla 1. Distribución de participantes por clasificación y país

País	EA	CN	DFT	TOTAL
Argentina	257	192	53	502
Chile	197	145	59	401
Colombia	320	232	155	707
México	30	21	7	58
Perú	100	16	8	124
Total	904	606	282	1792

EA: Enfermedad de Alzheimer, CN: Control sano, DFT: Demencia frontotemporal

3.2 Participantes

Los participantes fueron reclutados en once centros latinoamericanos que participan en ReDLat[9,19] - Argentina (3 centros), Colombia (3 centros), Chile (2 centros), México (2 centros) y Perú (1 centro)-. La muestra total (n = 1792) incluyó 904 participantes con EA, 282 con DFT y 606 CN (la información demográfica completa se proporciona en la Tabla 2). Todos los participantes dieron su consentimiento informado. Las Juntas de Revisión Institucional y el Comité Ejecutivo del consorcio ReDLat revisaron y aprobaron el presente estudio.

3.3 Evaluación clínica en todos los centros

Los diagnósticos clínicos se establecieron siguiendo los procedimientos estándar empleados en cada centro de investigación. En Colombia, los centros diagnostican a los pacientes a través de una conferencia de consenso realizada por un equipo multidisciplinario que incluye psiquiatras, neurólogos, neuropsicólogos y geriatras. En Chile, Perú, México y Argentina, los pacientes fueron diagnosticados por neurólogos conductuales y geriatras experimentados con aportes de los neuropsicólogos evaluadores. Cada centro aplicó un conjunto heterogéneo de medidas neuropsicológicas

para evaluar el cribado cognitivo, el funcionamiento frontal, la cognición social, los síntomas neuropsiquiátricos y el estado funcional. La batería neuropsicológica de cada país se seleccionó de acuerdo con sus respectivos procedimientos clínicos estándar y la disponibilidad de instrumentos. Las medidas específicas utilizadas en cada centro se describen en la Tabla 3. Como era de esperar, se detectó un elevado número de valores perdidos de forma no aleatoria entre las medidas neuropsicológicas, que se abordaron mediante el uso de tablas de conversión siguiendo los procedimientos recomendados y publicados previamente[20–23]. Independientemente de la batería específica empleada en cada centro, todos siguieron los criterios internacionales NINCDS-ADRDA para diagnosticar la EA[24] y los criterios publicados para diagnosticar las variantes conductuales[25] y lingüísticas[26] de la DFT.

3.4. Medidas neuropsicológicas

Utilizando métodos previamente publicados, se armonizaron[20–23] y normalizaron[27] las medidas utilizadas en todos los centros.

3.4.1 Evaluación cognitiva

Cada centro realizó un seguimiento del funcionamiento cognitivo general utilizando uno de tres tipos de escalas que comprenden el Examen del Estado Mental Mini-Mental (MMSE)[28], la Evaluación Cognitiva de Montreal (MoCA)[29] o el Examen Cognitivo de Addenbrooke (ACE III)[30] (Tabla 3). El MMSE es un instrumento clásico para evaluar dominios cognitivos, como la memoria verbal, la memoria de trabajo, el lenguaje y las funciones visoespaciales. Una puntuación por debajo de 24 puntos tiene una sensibilidad superior al 88,3% y una especificidad cercana al 87% para detectar deterioro cognitivo en

pacientes con demencia[26]. El MMSE se utilizó en ocho centros. El MoCA es una herramienta de evaluación cognitiva ampliamente utilizada, compuesta por 19 ítems que evalúan ocho dominios cognitivos, incluyendo habilidades ejecutivas, denominación, memoria, atención, lenguaje, abstracción, memoria diferida y orientación. Tiene un punto de corte de 26, una sensibilidad del 87% y una especificidad del 87%[27]. Este instrumento se utilizó en seis centros. El ACE III es un instrumento de evaluación cognitiva que evalúa las funciones cognitivas de atención, orientación, memoria, lenguaje, percepción visual y habilidades visoespaciales. El ACE III se utilizó en cinco centros.

3.4.2 Función ejecutiva

Cuatro países evaluaron la función ejecutiva utilizando la Evaluación Frontal INECO (IFS)[31]. La IFS abarca la fluidez verbal, el control inhibitorio, la velocidad de procesamiento, la memoria de trabajo y la flexibilidad cognitiva, y ha demostrado ser efectiva en la detección de disfunción ejecutiva en pacientes con demencia[31,32]. La puntuación máxima posible en la IFS es de 30 puntos.

3.4.3 Funcionalidad

Todos los países evaluaron las actividades básicas e instrumentales de la vida diaria utilizando el Cuestionario de Actividad Funcional de Pfeffer (FAQ)[33] o el índice de Barthel[34]. El FAQ evalúa el funcionamiento en actividades instrumentales como escribir cheques, pagar facturas, hacer compras y conducir. Consta de 10 preguntas y es completado por un informante familiarizado con el funcionamiento del paciente. El índice de Barthel evalúa las dificultades en las actividades básicas de la vida diaria, como vestirse, bañarse, arreglarse, usar el baño, la continencia intestinal y vesical y la movilidad.

El FAQ se utilizó en cinco centros, mientras que el índice de Barthel se utilizó en tres centros.

3.4.4 Síntomas neuropsiquiátricos

El Inventario Neuropsiquiátrico (NPI)[35] se utiliza para hacer un seguimiento de los síntomas neuropsiquiátricos, incluyendo delirios, alucinaciones, problemas de conducta y sueño, depresión, ansiedad y cambios en los patrones alimentarios en la demencia. Siete centros emplearon el NPI.

3.4.5 Cognición social

La Evaluación de Cognición Social y Emocional (SEA) en su forma corta (Mini-SEA)[36] se utilizó en cuatro centros. El Mini-SEA se compone de dos segmentos: teoría de la mente (ToM) y reconocimiento de emociones[36]. La ToM se evalúa a través del test de Faux-pas, que utiliza diez viñetas para hacer un seguimiento de la capacidad para detectar la adecuación social. El test de reconocimiento de emociones evalúa la capacidad para identificar emociones básicas utilizando las imágenes de Ekman.

3.4.6 Datos sociodemográficos.

Adicionalmente, se recabó información sociodemográfica disponible en los centros mencionados anteriormente, además de la edad, el sexo y la educación de los participantes. Los datos recabados adicionalmente corresponden a el estado civil, la cantidad de hijos, la cantidad de miembros del hogar, el rango de ingreso del hogar, el estado de ocupación, la nacionalidad y el lugar de residencia. Estas últimas dos variables fueron removidas de la muestra para evitar sesgos debido al desbalance de las

submuestras y de los diagnósticos entre centros y países. Por ejemplo, existen centros que solamente tienen un tipo de participante, centros con 2 tipos de participantes y centros que tienen una porción extremadamente minoritaria para uno de los diagnósticos.

Tabla 2. Información demográfica básica.

	EA	DFT	CN	Estadístico	One-way ANOVA P values
Sexo (%) *	M = 310 (17.3%) F = 594 (33.15%)	M = 141 (7.87%) F = 141 (7.87%)	M = 187 (10.43%) F = 419 (23.38%)	32.123	(P < 0.0001)
Edad +	81.58 (9.95%)	72.33 (9.14%)	73.65 (10.9%)	146.27	EA vs CN (P < 0.001) EA vs DFT (P < 0.001) CN vs DFT (P > 0.05)
Años de educación +	10.29 (4.88%)	12.77 (5.10%)	13.58 (4.7%)	81.133	EA vs CN (P < 0.001) EA vs DFT (P < 0.001) CN vs DFT (P > 0.05)

EA: Enfermedad de Alzheimer, CN: Control sano, DFT: Demencia frontotemporal

* Chi-Square test (Valor crítico); + ANOVA test. Media (desvío estándar)

Tabla 3. Evaluaciones por tipo de test y centro.

Centro	MMSE	MoCA	ACE-III	IFS	Mini-SEA	Barthel	Pfeffer	NPI	NPIC	Muestra(n)
Arg 1	0	3	31	34	0	0	0	34	0	54
Arg 2	248	0	0	0	0	0	0	0	0	249
Arg 3	0	63	101	181	114	0	38	23	22	199
Chi 1	1	111	1	0	0	0	0	0	0	127
Chi 2	271	177	176	173	159	0	155	229	223	274
Col 1	59	29	59	59	29	30	0	0	0	59
Col 2	324	35	0	34	0	343	0	13	11	454
Col 3	187	188	0	99	0	0	94	94	0	194
Mex 1	19	17	0	0	0	12	0	0	0	21
Mex 2	37	0	0	0	0	0	37	37	37	37
Per 1	124	0	124	124	124	0	124	124	0	124
Total	1270	623	492	704	426	385	448	554	293	1792

MMSE: Mini Mental State Examination; MoCA: Montreal Cognitive Assessment; ACE-III: Addenbrooke's Cognitive Examination; IFS: Ineco Frontal Screening; Mini-SEA: Mini-Social Cognition & Emotional Assessment; Barthel: Barthel scale; Pfeffer FAQ: Pfeffer Functional Activity Questionnaire; NPI: Neuropsychiatric Inventory; NPIC: Neuropsychiatric Inventory Caregiver. Arg: Argentina; Chi: Chile; Col: Colombia; Mex: México; Per: Perú.

3.5 Armonización entre países

Dada la heterogeneidad de las evaluaciones clínicas entre países, hubo una cantidad sustancial de datos faltantes. Para armonizar los datos disponibles y aumentar el número de individuos con medidas cognitivas homogéneas, aplicamos los siguientes procedimientos.

Armonización 1: Armonizamos las evaluaciones cognitivas globales breves utilizando tablas de equivalencia[20–23] para MMSE-MoCA y MMSE-ACE-III, obtenidas a partir de la aplicación de la igualación equipercentil, método utilizado para equiparar pruebas estandarizadas y que permite determinar puntuaciones de pruebas comparables de dos medidas diferentes sobre la base de sus correspondientes rangos percentiles. Este procedimiento permite la estimación de las puntuaciones de MoCA y ACE-III utilizando las puntuaciones de MMSE. También utilizamos las tablas de equivalencia para estimar las puntuaciones de MMSE utilizando las puntuaciones de MoCA y ACE III[20–23]. Siguiendo este enfoque, agregamos un total de siete nuevas variables convertidas y armonizadas y disminuimos el número de valores MMSE faltantes en 325. La media, mediana y desviación estándar para las puntuaciones originales de MMSE y las puntuaciones convertidas y armonizadas se proporcionan en la Tabla 4.

Tabla 4. Mediana, media y desvío estándar para puntajes totales de MMSE.

				MMSE conversion method								
	MMSE original			Van Steenoven			Lawton			Matías-Guiu		
	Mediana	Media	Ds	Mediana	Media	Ds	Mediana	Media	Ds	Mediana	Media	Ds
EA	22	20.74	5.45	22	20.78	5.36	22	20.77	5.34	22	20.86	5.4
DFT	23.5	22.24	6.23	23	22.35	6.1	23	22.33	6.08	24	22.47	6.04

MMSE: Mini Mental State Examination; EA: Enfermedad de Alzheimer, DFT: Demencia frontotemporal, Ds: Desvío estándar.

Adicionalmente y siguiendo procedimientos previamente publicados[37], las puntuaciones MMSE y MoCA se transformaron de la escala 0-30 a la escala 0-100 y se promediaron con la puntuación ACE para crear una nueva puntuación cognitiva por participante (*cognition*).

Por otra parte, se armonizaron el índice Barthel y el Pfeffer FAQ en la variable *functionality*. Este proceso incluyó la inversión del índice Barthel, ajustando su orientación para que un puntaje más alto indique una mejor funcionalidad, en concordancia con el Pfeffer FAQ, donde un puntaje más bajo denota una mejor condición. Posteriormente, se escalaron a 0-100 y se promediaron para obtener la nueva variable .

Armonización 2: Se implementó un método de escalado Min-Max[27] utilizando Scikit-learn[38]. Cada registro se transformó de la variable original en una nueva con un rango entre 0 y 1 (Eq1).

Eq1.

$$x_m = (x - x_{min}) / (x_{max} - x_{min})$$

Donde:

x_m es nuestro nuevo valor.

x es el valor original de la celda.

x_{min} es el valor mínimo de la variable x.

x_{max} es el valor máximo de la variable x.

Los métodos de escalado tienen distintas fortalezas y debilidades y, a menudo, se requieren o recomiendan para modelos lineales y SVM[27,38]. Este método de escalado se implementó con las medidas neuropsicológicas, las variables demográficas y las variables

obtenidas de la Armonización 1.

Después de estos procedimientos, el conjunto de datos final incluyó las variables convertidas (Armonización 1) y las variables normalizadas (Armonización 2). Los procedimientos de estandarización mediante puntaje z fueron realizados en un trabajo previo[18] en el cual el peor de los mejores cinco modelos incluyó estas variables, razón por la cual se omitió esta armonización en el presente.

3.6 Imputación de datos perdidos

3.6.1 Imputación univariada por media.

Las imputaciones univariadas funcionan bien cuando los datos perdidos son poco relevantes o escasos. Sin embargo, cuando los datos perdidos son de un tamaño considerable, este tipo de imputaciones introducen sesgos que difícilmente pueden ser pasados por alto. En términos generales, la imputación univariada conlleva una distorsión y disminución de la precisión de los resultados ya que al utilizar solamente una medida como valor de imputación de datos faltantes se ignora la variabilidad y la relación con otras variables existentes en los datos. Este problema se acentúa por la sensibilidad a valores atípicos o extremos cuando se usa la media para realizar imputaciones. Además, este tipo de técnicas asumen que la distribución de la variable es normal, lo cual es un requerimiento que no se cumple en muchos campos de estudio, y en particular en el caso que se aborda en este trabajo. A pesar de lo anterior, la simplicidad y rapidez de su aplicación resulta

conveniente en muchos casos para establecer este tipo de imputaciones como modelos base frente a métodos más sofisticados. Adicionalmente, estrategias de imputación más sofisticadas no garantizan un mejor desempeño en los modelos de clasificación. En el marco de este trabajo, aplicamos modelos sobre datos imputados por media para determinar si los modelos avanzados de imputación pueden ofrecernos mejores resultados en términos de la predicción de los modelos.

3.6.2 Estrategias de imputación avanzadas

Iterative Imputer de Scikit-learn

Esta técnica de imputación utiliza un modelo de regresión predefinido para imputar los valores faltantes de manera iterativa. En cada iteración, el modelo de regresión se ajusta a las variables observadas y se utiliza para predecir los valores faltantes. Luego, los valores imputados se sustituyen en el conjunto de datos y el proceso se repite hasta que se cumple un criterio de convergencia, como el número máximo de iteraciones o el cambio mínimo en los valores imputados entre iteraciones. Una de las ventajas de Iterative Imputer radica en su flexibilidad, ya que permite utilizar diferentes modelos de regresión para adaptarse a las características de los datos. Por defecto, utiliza una regresión lineal bayesiana, pero también se puede especificar otros modelos de regresión lineal y también modelos basados en árboles y clustering, entre otros. Además, Iterative Imputer es capaz de manejar datos faltantes en múltiples columnas simultáneamente, lo que lo hace adecuado para conjuntos de datos complejos con diferentes tipos de variables.

Multiple Imputation by Chained Equations (MICE)

MICE (Multiple Imputation by Chained Equations) es un método de imputación utilizado para manejar datos faltantes el cual realiza imputaciones secuenciales utilizando modelos de regresión. En cada iteración, se selecciona una variable dependiente y se imputa utilizando un modelo de regresión ajustado a las demás variables. Este proceso se repite para cada variable dependiente hasta completar todos los valores faltantes y se obtienen n conjuntos de datos definidos por el usuario. Al final, se pueden combinar los conjuntos de datos imputados para obtener una estimación final que captura la incertidumbre asociada con los valores faltantes. La principal ventaja de MICE radica en su capacidad para imputar datos faltantes de manera flexible y realista al incorporar la correlación entre las variables. Al utilizar modelos de regresión, MICE aprovecha la información disponible en las variables observadas para predecir los valores faltantes de forma más precisa. Además, MICE proporciona una forma de tratar diferentes tipos de variables, incluyendo variables categóricas y continuas, lo que lo hace aplicable a una amplia gama de conjuntos de datos.

La principal diferencia entre Iterative Imputer y MICE es que éste último requiere patrones de datos perdidos aleatorios (MAR) o completamente aleatorios (MCAR), mientras que Iterative Imputer no requiere un patrón específico. Además, iterative Imputer utiliza un modelo de Regresión predefinido mientras que MICE define modelos de regresión para cada variable. Para ambos métodos, la precisión de las imputaciones dependerá de la densidad de información en el conjunto de datos. Un conjunto de datos con variables completamente independientes y sin correlación no producirá imputaciones precisas.

El patrón de datos perdidos del conjunto de datos de este trabajo es principalmente no aleatorio (MNAR) debido a que las mayor cantidad de variables pérdidas se explican por la diferencia de protocolos entre cada uno de los centros que recolectaron los datos. Además, el material de archivo en formato papel y la tarea de data entry para la consolidación de los datasets de cada centro agregan una complejidad adicional. Sin embargo, la densidad de la información de los set de datos es alta ya que las variables expresan distintos niveles de correlación, lo cual puede colaborar en la obtención de buenos resultados.

En el marco de este trabajo se probaron los regresores Bayesian Ridge, Decision Trees, Extra Trees y KNN en conjunto con Iterative Imputer y MICE para determinar cuál de estos produce el mejor resultado para nuestro conjunto de datos. Para ello, se utilizó un subset de datos completo sobre el cual se realizaron las siguientes operaciones:

1- Se obtiene una muestra aleatoria del 10% de los índices del dataframe.
2- Se salvan los valores de esos índices para la primera variable y se introducen valores perdidos en el dataframe.
3- Se define y ejecuta una instancia del imputador para cada uno de los algoritmos mencionados arriba y se predicen los valores perdidos.
4- Se utiliza RMSE y MSE como medida de error y se registra el tiempo computacional para cada algoritmo.
5- Se repiten los pasos 1 a 4 1000 veces y se obtiene la media de RMSE y MSE para cada algoritmo.
6- Se repite 1 a 5 para las otras variables seleccionadas.

7- Se repite de 1 a 6 para distintos niveles de datos perdidos (40% y 80%).

Adicionalmente, se evaluó la incidencia de las mejores imputaciones utilizando Iterative Imputer y MICE en el resultado final de las predicciones sobre el set de validación. El objetivo de realizar imputaciones es obtener un mejor resultado posible para los modelos de clasificación.

3.7 Modelos de aprendizaje supervisado

La elección de modelos a entrenar estuvo basada por un lado, en los resultados reportados por otros trabajos[11,13,18,39–42] y, por otro lado, en una acercamiento insesgado respecto del problema y de los clasificadores que nos permitiera elegir el mejor modelo en términos del conjunto de datos específico.

3.7.1 Modelo baseline: Regresión logística

Primero probamos modelos de regresión logística (LR), ya que son eficientes, fácilmente interpretables y ampliamente utilizados. Un análisis exploratorio reveló clases superpuestas entre muchas de las variables. Como los datos no son linealmente separables, los modelos de regresión no producirían buenos resultados de clasificación. A continuación, presentamos los modelos de aprendizaje automático utilizados para superar las limitaciones de LR. Para los modelos de LR, se utilizaron algoritmos de Scikit-learn LR.

3.7.2 Random Forest

Utilizamos el algoritmo Random Forest (RF) de Scikit-learn sin bagging (agregación de bootstrap o remuestreo con reemplazo). RF evita mejor el sobreajuste[43] en contraste

con los modelos de LR. Se optimizaron los parámetros *max depth*, *n estimators*, *criterion*, *min samples split*, *min samples leaf* y *max features*.

3.7.3 Support Vector Machines

Utilizamos modelos de máquinas de vectores de soporte (SVM) para generar un modelo de clasificación binaria predictivo. SVM transforma el espacio de características para establecer una frontera de decisión lineal con márgenes amplios que resultan en una menor generalización y una menor propensión al sobreajuste[44]. Los modelos SVM RBF obtuvieron peores resultados en todas las instancias de optimización en contraste con el kernel polinómico, por lo que fue excluido de forma temprana. Para el kernel polinómico de Scikit-learn, optimizamos *C*, *gamma*, el grado del polinomio y el coeficiente. *C* y *gamma* son ambos hiperparámetros utilizados para la regularización; representan $L2^2$ y el coeficiente del kernel, respectivamente.

3.7.4 XGBoost

XGBoost implementa el algoritmo de gradient boosting para problemas de clasificación y regresión. Utiliza una estrategia de crecimiento basada en árboles con un criterio de división optimizado llamado "ganancia de información" para encontrar la mejor manera de dividir los nodos en cada árbol. Además, implementa técnicas de regularización para evitar el sobreajuste y mejorar la generalización del modelo. Puede manejar valores faltantes automáticamente y tratar características categóricas sin necesidad de codificación previa. Los parámetros optimizados fueron *max depth*, *learning rate*, *n estimators*, *colsample bytree* y *min child weight*.

3.7.5 Validación cruzada para ajuste de hiperparámetros y entrenamiento

Como primera medida separamos nuestro dataset en un conjunto de entrenamiento y otro conjunto de validación final (20%) realizando un muestreo estratificado, utilizando las variables sitio y diagnóstico como estratos.

Realizamos análisis de validación cruzada para controlar el ajuste de hiperparámetros para cada modelo de aprendizaje automático. Esto nos permitió buscar y probar recursivamente los mejores hiperparámetros para cada modelo mediante búsquedas bayesianas con parada temprana sobre 100 iteraciones. Para cada modelo, buscamos hiperparámetros aplicando validación cruzada estratificada, donde la muestra de entrenamiento se dividió en k=5 partes al azar, y cada subconjunto se usó para entrenar k-1 veces y una vez como prueba. Posteriormente, aplicamos una metodología bootstrapping, descrita más abajo y posteriormente hicimos predicciones en el conjunto de validación final mencionado al principio de esta sección (20% del conjunto de datos reservado como validación final de los modelos). Las proporciones de la muestra completa, de entrenamiento y validación final se pueden ver en la **Tabla 5**. Cabe destacar que la EA corresponde al 56.3% de las demencias diagnosticadas en Latinoamérica, mientras que la DFT solamente se ubica entre el 2.8% y 1.5% del total de las demencias diagnosticadas en el subcontinente[7]. En este marco, el desbalance de los casos en la muestra completa corresponde a un hecho ineludible de la realidad. De hecho, la muestra completa contempla entre 5 y 10 veces más casos que la proporción estimada para el universo latinoamericano.

Tabla 5. Muestreo estratificado. Proporciones.

	Muestra		
Diagnóstico	**Completa**	**Entrenamiento**	**Validación**
EA	0.512725	0.512376	0.513784
CN	0.328988	0.328383	0.330827
DFT	0.158287	0.159241	0.155388

EA: Enfermedad de Alzheimer, DFT: Demencia frontotemporal, CN: Control sano.

3.7.6 Selección secuencial de características

Se utilizaron algoritmos de selección secuencial de características, siguiendo procedimientos publicados[11], para reducir el espacio de características d dimensionales a un subespacio k, donde k < d, para seleccionar el subconjunto de características más predictivas. Aplicamos Selección Secuencial hacia adelante de Mlxtend[45].

3.7.7 Métricas de rendimiento

Se utilizaron siete métricas para evaluar el rendimiento (accuracy, precision, sensitivity, specificity, F1, ROC AUC, confusion matrix) en todos los modelos de aprendizaje supervisado y, dos métricas (MSE y RMSE) para evaluar los modelos de imputación.

4.Resultados

4.1 Performance de las estrategias de imputación avanzadas.

En términos del error medio por RMSE y MSE, los resultados muestran que la mejor técnica de imputación para este dataset es el la imputación iterativa utilizando el estimador Bayesian Ridge. Este imputador obtuvo los valores promedio más bajos de error de forma consistente para todas las variables y para los tres niveles de missing

testeados (RMSE promedio de 0.012 para 10% de valores perdidos, 0.01215 para 40% y 0.01313 para 80%). Previsiblemente, la performance de los imputadores empeora a medida que se aumenta la cantidad de valores perdidos. En este sentido, Bayesian Ridge fue el imputador que aumentó su error en el menor grado. El resumen de resultados para los diferentes niveles de datos perdidos generados aleatoriamente pueden observarse en las tablas a continuación (**Tabla 6a, Tabla 6b**, **Tabla 6c**).

Tabla 6a. Resultados de imputadores para 10% de valores perdidos generados aleatoriamente sobre subset completo de datos .

	RMSE				
Variable	**BR**	**DTR**	**ETR**	**KNR**	**MICE**
Años de educación	0.0224	0.02729	**0.02171**	0.02212	**0.0276**
Año de nacimiento	0.03302	0.0368	**0.03253**	0.03943	**0.0464**
Cognición	**0.00002**	0.00289	0.00105	**0.00319**	0.0008
Funcionalidad	0.01223	0.01475	**0.01108**	0.01502	**0.0153**
IFS	0.02517	**0.02547**	0.02449	0.02022	**0.0138**
MMSE	**0.00048**	0.0093	0.00689	**0.01315**	0.0061
MoCA	**0.00011**	0.0024	0.0025	**0.00542**	0.002
ACE III	**0.00018**	0.0046	0.00483	**0.00684**	0.0037
Resultado promedio	**0.01170**	**0.01544**	**0.01314**	**0.01567**	**0.01446**

RMSE: Error cuadrático medio. BR: Bayesian Ridge, DTR Decision Tree Regressor, ETR: Extra Trees Regressor, KNR: K-nearest Neighbors Regressor, MICE: Multiple Imputation by Chained Equations.

Tabla 6b. Resultados de imputadores para 40% de valores perdidos generados aleatoriamente sobre subset completo de datos .

	RMSE				
Variable	**BR**	**DTR**	**ETR**	**KNR**	**MICE**
Años de educación	0.02272	0.02823	**0.02208**	0.02326	**0.0286**
Año de nacimiento	0.03323	0.03916	**0.03313**	0.03835	**0.046**
Cognición	**0.00002**	0.00312	0.0012	**0.004**	0.0012
Funcionalidad	0.01236	0.01565	**0.0118**	**0.01597**	0.0159
IFS	0.02529	**0.02572**	0.02448	0.02102	**0.0156**
MMSE	**0.00051**	0.00973	0.00729	**0.01563**	0.0069
MoCA	**0.00012**	0.00292	0.0027	**0.00602**	0.0026
ACE III	**0.0002**	0.00547	0.00525	**0.00774**	0.0047
Resultado promedio	**0.01181**	**0.01625**	**0.01349**	**0.01650**	**0.01519**

RMSE: Error cuadrático medio. BR: Bayesian Ridge, DTR Decision Tree Regressor, ETR: Extra Trees Regressor, KNR: K-nearest Neighbors Regressor, MICE: Multiple Imputation by Chained Equations.

Tabla 6c. Resultados de imputadores para 80% de valores perdidos generados aleatoriamente sobre subset completo de datos .

	RMSE				
Variable	**BR**	**DTR**	**ETR**	**KNR**	**MICE**
Años de educación	0.02492	**0.03527**	**0.02446**	0.02467	0.0306
Año de nacimiento	**0.0343**	**0.05067**	0.03713	0.03841	0.0468
Cognición	**0.00003**	0.00467	0.00201	**0.00815**	0.0041
Funcionalidad	**0.01433**	**0.02256**	0.01485	0.01866	0.019
IFS	0.02554	**0.02596**	0.02379	0.02224	**0.0194**
MMSE	**0.00072**	0.01293	0.00882	**0.02248**	0.0113
MoCA	**0.00016**	0.00484	0.00346	**0.00936**	0.0054
ACE III	**0.00025**	0.00984	0.00707	**0.01163**	0.009
Resultado promedio	**0.01253**	**0.02084**	**0.01520**	**0.01945**	**0.01820**

RMSE: Error cuadrático medio. BR: Bayesian Ridge, DTR Decision Tree Regressor, ETR: Extra Trees Regressor, KNR: K-nearest Neighbors Regressor, MICE: Multiple Imputation by Chained Equations.

4.2 Performance de los clasificadores

En esta sección se exhiben los resultados principales de este trabajo, los modelos de imputación y los modelos de clasificación con los mejores resultados, explicitando previamente algunas características metodológicas adicionales.

4.2.1 Clasificación de EA frente a DFT con datos imputados por media

En primer lugar, clasificamos EA frente a DFT siguiendo los métodos especificados arriba y aplicando la técnica de bootstrapping en el conjunto de entrenamiento, lo que nos permitió evaluar el poder discriminativo medio de los modelos y su significancia estadística mediante la construcción de un p-value empírico. Para ello mezclamos aleatoriamente las etiquetas de las clases y contamos cuántas veces el modelo aplicado sobre las etiquetas mezcladas tuvo un puntaje mayor. Se entrenaron y evaluaron 5000 repeticiones con los mejores hiperparámetros encontrados para RF, SVM y XGBoost para cada una de las estrategias de imputación, para los cuales obtuvimos las métricas de evaluación y su intervalo de confianza al 95%. Adicionalmente, se utilizó la misma estrategia para evaluar la performance de un modelo baseline de regresión logística.

Posteriormente, se repitió el proceso con el modelo que arrojó los mejores resultados para el grupo de variables de evaluación neurocognitiva, eliminando las variables sociodemográficas, con el fin de observar la capacidad discriminativa de este conjunto de variables. Finalmente se probó, por un lado, un modelo en el cual las variables cognitivas fueron armonizadas en una sola variable para evaluar el diferencial de rendimiento que pudiera existir con esta estrategia y, por otro lado, un modelo para el subconjunto de datos completos para contrastar con los resultados obtenidos mediante la imputación.

El mejor modelo para discriminar entre pacientes con EA y DFT fue el modelo RF con datos imputados por media **(Fig. 1, Tabla 7)** e incluyó (por orden de importancia), año de nacimiento, función ejecutiva (IFS), cognición (ACE-III normalizado siguiendo el método de Matías-Guiu), cognición social (Mini-SEA total) y síntomas neuropsiquiátricos (NPI). Para evaluar la importancia de las variables se utilizó la técnica de permutación de Scikit-learn[38] y el algoritmo Shapley[46]. Para éste último las variables más importantes fueron (en orden) año de nacimiento, función ejecutiva (IFS), años de educación, cognición (ACE-III normalizado siguiendo el método de Matías-Guiu) e ingreso del hogar. Este modelo mostró un considerable poder discriminativo en el esquema de bootstrapping [ROC AUC = 0.869 (±0.009), accuracy = 0.806 (±0.011), sensibilidad = 0.826 (±0.011), especificidad = 0.738 (±0.012), F1 = 0.642 (±0.013), precisión = 0.57 (±0.014)] y una muy buena generalización [ROC AUC = 0.839, accuracy = 0.859, sensibilidad = 0.8, especificidad = 0.88, F1 = 0.727, precisión = 0.667].

El modelo RF que armonizó las variables de cognición **(Fig. 2, Tabla 8)** en una sola escala también tuvo resultados satisfactorios sin ver afectadas su performance [ROC AUC = 0.868 (±0.009), accuracy = 0.806 (±0.011), sensibilidad = 0.825 (±0.011), especificidad = 0.746 (±0.010), F1 = 0.645 (±0.013), precisión = 0.571 (±0.014)] ni su capacidad de generalización [ROC AUC = 0.839, accuracy = 0.8598, sensibilidad = 0.8, especificidad = 0.88, F1 = 0.7273, precisión = 0.6667]. Las variables más importantes fueron año de nacimiento, función ejecutiva (IFS), cognición, años de educación y síntomas neuropsiquiátricos (NPI). Utilizando shapley, las variables más importantes fueron las mismas, alternando el los lugares de cognición y años de educación, con excepción del quinto lugar donde aparece el ingreso del hogar. Adicionalmente se corrió una versión de

este modelo que consideró solamente las variables correspondientes a pruebas clínicas y obtuvo resultados inferiores al anterior en el esquema bootstrapping [ROC AUC = 0.821 (±0.011), accuracy = 0.784 (±0.01), sensibilidad = 0.653 (±0.013), especificidad = 0.824 (±0.01), F1 = 0.587 (±0.014), precisión = 0.537 (±0.014)] y en su capacidad de generalización [ROC AUC = 0.7346, accuracy = 0.785, sensibilidad = 0.64, especificidad = 0.83, F1 = 0.5818, precisión = 0.5333] **(Fig. 2, Tabla 8)**.

Por otro lado, el modelo que utilizó el subconjunto de datos completos **(Fig. 2, Tabla 8)** obtuvo resultados inferiores en términos de discriminación entre clases tanto para el bootstraping [ROC AUC = 0.843 (±0.01), accuracy = 0.833 (±0.01), sensibilidad = 0.89 (±0.008), especificidad = 0.57 (±0.014), F1 = 0.545 (±0.014), precisión = 0.552 (±0.014)] como para el conjunto de validación final [ROC AUC = 0.75, accuracy = 0.88, sensibilidad = 1, especificidad = 0.5, F1 = 0.667, precisión = 1]. Las mejores variables fueron función ejecutiva (IFS), cognición social (Mini SEA), año de nacimiento, cognición (MMSE normalizado siguiendo el método de Van-Steenoven), y años de educación.

4.2.2 Clasificación de EA frente a DFT con datos imputados por iterative imputer y MICE

El modelo RF con datos imputados con Iterative imputer y Bayesian Ridge **(Tabla 7)** alcanzó una performance equivalente en el esquema de bootstrapping [ROC AUC = 0.87 (±0.009), accuracy = 0.825 (±0.011), sensibilidad = 0.876 (±0.009), especificidad = 0.662 (±0.013), F1 = 0.642 (±0.013), precisión = 0.626 (±0.013)] pero insuficiente sobre el set de datos no vistos [ROC AUC = 0.737, accuracy = 0.832, sensibilidad = 0.91, especificidad = 0.56, F1 = 0.61, precisión = 0.667]. Las variables más importantes fueron cognición social (Mini SEA) , año de nacimiento, años de educación, función ejecutiva (IFS) y

miembros del hogar para ambas técnicas de evaluación, aunque miembros del hogar se ubica en tercera posición para shapley.

Por su parte, el modelo RF con datos imputados por MICE **(Tabla 7)** tuvo una peor performance en el esquema bootstrapping [ROC AUC = 0.801 (±0.011), accuracy = 0.775 (±0.012), sensibilidad = 0.827 (±0.010), especificidad = 0.606 (±0.013), F1 = 0.56 (±0.014), precisión = 0.522 (±0.014)] y su generalización fue también deficiente [ROC AUC = 0.743, accuracy = 0.841, sensibilidad = 0.93, especificidad = 0.56, F1 = 0.622, precisión = 0.7].

El modelo XGBoost con datos imputados por media tuvo resultados mixtos, obteniendo valores altos para algunas métricas en el esquema de bootstrapping [ROC AUC = 0.854, (±0.009), accuracy = 0.836 (±0.01), especificidad = 0.0.943 (±0.006)] pero escasa sensibilidad [0.503 (±0.014)]. Además, sus resultados de generalización no fueron buenos y confiables ya que exhibió una curva ROC AUC menor a la de RF y otras métricas mostraron puntajes muy elevados en contraste con el esquema de Bootstrapping **(Tabla A1)**. Los modelos restantes (regresión logística, SVM y el resto de los modelos XGBoost) también tuvieron peores resultados que los RF en todos los casos. Estos resultados pueden verse en la **Tabla A1**.

Tabla 7: Resultados principales para Random Forest según técnica de imputación.

Método de imputación	Conjunto	Técnica	Accuracy	ROC AUC	Precision	Sensitivity	Specificity	F1
Media	Entrenamiento	Bootstrapping	0.805 (± .0011)	0.866 (± 0.009)	0.569 (± 0.013)	0.738 (± 0.012)	0.826 (± 0.010)	0.641 (± 0.013)
	Validación	Out of sample	0.8505	0.819	0.6552	0.76	0.88	0.7037
Iterative Imputer	Entrenamiento	Bootstrapping	0.826 (± 0.011)	0.87 (0.009)	0.626 (± 0.013)	0.662 (± 0.013)	0.876 (± 0.009)	0.641 (± 0.013)
	Validación	Out of sample	0.8318	0.7373	0.6667	0.56	0.91	0.6087
MICE	Entrenamiento	Bootstrapping	0.775 (± 0.011)	0.801 (± 0.011)	0.523 (± 0.014)	0.607 (± 0.013)	0.827 (± 0.01)	0.560 (± 0.014)
	Validación	Out of sample	0.8411	0.7434	0.7	0.56	0.93	0.622

Tabla 8: Resultados adicionales para Random Forest e imputación por media.

Dataset	Conjunto	Técnica	Accuracy	ROC AUC	Precision	Sensitivity	Specificity	F1
Subset sin datos perdidos	Entrenamiento	Bootstrapping	0.833 (± 0.01)	0.843 (± 0.01)	0.552 (± 0.014)	0.57 (± 0.014)	0.89 (± 0.008)	0.545 (± 0.014)
	Validación	Out of sample	0.88	0.75	1	0.5	1	0.667
Cognitivas armonizadas	Entrenamiento	Bootstrapping	0.806 (± 0.011)	0.869 (± 0.009)	0.571 (± 0.014)	0.746 (± 0.012)	0.825 (± 0.011)	0.645 (± 0.013)
	Validación	Out of sample	0.8598	0.839	0.6667	0.8	0.88	0.7273
Solo pruebas clínicas y cog. armonizada	Entrenamiento	Bootstrapping	0.784 (± 0.01)	0.821 (± 0.011)	0.537 (± 0.014)	0.653 (± 0.013)	0.824 (± 0.01)	0.587 (± 0.014)
	Validación	Out of sample	0.785	0.7346	0.5333	0.64	0.83	0.5818

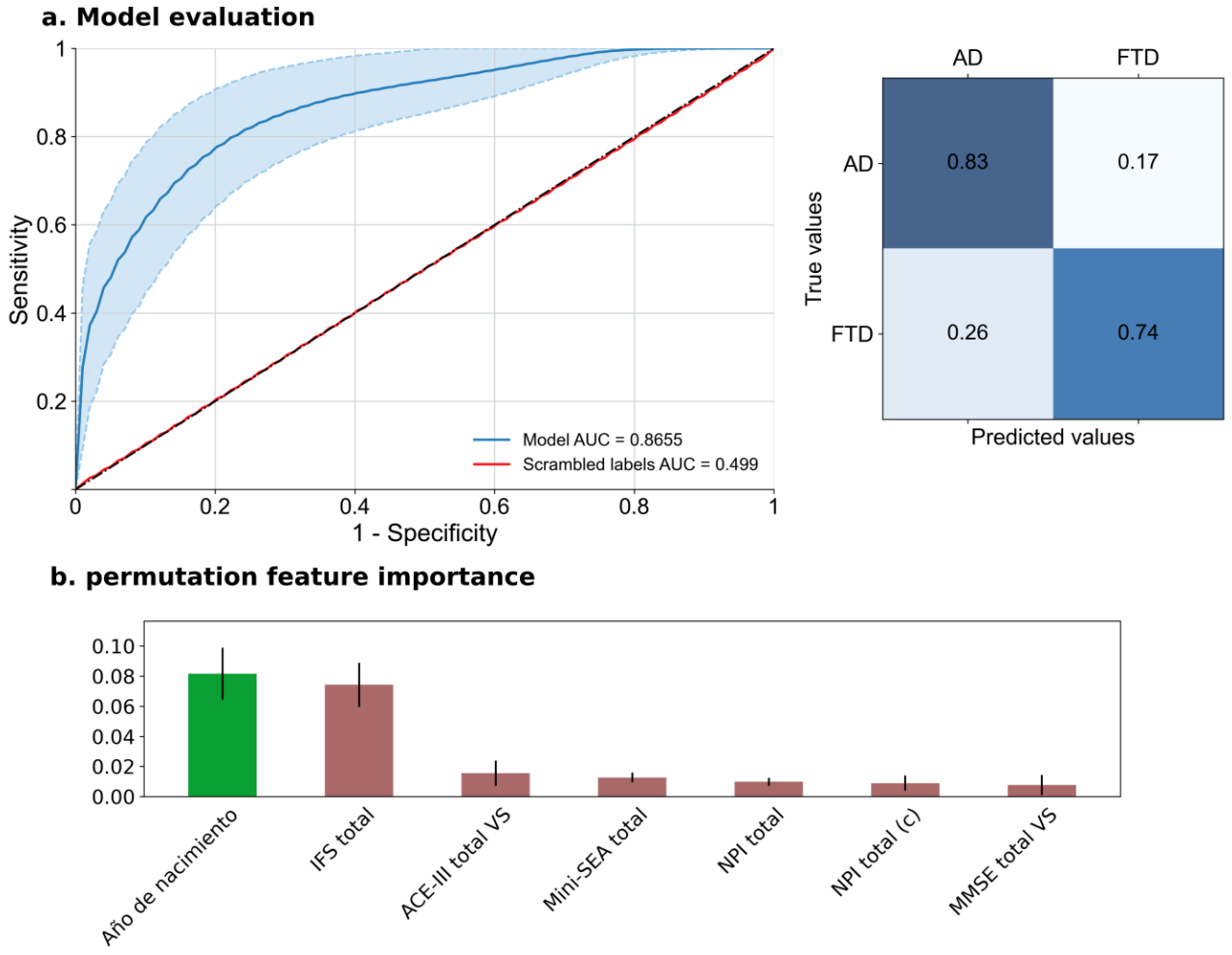

Fig 1. Resultados principales medios (5000 iteraciones), incluyendo la evaluación del modelo de Random Forest con datos imputados por media, la matriz de confusión y la importancia de las características bajo el esquema de permutación. (a) El gráfico superior izquierdo muestra la curva ROC con dos desviaciones estándar y su AUC. El gráfico superior derecho muestra la matriz de confusión en magnitudes relativas. (b) El gráfico inferior muestra la importancia de cada característica del modelo según los resultados de la permutación (*permutation feature importance*). Las más importantes fueron el año de nacimiento, las funciones ejecutivas (IFS), y en segundo lugar, el cribado cognitivo (medido con ACE-III), la cognición social (Mini-SEA total), y los síntomas neuropsiquiátricos (NPI).

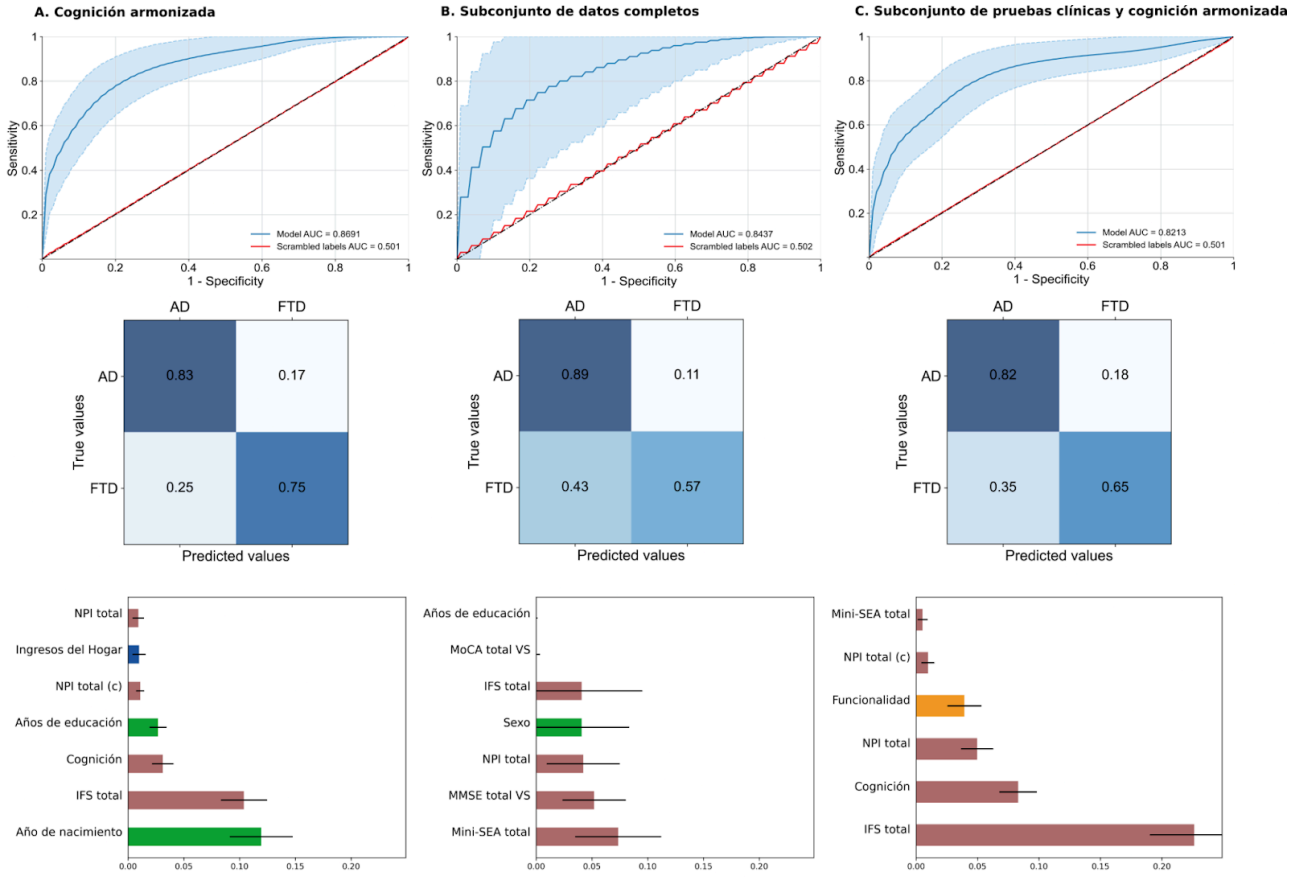

Fig 2: Resultados medios adicionales (5000 iteraciones) para Modelos de Random Forest con A) variables de cognición armonizadas, B) subconjunto de datos completos, y C) solo pruebas clínicas y variables de cognición armonizadas. Cada fila muestra un gráfico para cada uno de los tres modelos. La primera muestra las curvas ROC con dos desviaciones estándar y su AUC. La segunda muestra las matrices de confusión en magnitudes relativas. Finalmente, la tercera muestra las variables más importantes según (b) El gráfico inferior muestra la importancia de cada característica del modelo según los resultados de la permutación (5000 iteraciones, *permutation feature importance*). Para el modelo A) las más importantes fueron el año de nacimiento, las funciones ejecutivas (IFS Total), el cribado cognitivo (cognición armonizada y unificada), años de educación, los síntomas neuropsiquiátricos (NPI caregiver), ingresos del hogar, y los síntomas neuropsiquiátricos (NPI). Para el modelo B) las más importantes fueron la cognición social (Mini-SEA total), el cribado cognitivo (medido con MMSE normalizado mediante el método Van-Steenoven), los síntomas neuropsiquiátricos (NPI), sexo, y las funciones ejecutivas (IFS Total). Para el modelo C) las más importantes fueron las funciones ejecutivas (IFS Total), el cribado cognitivo (cognición armonizada y unificada), los síntomas neuropsiquiátricos (NPI), la funcionalidad, los síntomas neuropsiquiátricos (NPI caregiver) y la cognición social (Mini-SEA total) .

4.2.3 Selección Secuencial de Características.

Adicionalmente, se aplicó un algoritmo de selección secuencial de características (SFS) al conjunto de entrenamiento. Este algoritmo construye modelos agregando variables con mayor poder predictivo. El mejor modelo [ROC AUC = 0.813 (k=5 stratified folds)] se alcanzó en el quinto paso. Las características con mayor capacidad predictiva fueron la función ejecutiva (IFS), miembros del hogar, síntomas neuropsiquiátricos (NPI), cognición (ACE-III normalizado siguiendo el método de Matías-Guiu) y cognición social (Mini SEA). Año de nacimiento y sexo se encontraron entre las características menos potentes de la clasificación.

Tabla 9: Mejores predictores según elección secuencial de características.

Paso	Variables	Puntaje medio	Intervalo	Desvío estándar	Error estándar
1	IFS	0.75875	0.035641	0.015838	0.011199
2	..., Miembros del hogar	0.78749	0.039179	0.01741	0.012311
3	..., NPI caregiver	0.808031	0.048993	0.021771	0.015395
4	..., ACE III	0.806996	0.027988	0.012437	0.008795
5	... Mini SEA	0.813156	0.028004	0.012445	0.0088
6	..., NPI	0.813156	0.035995	0.015995	0.01131
7	..., Año de nacimiento	0.812127	0.038885	0.01728	0.012218
8	..., Sexo	0.812124	0.024299	0.010798	0.007635

"..." representa a todas las variables de los pasos anteriores.

4.2.4 Resultados complementarios: Modelos de aprendizaje automático para discriminar participantes sanos frente a pacientes.

Se optimizaron modelos RF con datos imputados por media para discriminar EA vs CN y DFT vs CN. Estos modelos alcanzaron altas puntuaciones (ambos modelos alcanzaron una ROC AUC =+0.9 en el esquema bootstrapping. EA vs CN obtuvo una ROC AUC = .9074 en el set de validación final y DFT vs CN de 0.818). Los mejores discriminadores de EA

frente a CN fueron el cribado cognitivo (ACE-III, MoCA y MMSE armonizadas mediante la metodología Van-Steenoven) y, en menor grado, los síntomas neuropsiquiátricos (NPI), la funcionalidad y la cognición social (Mini SEA). Por el contrario, los mejores discriminadores de DFT frente a CN fueron el funcionamiento ejecutivo (IFS), el cribado cognitivo (ACE-III y MMSE armonizadas mediante la metodología Van-Steenoven), y, en menor medida, el sexo , la funcionalidad y las medidas de síntomas neuropsiquiátricos (NPI). Los resultados completos pueden observarse en la **Tabla A2**.

5. Discusión

Este estudio implementó procedimientos de aprendizaje supervisado para discriminar entre pacientes con EA y DFT utilizando datos clínicos, cognitivos y demográficos de archivo recogidos en entornos clínicos multicéntricos, heterogéneos y subrepresentados a través de múltiples PAL. A pesar de la heterogeneidad y calidad de los datos, los enfoques de aprendizaje automático fueron eficaces para diferenciar los grupos. Un modelo de Random Forest sobre datos imputados por media demostró ser el más exitoso en la discriminación entre EA y DFT en datos no vistos (accuracy = 0.85 y ROC AUC = 0.82). Los factores más significativos para discriminar entre EA y DFT en PAL fueron el año de nacimiento, función ejecutiva (IFS), cognición (ACE-III normalizado siguiendo el método de Matías-Guiu), cognición social (Mini-SEA) y síntomas neuropsiquiátricos (NPI).

Si bien este trabajo puede verse como una profundización de una publicación realizada con anterioridad[18]; hasta donde sabemos, este es el primer estudio que utiliza modelos de aprendizaje automático para imputar datos perdidos y evaluar la capacidad predictiva

de los datos clínicos, cognitivos y demográficos recopilados de entornos clínicos multicéntricos, heterogéneos y subrepresentados para discriminar entre EA y DFT en PAL.

En la actualidad, existe una necesidad insatisfecha de estudiar el diagnóstico de la demencia en PAL, como lo destacan múltiples asociaciones internacionales[5,47]. Las técnicas basadas en datos empleadas aquí proporcionan un enfoque confiable para discriminar con precisión entre condiciones, en una región conocida por exhibir una alta heterogeneidad clínica, no contar con procedimientos sistemáticos para la evaluación clínica de la demencia y presentar un acceso limitado a biomarcadores[1,8,10,48–50]. Además, los resultados añaden nuevas pruebas que apoyan el uso de métodos de aprendizaje automático para la discriminación clínica de la demencia. Dichos métodos destacan el potencial discriminatorio de las medidas de funcionamiento ejecutivo y síntomas neuropsiquiátricos, además de las medidas cognitivas clásicas. Por otra parte, la cognición social mostró potencial discriminatorio cuando se utilizaron los datos completos y cuando la estrategia de imputación fue distinta a la media. En consecuencia de esto, y teniendo en cuenta resultados previamente reportados[18,37], sería plausible hipotetizar que la baja capacidad discriminatoria exhibida por la cognición social en modelos con datos imputados por media se debe a la pronunciada cantidad de datos perdidos y la pérdida de variabilidad específica que resulta de la mencionada estratégia de imputación. Por otra parte, la escala de funcionalidad no tuvo valores discriminativos elevados en la predicción de la EA y la DFT en nuestro modelo final, al igual que en resultados comunicados previamente[18,51,52]. La elevada dispersión de los valores de funcionalidad y la incapacidad para controlar otros posibles factores de confusión, como la gravedad de la enfermedad y los valores perdidos, podrían explicar el reducido poder

discriminativo. Nuestros hallazgos concuerdan con los de estudios realizados en países de renta alta[13,14], y muestran que en poblaciones subrepresentadas como las que se encuentran en PAL, la aplicación de modelos de aprendizaje automático sobre pruebas clínicas y cognitivas puede tener una alta precisión para discriminar la EA de otras demencias.

Por otro lado, los pacientes con DFT tienen una edad media de inicio de la enfermedad más temprana[53,54] y tienden a tener un mayor nivel educativo que los pacientes con EA[53] **(Tabla 2)**. Estos factores demográficos clásicos (edad y educación) tuvieron puntuaciones predictivas más altas que las variables cognitivas o clínicas en algunos de los modelos reportados en este estudio. No obstante, la selección secuencial de características para nuestro modelo final mostró que las variables clínicas tienen mayor capacidad de discriminación. Estos resultados coinciden con estudios anteriores que muestran que los factores demográficos son predictores menos eficaces que las medidas cognitivas a la hora de diferenciar la EA de la DFT[55]. En esta línea, el modelo que utilizó solamente pruebas cognitivas fue capaz de discriminar ambas clases muy por encima del azar, aunque con una pequeña caída en el rendimiento del mejor modelo reportado. Más aún, los resultados complementarios para el modelo de clasificación entre AD y CN no tuvo variables socio-demográficas entre los mejores predictores **(Fig A1)**.

Nuestros resultados también revelaron que las pruebas de cribado cognitivo (ACE III, MoCA y MMSE normalizados), la funcionalidad, el funcionamiento ejecutivo (IFS), la cognición social (Mini-SEA), y los síntomas neuropsiquiátricos (NPI), son los mejores predictores para diferenciar CN frente a EA y CN frente a DFT. Los mejores predictores

para la EA frente a los CN fueron principalmente el cribado cognitivo, y en segundo lugar, los síntomas neuropsiquiátricos, la funcionalidad, y la cognición social. En contraste, el funcionamiento ejecutivo fue más útil para diferenciar la DFT de los CN y el sexo tuvo una importancia de segundo orden **(Fig A1, Tabla A2)**. Los resultados concuerdan con estudios anteriores que muestran que, en comparación con los CN, los pacientes con EA y DFT tienden a presentar con mayor frecuencia alteraciones en los procesos cognitivos[56], en la cognición social[16] y en los síntomas neuropsiquiátricos (comparando CN frente a EA[51,57] y CN frente a DFT[51,57]). Adicionalmente, los niveles de precisión alcanzados por los procedimientos de aprendizaje automático en este estudio fueron comparables a los observados en estudios de aprendizaje automático que se sirven de neuroimágenes y de otros biomarcadores para la clasificación de EA y DFT[53].

Además, los resultados revelaron que el RF fue superior en la clasificación de casos de EA y DFT. El menor poder de clasificación de la regresión logística (en comparación con Random Forest y otros modelos de ML) podría explicarse por su menor precisión con conjuntos de datos multidimensionales tal como se ha reportado previamente[58,59]. RF es un modelo robusto a los datos ruidosos y complejos, capaz de captar relaciones no lineales entre los datos e interacciones complejas. Además puede manejar automáticamente las características irrelevantes o redundantes, es poco propenso al sobreajuste y requiere poco ajuste de hiperparámetros[60]. El modelo que redujo características redundantes, utilizando la variable de pruebas cognitivas armonizadas, no tuvo diferencias significativas en sus resultados, confirmando a nivel práctico la capacidad de RF en el manejo de características repetidas o irrelevantes **(Tabla 8)**. Por otro lado, los resultados del modelo que utilizó exclusivamente pruebas neuropsicológicas

mostraron que éstas son capaces de discriminar entre las clases de demencia muy por encima del azar **(Tabla 8)**. Finalmente, el modelo que utilizó el subset de datos completos no tuvo un rendimiento superior al modelo con datos imputados por media, presumiblemente por la escasa cantidad de casos con datos completos para realizar un entrenamiento adecuado del modelo **(Tabla 8)**.

En conjunto, los resultados revelan importantes conocimientos para el estudio del diagnóstico de la demencia en las poblaciones heterogéneas, subrepresentadas y de entornos de bajos recursos de PAL. (a) Los datos convencionales utilizados para la evaluación clínica de la demencia son una fuente confiable de información para discriminar la EA, la DFT y los controles a través de diferentes PAL, compensando parcialmente el limitado acceso regional a los biomarcadores; (b) se observaron puntajes de discriminación similares para EA y DFT, tal como se encontró en otros estudios que utilizan biomarcadores[1–3]; (c) los síntomas neuropsiquiátricos combinados, las funciones ejecutivas, el cribado cognitivo y la cognición social son medidas relevantes para caracterizar la demencia en PAL, (d) las estrategias de imputación pueden compensar la falta de armonización de protocolos y los datos perdidos, reteniendo una capacidad predictiva alta, evitando la pérdida excesiva de datos y limitando la variabilidad de los resultados, (e) los esquemas de imputación de datos por media para Ral-World-Clinical-Underrepresented-Data con un nivel moderado o alto de datos perdidos pueden resultar más productivos que otras metodologías más complejas, en términos de los resultados de clasificación obtenidos, y (d) las técnicas de machine-learning sobre Ral-World-Clinical-Underrepresented-Data podrían constituirse como un paso previo a los

estudios experimentales en aquellas regiones que necesitan de ellos pero que por razones económicas, sociales o logísticas no se se llevan a cabo.

6. Limitaciones y futuros trabajos

Nuestro trabajo tiene importantes limitaciones. Las diferencias entre los protocolos de los distintos centros dieron lugar a valores perdidos, lo que requirió la imputación de los valores faltantes. El proceso de recogida de datos entre centros no estaba armonizado, sólo se utilizaron puntuaciones globales y no se disponía de puntuaciones de gravedad. Además, la falta de esquemas sistematizados de recopilación y consolidación de datos clínicos en los distintos centros puede ser un factor adicional de distorsión. Estos son retos típicos cuando se utilizan datos multicéntricos. Abordamos este problema armonizando e imputando los datos que faltaban siguiendo los procedimientos publicados anteriormente[20–23,27,38,39]. Futuros trabajos podrían alternar distintas estrategias de imputación según la composición de datos perdidos y la especificidad de la variable a imputar.

Adicionalmente, probamos y verificamos la replicabilidad de nuestros resultados y la productividad de las variables clínicas y cognitivas, entrenando modelos complementarios que utilizaron datos completos, datos cognitivos armonizados, y pruebas clínicas y cognitivas solamente. Los resultados éstos produjeron puntuaciones predictivas y predictores similares, si bien el modelo que utilizó el subconjunto de datos completos mostró menor capacidad de generalización por el pequeño tamaño de la submuestra

resultante. Aunque estos resultados sugieren que los datos ausentes no afectaron ostensiblemente a los procesos de clasificación en nuestro estudio, futuros trabajos deberían probar la replicabilidad de nuestros modelos basados en mayores conjuntos de datos con más equilibrio entre diferentes centros y comparar esos resultados utilizando métodos de imputación de datos perdidos. Adicionalmente, nuevas estrategias de imputación podrían ser desarrolladas para contar con esquemas metodológicos específicos del campo y de las pruebas involucradas.

Por otro lado, la armonización de los protocolos clínicos sobre la demencia en América Latina podría beneficiar la investigación en la región y permitir la realización de estudios con poblaciones más grandes y heterogéneas desde el punto de vista cultural, social y genético. En esta línea, la capacidad de relevar y preservar los datos de los pacientes de forma sistemática y armonizada permitirá tener estimaciones más eficientes y confiables para discriminar los pacientes con DFT y EA e Impulsar la investigación de la demencia en otras regiones subrepresentadas, heterogéneas y con recursos limitados; y reducir los protocolos de evaluación o jerarquizar la importancia de las pruebas neuropsicológicas en función de su productividad diagnóstica y discriminativa.

Nuestro estudio también se vio limitado por un tamaño de muestra inferior al óptimo para los métodos utilizados. Sin embargo, el tamaño de la muestra fue comparable, y a menudo mayor, que el de otros estudios multicéntricos sobre la demencia[15,41,61,62]. Además, nuestra metodología aplicada nos permite integrar muestras futuras a nuestros datos actuales y probar nuestros modelos en nuevos datos no vistos. Otra posible limitación de nuestro estudio es el desequilibrio en el tamaño de la muestra de los dos grupos de

diagnóstico. Sin embargo, otros estudios en PAL han descrito una mayor prevalencia de EA (que representa el 55% de los casos de demencia atendidos en clínicas) que de DFT (que representa entre el 2,8% y el 1,5% de los casos de demencia)[2,3,7,8]. Estudios futuros podrían beneficiarse de aumentar el número de casos de DFT incluidos. Por otro lado, los datos socio-demográficos utilizados fueron muy limitados, con muchos datos perdidos y a menudo poco descriptivos. Futuros estudios podrían beneficiarse de incluir datos sociodemográficos relevantes, granulares y completos.

Finalmente, los diagnósticos de demencia en nuestro estudio se basaron en la experiencia clínica sin confirmación de biomarcadores, ya que el acceso a los biomarcadores tradicionales es limitado en PAL debido al costo. Aunque sería ideal para futuros estudios probar el poder de clasificación contra casos confirmados por biomarcadores, nuestros resultados indican que el uso de datos clínicos y neuropsicológicos podría apoyar el diagnóstico de demencia, particularmente en poblaciones subrepresentadas donde el acceso a biomarcadores es limitado.

7. Conclusiones

Se obtuvo una alta precisión de clasificación para la EA, DFT y CN mediante la combinación de procedimientos de aprendizaje automático aplicados a datos demográficos, cognitivos y conductuales de entornos clínicos heterogéneos, subrepresentados y de bajos recursos de PAL. Los resultados ponen de relieve la importancia de combinar la evaluación clínica convencional con pruebas cognitivas detalladas que rastreen la cognición social, el funcionamiento ejecutivo, los síntomas

conductuales y el cribado cognitivo para diagnosticar la EA y la DFT en PAL. Desarrollamos un enfoque metodológico robusto para aplicar transformaciones sobre Real-World-Clinical-Underrepresented-Data con el objetivo de clasificar subtipos de demencia (EA y DFT) y CN. Estos hallazgos apoyan el uso de métodos de aprendizaje automático para estudios multicéntricos que involucran regiones de bajos recursos con acceso restringido a biomarcadores para cohortes subrepresentadas. Adicionalmente, los resultados presentan evidencia acerca del uso de estrategias de imputación para evitar la pérdida excesiva de casos por datos perdidos. Los servicios de demencia en PAL podrían beneficiarse de la mejora de la precisión diagnóstica aplicando el enfoque metodológico actual. En este sentido, la capacidad de relevar y preservar los datos clínicos de forma sistemática y estandarizada permitirá tener mejores estimaciones sobre los pacientes con EA y DFT. Además, la armonización de protocolos permitirá evaluar más profundamente la productividad de las diferentes pruebas neuropsicológicas en función de la heterogeneidad y las particularidades de los países de la región. Finalmente, esta metodología podría ser aplicada y evaluada con datos similares de regiones con las mismas características para constituirse como un método para mitigar parcialmente la falta de estudios experimentales en regiones subrepresentadas y de escasos recursos.

8. Referencias

[1] Parra MA, Baez S, Sedeño L, Gonzalez Campo C, Santamaría‐García H, Aprahamian I, et al. Dementia in Latin America: paving the way toward a regional action plan. Alzheimers Dement 2021;17:295–313.

[2] Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet 2020;396:413–46.

[3] Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. Lancet 2017;390:2673–734. https://doi.org/10.1016/s0140-6736(17)31363-6.

[4] Dawson WD, Bobrow K, Ibanez A, Booi L, Pintado-Caipa M, Yamamoto S, et al. The necessity of diplomacy in brain health. Lancet Neurol 2020;19:972–4. https://doi.org/10.1016/S1474-4422(20)30358-6.

[5] Nichols E, Steinmetz JD, Vollset SE, Fukutaki K, Chalek J, Abd-Allah F, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7:e105–25. https://doi.org/10.1016/S2468-2667(21)00249-8.

[6] Ferri CP, Jacob KS. Dementia in low-income and middle-income countries: Different realities mandate tailored solutions. PLoS Med 2017;14:e1002271. https://doi.org/10.1371/journal.pmed.1002271.

[7] Custodio N, Wheelock A, Thumala D, Slachevsky A. Dementia in Latin America: Epidemiological Evidence and Implications for Public Policy. Front Aging Neurosci 2017;9. https://doi.org/10.3389/fnagi.2017.00221.

[8] Parra MA, Baez S, Allegri R, Nitrini R, Lopera F, Slachevsky A, et al. Dementia in Latin America: Assessing the present and envisioning the future. Neurology 2018;90:222–31. https://doi.org/10.1212/wnl.0000000000004897.

[9] Ibanez A, Yokoyama JS, Possin KL, Matallana DL, Lopera F, Nitrini R, et al. The Multi-Partner Consortium to Expand Dementia Research in Latin America (ReDLat): Driving Multicentric Research and Implementation Science. Front Neurol 2021;12:303.

[10] Parra MA, Orellana P, Leon T, Victoria CG, Henriquez F, Gomez R, et al. Biomarkers for dementia in Latin American countries: Gaps and opportunities. Alzheimers Dement 2022. https://doi.org/10.1002/alz.12757.

[11] Moguilner S, Birba A, Fittipaldi S, Gonzalez-Campo C, Tagliazucchi E, Reyes P, et al. Multi-feature computational framework for combined signatures of dementia in underrepresented settings. J Neural Eng 2022;19. https://doi.org/10.1088/1741-2552/ac87d0.

[12] Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, Trollor J, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep 2020;10:1–10.

[13] Garcia-Gutierrez et al. F. Diagnosis of Alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms. Int J Geriatr Psychiatry 2021;37.

[14] Gregory et al. CA. Can frontotemporal dementia and Alzheimer's disease be differentiated using a brief battery of tests? Int J Geriatr Psychiatry 1997;12.

[15] Battista P, Salvatore C, Castiglioni I. Optimizing Neuropsychological Assessments for Cognitive, Behavioral, and Functional Impairment Classification: A Machine Learning Study. Behav Neurol 2017;2017:1850909–1850909. https://doi.org/10.1155/2017/1850909.

[16] Baez S, Manes F, Huepe D, Torralva T, Fiorentino N, Richter F, et al. Primary empathy deficits in frontotemporal dementia. Front Aging Neurosci 2014;6:262. https://doi.org/10.3389/fnagi.2014.00262.
[17] Migeot JA, Duran-Aniotz CA, Signorelli CM, Piguet O, Ibanez A. A predictive coding framework of allostatic-interoceptive overload in frontotemporal dementia. Trends Neurosci 2022. https://doi.org/10.1016/j.tins.2022.08.005.
[18] Maito MA, Santamaría-García H, Moguilner S, Possin KL, Godoy ME, Avila-Funes JA, et al. Classification of Alzheimer's disease and frontotemporal dementia using routine clinical and cognitive measures across multicentric underrepresented samples: a cross sectional observational study. Lancet Reg Health - Am 2023;17:100387. https://doi.org/10.1016/j.lana.2022.100387.
[19] Ibanez A, Parra MA, Butler C. The Latin America and the Caribbean Consortium on Dementia (LAC-CD): From Networking to Research to Implementation Science. J Alzheimers Dis 2021;82:S379-s394. https://doi.org/10.3233/jad-201384.
[20] Kim R, Kim H-J, Kim A, Jang M-H, Kim HJ, Jeon B. Validation of the Conversion between the Mini-Mental State Examination and Montreal Cognitive assessment in Korean Patients with Parkinson's Disease. J Mov Disord 2018;11:30–4. https://doi.org/10.14802/jmd.17038.
[21] Matías-Guiu JA, Pytel V, Cortés-Martínez A, Valles-Salgado M, Rognoni T, Moreno-Ramos T, et al. Conversion between Addenbrooke's Cognitive Examination III and Mini-Mental State Examination. Int Psychogeriatr 2017;30:1227–33. https://doi.org/10.1017/s104161021700268x.
[22] Van Steenoven I, Aarsland D, Hurtig H, Chen‐Plotkin A, Duda JE, Rick J, et al. Conversion between Mini‐Mental State Examination, Montreal Cognitive Assessment, and Dementia Rating Scale‐2 scores in Parkinson's disease. Mov Disord 2014;29:1809–15. https://doi.org/10.1002/mds.26062.
[23] Roalf DR, Moberg PJ, Xie SX, Wolk DA, Moelter ST, Arnold SE. Comparative accuracies of two common screening instruments for classification of Alzheimer's disease, mild cognitive impairment, and healthy aging. Alzheimers Dement 2013;9:529–37. https://doi.org/10.1016/j.jalz.2012.10.001.
[24] Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 2007;6:734–46. https://doi.org/10.1016/S1474-4422(07)70178-3.
[25] Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 2011;134:2456–77. https://doi.org/10.1093/brain/awr179.
[26] Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. Neurology 2011;76:1006–14. https://doi.org/10.1212/WNL.0b013e31821103e6.
[27] Larose CDLDT. Data Mining and Predictive Analytics, 2nd Edition 2015.
[28] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.
[29] Nasreddine ZS, Phillips NA, BÃ©dirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. J Am Geriatr Soc 2005;53:695–9. https://doi.org/10.1111/j.1532-5415.2005.53221.x.
[30] Mathuranath PS, Nestor PJ, Berrios GE, Rakowicz W, Hodges JR. Addenbrooke's Cognitive Examination 2000. https://doi.org/10.1037/t75917-000.

[31] Torralva T, Roca M, Gleichgerrcht E, López P, Manes F. INECO Frontal Screening (IFS): A brief, sensitive, and specific tool to assess executive functions in dementia–CORRECTED VERSION. J Int Neuropsychol Soc 2009;15:777–86. https://doi.org/10.1017/s1355617709990415.
[32] Gleichgerrcht E, Roca M, Manes F, Torralva T. Comparing the clinical usefulness of the Institute of Cognitive Neurology (INECO) Frontal Screening (IFS) and the Frontal Assessment Battery (FAB) in frontotemporal dementia. J Clin Exp Neuropsychol 2011;33:997–1004. https://doi.org/10.1080/13803395.2011.589375.
[33] Pfeffer RI, Kurosaki TT, Harrah Jr C, Chance JM, Filos S. Measurement of functional activities in older adults in the community. J Gerontol 1982;37:323–9.
[34] Mahoney F, Barthel DW. Functional evaluation ; the Barthel index. A simple index of the independence useful in scoring improvement in the rehabilitation of the chronically ill. Md State Med J 1965;14:61–6.
[35] Cummings JL. The Neuropsychiatric Inventory: Assessing psychopathology in dementia patients. Neurology 1997;48:10S-16S. https://doi.org/10.1212/wnl.48.5_suppl_6.10s.
[36] Funkiewiez A, Bertoux M, de Souza LC, Lévy R, Dubois B. The SEA (Social Cognition and Emotional Assessment): A clinical neuropsychological tool for early diagnosis of frontal variant of frontotemporal lobar degeneration. Neuropsychology 2012;26:81–90. https://doi.org/10.1037/a0025318.
[37] Fittipaldi S, Legaz A, Maito M, Hernandez H, Altschuler F, Canziani V, et al. Heterogeneous factors influence social cognition across diverse settings in brain health and age-related diseases. Nat Ment Health 2024:1–13. https://doi.org/10.1038/s44220-023-00164-3.
[38] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[39] Donnelly-Kehoe PA, Pascariello GO, García AM, Hodges JR, Miller B, Rosen H, et al. Robust automated computational approach for classifying frontotemporal neurodegeneration: Multimodal/multicenter neuroimaging. Alzheimers Dement Diagn Assess Dis Monit 2019;11:588–98. https://doi.org/10.1016/j.dadm.2019.06.002.
[40] Gupta B A, Kahali. Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests. Alzheimers Dement 2020;1:e12049. https://doi.org/10.1002/trc2.12049.
[41] Grassi M, Perna G, Caldirola D, Schruers K, Loewenstein DA. A Clinically-Translatable Machine Learnin Algorithm for the Prediction of Alzheimers´s Disease Conversion in Individuals with Mild and Premild Cognitive Impairment. J Alzheimers Dis 2018;61:1555–73. https://doi.org/10.3233/jad-170547.
[42] Kwak S. LJY Oh DJ, Jeon YJ, Oh DY, Park SM, Kim H. Utility of Machine Learning Approach with Neuropsychological Tests in Predicting Functional Impairment of Alzheimer's Disease. J Alzheimers Dis 2022;3:1357–72. https://doi.org/10.3233/JAD-215244.
[43] Breiman L. Random Forest. Mach Learn 2001;45:5–32. https://doi.org/10.1023/a:1010933404324.
[44] Schölkopf B, Smola AJ. Learning with Kernels. The MIT Press; 2018.
[45] Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw 2018;3:638.
[46] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst., vol. 30, Curran Associates, Inc.; 2017.
[47] Miranda JJ, Barrientos-Gutiérrez T, Corvalan C, Hyder AA, Lazo-Porras M, Oni T, et al. Understanding the rise of cardiometabolic diseases in low-and middle-income

countries. Nat Med 2019;25:1667–79.
[48] Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol 2022;21:234–45. https://doi.org/10.1016/S1474-4422(21)00462-2.
[49] Baez S, M. GA, A. I. The Social Context Network Model in Psychiatric and Neurological Diseases. Curr Top Behav Neurosci 2016. https://doi.org/10.1007/7854_2016_443.
[50] Duran-Aniotz C, Orellana P, Leon Rodriguez T, Henriquez F, Cabello V, Aguirre-Pinto MF, et al. Systematic Review: Genetic, Neuroimaging, and Fluids Biomarkers for Frontotemporal Dementia Across Latin America Countries. Front Neurol 2021;12:663407. https://doi.org/10.3389/fneur.2021.663407.
[51] Santacruz Escudero JM, Beltrán J, Palacios Á, Chimbí CM, Matallana D, Reyes P, et al. Neuropsychiatric Symptoms as Predictors of Clinical Course in Neurodegeneration. A Longitudinal Study. Front Aging Neurosci 2019;11:176. https://doi.org/10.3389/fnagi.2019.00176.
[52] Park et al. LQ. Deficits in Everyday Function Differ in AD and FTD. Alzheimer Assoc Disord 2018;29:301–6.
[53] Borroni B, Alberici A, Agosti C, Premi E, Padovani A. Education plays a different role in Frontotemporal Dementia and Alzheimer's disease. Int J Geriatr Psychiatry 2008;23:796–800. https://doi.org/10.1002/gps.1974.
[54] Hodges JR, Piguet O. Progress and Challenges in Frontotemporal Dementia Research: A 20-Year Review. J Alzheimers Dis 2018;62:1467–80. https://doi.org/10.3233/jad-171087.
[55] Hutchinson AD, Mathias JL. Neuropsychological deficits in frontotemporal dementia and Alzheimer's disease: a meta-analytic review. J Neurol Neurosurg Psychiatry 2007;78:917–28.
[56] Arevalo-Rodriguez et al. I. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev 2015;3.
[57] Ismail Z, Smith EE, Geda Y, Sultzer D, Brodaty H, Smith G, et al. Neuropsychiatric symptoms as early manifestations of emergent dementia: Provisional diagnostic criteria for mild behavioral impairment. Alzheimers Dement 2016;12:195–202. https://doi.org/10.1016/j.jalz.2015.05.017.
[58] Bari Antor et al. M. A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. J Heal Eng 2021:9917919.
[59] Kharoubi R, Oualkacha K, Mkhadri A. The cluster correlation-network support vector machine for high-dimensional binary classification. J Stat Comput Simul 2019;89:1020–43. https://doi.org/10.1080/00949655.2019.1575382.
[60] Tan P-N, Steinbach M, Karpatne A, Kumar V. Introduction to data mining. 2019.
[61] Weakley. A, Williams. JA, Schmitter-Edgecombe. M, Cook. DJ. Neuropsychological test selection for cognitive impairment classification: A machine learning approach. J Clin Exp Neuropsychol 2015;37:899–916. https://doi.org/10.1080/13803395.2015.1067290.
[62] Gurevich P, Stuke H, Kastrup A, Stuke H, Hildebrandt H. Neuropsychological Testing and Machine Learning Distinguish Alzheimer's Disease from Other Causes for Cognitive Impairment. Front Aging Neurosci 2017;9. https://doi.org/10.3389/fnagi.2017.00114.

APÉNDICE

Tabla A1: Resultados para Regresión Logística, SVM y XGBoost

Modelo	Método de imputación	Conjunto	Técnica	Accuracy	ROC AUC	Precision	Sensitivity	Specificity	F1
LR	Media	Entrenamiento	Bootstrapping	0.81 (± 0.011)	0.814 (± 0.011)	0.762 (± 0.012)	0.287 (± 0.012)	0.971 (± 0.004)	0.414 (± 0.014)
		Validación	Out of sample	0.758	0.5817	0625	0.2	0.96	0.303
	Iterative Imputer	Entrenamiento	Bootstrapping	0.804 (± 0.011)	0.815 (± 0.011)	0.766 (± 0.011)	0.251 (± 0.012)	0.975 (± 0.004)	0.374 (± 0.013)
		Validación	Out of sample	0.757	0.7495	0.444	0.16	0.92	0.2353
	MICE	Entrenamiento	Bootstrapping	0.786 (± 0.011)	0.781 (±0.011)	0.662 (± 0.013)	0.199 (± 0.011)	0.967 (± 0.004)	0.302 (± 0.011)
		Validación	Out of sample	0.7664	0.5278	0.5	0.08	0.98	0.1379
SVM	Media	Entrenamiento	Bootstrapping	0.769 (± 0.012)	0.761 (± 0.012)	0.509 (± 0.014)	0.746 (± 0.012)	0.776 (± 0.011)	0.604 (± 0.013)
		Validación	Out of sample	0.8224	0.8146	0.5882	0.8	0.83	0.678
	Iterative Imputer	Entrenamiento	Bootstrapping	0.791 (± 0.011)	0.766 (± 0.011)	0.544 (± 0.013)	0.719 (± 0.012)	0.812 (± 0.011)	0.618 (± 0.013)
		Validación	Out of sample	0.6636	0.6137	0.3514	0.52	0.71	0.4194
	MICE	Entrenamiento	Bootstrapping	0.734 (± 0.012)	0.722 (± 0.012)	0.466 (± 0.014)	0.69 (± 0.013)	0.754 (±0.012)	0.555 (±0.014)
		Validación	Out of sample	0.7757	0.7563	0.5143	0.72	0.79	0.6
XGBoost	Media	Entrenamiento	Bootstrapping	0.836 (± 0.01)	0.854 (± 0.009)	0.708 (± 0.013)	0.531 (± 0.014)	0.931 (± 0.007)	0.604 (± 0.013)
		Validación	Out of sample	0.8972	0.8078	0.8889	0.64	0.98	0.7442
	Iterative Imputer	Entrenamiento	Bootstrapping	0.839 (± 0.011)	0.867 (± 0.009)	0.736 (±0.012)	0.503 (±0.014)	0.943 (± 0.006)	0.594 (± 0.013)
		Validación	Out of sample	0.7944	0.6295	0.6154	0.94	0.32	0.4211
	MICE	Entrenamiento	Bootstrapping	0.788 (± 0.011)	0.8 (± 0.011)	0.592 (± 0.014)	0.341 (± 0.013)	0.926 (± 0.007)	0.43 (± 0.014)
		Validación	Out of sample	0.8224	0.62	1	0.24	1	0.3871

Tabla A2 : Resultados para Random Forest, con datos imputados por media, para EA vs CN y DFT vs CN

Muestra	Conjunto	Técnica	Accuracy	ROC AUC	Precision	Sensitivity	Specificity	F1
EA vs CN	Entrenamiento	Bootstrapping	0.935 (± 0.007)	0.9769 (± 0.004)	0.961 (± 0.005)	0.932 (± 0.007)	0.94 (± 0.006)	0.946 (± 0.006)
	Validación	Out of sample	09265	0.9074	0.8913	1	0.81	0.9425
DFT vs CN	Entrenamiento	Bootstrapping	0.908 (± 0.008)	0.95 (± 0.006)	0.889 (± 0.009)	0.82 (± 0.01)	0.95 (± 0.006)	0.852 (± 0.01)
	Validación	Out of sample	0.8101	0.8181	0.6562	0.84	0.8	0.7368

EA: Enfermedad de Alzheimer, CN: Control sano, DFT: Demencia frontotemporal

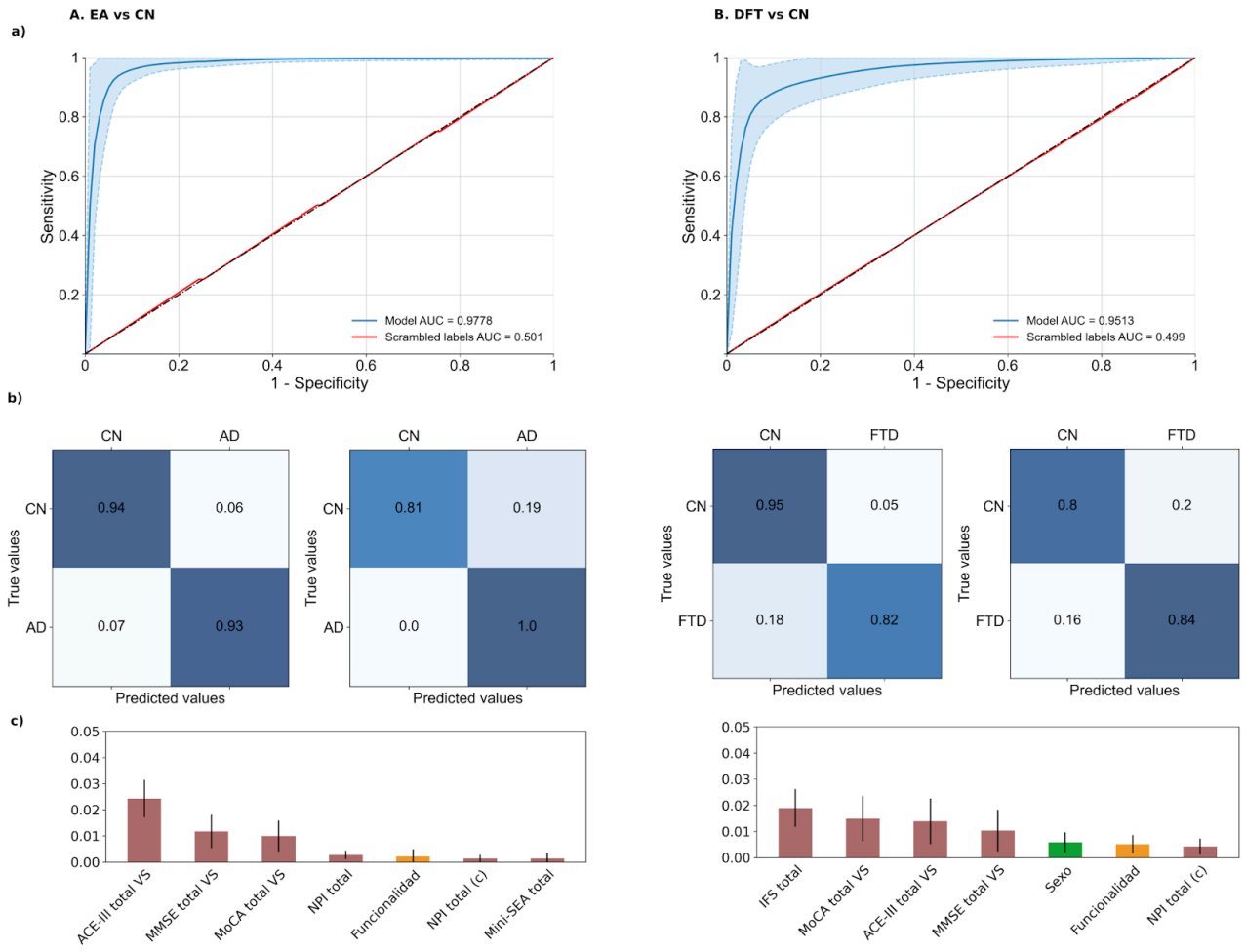

Fig A1: Resultados medios complementarios (5000 iteraciones) para Modelos de Random Forest entre participantes con enfermedad de Alzheimer y controles (**A. EA vs CN**) y participantes con demencia frontotemporal y controles (**B. DFT vs CN**). Los gráficos en **a)** muestran las curvas ROC con dos desviaciones estándar y su AUC. Los gráficos en **b)** muestran las matrices de confusión en magnitudes relativas para los conjuntos de test en bootstrapping a la izquierda y para los datos no vistos (validación) a la derecha. Los gráficos en **c)** muestran la importancia de cada característica de los modelos según los resultados de la permutación (5000 iteraciones, *permutation feature importance*). Para el **modelo A** las más importantes fueron el cribado cognitivo (ACE-III, MMSE y MoCA normalizados mediante el método Van-Steenoven), la funcionalidad, los síntomas neuropsiquiátricos (NPI) y la cognición social (Mini-SEA total). Para el modelo **B)** las más importantes fueron las funciones ejecutivas (IFS Total), el cribado cognitivo (MoCA, ACE-III y MMSE normalizados mediante el método Van-Steenoven), sexo, la funcionalidad, los síntomas neuropsiquiátricos (NPI) y la cognición social (Mini-SEA total). Para el modelo **B)** las más importantes fueron las funciones ejecutivas (IFS Total), el cribado cognitivo (MoCA, ACE-III y MMSE normalizados mediante el método Van-Steenoven), sexo, la funcionalidad y los síntomas neuropsiquiátricos (NPI total).