

UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS



DETECCIÓN DE FATIGA Y EXPLORACIÓN DE SU
ANTICIPACIÓN MEDIANTE INTELIGENCIA ARTIFICIAL Y
SEÑALES MULTIMODALES EN TRABAJADORES MINEROS

FELIPE ALONSO LIZAMA RODRÍGUEZ

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA

PROFESOR GUÍA: CARLA VAIRETTI

SANTIAGO, JUNIO DE 2026

Resumen

La fatiga laboral en la industria minera concentra aproximadamente el 8 % de las fatalidades laborales globales y motiva la búsqueda de sistemas de detección temprana basados en señales fisiológicas. Esta tesis evalúa algoritmos de aprendizaje de máquina para la detección de fatiga multidimensional en trabajadores mineros, en el marco del sistema *OmniFatigue Tracker*, mediante un trabajo experimental dividido en tres fases. La **Fase 1** implementa un *benchmark* comparativo de 14 modelos (5 clásicos y 9 de *Deep Learning*) sobre el dataset público FatigueSet (4 canales EEG, 12 sujetos), con análisis crítico de los esquemas de etiquetado del paper original; SVM alcanza Macro-F1 = 0,620 bajo el etiquetado por condición de intensidad. La **Fase 1-bis** aporta el *primer benchmark independiente* sobre el dataset MPD-DF (32 canales, 50 sujetos, etiquetado clínico): CatBoost alcanza F1 = 0,728, a la par del *benchmark* interno del paper (MSCNN-CAM, F1 = 0,705) y documentando el colapso de modelos *transformer* bajo desbalance extremo (94 %/6 %).

La **Fase 2** construye un experimento sobre la cohorte propietaria Sigma (16 canales OpenBCI + EmotiBit periférico, 10 sujetos efectivos sobre ciclos 7–11). Se triangulan siete esquemas de etiquetado y se adopta la fracción de épocas no-Wake estimada por YASA como **pseudo-etiqueta EEG-interna primaria**, justificada deductivamente por la replicación del fenómeno de *perceptual blindness* en mineros (concordancia nula con KSS: $\kappa = 0,001$, $\rho = 0,081$). El clasificador CatBoost sobre la pseudo-etiqueta alcanza Macro-F1 = 0,631 en validación LOSO, con generalización inter-sujeto estadísticamente significativa (control *shuffle within-subject*: $z = 5,84$ con sonda LDA y $z = 3,95$ con el propio CatBoost, $p_{\text{emp}} < 10^{-3}$).

Se evalúan ocho ablaciones independientes para superar este techo (*features* ricas con HRV/IMU *proper*, modelos secuenciales, *transfer learning* desde FatigueSet, ICA, fusión multimodal con seis configuraciones, geometría de Riemann, *foundation model* EEG CBraMod, y *foundation model* multimodal NormWear); ninguna mejora significativamente sobre el *baseline* EEG-único (Wilcoxon pareado $p > 0,05$ y *bootstrap* CI 95 % del delta que incluye cero). La convergencia de estas ocho direcciones de mejora (once contando las tres redes DL entrenables del *benchmark*) al mismo techo es consistente con un régimen donde el peso dominante del techo recae sobre el tamaño muestral (10 sujetos efectivos), más que sobre la representación de *features* o la arquitectura del modelo; es una lectura persuasiva, no deductiva.

Las dos hipótesis de trabajo se evalúan con respuesta matizada: H1 (fusión multimodal > unimodal) *no confirmada* en el régimen actual; H2 (transferibilidad de representaciones) con *evidencia direccional no concluyente*: la transferencia de pesos en Fase 1

sugiere una mejora preliminar, pero en la cohorte propietaria el *zero-shot* iguala al *baseline* en el nivel de azar (resultado nulo). Las contribuciones originales son cuatro: (i) pseudo-etiqueta EEG-interna validada para mineros sin *ground truth* subjetivo confiable; (ii) primer *benchmark* independiente de MPD-DF; (iii) resultado negativo robusto sobre fusión multimodal con estadística completa; (iv) pipeline reproducible de control de calidad integrando bitácora de analista. Las limitaciones se discuten explícitamente.

Abstract

Occupational fatigue in mining concentrates approximately 8 % of global workplace fatalities and motivates the search for early detection systems based on physiological signals. This thesis evaluates machine learning algorithms for multidimensional fatigue detection in mining workers, within the *OmniFatigue Tracker* system, through experimental work organized in three phases.

Phase 1 benchmarks 14 models (5 classical and 9 Deep Learning) on the public FatigueSet dataset (4 EEG channels, 12 subjects), with a critical analysis of the labeling schemes of the original paper; SVM reaches Macro-F1 = 0.620 under the intensity-condition labeling. **Phase 1-bis** contributes the *first independent benchmark* on the MPD-DF dataset (32 channels, 50 subjects, clinical labeling): CatBoost achieves F1 = 0.728, on par with the paper’s internal benchmark (MSCNN-CAM, F1 = 0.705) and documenting the collapse of transformer models under extreme class imbalance (94 %/6 %).

Phase 2 sets up an experiment on the proprietary Sigma cohort (16 OpenBCI EEG channels + peripheral EmotiBit, 10 effective subjects across cycles 7–11). Seven labeling schemes are triangulated and the fraction of non-Wake epochs estimated by YASA is adopted as the **primary EEG-internal pseudo-label**, deductively justified by replicating the *perceptual blindness* phenomenon in miners (null concordance with KSS: $\kappa = 0,001$, $\rho = 0,081$). A CatBoost classifier on the pseudo-label reaches Macro-F1 = 0.631 in LOSO, with statistically significant inter-subject generalization (within-subject shuffle control: $z = 5,84$ with an LDA probe and $z = 3,95$ with CatBoost itself, $p_{\text{emp}} < 10^{-3}$).

Eight independent ablations are evaluated to surpass this ceiling (richer features with proper HRV/IMU, sequential models, transfer learning from FatigueSet, ICA, multimodal fusion with six configurations, Riemannian geometry, EEG foundation model CBraMod, and multimodal foundation model NormWear); none significantly improves over the EEG-only baseline (paired Wilcoxon $p > 0,05$ and 95 % bootstrap CI of the delta including zero). The convergence of these eight improvement directions (eleven counting the three trainable DL networks of the benchmark) to the same ceiling is consistent with a regime where the dominant bottleneck is sample size (10 effective subjects) rather than feature representation or model architecture; this is a persuasive, not deductive, reading.

The two working hypotheses receive nuanced answers: H1 (multimodal > unimodal fusion) *not confirmed* in the current regime; H2 (transferability of representations) yields only *directional, inconclusive evidence*: Phase-1 weight transfer suggests a preliminary

gain, but on the proprietary cohort the zero-shot model matches the in-domain baseline at chance level (a null result). The four original contributions are: (i) an EEG-internal pseudo-label validated for miners without reliable subjective ground truth; (ii) the first independent benchmark of MPD-DF; (iii) a robust negative result on multimodal fusion with full statistics; (iv) a reproducible quality-control pipeline integrating the analyst’s signal-quality logbook. Limitations are discussed explicitly.

Agradecimientos

El desarrollo de esta tesis no habría sido posible sin el apoyo de muchas personas, a quienes quiero expresar mi gratitud.

A mi profesora guía, Carla Vairetti, le agradezco la confianza que depositó en este trabajo desde el comienzo y la claridad con que supo encauzar mis ideas cuando el rumbo no estaba del todo definido. A Sebastián Maldonado, por su mirada crítica y su rigor: varias de las decisiones técnicas de esta investigación se afinaron gracias a sus preguntas y observaciones.

A Sigma y a su equipo, por abrirme las puertas a un problema real y confiar en mí para trabajar con sus datos; buena parte de lo aprendido en este proceso nace de esa colaboración.

A mi familia, por sostenerme con paciencia y cariño a lo largo de todo el camino. A Vito, por acompañarme de cerca y darme calma en los días más exigentes. A mis amigos, por su aliento cuando más lo necesité. Y a los compañeros que conocí en la universidad, con quienes compartí el esfuerzo, las dudas y también las satisfacciones de esta etapa: gracias por hacer del trayecto algo mucho más llevadero.

Índice general

Resumen	III
Abstract	V
Agradecimientos	VII
1. Introducción	1
1.1. Contexto y Motivación	1
1.2. Planteamiento del Problema	3
1.3. Hipótesis	5
1.4. Objetivos	6
1.4.1. Objetivo General	6
1.4.2. Objetivos Específicos	6
1.5. Contribuciones	7
1.6. Estructura del Documento	9
2. Revisión Bibliográfica	10
2.1. Fundamentos Fisiológicos de la Fatiga	10
2.2. Modelos de Aprendizaje para Señales Fisiológicas	11
2.2.1. Clasificadores Clásicos sobre Características Diseñadas	12
2.2.2. Modelos Secuenciales y Convolucionales	13
2.3. Señales, Datasets y Etiquetado de la Fatiga	14
2.4. Transfer Learning y Adaptación de Dominio	15
2.5. Fusión Multimodal y Modelos Fundacionales para Señales Fisiológicas	16
2.5.1. Estrategias de Fusión para Bioseñales	17
2.5.2. Modelos Fundacionales Multimodales	18
2.5.3. Predicción Temporal con Anticipación	19
2.5.4. Robustez ante Modalidades Faltantes	20
2.6. Evaluación y Validación de Modelos de Fatiga	20
2.7. Síntesis y Brecha Detectada	22

3. Metodología y Diseño Experimental	25
3.1. Visión General y Estrategia de Investigación	25
3.2. Fase 1: Validación en Datasets Públicos y Transferencia	27
3.2.1. Datasets Utilizados	27
3.2.2. Selección de Arquitecturas y Benchmarking	28
3.2.3. Baselines de Machine Learning Clásico	29
3.2.4. Protocolo de Transfer Learning	32
3.3. Fase 1-bis: Validación en Dataset MPD-DF (32 canales)	33
3.3.1. Dataset MPD-DF	33
3.3.2. Diseño Experimental	33
3.3.3. Configuración Robusta para Modelos Transformer en Régimen Desbalanceado	34
3.4. Fase 2: Diseño Experimental OmniFatigue Tracker (Laboratorio)	35
3.4.1. Hipótesis Operacional	35
3.4.2. Alcance del Protocolo Ejecutado	35
3.4.3. Participantes y Entorno	35
3.4.4. Consideraciones Éticas	39
3.4.5. Adquisición de Señales (Multimodalidad)	39
3.4.6. Protocolo de Inducción y Etiquetado	43
3.4.7. Bitácora de Calidad de Señal y Filtrado	46
3.4.8. Representación de los Datos para el Modelado	48
3.4.9. Estrategia de Fusión Multimodal	49
3.4.10. Análisis Estadístico: Operacionalización de las Hipótesis	51
3.4.11. Implementación y Reproducibilidad	52
4. Resultados y Discusión	54
4.1. Resultados del Benchmark (Fase 1)	54
4.1.1. Esquema de Etiquetado y Balance de Clases	54
4.1.2. Rendimiento Comparativo de Arquitecturas	55
4.1.3. Análisis de Resultados	56
4.1.4. Comparación con Baselines Clásicos de Machine Learning	57
4.1.5. Validación por Permutación: ¿Qué Etiqueta Recupera Señal Real?	58
4.1.6. Validación de Transfer Learning	61
4.1.7. Discusión de la Fase 1	62
4.2. Resultados de Fase 1-bis: Dataset MPD-DF (32 canales, 50 sujetos)	62
4.2.1. ML Clásico: 32 canales vs. 16 canales	63

4.2.2.	Benchmark de Deep Learning: 32 canales	65
4.2.3.	Análisis del Colapso y Recuperación de Modelos Transformer . . .	65
4.2.4.	Ranking Consolidado: DL vs. ML Clásico	66
4.3.	Direcciones Derivadas de la Fase 1-bis	67
4.4.	Datos del Laboratorio y Esquema de Etiquetado (Fase 2)	67
4.4.1.	Alcance del Protocolo Ejecutado	68
4.4.2.	Cobertura de <i>Ground Truth</i>	69
4.4.3.	Bitácora de Calidad de Señal	70
4.4.4.	Estrategias de Etiquetado: Triangulación de Candidatos	70
4.4.5.	Validación: Tres Experimentos antes de Comprometer un Label Primario	71
4.4.6.	Decisión del Label Primario: la Pseudo-etiqueta EEG-interna de YASA	76
4.4.7.	Robustez del Label Primario: Sensibilidad al Threshold y Con- found TOD	79
4.4.8.	Cobertura Horaria y Estructura Temporal del Dataset	84
4.4.9.	Pipeline de Extracción de <i>Features</i> y Validación con LOSO . . .	85
4.4.10.	Interpretabilidad del Modelo Primario	88
4.4.11.	Comparación con Fase 1 y Fase 1-bis	90
4.4.12.	Ablación de Fusión Multimodal sobre la Cohorte Consolidada .	91
4.4.13.	Ablaciones con codificadores neurales: covarianza Riemann, foun- dation models EEG y multimodal	94
4.4.14.	Benchmark Comprehensivo de Modelos sobre la Pseudo-etiqueta YASA	99
4.4.15.	Anticipación con EmotiBit durante el bloque EMOTI	105
5.	Conclusiones	113
5.1.	Síntesis de los Hallazgos	113
5.2.	Conclusiones Principales	114
5.3.	Evaluación de las Hipótesis	121
5.4.	Limitaciones	123
5.5.	Plan de Trabajo Futuro	125
	Nomenclatura	128
	Referencias	129

Anexos	137
A. Resultados completos de Fase 1 (FatigueSet)	137
A.1. Etiquetado por condición de intensidad	137
A.2. Etiquetado por VAS (M0/M1/M2)	139
A.3. Ranking consolidado de modelos (intensidad binario)	140
B. Material suplementario de la Fase 2	141
B.1. Cobertura horaria y estructura temporal de los bloques	141
B.2. Ablaciones detalladas del experimento de anticipación EMOTI	143

Índice de figuras

1.1. Tendencia de accidentalidad minera en Chile	3
1.2. Correspondencia entre desafíos y contribuciones	8
2.1. Taxonomía de estrategias de fusión multimodal	17
2.2. Mecanismo de atención cruzada bidireccional	18
2.3. Evolución de los <i>foundation models</i> de bioseñales	19
3.1. El proceso KDD instanciado en la metodología	25
3.2. Flujo de datos de las tres fases experimentales	26
3.3. Validación cruzada <i>subject-aware</i> / LOSO	30
3.4. Entorno experimental del laboratorio Sigma	36
3.5. Montaje EEG de 16 canales del sistema Sigma	41
3.6. Montaje EEG de 16 canales (sistema 10-20)	41
3.7. Estación de monitoreo y adquisición en tiempo real	42
3.8. Generación y triangulación de las etiquetas de la Fase 2	47
3.9. Arquitectura de fusión multimodal tardía	50
4.1. Test de permutación por esquema de etiquetado (Fase 1)	59
4.2. Timeline ilustrativo de las estrategias de etiquetado	71
4.3. Sensibilidad del Macro-F1 al umbral τ y a <code>min_epochs</code>	81
4.4. Curvas ROC del label primario	87
4.5. Resumen SHAP global del modelo primario	89
4.6. Ablación por grupos de <i>features</i> del modelo primario	90
4.7. Forest plot de las ablaciones DL	97
4.8. <i>Benchmark</i> de 13 modelos sobre la pseudo-etiqueta YASA	101
4.9. Convergencia de los modelos DL entrenables (<i>fold sub_04</i>)	103
4.10. Curva de aprendizaje Macro-F1 vs. número de sujetos	105
5.1. Síntesis de las contribuciones originales	114
5.2. Roadmap de trabajo futuro	125

Índice de cuadros

2.1. Comparación de datasets públicos de fatiga por conducción o tarea sostenida.	15
3.1. Composición de la cohorte analizada (Fase 2)	38
3.2. Caracterización demográfica y clínica de la cohorte Sigma	39
3.3. Bloques de medición ejecutados en cada momento del día.	43
4.1. Benchmark DL sobre FatigueSet (etiquetado por intensidad)	55
4.2. Baselines clásicos vs. mejor DL sobre FatigueSet (intensidad)	57
4.3. Distribución de clases por esquema de etiquetado (FatigueSet)	59
4.4. Test de permutación sobre los esquemas válidos de FatigueSet	60
4.5. Comparación de rendimiento utilizando <i>Transfer Learning</i> con Deep4Net.	61
4.6. Clasificadores clásicos sobre MPD-DF (32 canales)	63
4.7. Comparación de densidad de canales: 32ch vs. subconjunto de 16ch (ML clásico, MPD-DF).	64
4.8. Benchmark DL sobre MPD-DF (32 canales)	65
4.9. Ranking consolidado de los 14 modelos sobre MPD-DF	66
4.10. Cobertura de <i>ground truth</i> y mediciones auxiliares (Fase 2)	69
4.11. Mejor modelo por etiqueta candidata en LOSO (Fase 2)	73
4.12. <i>Shuffle control within-subject</i> por etiqueta candidata	75
4.13. Comparación de los dos candidatos de etiqueta testables	76
4.14. Sensibilidad del label primario al threshold y al filtro de épocas	80
4.15. Control del confound de estado ocular	84
4.16. Métricas operacionales del label primario (AUC, sensibilidad, especificidad)	86
4.17. Comparación del rendimiento del mejor clasificador clásico entre fases.	90
4.18. Ablación de fusión multimodal (B0–B5) sobre la pseudo-etiqueta YASA	92
4.19. Ablaciones DL (Riemann, CBraMod, NormWear) vs. CatBoost B0	97
4.20. Benchmark comprehensivo de 13 modelos sobre la pseudo-etiqueta YASA	100
4.21. Estudio de viabilidad del dataset de pares EMOTI	107

4.22. CatBoost LOSO sobre los pares EMOTI	108
4.23. Transferencia FatigueSet $\rightarrow$ EMOTI	110
5.1. Veredicto sobre hipótesis (H) y objetivos específicos (OE) a la luz de la evidencia.	121
A.1. Clasificadores clásicos sobre FatigueSet (intensidad binaria)	137
A.2. Intensidad de 3 clases (FatigueSet)	138
A.3. Benchmark DL sobre FatigueSet (intensidad binaria)	138
A.4. Clasificadores clásicos sobre FatigueSet (etiquetado VAS)	139
A.5. Arquitecturas DL sobre FatigueSet (etiquetado VAS)	139
A.6. Ranking consolidado sobre FatigueSet (intensidad binaria)	140
B.1. Horas de EEG procesado por categoría de calidad	142
B.2. Distribución de duraciones por bloque corto con EEG	143
B.3. Ablación de modelo y <i>features</i> sobre Δ Rehacom (EMOTI)	144

Capítulo 1

Introducción

1.1. Contexto y Motivación

La fatiga laboral constituye uno de los factores de riesgo más relevantes para la seguridad en industrias de alta exigencia operacional. [Williamson et al. \(2011\)](#) revisaron, a partir de evidencia epidemiológica, tres mecanismos candidatos (la presión homeostática del sueño, las fluctuaciones circadianas y el tiempo sostenido en tarea), de los cuales solo la presión homeostática del sueño mostró evidencia clara de producir accidentes laborales. A nivel global, el *National Safety Council* estimó que el 13 % de los accidentes ocupacionales son atribuibles a fatiga y que el 97 % de los trabajadores presentan al menos un factor de riesgo de fatiga, con un costo para los empleadores de USD 136 mil millones anuales, del cual el 84 % corresponde a *presentismo* (es decir, trabajadores que asisten a sus funciones en estado de fatiga) y no a ausentismo ([National Safety Council, 2017](#)). A escala macroeconómica, [Hafner et al. \(2016\)](#) cuantificaron en cinco países OCDE (EE.UU., Reino Unido, Canadá, Alemania y Japón) que la privación de sueño cuesta a Estados Unidos entre USD 280 y 411 mil millones anuales (1,56–2,28 % del PIB), evidenciando que el problema trasciende el ámbito individual para convertirse en una carga sistémica.

La industria minera es particularmente vulnerable a este fenómeno. A nivel mundial, la minería concentra aproximadamente el 8 % de las fatalidades laborales, a pesar de representar una fracción menor del empleo total ([Cruz-Ausejo, Cama-Ttito, Flores Solano, Copez-Lonzoy & Vera-Ponce, 2024](#)). Los turnos rotativos extendidos, la exposición a

altura geográfica, la monotonía operacional y las condiciones ambientales adversas (oscuridad, ruido, polvo) configuran un escenario donde la fatiga se acumula de forma acelerada y sus consecuencias son potencialmente catastróficas (Bauerle, Sammarco, Dugdale & Dawson, 2021). Los Sistemas de Gestión de Riesgo de Fatiga (*Fatigue Risk Management Systems*, FRMS) han demostrado efectividad en sus componentes individuales (educación, regulación de turnos, monitoreo), pero la evidencia sobre su eficacia como sistema integrado es aún limitada (Sprajcer, Thomas, Sargent, Crowther, Boivin, Wong, Smiley & Dawson, 2022). En particular, Caldwell et al. (2019) identifican el monitoreo fisiológico en tiempo real como la frontera actual de estos sistemas, señalando que las estrategias reactivas, que detectan fatiga cuando el operador ya está comprometido, son insuficientes para entornos de alto riesgo.

En Chile, la minería es el pilar de la economía nacional y la seguridad de sus trabajadores ha sido una prioridad creciente. Según las estadísticas oficiales del Servicio Nacional de Geología y Minería, en 2022 se registraron 2.660 accidentes laborales en la industria extractiva minera, de los cuales 12 resultaron en víctimas fatales (0,45 %), 32 produjeron lesiones graves y 729 (27,4 %) generaron tiempo perdido para los trabajadores (SERNAGEOMIN, 2025; SERNAGEOMIN, 2023a). La tasa de frecuencia, medida como accidentes con tiempo perdido por millón de horas-hombre trabajadas, descendió sostenidamente de 1,53 en 2019 a 1,33 en 2022, y continuó reduciéndose a 1,22 en 2023 y 1,05 en 2024 (SERNAGEOMIN, 2023b; SERNAGEOMIN, 2025). La tasa de fatalidad ha experimentado una reducción sostenida en las últimas dos décadas, pasando de 45 fallecidos en 2010 a 11 en 2024 (SERNAGEOMIN, 2025). Esta mejora se atribuye a la incorporación de nuevas políticas y procedimientos de seguridad, así como a la monitorización visual por parte de terceros. No obstante, la Política Nacional Minera 2050 establece como Meta 23 la aspiración de cero fatalidades, alineada con estándares internacionales, y como Meta 24 la reducción de un 75 % en la tasa de severidad y un 45 % en la frecuencia de accidentes al año 2030 (Ministerio de Minería, 2022). Alcanzar estas metas (Figura 1.1) requiere complementar las estrategias regulatorias existentes con tecnologías de detección temprana que permitan intervenir *antes* de que el estado de fatiga comprometa la seguridad del trabajador.

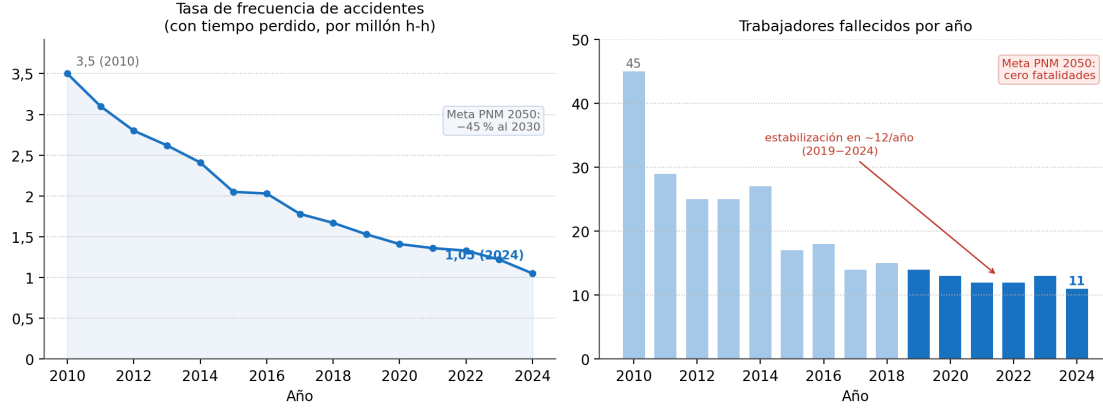


Figura 1.1: Evolución de la seguridad en la minería chilena, 2010–2024 (SERNAGEOMIN). **Izquierda:** la tasa de frecuencia de accidentes con tiempo perdido (por millón de horas-hombre) cae de forma sostenida de 3,5 (2010) a 1,05 (2024); la Política Nacional Minera 2050 (Meta 24) exige una reducción adicional del 45 % al 2030. **Derecha:** los trabajadores fallecidos descienden de 45 (2010) a 11 (2024), pero la reducción se desacelera y se estabiliza en torno a ~12 al año desde 2019, sin alcanzar la meta de cero fatalidades (Meta 23). Esa meseta sobre cero, que las estrategias regulatorias convencionales no han cerrado, motiva las tecnologías de detección temprana de fatiga evaluadas en esta tesis.

1.2. Planteamiento del Problema

A pesar de la evidencia que vincula la fatiga con accidentes laborales, los sistemas de detección multimodal de fatiga en el contexto específico de la minería son escasos y presentan limitaciones significativas. La industria minera dispone de soluciones comerciales desplegadas operacionalmente, como sistemas basados en oculometría infrarroja (Anderson, Chang, Sullivan, Ronda & Czeisler, 2013) o EEG monocanal integrado en cascos, pero estos operan con una sola modalidad, funcionan de forma reactiva y no publican sus arquitecturas ni datos de entrenamiento, lo que impide su reproducibilidad científica. En el ámbito académico, la mayoría de los estudios se circunscriben a una sola señal fisiológica y a poblaciones no representativas del sector industrial (Hooda, Joshi & Shah, 2021). Recientemente, Pan et al. (2025) propusieron un sistema de fusión multimodal para mineros que integra ECG, EDA, presión arterial, SpO2, temperatura cutánea e imágenes faciales mediante el modelo *Transformer+*, que procesa un flujo fisiológico y uno facial ($accuracy = 93,15\% \pm 3,54$ en 4 categorías), y Chen et al. (2025) demostraron la fusión cross-modal de EEG y fNIRS en conductores mineros con atención bidireccional (97,92% de $accuracy$). Sin embargo, ambos estudios se limitan

a validación en laboratorio con datasets propietarios no publicados, abordan exclusivamente la somnolencia (sin considerar la fatiga física ni cognitiva), y no evalúan el comportamiento del sistema ante la pérdida de sensores.

Desde la perspectiva fisiológica, la fatiga tiene una base mecanicista bien caracterizada. El modelo de dos procesos de [Borbély \(1982\)](#) describe cómo la presión homeostática del sueño (Proceso S) se acumula progresivamente durante la vigilia, mientras que el ritmo circadiano (Proceso C) modula la propensión al sueño de forma cíclica. Esta acumulación se manifiesta en biomarcadores electrofisiológicos específicos: [Lal & Craig \(2001\)](#) vinculan los incrementos en las bandas theta y delta del electroencefalograma (EEG) con la transición hacia estados de fatiga. Estos fundamentos, que se desarrollan en detalle en el Capítulo 2, justifican el uso de señales fisiológicas, y del EEG en particular, como base para sistemas de detección automática.

Sin embargo, la implementación de estos principios como un sistema operacional para minería enfrenta cuatro desafíos concretos:

- **Escasez de datos etiquetados:** La adquisición de datos fisiológicos en contextos industriales es costosa y requiere protocolos prolongados, lo que limita el volumen de datos disponibles para entrenar modelos de aprendizaje de máquina. Los dos estudios multimodales en minería ([Pan, Tong, Wei & Teng, 2025](#); [Chen, Liu, Lu & Chai, 2025](#)) utilizan datasets propietarios no publicados, y no se encontró ningún dataset público de fatiga en contextos de minería o industria pesada. El dataset ocupacional más cercano es MEFAR ([Derdiyok, Patlar Akbulut & Catal, 2024](#)), que registra EEG monocanal (NeuroSky MindWave, EEG raw a 512 Hz pero distribuido como “*attention*” y “*meditation*” agregados a 1 Hz) y señales periféricas de 23 trabajadores de cuatro grupos ocupacionales (académicos, técnicos, ingenieros informáticos y personal de cocina), un contexto distante de las condiciones operativas mineras y con una fidelidad de EEG muy inferior a la requerida por arquitecturas multicanal. La presente tesis trabaja con una cohorte propia de 11 trabajadores mineros (10 efectivos tras filtros de calidad), tamaño consistente con la literatura ocupacional pero que constituye la principal limitación estructural del trabajo y se discute explícitamente en el capítulo de Conclusiones.
- **Heterogeneidad de señales:** Un sistema multimodal debe integrar señales con

frecuencias de muestreo dispares (EEG a 125–256 Hz, fotopletimografía a 25–64 Hz, actigrafía a 25 Hz), lo que requiere estrategias de fusión que respeten las escalas temporales de cada modalidad.

- **Fatiga multidimensional:** Los sistemas existentes se enfocan en la somnolencia como única dimensión de fatiga. En entornos mineros, la fatiga física (por carga muscular sostenida) y la fatiga cognitiva (por tareas repetitivas de alta consecuencia) coexisten y requieren ser detectadas de forma integrada.
- **Robustez operacional:** En entornos mineros, la pérdida de señal por sudoración, desconexión de electrodos o agotamiento de batería es inevitable. Un sistema práctico debe tolerar la ausencia parcial de modalidades sin degradar catastróficamente su rendimiento.

1.3. Hipótesis

En función de la evidencia revisada, se plantean las siguientes hipótesis de trabajo:

- a) **Detección multimodal:** La integración del EEG cortical con señales fisiológicas periféricas (PPG, EDA, IMU, temperatura) y con señales derivadas del propio EEG (parpadeo desde canales frontales, EMG facial desde canales temporales) permite clasificar estados de fatiga con un rendimiento superior al de cualquier modalidad individual.
- b) **Transferibilidad de representaciones:** Las representaciones latentes de fatiga aprendidas en un dominio (datasets públicos) son transferibles a un dominio distinto (datos de laboratorio con trabajadores mineros), reduciendo la necesidad de grandes volúmenes de datos etiquetados.

Una precisión necesaria sobre el alcance de estas hipótesis en la cohorte propietaria (Sigma): el *perceptual blindness* documentado en trabajadores mineros (la baja concordancia entre fatiga auto-reportada y desempeño objetivo) impide disponer de un *ground truth* de fatiga subjetiva confiable (se replica en la Sección 4.4.5). Por ello, la “fatiga” se operacionaliza en la Fase 2 mediante una **pseudo-etiqueta EEG-interna** derivada de un *sleep stager* validado (YASA), y los objetivos se evalúan contra ese marcador, no

contra fatiga vivenciada ni desempeño cognitivo. Esta distinción (desarrollada y acotada en la Sección 4.4.6) es deliberada y condiciona toda interpretación posterior: el sistema aprende a predecir un marcador electroencefalográfico de adormecimiento, no la experiencia subjetiva de fatiga del trabajador.

1.4. Objetivos

1.4.1. Objetivo General

Diseñar e implementar algoritmos de inteligencia artificial basados en aprendizaje de máquina para la detección de fatiga multidimensional (física, cognitiva y somnolencia) a partir de señales biológicas multimodales, en el marco del proyecto *OmniFatigue Tracker*, evaluando su rendimiento tanto en datasets públicos como en datos propietarios adquiridos en un laboratorio de realidad mixta de Sigma con simulación de minería.

1.4.2. Objetivos Específicos

- a) Caracterizar el dataset adquirido por Sigma (EEG 16 canales del sistema OpenBCI Cyton+Daisy, señales periféricas del dispositivo EmotiBit) e implementar un *pipeline* reproducible de control de calidad que integre la bitácora de revisión manual del analista de señal.
- b) Diseñar y validar un esquema de etiquetado robusto para el dataset Sigma mediante triangulación de fuentes (auto-reporte KSS y SOFI-SM, tiempos de reacción Rehacom, label EEG-driven derivado del sleep stager YASA, consensus voting y clustering), comparando su rendimiento mediante un *benchmark* LOSO y un control de *shuffle within-subject*.
- c) Evaluar y comparar arquitecturas de *Deep Learning* y clasificadores clásicos de *Machine Learning* para la detección de fatiga con EEG, identificando las características espectrales y temporales más discriminantes.
- d) Cuantificar la contribución de cada modalidad fisiológica al rendimiento de clasificación mediante fusión tardía y ablación por modalidad, incluyendo las derivaciones EOG y EMG obtenidas del propio EEG.

- e) Validar la transferibilidad de las representaciones de fatiga aprendidas entre dominios (datasets públicos FatigueSet y MPD-DF, datos propietarios de Sigma), evaluando estrategias de *Transfer Learning* y *fine-tuning*.

1.5. Contribuciones

Las contribuciones originales de la presente tesis son cuatro:

- a) **Pseudo-etiqueta EEG-interna validada para detección de fatiga en mineros sin *ground truth* subjetivo confiable.** Se construye y valida una etiqueta supervisada derivada del *sleep stager* AASM-trained YASA Vallat & Walker (2021) aplicada *off-label* a registros EEG diurnos cortos de trabajadores mineros, con justificación deductiva basada en la replicación del fenómeno de *perceptual blindness* reportado por Purto et al. (2024). La pseudo-etiqueta alcanza Macro-F1 = 0,631 en LOSO con CatBoost sobre 10 sujetos, con generalización inter-sujeto estadísticamente significativa (control *shuffle within-subject*: $z = 5,84$ con sonda LDA y $z = 3,95$ con el propio CatBoost, $p_{\text{emp}} < 10^{-3}$). La distinción terminológica respecto del concepto de *fatiga subjetiva* es deliberada: la pseudo-etiqueta es un marcador EEG-interno, no una medida de fatiga vivenciada.
- b) **Primer *benchmark* independiente del dataset MPD-DF.** Se evalúan 14 modelos (5 clásicos + 9 *deep learning*) sobre el dataset MPD-DF (Li, Fu, Tang, Zhou, Chen, Zhou, Chen & Luo, 2026); CatBoost alcanza Macro-F1 = 0,728 sobre 50 sujetos y 32 canales, a la par del *benchmark* interno del paper original (MSCNN-CAM, F1 = 0,705; Macro-F1 = 0,730 al reproducir bajo la misma binarización del paper) y demostrando que el *ML clásico* con *features handcrafted* es competitivo con las arquitecturas DL evaluadas en el régimen de cohortes pequeñas de fatiga ocupacional.
- c) **Resultado negativo robusto sobre fusión multimodal en régimen de cohorte pequeña.** Sobre la cohorte Sigma se evalúan ocho ablaciones independientes (*features* ricas con HRV/IMU *proper*, modelos secuenciales CNN 1D, *transfer learning*, ICA, fusión multimodal con seis configuraciones, geometría de Riemann, *foundation model* EEG CBraMod y *foundation model* multimodal NormWear) sin

que ninguna mejore significativamente sobre el *baseline* CatBoost B0 (EEG-only, $F1 = 0,602 \pm 0,105$). El resultado se reporta con estadística completa (Wilcoxon pareado, *bootstrap* CI 95 %, tamaño del efecto Cohen's *d*) y se interpreta como consistente con un régimen donde el cuello de botella dominante es el tamaño muestral, más que la calidad de la representación (lectura persuasiva, no deductiva; once direcciones convergentes contando las tres redes DL entrenables del *benchmark* comprehensivo).

- d) **Pipeline reproducible de control de calidad con bitácora de analista.** Se integra la bitácora estructurada de calidad de señal del analista de Sigma (ciclos 7 a 12, 474 sesiones) con el *pipeline* de filtrado y modelado, una práctica de *control de calidad* pocas veces aplicada en estudios EEG ocupacionales publicados. El cruce reduce el universo de entrenamiento de 123,9 a 89,7 horas en categoría *Usable* (72 %) y separa señal genuina de ruido por mal posicionamiento del casco, electrodos sueltos o artefactos de movimiento.

Cada contribución responde a uno o más de los desafíos planteados, como resume la Figura 1.2.

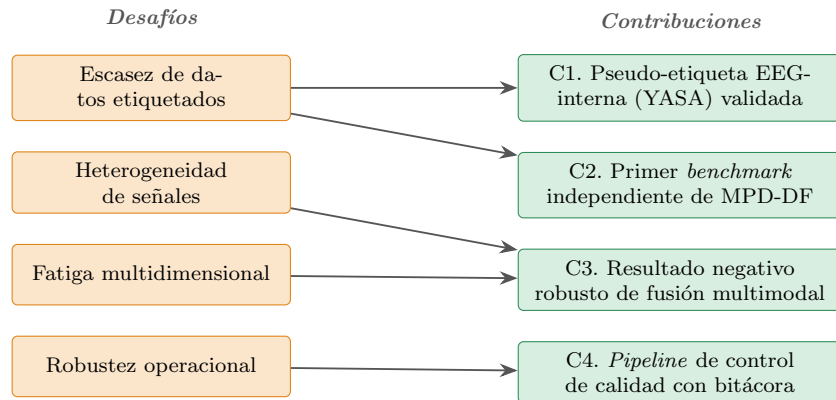


Figura 1.2: Correspondencia entre los cuatro desafíos del problema y las cuatro contribuciones originales de la tesis. La escasez de datos etiquetados motiva la pseudo-etiqueta EEG-interna y el *benchmark* independiente; la heterogeneidad de señales y la naturaleza multidimensional de la fatiga se abordan mediante la evaluación de fusión multimodal; y la robustez operacional, mediante el *pipeline* de control de calidad.

1.6. Estructura del Documento

El presente documento se organiza de la siguiente manera. El Capítulo 2 presenta la revisión bibliográfica, analizando el estado del arte en detección de fatiga, las señales y *features* utilizadas, las arquitecturas de clasificación, y las estrategias de fusión multimodal. El Capítulo 3 describe la metodología experimental, incluyendo los datasets utilizados, las arquitecturas evaluadas, el protocolo de adquisición de datos en laboratorio y el esquema de etiquetado adoptado; cierra con una sección de *Análisis Estadístico* que operacionaliza los criterios de evaluación de las hipótesis declaradas en este capítulo. El Capítulo 4 presenta los resultados experimentales de las tres fases del trabajo (FatigueSet, MPD-DF y Sigma), el análisis del esquema de etiquetado, las ablaciones convergentes y la fusión multimodal. Finalmente, el Capítulo 5 sintetiza los hallazgos, responde a las hipótesis, declara las limitaciones y propone las direcciones de iteración futura. Las cuatro contribuciones originales de la tesis (Sección 1.5) se traducen al final del documento en un Resumen y un Abstract complementarios.

Capítulo 2

Revisión Bibliográfica

La detección temprana de la fatiga se ha convertido en un área de creciente interés científico y tecnológico, especialmente en industrias de alto riesgo como la minería. La convergencia de sensores vestibles, el internet de las cosas (IoT) y la inteligencia artificial ha abierto nuevas fronteras para el monitoreo fisiológico en tiempo real. Esta revisión bibliográfica analiza el estado del arte en la detección de fatiga mediante aprendizaje de máquina, explorando los enfoques metodológicos, las señales biológicas utilizadas y las técnicas de modelado. El objetivo es contextualizar la presente investigación, identificar las brechas existentes en la literatura y fundamentar las decisiones metodológicas adoptadas para el desarrollo del sistema *OmniFatigue Tracker*.

2.1. Fundamentos Fisiológicos de la Fatiga

La fatiga no constituye un estado único, sino un fenómeno **multidimensional** cuyas componentes tienen orígenes y manifestaciones distintas. La literatura ocupacional distingue al menos tres dimensiones: la **fatiga física**, asociada a la carga muscular y metabólica sostenida; la **fatiga cognitiva**, derivada de la demanda atencional y de control ejecutivo prolongada; y la **somnolencia**, ligada a la presión homeostática del sueño y al ritmo circadiano. Estas dimensiones pueden coexistir y disociarse: un trabajador puede exhibir somnolencia elevada sin fatiga muscular, o agotamiento físico con la vigilancia preservada. Esta heterogeneidad motiva un enfoque de detección que no asuma una sola dimensión, como el que adopta el sistema *OmniFatigue Tracker*.

El sustrato de la somnolencia está bien caracterizado por el **modelo de dos proce-**

sos de Borbély (1982), que descompone la propensión al sueño en dos componentes interactuantes: el *Proceso S* (presión homeostática del sueño), que se acumula de forma monótona durante la vigilia y se disipa durante el sueño, y el *Proceso C* (ritmo circadiano), que modula cíclicamente la alerta a lo largo del día. La acumulación del Proceso S durante una jornada laboral predice un aumento progresivo de la fatiga, base teórica de los esquemas de etiquetado por momento del día (AM/MED/PM) considerados en esta tesis.

A nivel electrofisiológico, la transición hacia estados de fatiga se manifiesta en cambios espectrales medibles del electroencefalograma (EEG). En su revisión, Lal & Craig (2001) señalan que los incrementos de potencia en las bandas *theta* (4–8 Hz) y *delta* (0,5–4 Hz), junto con cambios en *alpha* (8–13 Hz), están fuertemente vinculados a la disminución de la vigilancia, e identifican al EEG como uno de los indicadores psicofisiológicos más prometedores. Complementariamente, Klimesch (1999) documentó el papel de las oscilaciones *alpha* y *theta* en el rendimiento cognitivo y la memoria de trabajo, lo que respalda su uso como biomarcadores de carga mental. Sobre esta base, ciertos índices espectrales compuestos se han consolidado como marcadores prácticos de somnolencia: el ratio $(\theta + \alpha)/\beta$, evaluado por Jap et al. (2009) entre varios algoritmos de detección de fatiga al volante, y el ratio θ/α , que aumenta a medida que disminuye la alerta. Estos fundamentos justifican tres decisiones metodológicas de la presente tesis: (i) el uso del EEG como señal central de detección; (ii) la extracción de *features* espectrales explícitas (potencias por banda y ratios como $(\theta + \alpha)/\beta$ y θ/α) además de las representaciones aprendidas *end-to-end*; y (iii) el tratamiento de la fatiga como fenómeno multidimensional al diseñar el esquema de etiquetado y la fusión multimodal.

2.2. Modelos de Aprendizaje para Señales Fisiológicas

Los enfoques de detección de fatiga abarcan desde clasificadores clásicos sobre características diseñadas manualmente hasta redes profundas que aprenden representaciones de la señal cruda. La revisión comprehensiva de Hooda et al. (2021), que sintetiza 50 estudios, organiza el campo en cuatro familias (modelos matemáticos, basados en reglas, *machine learning* clásico y *deep learning*) y constata que la mayoría se circunscribe a

clasificación unidimensional y *offline*, con escasez de enfoques predictivos, multimodales y *multi-task*. Esta sección recorre las familias de modelos más relevantes para señales fisiológicas, integrando los estudios empíricos que mejor las ilustran.

2.2.1. Clasificadores Clásicos sobre Características Diseñadas

Los clasificadores clásicos (árboles potenciados, *Random Forest*, máquinas de soporte vectorial, KNN) operan sobre características derivadas del conocimiento del dominio, tales como potencias espectrales, entropías y estadísticos temporales. Constituyen una referencia robusta, interpretable y eficiente en datos, virtudes especialmente valiosas cuando el tamaño muestral es acotado. El *gradient boosting* sobre árboles (XGBoost, Chen & Guestrin (2016); CatBoost, Prokhorenkova et al. (2018)), el *Random Forest* (Breiman, 2001) y las máquinas de soporte vectorial (Cortes & Vapnik, 1995) concentran gran parte de los mejores resultados en la literatura de fatiga.

Tres estudios ilustran su efectividad a través de modalidades y dimensiones distintas. En fatiga física, Bustos et al. (2023) clasificaron cuatro niveles de la escala de Borg en bomberos durante ejercicio incremental, comparando KNN, árboles potenciados, SVM, redes neuronales y *Random Forest* sobre características de frecuencia cardíaca, respiración y temperatura; XGBoost obtuvo el mejor desempeño, con una *accuracy* promedio de 82 % bajo validación *leave-one-subject-out*. Resulta metodológicamente significativo que los autores privilegien deliberadamente esa cifra LOSO sobre el 88 % del esquema 10-fold estratificado, por ser la métrica honesta de generalización inter-sujeto y evitar la fuga de sujetos entre entrenamiento y prueba. En la misma dimensión física, Antwi-Afari et al. (2023) clasificaron cuatro niveles de fatiga en operarios de la construcción con una plantilla instrumentada (16 sensores de presión y acelerometría), donde *Random Forest* alcanzó 86 % de *accuracy* bajo validación cruzada de 10 particiones. Sobre EEG, Hu & Min (2018) detectaron fatiga binaria en conductores con un *Gradient-Boosting Decision Tree* alimentado por entropías no lineales (la *fuzzy entropy* fue la más discriminante), logrando una *accuracy* cercana a 94 % incluso con un solo canal.

La consistencia de estos resultados, en señales y dimensiones heterogéneas, anticipa uno de los hallazgos centrales de esta tesis: cuando se dispone de características espectrales bien diseñadas y el tamaño muestral es limitado, los clasificadores clásicos igualan o superan a alternativas más complejas, además de ofrecer mayor transparencia para

auditar las decisiones del modelo.

2.2.2. Modelos Secuenciales y Convolucionales

Las redes profundas buscan aprender representaciones directamente de la señal, capturando su estructura temporal (modelos recurrentes) o espacial (convolucionales). Las redes *Long Short-Term Memory* (LSTM) de Hochreiter & Schmidhuber (1997) y sus variantes, como las *Gated Recurrent Units* (GRU), modelan dependencias temporales de largo alcance mitigando el desvanecimiento del gradiente, mientras que las redes convolucionales (CNN) explotan la topografía espacial de los montajes EEG.

En contexto minero, Xu et al. (2023) abordaron la detección triclase de fatiga proyectando la potencia espectral relativa de 32 canales en mapas topográficos de 9×9 y clasificándolos con una CNN compacta, alcanzando 94,5 % de *accuracy* y superando a LSTM, RNN y DBN; su análisis confirmó el patrón fisiológico esperado (aumento de δ y θ y descenso de α al transitar hacia la fatiga). En el plano multimodal, Cao et al. (2025) combinaron un *backbone* ResNet-18 para imagen facial con una LSTM para series fisiológicas sobre el dataset DROZY (Massoz, Langohr, François & Verly, 2016), y mostraron que una fusión *acoplada*, que multiplica cruzadamente los *embeddings* de cada modalidad, supera a la concatenación simple, alcanzando 98,4 % de *accuracy* en condiciones de laboratorio con privación de sueño. Antes, Du et al. (2017) ya habían empleado un *Deep Autoencoder* para aprender una representación conjunta de EEG temporal y EOG frontal y estimar de forma continua el índice PERCLOS (correlación $r = 0,85$), evidenciando que la fusión profunda captura interacciones EEG–EOG que las fusiones lineales no modelan. Por su parte, Yuan et al. (2023) combinaron un extractor convolucional con atención (CBAM) y una GRU para regresión de fatiga; su principal aporte, sin embargo, reside en la adaptación de dominio entre datasets, que se analiza en la Sección 2.4.

Una tercera familia, los modelos basados en mecanismos de atención y *Transformers* (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, 2017), ha ganado protagonismo en la fusión multimodal y en los modelos fundacionales; por su centralidad en el estado del arte actual, se tratan en detalle en la Sección 2.5. En conjunto, estos trabajos muestran que las arquitecturas profundas extraen señal útil, pero a costa de mayores requerimientos de datos y, con frecuencia, de validación en laboratorio

con muestras pequeñas, lo que condiciona su generalización a entornos operacionales.

2.3. Señales, Datasets y Etiquetado de la Fatiga

La detección de fatiga se apoya en señales de naturaleza diversa, cuya elección condiciona tanto el poder discriminante como la viabilidad de despliegue. El EEG cortical es un marcador directo de la actividad del sistema nervioso central y concentra la mayor capacidad discriminante; las señales oculares (EOG, PERCLOS) son sensibles específicamente a la somnolencia; y las señales periféricas del sistema nervioso autónomo (frecuencia cardíaca y su variabilidad, actividad electrodérmica, acelerometría, temperatura) reflejan la respuesta del organismo a la carga y son las más cómodas de registrar de forma continua y no intrusiva.

Un desafío transversal a todos estos estudios es el **etiquetado**, esto es, la definición del *ground truth* de fatiga. Gran parte de la literatura recurre al auto-reporte (KSS, escala de Borg) o a métricas conductuales (tiempo de reacción, PERCLOS), ambas con limitaciones conocidas: el auto-reporte introduce sesgo subjetivo y baja resolución temporal, mientras que las métricas conductuales son ruidosas y solo capturan manifestaciones tardías de la fatiga. El dataset MPD-DF [Li et al. \(2026\)](#) representa el avance más reciente en esta dirección: con 50 sujetos, 32 canales EEG, ECG, EOG dual y esfuerzo respiratorio registrados durante 2 horas de conducción simulada, incorpora un etiquetado clínico de cinco niveles anotado por un médico especialista en sueño según criterios de *sleep staging* (vigilia $\rightarrow$ N1 $\rightarrow$ N2). Su validación técnica con MSCNN-CAM (*Multi-scale CNN with Channel Attention*) obtuvo un F1 de 0,705 en clasificación binaria con EEG. Es notable, no obstante, que los propios autores adviertan que dicho etiquetado “is not modality-agnostic but rather reflects characteristics inherent to EEG signals”, confiriendo una ventaja inherente a los clasificadores EEG sobre otras modalidades, una advertencia directamente pertinente para el etiquetado por intensidad EEG adoptado en esta tesis. La Tabla 2.1 sitúa MPD-DF en el contexto de los datasets públicos disponibles.

Cuadro 2.1: Comparación de datasets públicos de fatiga por conducción o tarea sostenida.

Dataset	N	Canales	Modalidades	Labels
FatigueSet (Kalanadhabhatta, Min, Montanari & Kawsar, 2022)	12	4	EEG	Intensidad / VAS
SADT (Cao, Chuang, King & Lin, 2019)	27	32	EEG	Tiempo de reacción
SEED-VIG	23	17	EEG + EOG	PERCLOS
DROZY (Massoz, Langohr, François & Verly, 2016)	14	32	EEG+EOG+ECG	KSS + PVT
MPD-DF	50	32	EEG+ECG+EOG+Resp	Médico (5 niveles)

MPD-DF se distingue por combinar la mayor cantidad de sujetos, la mayor densidad de canales EEG, cuatro modalidades fisiológicas y el único esquema de etiquetado clínico basado en *sleep staging*. Al momento de esta tesis, no existen benchmarks independientes publicados sobre este dataset.

2.4. Transfer Learning y Adaptación de Dominio

Un supuesto implícito en la mayoría de los estudios revisados es que los datos de entrenamiento y evaluación provienen del mismo dominio: mismo protocolo, población y dispositivo de adquisición. Sin embargo, transferir un detector de fatiga entrenado en un dataset de laboratorio (ej. *FatigueSet*) a un entorno industrial (ej. minería subterránea) constituye un problema de **adaptación de dominio** (*Domain Adaptation*), donde la distribución de los datos fuente y objetivo difiere (Ben-David, Blitzer, Crammer, Kulesza, Pereira & Vaughan, 2009).

Ben-David et al. (2009) formalizaron este problema demostrando que el error de un clasificador en el dominio objetivo está acotado por tres factores: (1) el error en el dominio fuente, (2) la $\mathcal{H}\Delta\mathcal{H}$ -divergencia entre las distribuciones de ambos dominios, y (3) un término de error ideal conjunto que refleja la dificultad intrínseca de la tarea. Esta cota implica que la transferencia será exitosa si las representaciones latentes de fatiga son compartidas entre dominios (es decir, si la $\mathcal{H}\Delta\mathcal{H}$ -divergencia es baja) independientemente de las diferencias superficiales en protocolo o equipo.

Desde la perspectiva de redes profundas, Yosinski et al. (2014) demostraron empíricamente que las capas tempranas de las CNNs aprenden *features* generales (filtros espectrales básicos, detectores de bordes), mientras que las capas superiores se especializan progresivamente en la tarea de entrenamiento. Este gradiente de generalidad-

a-especificidad justifica la estrategia de *fine-tuning*: congelar las capas tempranas (que capturan patrones EEG universales) y re-entrenar las capas superiores sobre el dominio objetivo, reduciendo así la necesidad de datos etiquetados en el nuevo contexto.

Wang & Deng (2018) organizaron los métodos de *Domain Adaptation* profunda en tres familias: (1) **basados en discrepancia**, que minimizan métricas de distancia entre distribuciones como *Maximum Mean Discrepancy* (MMD) o CORAL; (2) **adversariales**, que entrenan un discriminador de dominio para forzar representaciones invariantes; y (3) **basados en reconstrucción**, que usan autoencoders para aprender representaciones compartidas. En el contexto específico de la fatiga con EEG, Yuan et al. (2023) aplicaron la primera familia (MMD) para alinear distribuciones entre datasets heterogéneos, alcanzando una *accuracy* del 66,2 % con solo el 10 % de muestras etiquetadas en el dominio objetivo, un resultado que, según los autores, “llena un vacío en el campo de la detección de fatiga”.

Los modelos fundacionales revisados más adelante (BENDR, LaBraM, BIOT) representan una forma implícita de *Transfer Learning*: el pre-entrenamiento auto-supervisado a gran escala genera representaciones generalizables que luego se adaptan a tareas específicas mediante *fine-tuning*. No obstante, como se demostrará en el Capítulo 4, estos modelos colapsan con montajes de baja densidad (4 canales EEG), lo que motiva la exploración de estrategias de *fine-tuning* más conservadoras combinadas con *features* espectrales explícitas.

2.5. Fusión Multimodal y Modelos Fundacionales para Señales Fisiológicas

Si bien algunos de los trabajos revisados ya exploran la fusión profunda de modalidades (por ejemplo, el *Deep Autoencoder* de Du et al. (2017) o el acoplamiento cruzado de Cao et al. (2025)), la mayoría se apoya en una sola modalidad fisiológica o en fusiones superficiales (concatenación de *features* o votación). La literatura reciente (2023–2025) ha convergido en que las arquitecturas basadas en **atención inter-modal** (*Cross-Modal Attention*) superan consistentemente a las estrategias de fusión temprana y tardía cuando las señales de entrada operan a frecuencias de muestreo heterogéneas (Zhao, Zhang & Geng, 2024).

2.5.1. Estrategias de Fusión para Bioseñales

Zhao et al. (2024) argumentan que la taxonomía clásica de fusión multimodal (temprana, tardía, híbrida tensorial) “ya no es adecuada para la era moderna del *deep learning*” y proponen una nueva clasificación en cinco familias basadas en la técnica dominante: (1) *Encoder-Decoder*, (2) mecanismos de atención, (3) redes neuronales de grafos (GNN), (4) redes generativas, y (5) métodos basados en restricciones. Para señales fisiológicas con tasas de muestreo dispares (como EEG a 256 Hz, PPG a 64 Hz y actigrafía a 25 Hz), la familia de *Encoder-Decoder* con atención cruzada ofrece la mejor alternativa: la fusión temprana fuerza un re-muestreo destructivo, mientras que la fusión tardía ignora las correlaciones fisiológicas cruzadas entre el sistema nervioso central y el autónomo. La Figura 2.1 resume esta taxonomía.

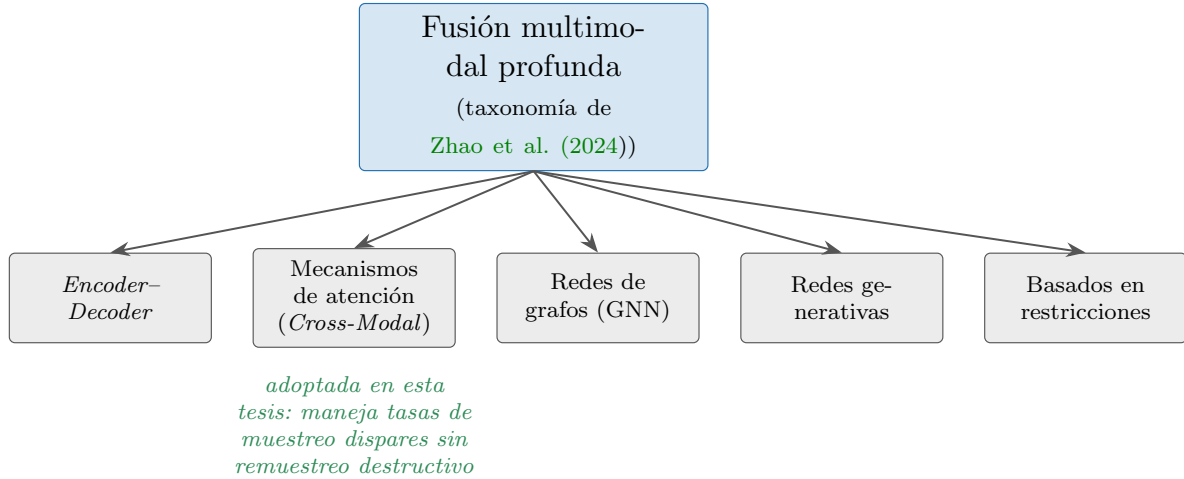


Figura 2.1: Taxonomía de fusión multimodal profunda en cinco familias Zhao et al. (2024), que reemplaza la dicotomía clásica temprana/tardía. La rama de **atención cruzada** (*Cross-Modal Attention*) es la más adecuada para bioseñales con frecuencias de muestreo heterogéneas y se adopta como dirección de fusión aprendida (vs. la concatenación tabular usada en la Fase 2 por restricción de tamaño muestral).

Los mecanismos de *Cross-Modal Attention* resuelven ambas limitaciones al proyectar cada modalidad en un espacio latente compartido, donde las consultas (*Query*) de una modalidad atienden selectivamente a las claves (*Key*) y valores (*Value*) de las demás. Chen et al. (2025) demostraron esta estrategia en el dominio de fatiga minera mediante un mecanismo bidireccional de atención cruzada EEG-fNIRS, donde cada modalidad genera consultas que atienden a la otra, permitiendo que señales con resoluciones temporales dispares intercambien información en el espacio latente sin requerir alineación

temporal punto a punto. Complementariamente, [Guo et al. \(2025\)](#) propusieron MMA-Net, que combina ramas CNN-LSTM independientes por modalidad (EEG, EDA, PPG) con un módulo de atención que pondera dinámicamente la contribución de cada sensor según su relevancia para la clasificación. La Figura 2.2 ilustra el mecanismo bidireccional.

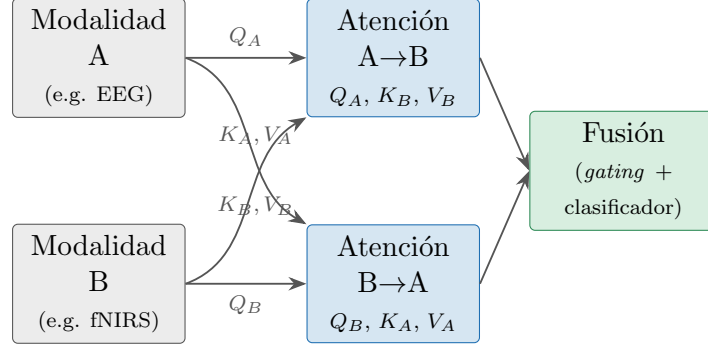


Figura 2.2: Atención cruzada bidireccional (*Cross-Modal Attention*). Cada modalidad genera consultas (Q) que atienden a las claves y valores (K, V) de la otra: la dirección $A \rightarrow B$ contextualiza A con información de B y viceversa. Permite que señales con resoluciones temporales dispares intercambien información en un espacio latente compartido, sin remuestreo conjunto, antes de la fusión final.

2.5.2. Modelos Fundacionales Multimodales

El paradigma de modelos fundacionales (*Foundation Models*), exitoso en procesamiento de lenguaje natural, se ha extendido recientemente a las señales fisiológicas. Mientras que los primeros modelos para EEG (BENDR, LaBraM) se limitaban a una sola modalidad, una nueva generación de modelos fundacionales multimodales (2024–2025) acepta combinaciones heterogéneas de bioseñales:

- **PhysioOmni** [Jiang et al. \(2025\)](#): pre-entrenado sobre EEG, ECG, EMG y EOG mediante un *tokenizador desacoplado* que separa patrones homogéneos (compartidos entre modalidades) de patrones heterogéneos (específicos de cada sensor). Incorpora un entrenamiento de ajuste fino resiliente que alinea las representaciones en un espacio latente común, permitiendo inferencia robusta ante la ausencia de sensores.
- **NormWear** [Luo et al. \(2024\)](#): diseñado para cinco modalidades de sensores vestibles (PPG, ECG, EEG, GSR e IMU), utiliza una representación multi-escala

basada en *Continuous Wavelet Transform* (CWT) con *channel-aware attention*. Combina codificadores intra-sensor e inter-sensor vía un *token* [CLS] compartido, y soporta inferencia *zero-shot* y *few-shot* mediante alineación de modalidades.

- **BIOT** Yang et al. (2023): tokeniza cada canal de bioseñal (EEG, ECG) segmentándolo y aplicando *Fast Fourier Transform* (FFT) por segmento, y emplea un *Transformer* con atención lineal para aprendizaje cruzado entre datasets heterogéneos.

La evolución de estos modelos (Figura 2.3) muestra una transición clara desde codificadores EEG unimodales hacia modelos multimodales con manejo de modalidades faltantes.

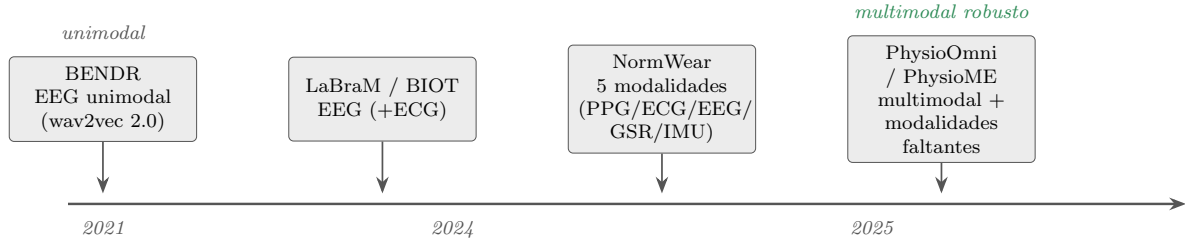


Figura 2.3: Línea de tiempo de los *foundation models* de EEG/bioseñales (2021–2025): de codificadores EEG unimodales (BENDR, LaBraM) a modelos multimodales (NormWear) y, finalmente, a modelos robustos ante modalidades faltantes (PhysioOmni, PhysioME). Esta progresión motiva su evaluación como codificadores pre-entrenados en el régimen de datos limitado de la presente tesis.

Estos modelos abordan dos desafíos críticos para el despliegue industrial: (1) la **escasez de datos etiquetados**, ya que sus encoders pre-entrenados pueden congelarse y ajustarse con adaptadores ligeros (ej. LoRA) sobre datasets pequeños del dominio objetivo; y (2) la **variabilidad en configuraciones de sensores**, ya que técnicas de *channel embedding* basadas en coordenadas espaciales permiten transferir desde montajes de 64 canales a configuraciones de 4 canales sin modificar la arquitectura.

2.5.3. Predicción Temporal con Anticipación

La detección reactiva de fatiga, que alerta cuando el operador ya está fatigado, es insuficiente para entornos de alto riesgo. La predicción anticipada requiere modelos capaces de ingerir secuencias extensas de bioseñales para extrapolar el estado futuro. Gu &

Dao (2024) propusieron Mamba, un *Selective State-Space Model* (SSM) que procesa secuencias con complejidad lineal $O(N)$, superando la limitación cuadrática $O(N^2)$ de los *Transformers* convencionales. Para una ventana de anticipación de 30 minutos con 4 sensores a distintas frecuencias, los SSMs permiten mantener un campo receptivo extenso sin sacrificar resolución temporal, filtrando selectivamente el ruido de alta frecuencia mientras preservan las tendencias fisiológicas lentas indicativas de fatiga inminente. La literatura indica que, al extender el horizonte de predicción de 5 a 30 minutos, la precisión puntual se degrada inevitablemente, transformando la tarea de una clasificación binaria exacta a una **estratificación probabilística de riesgo**. Las funciones de pérdida ponderadas por incertidumbre permiten optimizar el *Valor Predictivo Positivo* (VPP), asegurando que las alarmas tempranas sean biológicamente válidas aun cuando el minuto exacto del episodio varíe.

2.5.4. Robustez ante Modalidades Faltantes

En entornos mineros operacionales, la pérdida de sensores por sudoración, desconexión o agotamiento de batería es inevitable. Los modelos convencionales fallan catastróficamente cuando una modalidad desaparece en tiempo de inferencia. Las soluciones recientes incluyen: (1) **enrutamiento suave** mediante *Mixture-of-Experts*, que redistribuye pesos hacia los sensores disponibles; (2) **imputación generativa**, donde redes adversarias sintetizan señales sustitutas; y (3) **alineación de prototipos**, donde modelos como PhysioOmni (Jiang, Fu, Ding & Guan, 2025) y PhysioME (Lee, Lee, Jung & Kim, 2025) proyectan todas las modalidades en un espacio semántico compartido, de modo que la ausencia de un sensor no distorsiona la representación global. PhysioME emplea un decodificador de restauración que reconstruye tokens faltantes a partir del contexto de las modalidades observadas, logrando degradación graceful en lugar de fallo catastrófico.

2.6. Evaluación y Validación de Modelos de Fatiga

La evaluación de un detector de fatiga plantea desafíos metodológicos que determinan si las cifras reportadas reflejan una capacidad de generalización real o un optimismo artificial. El más insidioso es la **fuga de sujeto** (*subject leakage*): cuando segmentos

de un mismo individuo aparecen simultáneamente en los conjuntos de entrenamiento y prueba, el modelo puede memorizar rasgos idiosincráticos de la persona (su EEG basal, su anatomía) en lugar de aprender marcadores de fatiga generalizables. [Saeb et al. \(2017\)](#) demostraron que la validación que mezcla registros de un mismo sujeto (*record-wise*) infla sistemáticamente el desempeño respecto a una que separa sujetos (*subject-wise*), a veces en decenas de puntos porcentuales, produciendo modelos que no generalizan a personas nuevas. Esta evidencia justifica que tanto los estudios revisados (por ejemplo, la cifra LOSO que [Bustos et al. \(2023\)](#) privilegian) como la presente tesis adopten esquemas *leave-one-subject-out* o de validación cruzada estratificada por grupos, en los que ningún sujeto comparte datos entre particiones.

La elección de métricas es igualmente crítica cuando las clases están desbalanceadas, situación habitual en fatiga, donde los estados severos suelen ser minoritarios. La *accuracy* global resulta engañosa: un clasificador que prediga siempre la clase mayoritaria puede alcanzar una *accuracy* alta sin detectar un solo caso de fatiga. En su análisis sistemático de medidas de desempeño, [Sokolova & Lapalme \(2009\)](#) muestran que el F_1 por clase y su promedio (Macro- F_1), o la exactitud balanceada, reflejan mucho mejor el comportamiento sobre las clases minoritarias, al ponderar por igual a todas las clases con independencia de su frecuencia. Por ello, esta tesis reporta Macro- F_1 y exactitud balanceada como métricas primarias, en lugar de la *accuracy* global.

Más allá de la corrección estadística, el contexto operacional impone una **lente de seguridad**. En la detección de fatiga en faena los errores no son simétricos: un falso negativo (no detectar a un trabajador fatigado) puede derivar en un accidente, mientras que un falso positivo (una alerta innecesaria) tiene un costo mucho menor. Esta asimetría motiva reportar la **sensibilidad** (proporción de estados de fatiga correctamente detectados) y la **especificidad** de forma separada, junto con el área bajo la curva ROC (AUC-ROC), que resume la capacidad discriminante del modelo a lo largo de todos los umbrales de decisión y es relativamente robusta al desbalance de clases. Estas métricas permiten calibrar el umbral operativo según la tolerancia al riesgo de la organización, privilegiando la sensibilidad cuando el costo de un falso negativo es elevado.

Por último, en aplicaciones de seguridad la **interpretabilidad** no es un adorno sino un criterio de validación: un modelo cuyas decisiones se apoyan en marcadores fisiológicamente plausibles (por ejemplo, el aumento de potencia en *theta*) es más confiable que

otro de desempeño similar pero sustentado en artefactos espurios. Los valores SHAP (Lundberg & Lee, 2017), fundamentados en la teoría de juegos cooperativos, atribuyen a cada característica su contribución marginal a la predicción y permiten verificar que el modelo aprende patrones coherentes con la fisiología de la fatiga descrita en la Sección 2.1. Esta tesis emplea SHAP con ese propósito, integrando la validación cuantitativa con la plausibilidad fisiológica.

2.7. Síntesis y Brecha Detectada

Antes de identificar las brechas, es necesario contextualizar el estado de la detección de fatiga en la industria minera. En el ámbito comercial, existen soluciones desplegadas operacionalmente: sistemas de oculometría infrarroja como Optalert, cuyo índice JDS ha sido validado de forma independiente como predictor de somnolencia (Anderson, Chang, Sullivan, Ronda & Czeisler, 2013); sistemas de EEG monocanal integrado en cascos de seguridad (SmartCap); y cámaras *Driver Monitoring Systems* (Seeing Machines, Caterpillar DSS) que detectan cierre palpebral y cabeceo. Sin embargo, estas soluciones son propietarias, operan con una sola modalidad, funcionan de forma reactiva (alertan cuando el operador ya muestra signos de somnolencia) y no publican sus algoritmos ni datos de entrenamiento, lo que impide su evaluación científica independiente.

En el ámbito académico, los estudios de fatiga específicos para mineros son escasos. Xu et al. (2023) demostraron la viabilidad de clasificar fatiga en mineros con EEG multicanal y CNNs (94,5 % de *accuracy*), pero sin fusión multimodal. Más recientemente, Pan et al. (2025) propusieron un sistema que fusiona seis señales periféricas (ECG, EDA, presión arterial, SpO₂, temperatura, imagen facial) mediante el modelo *Transformer+*, que procesa un flujo fisiológico y uno facial (*accuracy* = 93,15 % $\pm$ 3,54 sobre 4 categorías de fatiga), y Chen et al. (2025) presentaron una red de fusión *cross-modal* EEG-fNIRS con atención bidireccional (97,92 % de *accuracy*). Estos avances validan la dirección multimodal pero comparten tres limitaciones: (1) datasets propietarios no publicados, (2) validación exclusivamente en laboratorio, y (3) enfoque exclusivo en somnolencia, sin abordar la fatiga física ni cognitiva.

A partir de esta revisión, se identifican cuatro brechas principales:

- a) **Generalización inter-dominio:** Como se analizó en la Sección 2.4, la gran mayoría de los modelos de detección de fatiga depende de datos recolectados y validados en el mismo dominio (*intra-subject* o *intra-dataset*). Desde el marco teórico de Ben-David, Blitzer, Crammer, Kulesza, Pereira & Vaughan (2009), la transferencia entre dominios es viable cuando la $\mathcal{H}\Delta\mathcal{H}$ -divergencia entre las distribuciones fuente y objetivo es baja. Sin embargo, solo Yuan, Yue, Xiong, Jiang, Zan & Li (2023) han validado empíricamente la transferencia cross-dataset para fatiga con EEG, y únicamente entre protocolos de conducción simulada. La transferencia laboratorio $\rightarrow$ industria (con cambios simultáneos de población, tarea, equipo y entorno) permanece inexplorada. Esto fundamenta la estrategia de *Transfer Learning* de la Fase 1 de esta tesis.
- b) **Fusión multimodal para fatiga multidimensional:** Los sistemas multimodales existentes para minería (Pan, Tong, Wei & Teng, 2025; Chen, Liu, Lu & Chai, 2025) abordan exclusivamente la somnolencia. Ningún sistema integra la detección simultánea de fatiga física (vía EMG, actigrafía), cognitiva (vía EEG espectral, tiempos de reacción) y somnolencia (vía EOG, HRV) en un *pipeline* unificado. Los *State-Space Models* (Gu & Dao, 2024) y la atención inter-modal (Chen, Liu, Lu & Chai, 2025; Guo, Yang & Wu, 2025) son las arquitecturas más prometedoras para esta integración.
- c) **Ausencia de datos públicos en minería:** No existe un dataset público de fatiga con señales fisiológicas multimodales en contextos de minería o industria pesada. Los estudios más recientes (Pan, Tong, Wei & Teng, 2025; Chen, Liu, Lu & Chai, 2025) emplean datos cerrados, lo que impide la reproducibilidad y la comparación entre enfoques. El dataset ocupacional público más cercano, MEFAR (Derdoyok, Patlar Akbulut & Catal, 2024), registra EEG monocanal (NeuroSky MindWave, distribuido como métricas agregadas de *attention* y *meditation* a 1 Hz, no la señal cruda de 512 Hz) y señales periféricas de trabajadores de cuatro grupos ocupacionales (académicos, técnicos, ingenieros informáticos y personal de cocina), un contexto distante de la operación minera y con una fidelidad de EEG que restringe severamente su utilidad para modelos multicanal.
- d) **Robustez operacional:** Ningún sistema revisado para fatiga en contextos in-

dustriales aborda explícitamente la pérdida dinámica de sensores. Los modelos fundacionales multimodales con manejo de modalidades faltantes (Jiang, Fu, Ding & Guan, 2025; Lee, Lee, Jung & Kim, 2025) representan la frontera del conocimiento, pero aún no han sido validados en entornos operacionales hostiles.

Estas brechas fundamentan la estrategia metodológica de esta tesis: (1) validar la transferibilidad de las representaciones de fatiga entre dominios mediante *Transfer Learning* en datasets públicos, (2) construir un *pipeline* de fusión multimodal que integre señales de las tres dimensiones de fatiga, (3) generar un dataset propietario con señales fisiológicas de trabajadores mineros en tareas simuladas con *ground truth* multi-instrumento, y (4) evaluar la robustez del sistema ante la ausencia parcial de modalidades. Esto permitirá robustecer el sistema *OmniFatigue Tracker* frente a la escasez de datos etiquetados y las condiciones adversas típicas de la operación minera.

Alcance de lo revisado frente a lo implementado. Conviene anticipar qué fracción de este estado del arte se implementa en el trabajo experimental y qué fracción queda como horizonte de iteración, dado el régimen de datos disponible (Capítulo 3). **Se implementan y evalúan:** los clasificadores clásicos sobre *features* espectrales, las arquitecturas CNN/híbridas con atención (ShallowFBCSPNet, EEGNet, ATCNet), la geometría de Riemann, los *foundation models* como *linear probe* (CBraMod EEG y NormWear multimodal) y la fusión multimodal por concatenación tardía. **Quedan como trabajo futuro** (Capítulo 5), por requerir cohortes mayores o grabación continua que el protocolo ejecutado no produce: la fusión por *Cross-Modal Attention* aprendida, los *Selective State-Space Models* (Mamba) para anticipación temporal, y el manejo explícito de modalidades faltantes (*modality dropout*, PhysioOmni/PhysioME); este último, asociado al desafío de robustez operacional, se aborda en esta tesis solo de forma parcial (imputación nativa de *features* faltantes en el *gradient boosting*), no mediante arquitecturas dedicadas. La revisión de estas líneas no implementadas se mantiene porque fundamenta la justificación de las decisiones metodológicas y el plan de trabajo futuro, no como descripción de lo entregado.

Capítulo 3

Metodología y Diseño Experimental

3.1. Visión General y Estrategia de Investigación

El objetivo final del proyecto *OmniFatigue Tracker* es la predicción multimodal de fatiga en entornos mineros reales. Sin embargo, la recolección de datos fisiológicos de alta calidad en entornos operativos y de laboratorio es costosa y requiere tiempos prolongados. Metodológicamente, el trabajo se estructura como un proceso de *Knowledge Discovery in Databases* (KDD), cuyas cinco etapas se instancian en las herramientas concretas de esta tesis (Figura 3.1).

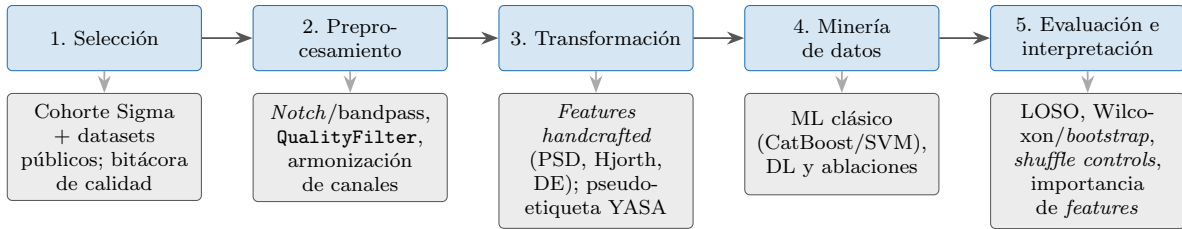


Figura 3.1: La metodología como proceso *Knowledge Discovery in Databases* (KDD): las cinco etapas canónicas (fila superior) y su instanciación concreta en esta tesis (fila inferior). El *pipeline* de las tres fases (Figura 3.2) es la realización operativa de este marco.

Para mitigar el riesgo asociado a la escasez de datos iniciales y asegurar la robustez de los algoritmos antes del despliegue en campo, se ha adoptado una estrategia de investigación del tipo *"Divide y Vencerás"*. Esta estrategia divide la metodología en dos fases claras:

- a) **Fase 1 (Validación Computacional y Transfer Learning):** Acotar el proble-

ma a una dimensión (fatiga mental/atencional) y una modalidad (EEG) utilizando datasets públicos. El objetivo es validar arquitecturas de *Deep Learning* y comprobar la hipótesis de que es posible transferir aprendizaje (*Transfer Learning*) de un dominio a otro. Esto es crítico para fundamentar el uso futuro de modelos pre-entrenados o modelos fundacionales cuando se disponga de la data real (que será limitada en volumen).

- b) **Fase 2 (Validación Experimental Multimodal):** Despliegue del protocolo completo en laboratorio de realidad mixta con trabajadores mineros, integrando señales periféricas y EEG en el marco del proyecto *OmniFatigue Tracker*.

A continuación, se detallan los diseños experimentales para ambas fases. La Figura 3.2 resume el flujo de datos completo de las tres etapas del trabajo (Fase 1, Fase 1-bis y Fase 2).

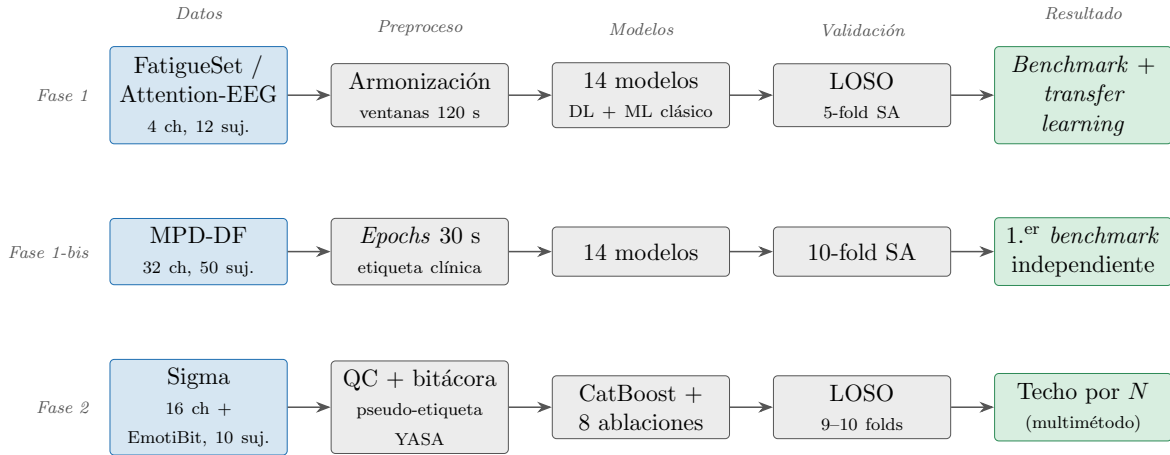


Figura 3.2: Flujo de datos *end-to-end* de las tres fases. La complejidad crece de izquierda a derecha (densidad de canales, número de sujetos, multimodalidad) y cada fase culmina en una contribución: validación + transferencia (Fase 1), primer *benchmark* independiente de MPD-DF (Fase 1-bis) y caracterización del techo por tamaño muestral sobre la cohorte propietaria (Fase 2). SA: *subject-aware*.

3.2. Fase 1: Validación en Datasets Públicos y Transferencia

Antes de la implementación física, se diseñó un experimento *in-silico* para evaluar la viabilidad de utilizar representaciones latentes de EEG aprendidas en un contexto para predecir fatiga en otro.

3.2.1. Datasets Utilizados

Se seleccionaron dos conjuntos de datos públicos que permiten emular la escasez de datos etiquetados (*Few-Shot Learning*):

- a) **FatigueSet (Origen):** Dataset multimodal enfocado en fatiga mental [Kalanadhabhatta et al. \(2022\)](#). Se utilizaron los 4 canales de EEG disponibles, remuestreados para coincidir con la frecuencia del dataset destino.
- b) **Mental Attention EEG (Destino):** Se utilizó el conjunto de datos "*EEG Data for Mental Attention State Detection*" [Cigdem \(2019\)](#), que recopila registros EEG de estados de atención mental (*focused*, *unfocused*, *drowsy*) y es de uso frecuente en la literatura de clasificación de estados cognitivos con EEG.
 - **Adquisición:** Datos capturados mediante el dispositivo *EMOTIV EPOC* con 14 electrodos (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) a 128 Hz, ubicados según el sistema internacional 10-20. Para la armonización con FatigueSet se seleccionó un subconjunto de 4 canales (AF3, AF4, T7, T8), las posiciones del EMOTIV más cercanas a los electrodos del dispositivo Muse 2 (AF7, AF8, TP9, TP10).
 - **Clases Originales:** Concentrado (*Focused*), Desconcentrado (*Unfocused*) y Somnoliento (*Drowsy*).
 - **Pre-procesamiento Específico:** Para homologar este problema a una detección binaria de riesgo, se agruparon las clases *Unfocused* y *Drowsy* bajo la etiqueta **Fatiga/Riesgo**, mientras que *Focused* se etiquetó como **Estado Normal/Alerta**.

Ambos datasets fueron armonizados espacialmente (seleccionando canales frontales y temporales equivalentes) y temporalmente (ventanas de 120 segundos). FatigueSet utiliza una diadema Muse 2 de 4 canales, mientras que *Attention EEG* emplea un EMOTIV EPOC de 14 canales del cual se seleccionó un subconjunto de 4 canales para la comparación. Esta limitación a 4 canales restringe la densidad espacial disponible para las arquitecturas de *Deep Learning*. Las implicancias de esta limitación se discuten en la Sección 4.1.7.

3.2.2. Selección de Arquitecturas y Benchmarking

Con el fin de situar el rendimiento de diferentes familias de modelos en la tarea de detección binaria de fatiga con EEG de baja densidad, se diseñó un *benchmark* exhaustivo que abarca tres categorías de arquitecturas:

- a) **CNNs clásicas para EEG:** Modelos compactos diseñados específicamente para señales de *Brain-Computer Interfaces* (BCI), que operan directamente sobre las series temporales crudas:
 - **ShallowFBCSPNet** Schirrmeister et al. (2017): Red superficial inspirada en los filtros FBCSP, optimizada para extraer potencia en bandas espectrales.
 - **EEGNet** Lawhern et al. (2018): Arquitectura ligera con convoluciones separables en profundidad, diseñada para generalizar entre distintos paradigmas BCI.
 - **Deep4Net** Schirrmeister et al. (2017): CNN profunda con cuatro bloques convolucionales y *max-pooling* temporal progresivo.
- b) **Modelos híbridos con atención temporal:** Arquitecturas que combinan convoluciones con mecanismos de atención o redes temporales (*Temporal Convolutional Networks*, TCN):
 - **ATCNet** Altaheri et al. (2023): Combina un bloque convolucional con *multi-head attention* y TCN en cascada.
 - **EEGTCNet** Ingolfsson et al. (2020): Combinación de EEGNet con un módulo TCN para capturar dependencias temporales de mayor alcance, orientada a implementación embebida.

- **EEGITNet** Salami et al. (2022): Arquitectura *Inception*-TCN que aplica convoluciones multi-escala antes de la agregación temporal.
 - **EEGConformer** Song et al. (2023): *Convolutional Transformer* que integra un *stem* convolucional con un codificador *Transformer* completo.
- c) **Modelos fundacionales pre-entrenados:** Arquitecturas basadas en *Transformers* pre-entrenados con aprendizaje auto-supervisado sobre grandes corpus de EEG:
- **BENDR** Kostas et al. (2021): Modelo inspirado en wav2vec 2.0 que aprende representaciones contextuales de EEG mediante predicción contrastiva.
 - **LaBraM** Jiang et al. (2024): *Large Brain Model* pre-entrenado con *masked token prediction* sobre aproximadamente 2.500 horas de EEG heterogéneo provenientes de ~ 20 datasets; disponible en variantes Base (5,8M), Large (46M) y Huge (369M parámetros).

Todos los modelos fueron entrenados bajo las mismas condiciones experimentales para garantizar una comparación justa: validación cruzada *subject-aware* (**StratifiedGroupKFold** agrupando por sujeto, de modo que ningún sujeto aparece simultáneamente en entrenamiento y prueba; Figura 3.3), 30 épocas máximas, *early stopping* con paciencia de 5 épocas, tasa de aprendizaje de 10^{-3} , y *batch size* de 16. Los *kernels* temporales de los modelos con atención fueron reescalados automáticamente para compensar la frecuencia de muestreo del dataset. En la Fase 2, con 10 sujetos efectivos, esta validación se lleva al límite *Leave-One-Subject-Out* (LOSO) puro, con un sujeto distinto en prueba por cada *fold*.

3.2.3. Baselines de Machine Learning Clásico

Como complemento al *benchmark* de *Deep Learning*, se evaluaron cinco clasificadores clásicos operando sobre *features* espectrales *handcrafted* extraídas de las mismas ventanas de 120 s. Este enfoque permite evaluar si la ingeniería de características explícita puede igualar o superar el aprendizaje *end-to-end* de las redes neuronales con montajes EEG de baja densidad.

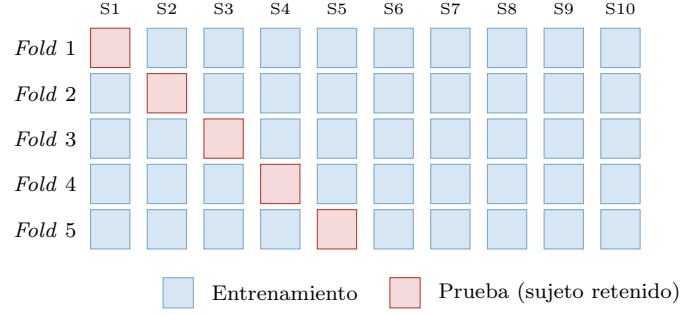


Figura 3.3: Validación cruzada *subject-aware*: en cada *fold* un sujeto completo (columna roja) se reserva para prueba y el resto se usa para entrenamiento, garantizando que no haya fuga de información inter-sujeto. Con N sujetos efectivos y un sujeto por *fold*, el esquema equivale a *Leave-One-Subject-Out* (LOSO).

Extracción de Características

Para cada ventana se extrajeron 19 *features* por canal EEG:

- **Potencias en bandas (5):** Potencia relativa en delta (0,5–4 Hz), theta (4–8 Hz), alfa (8–13 Hz), beta (13–30 Hz) y gamma (30–45 Hz), calculadas mediante PSD de Welch con ventana de 2 s. La elección de estas bandas sigue las convenciones de la neurofisiología clínica; en particular, [Klimesch \(1999\)](#) revisó la evidencia de que las oscilaciones alfa y theta reflejan el rendimiento cognitivo y la carga de memoria, lo que justifica su uso como biomarcadores cognitivos.
- **Ratios espectrales (2):** θ/α (aumenta con somnolencia) y $(\theta + \alpha)/\beta$ (disminuye con alerta; evaluado por [Jap et al. \(2009\)](#) como algoritmo espectral de detección de fatiga), métricas ampliamente utilizadas en la detección de somnolencia basada en EEG.
- **Entropía diferencial (5):** $DE = \frac{1}{2} \ln(2\pi e \cdot \sigma_{\text{banda}}^2)$, una por banda frecuencial. La entropía diferencial fue propuesta por [Shi et al. \(2013\)](#) como *feature* para la estimación de vigilancia con EEG, demostrando equivalencia matemática con el logaritmo del espectro de energía y superando a descriptores como la dimensión fractal, los parámetros AR y la entropía muestral.
- **Parámetros de Hjorth (3):** Actividad (varianza), movilidad (frecuencia media del espectro) y complejidad (ancho de banda), propuestos originalmente por

Hjorth (1970) como descriptores compactos del EEG en el dominio temporal, sin necesidad de transformada de Fourier.

- **Estadísticos temporales (4):** Media, varianza, asimetría (*skewness*) y curtosis.

Adicionalmente, se calcularon 2 *features* inter-canal: **asimetría alfa frontal** ($\log P_{\alpha, \text{der}} - \log P_{\alpha, \text{izq}}$ entre pares frontales) y **asimetría alfa temporal** (entre pares temporales). En total, el vector de *features* tiene $C \times 19 + 2$ dimensiones, donde C es el número de canales (78 para 4 canales en Fase 1, 306 para 16 canales en Fase 2).

Clasificadores

- **SVM (RBF):** *Support Vector Machine* con *kernel* de función de base radial (Cortes & Vapnik, 1995). Busca el hiperplano de máximo margen en un espacio de alta dimensión inducido por el *kernel* gaussiano $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$. Regularización $C = 1,0$, $\gamma = \text{scale}$ (inverso de $n_{\text{features}} \times \text{varianza}$). SVM es particularmente efectivo en espacios de alta dimensión con muestras limitadas.
- **XGBoost:** *eXtreme Gradient Boosting* (Chen & Guestrin, 2016), un ensamble de árboles de decisión entrenados secuencialmente mediante *gradient boosting*. Cada árbol corrige los errores residuales del anterior, con regularización L1/L2 sobre los pesos de las hojas para evitar sobreajuste. Hiperparámetros: 200 árboles, profundidad máxima 6, tasa de aprendizaje 0,1. XGBoost maneja valores faltantes de forma nativa, lo que lo hace robusto ante cobertura parcial de sensores.
- **CatBoost:** Algoritmo de *gradient boosting* (Prokhorenkova, Gusev, Vorobev, Dorogush & Gulin, 2018) con manejo nativo de variables categóricas y *ordered boosting* para reducir el sesgo de predicción durante el entrenamiento. Utiliza pesos de clase balanceados automáticos (`auto_class_weights='Balanced'`). Hiperparámetros: 200 iteraciones, profundidad 6, tasa de aprendizaje 0,1.
- **LDA:** *Linear Discriminant Analysis*, un clasificador lineal que proyecta los datos sobre un eje que maximiza la separación entre clases (ratio de varianza inter-clase / intra-clase). Asume distribuciones gaussianas con covarianza común. Es el clasificador más simple del *benchmark* y sirve como referencia de rendimiento lineal.

- **Random Forest:** Ensamble de 200 árboles de decisión (Breiman, 2001) entrenados sobre subconjuntos aleatorios de muestras (*bagging*) y *features* (*feature subsampling*). Reduce la varianza del modelo individual mediante promediado. Genera *feature importance* basada en la reducción de impureza de Gini.

Todos los clasificadores se entrenaron con pesos de clase balanceados para compensar el desbalance residual, normalización *StandardScaler* por *fold*, y la misma estrategia LOSO utilizada en los modelos DL. Las métricas reportadas son Macro-F1, *Balanced Accuracy* y *Accuracy*. Para la lectura operacional del label primario de Fase 2 se incorporan además dos métricas con lente de seguridad: el **AUC-ROC**, que evalúa la capacidad de *ordenar* el riesgo de fatiga de forma independiente del umbral de decisión, y el par **sensibilidad–especificidad** sobre la clase fatiga. Esta distinción es relevante porque, en un contexto de seguridad operacional, los dos tipos de error tienen costos asimétricos: un falso negativo (un operador fatigado no detectado) implica un riesgo de seguridad, mientras que un falso positivo (una falsa alarma) implica solo un costo operativo menor.

3.2.4. Protocolo de Transfer Learning

Para validar la hipótesis de reutilización de conocimiento, se diseñaron tres escenarios de entrenamiento:

- a) **Línea Base (FatigueSet):** Entrenamiento y validación *subject-independent* (StratifiedGroupKFold) sobre FatigueSet.
- b) **Entrenamiento desde Cero (Scratch):** Entrenar Deep4Net en el dataset de Atención inicializando los pesos aleatoriamente.
- c) **Transfer Learning (Fine-Tuning):** Utilizar los pesos aprendidos en FatigueSet como inicialización para entrenar en el dataset de Atención, permitiendo el ajuste fino de las capas.

3.3. Fase 1-bis: Validación en Dataset MPD-DF (32 canales)

Para evaluar si la ventaja de los modelos DL sobre ML clásico emerge con mayor resolución espacial, se incorporó el dataset público MPD-DF (*Multimodal Phenotyping Dataset of Driving Fatigue*) (Li, Fu, Tang, Zhou, Chen, Zhou, Chen & Luo, 2026), publicado en enero de 2026 y sin benchmarks externos previos.

3.3.1. Dataset MPD-DF

MPD-DF contiene registros de 50 sujetos (35 mujeres, 15 hombres, 23–59 años) realizando conducción simulada monótona durante 2 horas, en sesiones programadas entre 13:00 y 16:30 para aprovechar el valle circadiano post-almuerzo. Las señales registradas incluyen EEG de 32 canales (sistema 10-10, LiveAmp, Brain Products, 500 Hz), ECG (Lead II, 1024 Hz), EOG dual (256 Hz) y esfuerzo respiratorio (32 Hz), todas capturadas con un Grael 4K PSG.

Las anotaciones de fatiga fueron realizadas por un médico especialista en sueño del Hospital Huashan (Fudan University), utilizando criterios de *sleep staging* en épocas de 30 segundos. El esquema define cinco niveles: vigilia (nivel 0, beta dominante), fatiga leve (nivel 1, emergencia de alfa), fatiga moderada (nivel 2, alfa difusa), N1 NREM (nivel 3) y N2 NREM (nivel 4, K-complexes y husos de sueño). Los niveles 8 y 9 marcan artefactos.

Circularidad del etiquetado: Los labels se derivan de patrones EEG, lo que confiere una ventaja inherente a los clasificadores basados en EEG. El propio paper reconoce que “the proposed labeling approach is not modality-agnostic but rather reflects characteristics inherent to EEG signals” (Li, Fu, Tang, Zhou, Chen, Zhou, Chen & Luo, 2026). Esta limitación debe considerarse al interpretar los resultados.

3.3.2. Diseño Experimental

Se preprocesaron las señales EEG remuestreándolas a 200 Hz y segmentándolas en ventanas de 30 s con *stride* de 15 s, coincidiendo con la granularidad de las anotaciones médicas. La clasificación binaria mapea los niveles 0–1 como alerta y 2–4 como fatiga,

generando 24.782 ventanas (23.328 alerta / 1.454 fatiga, razón 94 %/6 %).

Se evaluaron dos configuraciones de canales:

- **32 canales** (montaje 10-10 completo): 610 *features* handcrafted (19 por canal + 2 asimetría).
- **16 canales** (subconjunto que replica el montaje de la Fase 2): Fp1, Fp2, C3, C4, T5, T6, O1, O2, Fz, Cz, F3, F4, T3, T4, P3, P4. 306 *features*. Permite comparación directa con los datos de laboratorio de Fase 2.

Se aplicaron los mismos 9 modelos DL y 5 clasificadores clásicos de Fase 1, con *StratifiedGroupKFold* de 10 folds (ML clásico) y 5 folds (DL, por limitaciones de cómputo).

3.3.3. Configuración Robusta para Modelos Transformer en Régimen Desbalanceado

Para modelos Transformer y *foundation models* pre-entrenados (BENDR, EEGConformer, LaBraM), el régimen de entrenamiento con 94 %/6 % de desbalance presenta un riesgo conocido de colapso de predicciones: el modelo predice solo la clase mayoritaria. Cinco factores interactuantes identifican este régimen de riesgo: (1) desbalance extremo donde el 78 % de los batches de tamaño 4 no contienen muestras de la clase minoritaria; (2) sobreparametrización masiva (BENDR: 157M parámetros, $7.144\times$ la cantidad de muestras de entrenamiento); (3) tasa de aprendizaje $10-100\times$ superior a la recomendada para transformers; (4) ausencia de *gradient clipping*; y (5) BENDR entrenado desde cero sin los pesos pre-entrenados que requiere por diseño. El fenómeno está formalizado como *Minority Collapse* por Fang et al. (2021) y replicado independientemente por EEG-FM-Bench (Xiong, Li, Li, Zhu & Jiang, 2025) sobre las mismas arquitecturas.

La configuración robusta empleada en el *benchmark* incluye:

- **Gradient clipping** ($\text{max_norm} = 1,0$) para prevenir explosión de gradientes.
- **Tasa de aprendizaje reducida:** 10^{-4} para EEGConformer.
- **Channel mapping para BENDR:** Remapeo de los 32 canales al formato de 20 canales (19 EEG 10-20 + canal SCALE) esperado por los pesos pre-entrenados.

Esta configuración refleja la práctica recomendada por la literatura para mitigar el riesgo de *Minority Collapse* en arquitecturas grandes bajo régimen desbalanceado.

3.4. Fase 2: Diseño Experimental OmniFatigue Tracker (Laboratorio)

Esta fase corresponde a la captura de datos propietarios en colaboración con Sigma-Corp. El diseño busca generar el dataset *OmniFatigue* que alimentará los modelos multimodales finales.

3.4.1. Hipótesis Operacional

El desarrollo de algoritmos basados en el aprendizaje de representaciones multimodales permite predecir fatiga multidimensional con un rendimiento superior al de los modelos unimodales, aprovechando la complementariedad entre la actividad cortical EEG y los marcadores fisiológicos periféricos del EmotiBit.

3.4.2. Alcance del Protocolo Ejecutado

El dataset Sigma sobre el cual se construye esta tesis se adquirió en un laboratorio de minería simulada y comprende bloques discretos de medición (~ 20 minutos de EEG continuo por momento) en tres momentos del día (AM, MED, PM) sobre una cohorte de 11 trabajadores reclutados a lo largo de los ciclos 6 a 11 del estudio; tras el filtrado de calidad, el análisis se concentra en 10 sujetos efectivos de los ciclos 7–11 (la composición se formaliza en la Tabla 3.1). El carácter discreto de los bloques (frente a un registro continuo de jornada completa) acota el tipo de fatiga capturable y constituye una limitación de alcance relevante para la interpretación de los resultados, que se discute en el Capítulo de conclusiones.

3.4.3. Participantes y Entorno

La cohorte total consta de **11 trabajadores adultos, todos varones**, identificados con códigos `sub_01` a `sub_11`. Cada trabajador participó en uno o más ciclos (1 a 5 ciclos

por sujeto, no necesariamente consecutivos). El análisis efectivo considera **10 sujetos**: sub_02 solo participó en los ciclos 5 y 6, sin bitácora estructurada de calidad de señal y con un subconjunto en formato legacy (BrainFlow CSV en el ciclo 5), por lo que queda fuera del filtrado por calidad utilizado para construir el dataset de entrenamiento. El protocolo se ejecutó en un laboratorio de realidad mixta de Sigma, diseñado para simular condiciones de minería subterránea con control de luminosidad y ruido ambiental (Figura 3.4).



(a) Laboratorio de simulación con equipamiento operacional minero.



(b) Trabajador ejecutando la tarea cognitiva con montaje EEG de 16 canales.

Figura 3.4: Entorno experimental del laboratorio de realidad mixta de Sigma. (a) Sala de simulación con equipamiento que reproduce condiciones operacionales mineras. (b) Trabajador instrumentado con el gorro EEG de 16 canales durante la ejecución del protocolo en la consola Rehacom. El rostro del participante se difumina para preservar el anonimato.

Criterios de inclusión y exclusión. La cohorte de análisis de la Fase 2 se construyó con tres objetivos: (i) garantizar la generalización inter-sujeto, evaluada mediante validación LOSO sin fuga de identidad entre particiones; (ii) asegurar la fidelidad de la señal EEG; y (iii) mantener una cohorte homogénea y reproducible, idéntica para todos los modelos del *benchmark*. Un bloque EEG se incluye si cumple simultáneamente:

- calidad *Usable* (Buena o Excelente) según la bitácora de calidad (Sección 3.4.7), descartando las sesiones marcadas como *Mala* y los canales o segmentos anotados como ruidosos;

- *label* YASA válido, lo que exige un número suficiente de épocas de 30 s para el *sleep staging*;
- tipo de bloque OA, OC o REA (bloques cortos con EEG cortical); los bloques EMOTI y P300 se reservan para los análisis periféricos y no integran la cohorte de fatiga YASA.

Se excluyen explícitamente: **sub_02** (participó solo en los ciclos 5 y 6, sin bitácora estructurada de calidad y con un subconjunto en formato *legacy* BrainFlow CSV) y los ciclos previos al 7 (el Ciclo 5 corresponde a un protocolo aún en consolidación y la bitácora del Ciclo 6 está en prosa libre, no parseable de forma reproducible). El resultado es una cohorte de 10 sujetos efectivos en los ciclos 7–11.

La Tabla 3.1 formaliza la composición de la cohorte analizada: los 10 sujetos efectivos, los ciclos en que participó cada uno y el número de bloques EEG *Usable* (OA, OC y REA) que aporta cada sujeto. La participación es heterogénea (cada trabajador aparece en uno o dos ciclos, no necesariamente consecutivos) y la distribución de bloques es marcadamente desigual: **sub_09** aporta apenas seis bloques, insuficientes para representar ambas clases, por lo que su *fold* LOSO queda mono-clase y se excluye del promedio, dejando 9 *folds* efectivos. Esta desigualdad inter-sujeto es una característica estructural de la cohorte que condiciona la varianza de las estimaciones LOSO.

Cuadro 3.1: Composición de la cohorte analizada en la Fase 2: 10 sujetos efectivos (sub_02 queda excluido por el filtrado de calidad) en los ciclos 7–11. El recuento de bloques EEG *Usable* (OA, OC y REA con *label* YASA válido) corresponde a la reproducción con el *pipeline* actual; los *benchmarks* reportados en los Capítulos 4 y 5 se computaron sobre el *snapshot* congelado de 849 bloques-sesión (clase 400/449), con una diferencia menor (± 2 bloques) atribuible a una regeneración posterior del parquet de *features*, sin efecto sobre las conclusiones.

Sujeto	Ciclo(s)	Bloques EEG <i>Usable</i>
sub_01	8	57
sub_03	7, 10	138
sub_04	7, 10	134
sub_05	7, 10	150
sub_06	8	68
sub_07	8, 11	118
sub_08	9	41
sub_09	9	6
sub_10	9, 11	70
sub_11	11	69
Total	7–11	851

Caracterización de la cohorte. La Tabla 3.2 resume las variables demográficas y clínicas de los 10 sujetos analizados (fuente: ficha clínica del estudio Sigma). La cohorte es homogénea en sexo (todos varones, lo que acota la generalización a población femenina) y en patrón de turno (sistema 5×2). Es relevante para un estudio de fatiga que **8 de los 10 trabajadores presentan mala calidad de sueño** (PSQI > 5), lo que contextualiza la propensión al adormecimiento que captura la pseudo-etiqueta de la Fase 2.

Cuadro 3.2: Caracterización demográfica y clínica de la cohorte analizada ($N = 10$; valores agregados y anonimizados). Donde la cobertura es parcial se indica el n efectivo.

Variable	Valor
Sexo	10/10 masculino
Edad (años)	$35,4 \pm 14,4$ (rango 22–63; $n = 6$)
IMC (kg/m^2)	$27,1 \pm 3,2$ (rango 22–31; $n = 9$)
Años de profesión	$15,1 \pm 14,9$ (rango 1–43)
Turno	10/10 sistema 5×2
Escolaridad	Media completa (4), Técnica completa (2), Básica incompleta (2), Media incompleta (1), Básica completa (1)
Tabaquismo	4 fumadores / 6 no fumadores
AUDIT-C (alcohol)	$3,2 \pm 2,2$ (rango 1–8)
PSQI (calidad de sueño)	$6,8 \pm 2,6$; 8/10 con PSQI > 5 (mala calidad)

La fecha de nacimiento no está registrada para cuatro sujetos, por lo que la edad cubre $n = 6$; un registro de IMC con valor nulo (error de carga) se excluye, dejando $n = 9$. Las demás variables tienen cobertura completa. El desbalance de sexo (todos varones) es una limitación de generalización que se retoma en el Capítulo 5.

3.4.4. Consideraciones Éticas

Los datos analizados en esta tesis fueron adquiridos por Sigma, responsable del estudio, con consentimiento informado de los participantes. El acceso del autor a los datos se rige por un acuerdo de confidencialidad (*NDA*) suscrito con Sigma. Para el presente trabajo los registros fueron seudonimizados (identificadores **sub_01** a **sub_11**) y el documento no incluye datos personales identificables. Esta tesis constituye, por tanto, un **análisis secundario** de datos preexistentes: el autor no realizó reclutamiento ni intervención directa sobre los participantes.

3.4.5. Adquisición de Señales (Multimodalidad)

Se capturan dos modalidades de hardware en paralelo:

- **Sistema OpenBCI Cyton+Daisy:** 16 canales EEG a 125 Hz con electrodos secos, mapeados al sistema internacional 10-20 (Fp1, Fp2, C3, C4, T5, T6, O1, O2, Fz, Cz, F3, F4, T3, T4, P3, P4). Adicionalmente se incluye un canal de acelerómetro del casco, 5 canales digitales, 3 canales analógicos y un canal de marcadores. Los canales frontales (Fp1, Fp2) permiten derivar señal de parpadeo (*EOG* indirecto) mediante filtrado pasa-banda 0,1–5 Hz; los canales temporales (T3, T4) permiten derivar EMG facial mediante filtrado pasa-altas 20 Hz con *notch* a 50 Hz. Estas derivaciones se obtienen sin sensores adicionales, siguiendo la metodología de [Roy et al. \(2014\)](#). El diccionario oficial Sigma documenta el EMG como “Electromiografía obtenida de canales temporales del EEG”, confirmando que no son sensores independientes.
- **Dispositivo EmotiBit (señales periféricas):** 19 streams LSL independientes:
 - *Cardiovascular:* PPG verde, rojo e infrarrojo a 25 Hz; HeartRate e IBI con muestreo irregular. Permiten cálculo de variabilidad cardíaca (HRV).
 - *SNA (Sistema Nervioso Autónomo):* EDA a 15 Hz; SCR (Skin Conductance Response) amplitud, *rise time* y *frequency* con muestreo irregular; temperatura cutánea a 7,5 Hz (sensor de contacto y termopila infrarroja).
 - *Movimiento (IMU 9 ejes):* acelerómetro, giroscopio y magnetómetro a 25 Hz por eje.

La cobertura efectiva de cada hardware no es uniforme entre bloques. Una asimetría relevante es que el bloque EMOTI graba solo EmotiBit, sin EEG: los 341 archivos *_EMOTI.xdf contienen exclusivamente los 19 *streams* del EmotiBit, sin ningún stream del OpenBCI. El contenido específico de este bloque no está documentado en el protocolo de adquisición; la señal de interés capturada es la respuesta autónoma (PPG, EDA, IMU, temperatura) durante un período prolongado.

Para los bloques OA, OC y REA, ambos tipos de hardware graban en paralelo en la mayoría de los archivos; las excepciones corresponden a fallos puntuales de *setup* o desconexiones de hardware. El Ciclo 5 (excluido del análisis) presenta una cobertura conjunta menor, consistente con un protocolo aún en consolidación al inicio del estudio.



(a) Gorro EEG de 16 canales montado sobre la cabeza de calibración usada en la preparación.



(b) Gorro EEG junto al diagrama oficial del montaje 10-20 (cf. Figura 3.6).

Figura 3.5: Montaje EEG del sistema Sigma (OpenBCI Cyton+Daisy, 16 canales). Los electrodos siguen el sistema internacional 10-20; los canales frontales y temporales se reutilizan para derivar EOG y EMG sin sensores adicionales.

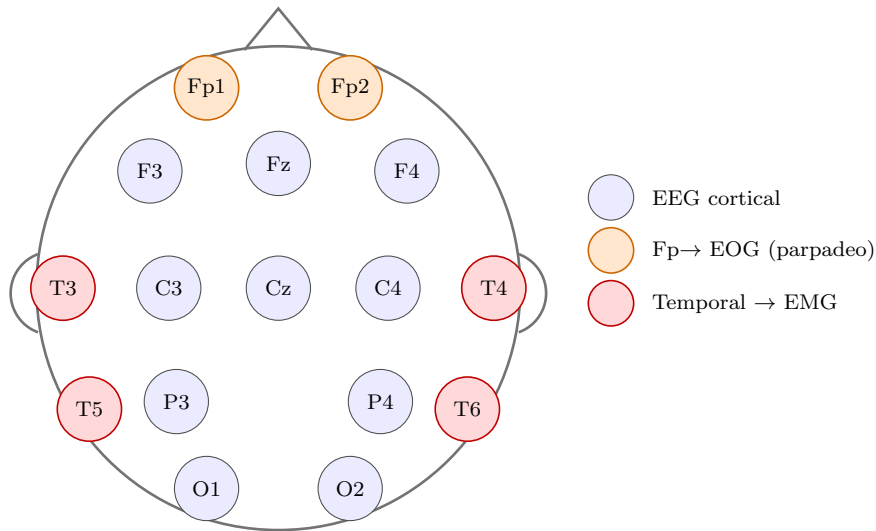


Figura 3.6: Esquema del montaje EEG de 16 canales sobre el sistema internacional 10-20 (redibujado del procedimiento oficial de sensorización Sigma). Los electrodos frontales Fp1/Fp2 se reutilizan para derivar EOG (parpadeo) y los temporales T3/T4/T5/T6 para EMG facial, sin sensores adicionales. Nomenclatura 10-20: T3/T4/T5/T6 equivalen a T7/T8/P7/P8 en notación 10-10.

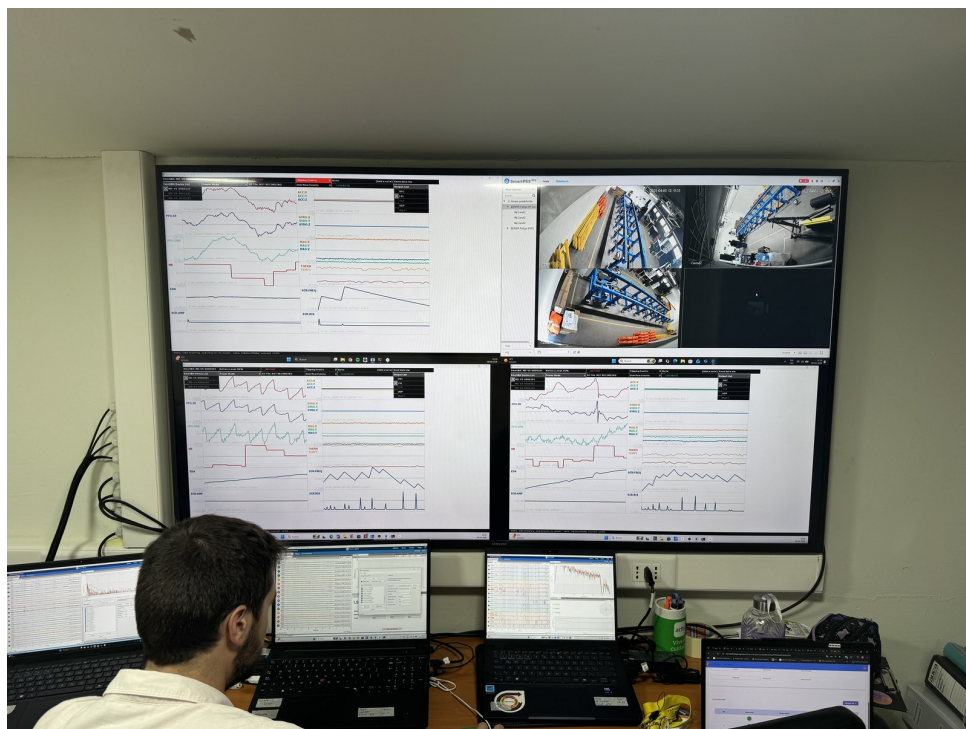


Figura 3.7: Estación de monitoreo y adquisición en tiempo real. El analista supervisa los *streams* fisiológicos sincronizados (EEG, PPG, EDA, IMU) y las cámaras del laboratorio, registrando incidencias de calidad de señal en la bitácora que alimenta el filtrado posterior (Sección 3.4.7).

3.4.6. Protocolo de Inducción y Etiquetado

Cada día de evaluación contempla tres momentos (AM, MED, PM) con una estructura idéntica. En cada momento se graban entre 3 y 5 bloques de EEG según la disponibilidad del protocolo (Tabla 3.3). Los bloques OA, OC y REA se ejecutan de forma consecutiva, con pausas cortas (10 a 60 segundos) entre uno y otro durante las cuales el operador detiene y reinicia la grabación pero el sujeto no se quita el casco. La duración total “puesto el casco” por momento es 19 a 28 minutos.

Cuadro 3.3: Bloques de medición ejecutados en cada momento del día.

Momento	Bloques EEG / multimodal	Cuestionarios
AM (pre-turno)	OA, OC, REA (EEG); EMOTI (solo Emoti-Bit)	KSS, SOFI-SM
MED (mediodía)	OA, OC, REA (EEG)	KSS, SOFI-SM
PM (post-turno)	OA, OC, REA (EEG); EMOTI (solo Emoti-Bit)	KSS, SOFI-SM

Los bloques son:

- **OA (Ojos Abiertos, resting puro), ~5 min:** sujeto mira un punto fijo en la pantalla, línea base de alerta visual.
- **OC (Ojos Cerrados, resting puro), ~5 min:** sujeto cierra los ojos quieto, prueba clásica del ritmo alfa (8–12 Hz, “reposo despierto”).
- **REA (Rehacom REA1, carga cognitiva), ~10 min:** EEG grabado mientras el sujeto realiza el módulo REA1 del software Rehacom (Capacidad de respuesta). El test entrena la velocidad y precisión de reacción ante estímulos visuales y acústicos mediante tareas de reacción simple, elección simple y elección múltiple. La equivalencia entre el bloque “REA” del XDF y el módulo Rehacom REA1 se confirma por (i) las ocho categorías Rehacom documentadas en el Diccionario de variables del proyecto (REA1, REVE, AUFM, CALC, VRO1, RO3D, PLAN, LODE), (ii) los informes médicos PDF del proyecto, cuyos nombres de archivo incluyen explícitamente el módulo (*e.g.*, ...REA1 nivel intermedio pm.pdf),

y (iii) la coincidencia de duración (~ 10 min) y horario con los registros RehaCom de la API. Este bloque no es resting puro: el sujeto está bajo carga cognitiva durante toda la grabación.

- **EMOTI (bloque sin EEG), ~ 80 a 100 min:** grabación de EmotiBit únicamente, sin EEG. El contenido específico de este bloque no está documentado en el protocolo de adquisición. Una verificación empírica posterior (Sección 3.4.6) acota la naturaleza de este bloque a registro periférico continuo durante actividad sostenida de mediana intensidad.

Verificación empírica del bloque EMOTI. Dado que el protocolo de adquisición no documenta la tarea ejecutada durante el bloque EMOTI, se realizó una verificación cuantitativa sobre los archivos disponibles para acotar la naturaleza de este bloque. Tres líneas de evidencia independientes apuntan al mismo escenario: registro continuo del EmotiBit mientras el sujeto realiza una actividad sostenida de mediana intensidad. (i) **Firma IMU del acelerómetro:** sobre 199 sesiones-bloque distribuidas en los 11 sujetos, la desviación estándar de la magnitud del acelerómetro durante EMOTI ($\bar{\sigma} = 0,072 g$, $n = 61$) es $14\times$ superior a OC ($0,005 g$, sujeto sentado quieto con ojos cerrados), $8\times$ superior a OA ($0,009 g$) y $3\times$ superior a REA ($0,024 g$, sujeto sentado realizando test cognitivo). La magnitud media se mantiene cerca de $1 g$, descartando ejercicio aeróbico continuo y compatible con la firma típica de trabajo manual de cuerpo relativamente estable (literatura accelerometry: $0,03\text{--}0,10 g$). (ii) **Frecuencia cardíaca:** sobre 136 sesiones, la mediana de HR durante EMOTI ($76,5$ bpm, IQR $[68,3, 84,7]$, $n = 38$) está $+15$ bpm sobre OC ($61,3$ bpm) y $+11$ bpm sobre OA ($65,4$ bpm), dentro del rango reportado en la literatura para trabajo manual ligero sostenido ($70\text{--}85$ bpm), y por debajo del umbral de ejercicio aeróbico (≥ 100 bpm). (iii) **Trayectoria temporal:** la actividad IMU promedio por minuto durante EMOTI se mantiene plana y elevada ($\sim 0,06\text{--}0,07 g$) por 50 minutos o más, sin picos ni caídas características de ejercicio aeróbico ni de actividad intermitente. El análisis completo (scripts, datos crudos y figuras) se documenta en el material suplementario. Como consecuencia, los 341 archivos `*_EMOTI.xdf` (~ 450 a 600 horas de señal periférica multimodal) constituyen el subcorpus más extenso del dataset capturado durante actividad laboral simulada, y son la base de los experimentos de anticipación discutidos en el Capítulo 4.

Para fines del análisis EEG, los tres primeros bloques (OA, OC y REA) se agrupan como “bloques cortos con EEG” por su duración acotada y por contar con los 16 canales del OpenBCI. Esta agrupación es operativa, no conceptual: OA y OC capturan estados resting, mientras que REA captura EEG bajo carga cognitiva controlada. La distinción es relevante al interpretar las distribuciones de etapas YASA por bloque (Sección 4.4.8).

Ground Truth disponible.

- *Psicométrico cognitivo*: el módulo REA1 de Rehacom (Capacidad de respuesta, ver Diccionario Sigma fila 214) se aplica una vez al inicio de cada momento. Mide tiempos de reacción a estímulos simples, de elección simple y de elección múltiple, con métricas como tiempo de reacción mediana, total de aciertos, número de omisiones, errores de tecla y otras. **No se aplicó PVT** (Psychomotor Vigilance Task (Dinges, Pack, Williams, Gillen, Powell, Ott, Aptowicz & Pack, 1997), considerado el gold-standard en estudios de fatiga); Rehacom se utiliza como proxy del tiempo de reacción.
- *Subjetivo*: la escala KSS (Karolinska Sleepiness Scale, 1 a 9) (Åkerstedt & Gillberg, 1990) y el inventario SOFI-SM (*Swedish Occupational Fatigue Inventory*, versión modificada) (Åhsberg, Gamberale & Gustafsson, 2000) se aplican al inicio de cada momento.
- *Físico*: en momentos AM y PM se mide lactato sanguíneo, escala BORG de esfuerzo percibido, dinamometría de mano (handgrip, 3 intentos) y altura de salto CMJ (Counter Movement Jump). En MED se mide solo lactato, BORG y handgrip.

Tareas físicas. El protocolo de adquisición contemplaba tareas físicas de simulación operacional minera (armado/desarmado de correa transportadora y mantención de tablero eléctrico). Estas tareas se ejecutaron durante el protocolo, pero *no de forma concurrente* con los bloques EEG analizados en esta tesis (OA y OC en *resting*, REA bajo carga cognitiva del test Rehacom REA1). En consecuencia, la fatiga capturada durante dichos bloques depende del propio transcurso de la jornada laboral del trabajador minero (presión homeostática del sueño según el Proceso S de Borbély (1982)) y de la carga cognitiva del test, no de la carga física inmediata.

Generación de la pseudo-etiqueta y triangulación

A diferencia de la Fase 1 (donde la etiqueta se reduce al factor experimental manipulado), el protocolo Sigma carece de un *ground truth* de fatiga único e incuestionable: ninguna de las fuentes anteriores mide directamente el estado neurofisiológico de fatiga durante los bloques de EEG. Por ello la etiqueta no se elige a priori, sino que se construyen **siete estrategias de etiquetado** agrupadas en cinco familias (diseño temporal, subjetiva, de desempeño, EEG-interna y multi-fuente) y se triangulan empíricamente antes de comprometer una como primaria (la evaluación completa se presenta en la Sección 4.4.5). La Figura 3.8 esquematiza el proceso.

La estrategia adoptada como **pseudo-etiqueta EEG-interna primaria** es la derivada de YASA Vallat & Walker (2021), un estadificador de sueño entrenado bajo criterios AASM. YASA se aplica sobre las épocas de 30 s de los bloques cortos con EEG (OA, OC, REA), produciendo un hipnograma por sesión (W, N1, N2, N3, REM); la **fracción de épocas no-Wake** por sesión, umbralizada en $\tau = 0,30$, define la etiqueta `yasa_drowsy_fraction`. La elección de una etiqueta EEG-interna (en lugar del auto-reporte subjetivo) se justifica deductivamente por la replicación del *perceptual blindness* en mineros: la concordancia entre KSS y los marcadores objetivos es prácticamente nula (Sección 4.4.5), por lo que usar el auto-reporte como *ground truth* introduciría ruido sistemático. La etiqueta se trata como marcador EEG-interno, no como medida de fatiga subjetiva.

3.4.7. Bitácora de Calidad de Señal y Filtrado

Un analista de Sigma revisa manualmente cada grabación EEG y registra, por sesión (sujeto, ciclo, día, momento), la calidad general (Excelente / Buena / Regular / Mala), los canales con ruido a excluir, los segmentos de tiempo a rechazar y notas de setup (*e.g.*, “placa gorro semiseco bota conexión”, “movimiento durante grabación”). Esta bitácora cubre los ciclos 6 a 12 y constituye un *ground truth* secundario sobre la fidelidad de la señal. Su uso es obligatorio en el preprocesamiento: cualquier sesión marcada como *Mala* se excluye del entrenamiento, y los canales o segmentos anotados se descartan antes de la extracción de *features*.

El formato de la bitácora cambió entre ciclos: Ciclo 6 con prosa libre (no parseable de

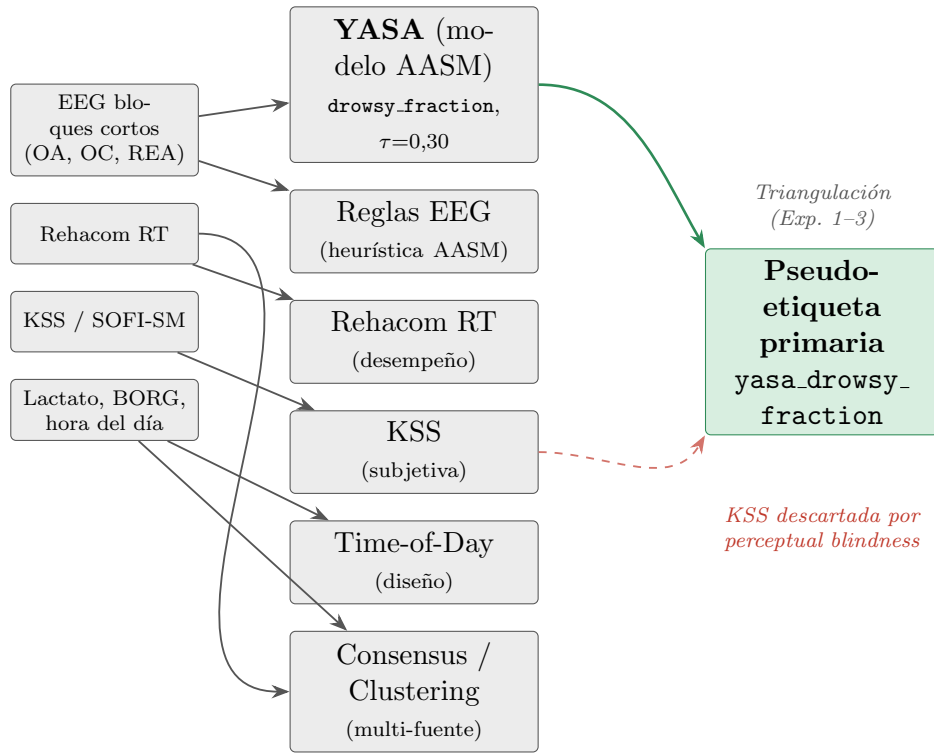


Figura 3.8: Construcción y triangulación de las siete estrategias de etiquetado (cinco familias). Las fuentes de *ground truth* alimentan estrategias candidatas que se evalúan empíricamente (Sección 4.4.5); se adopta como primaria la pseudo-etiqueta EEG-interna de YASA, mientras que el auto-reporte subjetivo (KSS) se descarta por el *perceptual blindness* documentado en mineros.

forma reproducible), Ciclo 7 con flags de reactividad semi-estructuradas (parseado mediante una rutina dedicada de extracción), y Ciclos 8 a 12 con formato consolidado. El filtrado final cubre 474 sesiones a nivel de momento (`sujeto, ciclo, día, momento`) en los ciclos 7 a 12, de las cuales el 80 % queda en categoría *Usable* (Buena o Excelente). Al contar los bloques cortos con EEG por separado (OA, OC y REA dentro de cada momento), la cohorte con *label* YASA válido empleada en la Fase 2 asciende a **849 sesiones-bloque** *Usable*, correspondientes a 10 sujetos efectivos de los ciclos 7–11 (Tabla 3.1); salvo indicación contraria, las cifras de sesiones de la Fase 2 se refieren a esta unidad de bloque.

3.4.8. Representación de los Datos para el Modelado

El *benchmark* de la Fase 2 compara dos regímenes de representación de entrada sobre la **misma** cohorte (Tabla 3.1), de modo que las diferencias de desempeño se atribuyen a la representación y no a los datos. Sea $N = 849$ el número de sesiones-bloque *Usable* con *label* YASA válido.

Representación tabular (*features handcrafted*). Cada sesión-bloque se resume en un vector de características espectrales y multimodales agregadas a nivel de bloque (detalladas por modalidad en la Sección 3.4.9). La cohorte se representa como una matriz

$$X^{(\text{tab})} \in \mathbb{R}^{N \times P},$$

donde cada fila $\mathbf{x}_i \in \mathbb{R}^P$ es el vector de P *features* del bloque i . Esta es la representación que consumen los clasificadores clásicos (LDA, *Logistic Regression*, *Random Forest*, XGBoost, CatBoost).

Representación de señal cruda (*epochs*). Los modelos de *deep learning* entrenables *end-to-end* (ShallowFBCSPNet, EEGNet, ATCNet) y los codificadores fundacionales operan sobre la señal EEG segmentada en *epochs* de 30 s remuestreados a 125 Hz, esto es, un tensor por bloque de la forma (*epochs* $\times$ canales $\times$ muestras). Como la etiqueta se define a nivel de sesión-bloque, las predicciones por *epoch* se agregan a una predicción de bloque por **media de probabilidades**, preservando la unidad de evaluación común a ambos regímenes.

En ambos casos la etiqueta es el vector binario $\mathbf{y} \in \{0, 1\}^N$ de la pseudo-etiqueta

primaria `yasa_drowsy_fraction` ($y_i = 1$ si la fracción de épocas no-Wake $\geq 0,30$), y la validación es *leave-one-subject-out* sobre los 9 *folds* efectivos (se excluye el *fold* mono-clase de `sub_09`).

3.4.9. Estrategia de Fusión Multimodal

La adquisición de datos contempla múltiples fuentes (EEG cortical, EOG y EMG derivados desde el mismo sistema OpenBCI, señales periféricas del EmotiBit, marcadores cognitivos y subjetivos). La estrategia de fusión adoptada en esta tesis es **fusión tardía por concatenación de *features*** (*late fusion*) operando sobre un clasificador único de *gradient boosting*, en lugar de la arquitectura de *Cross-Modal Attention* con *Selective State-Space Models* discutida en la literatura (Capítulo 2). La elección se sustenta en tres consideraciones del régimen de datos disponible:

- a) **Tamaño muestral.** Con 10 sujetos efectivos y ~ 849 sesiones-bloque bajo filtro de calidad, el régimen es órdenes de magnitud menor que el requerido por arquitecturas *transformer*-based con atención inter-modal, que típicamente se entrenan con miles de pares para alcanzar generalización inter-sujeto estable.
- b) **Heterogeneidad temporal.** Los bloques EEG cortos (5–10 min) impiden el aprendizaje de dependencias temporales prolongadas que justificarían *State-Space Models*. La anticipación a 30 minutos, originalmente proyectada con *Mamba*, requeriría grabaciones continuas que el protocolo ejecutado no produce salvo en los bloques EMOTI (Sección 3.4.6).
- c) **Manejo nativo de *features* faltantes.** CatBoost y XGBoost incorporan direcciones por defecto aprendidas por nodo del árbol (Chen & Guestrin, 2016), permitiendo entrenar con la heterogeneidad de cobertura observada (EmotiBit $\sim 92\%$, Rehacom $\sim 78\%$, físicos $\sim 35\text{--}44\%$) sin imputación destructiva.

Bloques de *features* por modalidad. El esquema de fusión organiza las *features* en cinco bloques que pueden activarse independientemente para construir ablaciones:

- **EEG cortical (24–25 *features*):** potencias en bandas $\delta, \theta, \alpha, \beta, \gamma$ y ratios espectrales (*jap_index*, α/θ) sobre 5 canales fronto-centrales (F3, F4, Fz, C3, C4),

agregadas como media y desviación estándar a través de ventanas de 2 s con 50 % *overlap*.

- **EOG derivado del EEG (5 *features*):** señal de parpadeo extraída de los canales frontales Fp1 y Fp2 mediante filtro pasa-banda 0,1–5 Hz, siguiendo la metodología de Roy et al. (2014). Las *features* incluyen tasa de parpadeo por minuto, amplitud media y máxima, intervalo inter-parpadeo medio y coeficiente de variación.
- **EMG derivado del EEG (5 *features*):** actividad muscular masticatoria extraída de los canales temporales T3 y T4 mediante filtro pasa-altas 20 Hz con *notch* a 50 Hz. Se calcula RMS y valor absoluto medio por canal, más la frecuencia mediana del espectro EMG.
- **EmotiBit periférico (22 *features*):** *heart rate* (6 estadísticos), EDA (5 estadísticos), magnitud del acelerómetro IMU (7 *features* incluyendo bandas espectrales y *jerk*), y temperatura cutánea (4 *features*).
- **Cognitivo-subjetivo (6 *features*):** escalas KSS y SOFI-SM (auto-reporte previo al momento) y tiempos de reacción del módulo Rehacom REA1 (medianas *mean_q1*, *mean_q2*, *mean_q3* y total de errores).

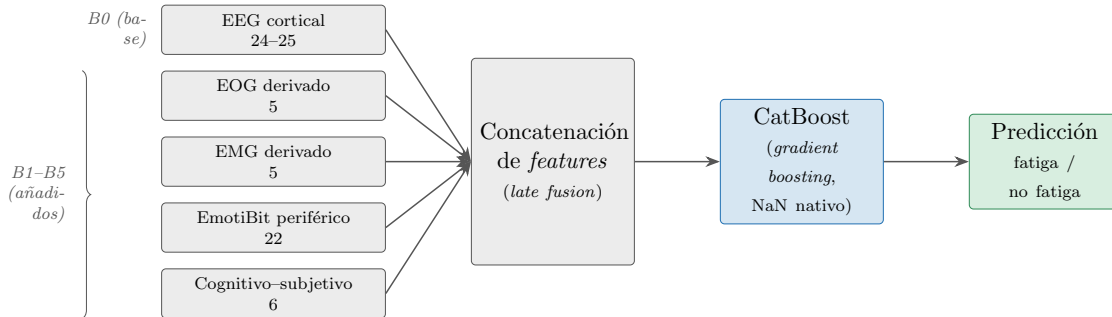


Figura 3.9: Esquema de fusión multimodal tardía (*late fusion*). Los cinco bloques de *features* se concatenan y alimentan un único clasificador de *gradient boosting* (CatBoost, con manejo nativo de valores faltantes). Las ablaciones B0–B5 activan incrementalmente los bloques: B0 usa solo EEG cortical; B1–B5 añaden cada modalidad adicional para medir su aporte sobre el *baseline* unimodal.

Ablaciones de modalidad. Los bloques se combinan en seis configuraciones experimentales (B0–B5) que sistematizan la contribución incremental de cada modalidad sobre el baseline EEG-único; los resultados se presentan en la Sección 4.4.12 del Capítulo de Resultados. La justificación de esta estructura es responder explícitamente a la pregunta de si las modalidades adicionales aportan información discriminante más allá del EEG cortical, antes de invertir en arquitecturas de fusión más complejas. La arquitectura de *Cross-Modal Attention* con *Selective State-Space Models* se mantiene como dirección de trabajo futuro (Capítulo 5) condicional a la disponibilidad de cohortes ampliadas y grabaciones continuas.

3.4.10. Análisis Estadístico: Operacionalización de las Hipótesis

Cada hipótesis declarada en el Capítulo 1 se operacionaliza mediante un criterio estadístico explícito sobre la métrica primaria del experimento (Macro-F1 bajo validación LOSO o equivalente *subject-aware*). Los criterios se declaran a continuación para evitar ambigüedad interpretativa en el Capítulo de Resultados.

H1: Fusión multimodal > unimodal. La hipótesis se evalúa sobre las seis ablaciones B0–B5 declaradas arriba. Para cada modalidad adicional ($k = 1, 2, 3, 5$) se computa el *delta* pareado por *fold* $\Delta_k = F_1(B_k) - F_1(B_0)$ sobre los 10 *folds* LOSO. **H1 se considera respaldada** para la modalidad k si simultáneamente: (i) el *test* pareado de Wilcoxon de dos colas sobre los 10 deltas rechaza $H_0 : \text{mediana}(\Delta_k) = 0$ a $\alpha = 0,05$; y (ii) el intervalo de confianza *bootstrap* percentil 95 % sobre la media de Δ_k no incluye cero por su límite inferior. Adicionalmente se reporta el *paired t-test* sobre los mismos 10 *folds* como verificación cruzada paramétrica. **H1 se considera no respaldada** si ambas pruebas fallan a $\alpha = 0,05$, lo cual debe leerse como “no se detectó mejora bajo la cohorte y el tamaño muestral disponible” y no como “la fusión no aporta nunca”.

H2: Transferibilidad de representaciones. La hipótesis se evalúa en dos escenarios: (a) *transfer learning* cross-dataset público (FatigueSet $\rightarrow$ Attention-EEG) y (b) *transfer learning* público $\rightarrow$ Sigma (FatigueSet $\rightarrow$ EMOTI). Para cada escenario se compara una configuración con pesos transferidos vs. una configuración entrenada des-

de cero (*from scratch*) sobre el dominio objetivo, manteniendo idénticas las *features* de entrada y el clasificador. **H2 se considera respaldada en su versión débil** si la configuración con transferencia iguala o supera la línea base entrenada desde cero (Macro-F1 con CI 95 % superpuestos o superiores). **H2 se considera respaldada en su versión fuerte** si la diferencia es estadísticamente significativa ($p < 0,05$ con *paired t-test* sobre *folds*) o si el CI 95 % del delta excluye cero por su límite inferior.

Control de comparaciones múltiples. A lo largo del Capítulo de Resultados se reportan aproximadamente ~ 60 comparaciones ($4 \text{ modelos} \times 7 \text{ esquemas de etiquetado en Fase 2} + 4 \text{ modelos} \times 6 \text{ ablaciones de fusión} + 4 \text{ modelos} \times 3 \text{ thresholds} \times 3 \text{ min_epochs}$ en el análisis de sensibilidad del label primario). Las conclusiones primarias se respaldan adicionalmente con *shuffle controls* (z -scores sobre etiquetas barajadas *within-subject*), *bootstrap* CIs y tamaño del efecto (Cohen’s d pareado). Las afirmaciones específicas que dependen de un único $p < 0,05$ se etiquetan como “direccionales” en el Capítulo de Resultados, a menos que sobrevivan corrección de Bonferroni laxa ($\alpha_{\text{adj}} = 0,05/60 \approx 8 \cdot 10^{-4}$).

Reporte de varianza. Toda métrica reportada sigue la convención $\mu \pm \sigma$, donde σ es la desviación estándar inter-*fold* (no el error estándar de la media). Esta convención se explicita en cada tabla del capítulo de Resultados.

3.4.11. Implementación y Reproducibilidad

Todo el procesamiento se implementó en Python. Las bibliotecas principales y sus versiones son: `numpy` 2.2, `scipy` 1.15, `pandas` 2.3 y `scikit-learn` 1.7 para el manejo de datos y el aprendizaje de máquina clásico; `mne` 1.12 y `pyxdf` 1.17 para la carga y el pre-procesamiento de EEG; `YASA` 0.7 [Vallat & Walker \(2021\)](#) para la estadificación AASM que genera la pseudo-etiqueta primaria; `CatBoost` 1.2 y `XGBoost` 3.2 para los clasificadores de *gradient boosting*; `PyTorch` 2.1 con `braindecode` (1.2 para los modelos entrenables ShallowFBCSPNet, EEGNet y ATCNet; 1.5 para los *foundation models* CBraMod, NormWear y BENDR) y `skorch` 1.4 para el aprendizaje profundo; `pyRiemann` 0.11 para la geometría de Riemann; y `cleanlab` 2.9 [Northcutt et al. \(2021\)](#) para la auditoría de ruido de etiqueta. Todos los experimentos fijan una semilla aleatoria global (`seed = 42`)

en `numpy`, `torch` y `random`, y se ejecutaron en CPU para garantizar determinismo. El código de todos los experimentos y los artefactos de resultados (métricas, etiquetas y figuras) se versionan en el repositorio del proyecto, de modo que cada cifra reportada en esta tesis es reproducible a partir de los datos de origen.

Capítulo 4

Resultados y Discusión

4.1. Resultados del Benchmark (Fase 1)

Como parte del cumplimiento de los objetivos específicos y para fundamentar la selección de arquitectura para el sistema *OmniFatigue Tracker*, se presentan los resultados del *benchmark* comparativo sobre el dataset *FatigueSet* (Kalanadhabhatta, Min, Montanari & Kawsar, 2022).

4.1.1. Esquema de Etiquetado y Balance de Clases

El *benchmark* adopta como etiqueta primaria la **condición de intensidad de actividad física** del propio diseño de *FatigueSet*. Los autores manipularon experimentalmente la carga física mediante un contrabalanceo *Latin Square* (*low/medium/high*, $\sim 5/7/9$ km/h), de modo que constituye el factor controlado del estudio y la dimensión de fatiga con mayor respaldo causal. Su binarización (Low vs. Medium+High) produce una distribución balanceada de $\sim 33\%/67\%$ **sin folds mono-clase**, lo que permite una validación cruzada bien condicionada. Como medidas complementarias de auto-reporte se derivan dos esquemas adicionales para triangular (Sección 4.1.5): VAS de fatiga física y VAS de fatiga mental.

Aun con el etiquetado balanceado persiste un desbalance moderado ($67\%/33\%$), problema ampliamente documentado en la literatura de aprendizaje automático (He & Garcia, 2009). Siguiendo las recomendaciones de Sokolova & Lapalme (2009) para clasificación con clases desbalanceadas, se reportan el *Macro-F1* y la *Balanced Accuracy*

como métricas principales (ponderan equitativamente ambas clases) y se analiza la varianza inter-*fold* como indicador de robustez. La validación es 5-*fold subject-aware* (`StratifiedGroupKFold`, `n_splits = 5`): con 12 sujetos en `FatigueSet`, cada *fold* de test mantiene $\sim 2\text{--}3$ sujetos, reduciendo la varianza inter-*fold* respecto a una LOSO estricta de 12 *folds* con un único sujeto por *fold*.

4.1.2. Rendimiento Comparativo de Arquitecturas

La Tabla 4.1 resume el rendimiento de las nueve arquitecturas de *Deep Learning* evaluadas bajo el etiquetado por intensidad. Se reportan media $\pm$ desviación estándar a través de los 5 *folds subject-aware*. Los modelos se agrupan en tres categorías: CNNs clásicas, modelos híbridos con atención temporal, y modelos fundacionales pre-entrenados.

Cuadro 4.1: Benchmark comparativo de arquitecturas de *Deep Learning* sobre `FatigueSet` (4 canales EEG, etiquetado por intensidad binario, validación 5-*fold subject-aware* `StratifiedGroupKFold`). La columna “Col.” indica el número de *folds* con colapso de predicciones; los modelos marcados con $\dagger$ colapsan en al menos un *fold* (ver Sección 4.1.3).

Categ.	Modelo	Macro-F1	Bal. Acc	Acc	Col.
CNNs	ShallowFBCSPNet	0,506 $\pm$ 0,073	0,540 $\pm$ 0,089	0,525 $\pm$ 0,074	0/5
	Deep4Net †	0,500 $\pm$ 0,028	0,505 $\pm$ 0,023	0,566 $\pm$ 0,061	1/5
	EEGNet	0,494 $\pm$ 0,033	0,514 $\pm$ 0,048	0,557 $\pm$ 0,075	0/5
Híbridos atención	ATCNet	0,544 $\pm$ 0,066	0,566 $\pm$ 0,060	0,584 $\pm$ 0,077	0/5
	EEGTCNet	0,478 $\pm$ 0,080	0,511 $\pm$ 0,056	0,591 $\pm$ 0,092	0/5
	EEGITNet	0,456 $\pm$ 0,050	0,506 $\pm$ 0,061	0,517 $\pm$ 0,076	0/5
	EEGConformer †	0,436 $\pm$ 0,041	0,504 $\pm$ 0,015	0,673 $\pm$ 0,041	3/5
Fundac.	LaBraM †	0,470 $\pm$ 0,058	0,513 $\pm$ 0,023	0,653 $\pm$ 0,050	2/5
	BENDR †	0,240 $\pm$ 0,011	0,500 $\pm$ 0,000	0,316 $\pm$ 0,020	5/5

4.1.3. Análisis de Resultados

Modelos fiables: CNNs y modelos híbridos con atención

El mejor modelo de *Deep Learning* es **ATCNet** (Macro-F1 = 0,544), cuyo mecanismo de atención multi-cabeza aporta una ventaja modesta sobre las CNNs clásicas (ShallowFBCSPNet 0,506, EEGNet 0,494) cuando la etiqueta está alineada con el factor experimental. Las CNNs se agrupan en torno a *Macro-F1* $\approx$ 0,50 con varianzas moderadas, lo que sugiere que, con solo 4 canales EEG, la capacidad de estas arquitecturas para extraer características discriminantes de la fatiga alcanza un techo similar independientemente de la profundidad de la red. ShallowFBCSPNet mantiene la menor desviación estándar entre las CNNs ($\pm$ 0,073), lo que la posiciona como la opción más estable. El resto de los modelos híbridos (EEGTCNet, EEGITNet) no supera a las CNNs, consistente con que un montaje de 4 canales restringe la información espacial disponible para los mecanismos de atención.

Diagnóstico del colapso de predicciones en modelos fundacionales

Tres modelos (BENDR, LaBraM y EEGConformer) presentan **colapso de predicciones**: predicen exclusivamente una clase en una fracción de los *folds* (columna “Col.” de la Tabla 4.1, verificada mediante conteo de clases únicas predichas y matrices de confusión por *fold*). El diagnóstico es concluyente precisamente porque el etiquetado por intensidad está **balanceado** (33 %/67 %, sin *folds* mono-clase): el colapso no puede atribuirse a un artefacto del particionado, sino a una limitación intrínseca de estas arquitecturas con 4 canales.

- **BENDR** colapsa en los 5 *folds* (Bal. Accuracy = 0,500 exacta, Macro-F1 = 0,240): nunca discrimina, prediciendo sistemáticamente una sola clase.
- **EEGConformer** colapsa en 3 de 5 *folds*; su Accuracy aparente (0,673) refleja la predicción de la clase mayoritaria, mientras que su *Macro-F1* (0,436) revela la ausencia de discriminación.
- **LaBraM** colapsa en 2 de 5 *folds* (Macro-F1 = 0,470), el menos afectado de los tres pero aún por debajo de las CNNs clásicas.

La brecha entre Accuracy y *Macro-F1* en estos modelos (p. ej. EEGConformer 0,673 vs. 0,436) ilustra cómo una métrica *naive* puede enmascarar la ausencia de señal incluso bajo un desbalance moderado. En consecuencia, la complejidad paramétrica adicional de los *Transformers* y modelos fundacionales no se traduce en ventaja cuando el montaje se restringe a 4 canales EEG.

4.1.4. Comparación con Baselines Clásicos de Machine Learning

Para determinar si el techo de rendimiento observado en los modelos DL (Macro-F1 $\approx$ 0,50) es una limitación intrínseca de las arquitecturas o una consecuencia de la información disponible en los datos, se entrenaron cinco clasificadores clásicos de *Machine Learning* bajo las mismas condiciones experimentales: validación cruzada 5-fold *subject-aware* (StratifiedGroupKFold, misma semilla aleatoria), pesos de clase balanceados por frecuencia inversa, y las mismas métricas de evaluación.

Se extrajeron 78 *features handcrafted* de cada ventana de EEG: 19 por canal (potencias en bandas δ , θ , α , β , γ vía PSD de Welch; ratios espectrales θ/α y $(\theta + \alpha)/\beta$; entropía diferencial por banda; parámetros de Hjorth: actividad, movilidad, complejidad; media, varianza, asimetría y curtosis) más 2 inter-canal (asimetría alfa frontal AF7–AF8 y temporal TP9–TP10). Todas las *features* fueron estandarizadas (media cero, varianza unitaria) antes de la clasificación.

Cuadro 4.2: Comparación de baselines clásicos de ML con el mejor modelo de *Deep Learning* sobre FatigueSet, etiquetado por intensidad (mismas condiciones: 5-fold *subject-aware* StratifiedGroupKFold, pesos balanceados). Ningún clasificador clásico presenta colapso de predicciones.

Tipo	Modelo	Macro-F1	Bal. Acc	Acc
ML clásico (espectrales)	SVM (RBF)	0,620 $\pm$ 0,074	0,642 $\pm$ 0,088	0,661 $\pm$ 0,042
	XGBoost	0,589 $\pm$ 0,047	0,591 $\pm$ 0,033	0,700 $\pm$ 0,025
	CatBoost	0,580 $\pm$ 0,070	0,585 $\pm$ 0,056	0,682 $\pm$ 0,034
	LDA	0,579 $\pm$ 0,121	0,578 $\pm$ 0,111	0,657 $\pm$ 0,123
	Random Forest	0,526 $\pm$ 0,087	0,544 $\pm$ 0,065	0,675 $\pm$ 0,048
DL (ref.)	ATCNet	0,544 $\pm$ 0,066	0,566 $\pm$ 0,060	0,584 $\pm$ 0,077

Los resultados revelan una **superioridad consistente** de los clasificadores clásicos sobre el mejor modelo de *Deep Learning*: los cinco baselines con *features handcrafted* superan a ATCNet (0,544), y SVM lidera con un *Macro-F1* de 0,620 ($\Delta = +0,076$ sobre ATCNet). A diferencia de los modelos fundacionales, ningún clasificador clásico colapsa bajo el etiquetado balanceado. Esto sugiere que, con solo 4 canales EEG, la ingeniería explícita de características espectrales captura información discriminante que las convoluciones *end-to-end* no logran extraer automáticamente de la señal cruda. Este hallazgo tiene una implicancia directa para el proyecto: el techo de rendimiento en Fase 1 no es atribuible a la arquitectura del modelo (ni convoluciones profundas, ni *Transformers*, ni clasificadores lineales lo superan ampliamente), sino a la **información contenida en un montaje de 4 canales EEG**. Este resultado motivó la hipótesis de que ampliar el montaje a 16 canales y agregar modalidades complementarias (EOG, EMG, EmotiBit periférico, métricas cognitivas) podría superar las limitaciones de un montaje unimodal de baja densidad, hipótesis evaluada empíricamente en la Sección 4.4.12 sobre la cohorte Sigma.

4.1.5. Validación por Permutación: ¿Qué Etiqueta Recupera Señal Real?

Los resultados anteriores establecen que, bajo el etiquetado por intensidad, los clasificadores clásicos superan a las arquitecturas DL (SVM 0,620 > ATCNet 0,544). Quedan, sin embargo, dos preguntas que un *Macro-F1* por encima del azar no responde por sí solo: (i) **¿la señal es real**, o el modelo explota estructura espuria del particionado (p. ej. memorización de la identidad del sujeto)?; y (ii) **¿importa el esquema de etiquetado**, o cualquier dimensión de fatiga rinde igual? Para responder ambas con una prueba directa (no por inferencia) se aplica un **test de permutación** (*shuffle within-subject*): se baraja la etiqueta *dentro* de cada sujeto, rompiendo la relación EEG-fatiga pero preservando la firma individual, y se re-entrena. Si el F1 real no supera la distribución de F1 barajados, el modelo no captura señal de fatiga. El procedimiento es idéntico al *shuffle control* de la Fase 2 (Sección 4.4.7), lo que unifica el criterio de rigor entre fases. Se evalúan los tres esquemas de etiquetado derivables del diseño de FatigueSet: (i) **condición de intensidad** (Latin Square *low/medium/high*), el factor manipulado

experimentalmente; (ii) **VAS física** ($\Delta \text{VAS}_{\text{phys}} \geq 5$ vs. M_0); y (iii) **VAS mental** ($\Delta \text{VAS}_{\text{ment}} \geq 10$ vs. M_0). La Tabla 4.3 compara sus distribuciones de clase; las tres están razonablemente balanceadas y libres de *folds* mono-clase.

Cuadro 4.3: Distribución de clases según esquema de etiquetado válido en FatigueSet (12 sujetos $\times$ 3 sesiones).

Esquema	Clases	Distribución	Mono-clase
Intensidad binario	2 (Low / Fatigued)	33 % / 67 %	0 de 12
Intensidad 3 clases	3 (Low / Med / High)	33 % / 33 % / 33 %	0 de 12
VAS física, $\Delta \geq 5$	2 (Alerta / Fatiga)	56 % / 44 %	0 de 12
VAS mental, $\Delta \geq 10$	2 (Alerta / Fatiga)	74 % / 26 %	1 de 12 (P07)

Resultados del test de permutación. La Figura 4.1 y la Tabla 4.4 contrastan, para cinco clasificadores y $N = 200$ permutaciones por esquema, el *Macro-F1* real frente a su nulo barajado.¹

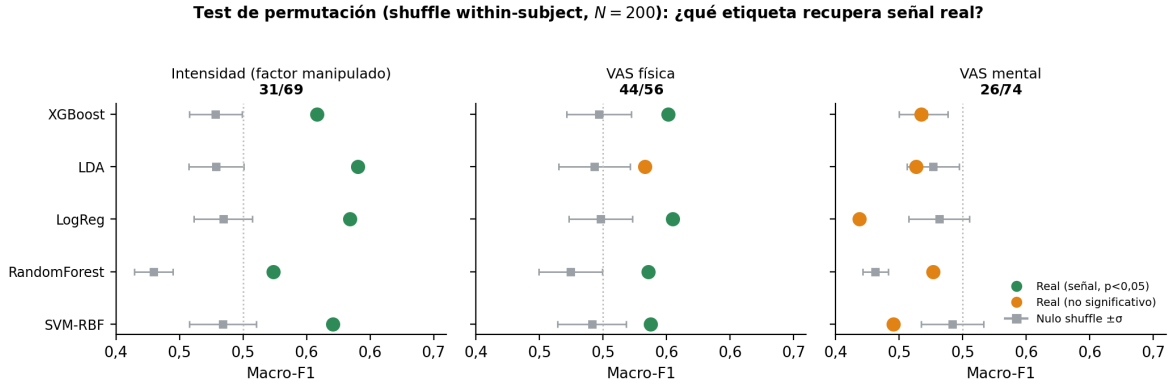


Figura 4.1: Test de permutación por esquema (*shuffle within-subject*, $N = 200$). Cada panel muestra, para cinco clasificadores, el *Macro-F1* real (círculo verde si supera su nulo con $p < 0,05$ y $F1 > 0,5$; naranja si no) frente a la distribución de $F1$ barajados (cuadrado gris, media $\pm \sigma$). La línea punteada marca el azar (0,5). El gradiente intensidad $\rightarrow$ VAS física $\rightarrow$ VAS mental evidencia que el esquema de etiquetado determina cuánta señal real es recuperable.

¹El test emplea los clasificadores con hiperparámetros por defecto, sin la optimización de la Tabla 4.2; por ello los $F1$ absolutos son levemente inferiores al titular (SVM = 0,620). Lo determinante es el contraste real–shuffle (z), invariante al nivel absoluto, el mismo criterio aplicado en la Fase 2, donde el *baseline* LDA (0,595) acompaña al titular CatBoost (0,631).

Cuadro 4.4: Test de permutación (*shuffle within-subject*, $N = 200$) sobre los tres esquemas válidos de FatigueSet. Se reporta el rango entre los cinco clasificadores. $z = (F1_{\text{real}} - \overline{F1}_{\text{shuffle}})/\sigma_{\text{shuffle}}$; un esquema “captura señal” si $p < 0,05$ con F1 sobre el azar.

Esquema	Balance	Macro-F1 real	z (rango)	Modelos con señal
Intensidad (manipulado)	33/67	0,524–0,590	+3,3 a +6,2	5/5 ($p = 0,005$)
VAS física, $\Delta \geq 5$	44/56	0,533–0,555	+1,4 a +2,4	4/5 ($p < 0,05$)
VAS mental, $\Delta \geq 10$	26/74	0,419–0,477	–2,6 a +4,5*	0/5

* El único $z > 0$ en VAS mental (Random Forest) es espurio: su *Macro-F1* (0,477) queda bajo el azar; el z alto surge solo porque su nulo barajado es igualmente degenerado (predice la clase mayoritaria).

Dos conclusiones directas. (A) El etiquetado importa. El nivel de señal recuperable depende fuertemente del esquema, en un gradiente nítido: la **intensidad** (factor manipulado) es señal robusta en los 5 modelos ($z = 3,3\text{--}6,2$, $p = 0,005$); la **VAS física** es señal moderada (4/5 modelos, $z = 1,7\text{--}2,4$); y la **VAS mental** cae al azar (4/5 modelos con $z \leq 0$). Este gradiente intensidad > VAS física > VAS mental refleja la estructura del protocolo: el factor controlado experimentalmente es la carga física, por lo que es la dimensión más predecible desde EEG, mientras que la fatiga mental (secundaria) es naturalmente más ruidosa. **(B) Donde hay señal, es genuina.** Bajo intensidad, los cinco modelos superan su nulo barajado por entre 3 y 6 desviaciones estándar con $p = 0,005$ (el mínimo alcanzable con 200 permutaciones), lo que descarta que el rendimiento sea memorización de la identidad de sujeto o un artefacto del particionado.

La métrica *naive* engaña bajo desbalance. El caso de la VAS mental (26 %/74 %) ilustra un riesgo recurrente: Random Forest alcanza una **Accuracy de 0,735** (aparentemente alta) pero su **Macro-F1 es 0,477, por debajo del azar** (0,5); el modelo simplemente predice la clase mayoritaria. Este contraste Accuracy-vs-Macro-F1 confirma la elección de *Macro-F1* como métrica principal (Sección 4.1.1) y advierte que una etiqueta desbalanceada puede inflar una métrica ingenua sin que exista discriminación real, el mismo fenómeno observado en los modelos fundacionales colapsados (Sección 4.1.3). Los resultados completos por modelo y esquema se reportan en el Apéndice A (Tablas A.1, A.2, A.3, A.4, A.5, A.6).

Para contextualizar el rendimiento SVM = 0,620 frente a la literatura: Xu et al. (2023) reportaron 94,5 % accuracy con 32 canales (8× más resolución espacial); Pan et al. (2025) alcanzaron 93 % con fusión de 6 modalidades sin EEG; Cao et al. (2025) reportaron 98,4 % con EEG+ECG+video pero con privación aguda de sueño en laboratorio (condición lejana a operación minera real). El rendimiento inferior aquí es consistente con la limitación espacial del montaje de 4 canales.

Implicación para la Fase 2 (Sigma). El esquema de etiquetado por bloques temporales (AM/MED/PM) adoptado en Fase 2 **no es análogo** al de FatigueSet: se justifica por la acumulación homeostática de presión del sueño en jornadas de 8 horas (Cajochen, Borbély; ver Sección 4.4.4), mientras que el de FatigueSet se justifica por el contrabalanceo de intensidad física en sesiones de ~21 min. Ambos son válidos en su contexto pero las justificaciones bibliográficas son distintas y no deben confundirse. Adicionalmente, la triangulación de Fase 1 motivó el diseño de la triangulación equivalente de Fase 2 (Sección 4.4.5) sobre las siete estrategias candidatas de Sigma.

4.1.6. Validación de Transfer Learning

Complementariamente al *benchmark*, se evaluó si el conocimiento adquirido en *FatigueSet* mejora el rendimiento en un dataset nuevo (*Attention EEG*). Los resultados comparativos se muestran en la Tabla 4.5.

Cuadro 4.5: Comparación de rendimiento utilizando *Transfer Learning* con Deep4Net.

Estrategia de Entrenamiento	Balanced Accuracy	Macro-F1
Entrenamiento desde cero (<i>Scratch</i>)	0,59	0,57
Transfer Learning (Pre-entrenado)	0,62	0,61

Se observa una mejora consistente tanto en *Balanced Accuracy* (+3 puntos porcentuales) como en *Macro-F1* (+4 puntos porcentuales) al utilizar pesos pre-entrenados. Este resultado es consistente con los hallazgos de Yosinski, Clune, Bengio & Lipson (2014), quienes demostraron que las capas tempranas de redes profundas aprenden *features* generales transferibles entre dominios, mientras que las capas superiores se especializan. Desde la perspectiva teórica de Ben-David, Blitzer, Crammer, Kulesza, Pereira &

Vaughan (2009), la mejora sugiere que la divergencia entre los dominios fuente (*Fatigue-Set*, fatiga física inducida por caminata en cinta) y objetivo (*Attention-EEG*, estados de concentración mental) es suficientemente baja para que las capas tempranas transfieran ventaja. Ahora bien, dada la distancia entre ambos constructos (fatiga física vs. atención sostenida), lo más plausible es que lo transferido no sea una *representación latente de la fatiga* en sentido estricto, sino una *representación general del EEG* (estructura espectral basal y contraste reposo vs. actividad) común a tareas cognitivo-motoras; esta lectura, más conservadora, es igualmente valiosa frente a la escasez de datos, pues justifica el pre-entrenamiento cross-dominio como estrategia de inicialización.

4.1.7. Discusión de la Fase 1

Los resultados de la Fase 1 sobre FatigueSet sostienen cinco lecturas que se retoman, integradas con las demás fases, en las conclusiones globales (Capítulo 5): (i) la viabilidad de clasificar fatiga con EEG de baja densidad bajo el etiquetado alineado con el factor experimental ($SVM = 0,620 > ATCNet = 0,544$); (ii) la superioridad consistente del *ML* clásico con *features* espectrales sobre el *deep learning end-to-end*; (iii) el colapso intrínseco de los modelos fundacionales (BENDR, LaBraM, EEGConformer) con solo 4 canales; (iv) la dependencia de la señal recuperable respecto del esquema de etiquetado, evidenciada por el test de permutación (intensidad robusta, VAS física moderada, VAS mental en azar); y (v) una tendencia direccional de *transfer learning* cross-dominio (+4 pp, preliminar). Estas lecturas motivan directamente el diseño de la Fase 2.

4.2. Resultados de Fase 1-bis: Dataset MPD-DF (32 canales, 50 sujetos)

Esta sección presenta los resultados del primer *benchmark* independiente sobre el dataset MPD-DF (Li, Fu, Tang, Zhou, Chen, Zhou, Chen & Luo, 2026), evaluando los mismos 14 modelos de Fase 1 sobre un montaje de 32 canales EEG con etiquetado clínico. Constituye, hasta donde alcanza nuestra revisión al cierre de esta tesis (junio de 2026), la primera validación externa de este dataset publicado en enero de 2026.

4.2.1. ML Clásico: 32 canales vs. 16 canales

Cuadro 4.6: Clasificadores clásicos sobre MPD-DF, 32 canales (610 *features* por ventana de 30 s). 10-fold StratifiedGroupKFold, 50 sujetos. Desbalance 94%/6% bajo nuestra binarización (niveles 0-1 vs 2-4).

Modelo	Macro-F1	Bal. Accuracy	Accuracy
CatBoost	0,728 $\pm$ 0,088	0,760 $\pm$ 0,086	0,942 $\pm$ 0,036
XGBoost	0,686 $\pm$ 0,091	0,642 $\pm$ 0,074	0,952 $\pm$ 0,041
SVM	0,670 $\pm$ 0,058	0,694 $\pm$ 0,057	0,929 $\pm$ 0,042
LDA	0,667 $\pm$ 0,092	0,691 $\pm$ 0,093	0,930 $\pm$ 0,048
RandomForest	0,632 $\pm$ 0,066	0,592 $\pm$ 0,045	0,947 $\pm$ 0,045

CatBoost alcanza Macro-F1 = 0,728, *del mismo orden* que el *benchmark* de aprendizaje profundo publicado para MPD-DF (MSCNN-CAM, F1 = 0,705), logrado con *features* handcrafted y un clasificador tree-based. La alta *Accuracy* (>0,92 en todos los modelos) refleja parcialmente el desbalance extremo: un clasificador que predijera siempre “alerta” obtendría 94% de accuracy pero 0,50 de *Balanced Accuracy*.

Comparabilidad con el *benchmark* del paper. La comparación con el F1 = 0,705 del paper exige cuatro precisiones. **(i) Métrica:** ambas cifras son Macro-F1. El paper no lo declara textualmente, pero su F1 = 0,705 no coincide con la media armónica de la *Precision* (0,760) y el *Recall* (0,699) que reporta (que sería 0,728), lo que descarta una F1 de la clase positiva y es consistente con un macro-promedio bajo desbalance. **(ii) Binarización:** nuestro *pipeline* titular agrupa niveles 0-1 vs 2-4 (desbalance $\approx 94/6$), mientras que el paper colapsa nivel 0 vs niveles 1-4 ($\approx 78/22$): no es exactamente el mismo problema de clasificación. Para una comparación directa se reentrenó CatBoost bajo **la binarización del paper** (0 vs 1-4), obteniendo Macro-F1 = **0,730 $\pm$ 0,063**, a la par del MSCNN-CAM. **(iii) Protocolo:** el paper no detalla su validación cruzada; el nuestro es StratifiedGroupKFold de 10 *folds* agrupado por sujeto (sin fuga inter-sujeto). **(iv) Representación:** *features handcrafted* (610 por ventana de 30 s) frente a una CNN *end-to-end*. En consecuencia, el delta no debe leerse como superioridad estadística (queda dentro de la dispersión inter-*fold* y el paper no

reporta desviación); la lectura defendible es que el ML clásico con *features* handcrafted alcanza un rendimiento *del mismo orden* que el *benchmark* de aprendizaje profundo publicado sobre este dataset, evidencia coherente con que el DL no domina en este régimen.

Cuadro 4.7: Comparación de densidad de canales: 32ch vs. subconjunto de 16ch (ML clásico, MPD-DF).

Modelo	F1 (32ch)	F1 (16ch)	Δ
CatBoost	0,728	0,711	−0,017
LDA	0,667	0,691	+0,024
XGBoost	0,686	0,684	−0,002
SVM	0,670	0,668	−0,002
RandomForest	0,632	0,654	+0,022

La diferencia entre 32 y 16 canales es mínima para ML clásico (1–2 pp), lo que indica que los canales adicionales del sistema 10-10 (FC, CP, TP, PO) aportan poco valor a las *features* handcrafted basadas en potencias espectrales.

4.2.2. Benchmark de Deep Learning: 32 canales

Cuadro 4.8: Benchmark DL sobre MPD-DF, 32 canales. Para EEGConformer y BENDR se aplica la configuración robusta descrita en la Sección 3.3.3 (LR= 10^{-4} + *gradient clipping* para EEGConformer; pesos pre-entrenados + *channel mapping* para BENDR). Las filas etiquetadas (*default*) corresponden al entrenamiento sin esa configuración robusta y se incluyen como referencia del régimen de colapso.

Modelo	Macro-F1	Bal. Acc.	Folds	Colapso
EEGITNet	0,690 $\pm$ 0,094	0,750 $\pm$ 0,052	5	0/5
ATCNet	0,667 $\pm$ 0,068	0,721 $\pm$ 0,062	5	0/5
LaBraM	0,653 $\pm$ 0,105	0,771 $\pm$ 0,024	2*	0/2
Deep4Net	0,642 $\pm$ 0,091	0,716 $\pm$ 0,088	10	0/10
EEGNetv4	0,642 $\pm$ 0,101	0,714 $\pm$ 0,113	10	0/10
EEGConformer	0,640 $\pm$ 0,091	0,727 $\pm$ 0,094	5	0/5
EEGTCNet	0,624 $\pm$ 0,036	0,691 $\pm$ 0,064	5	0/5
ShallowFBCSPNet	0,587 $\pm$ 0,096	0,686 $\pm$ 0,115	10	0/10
EEGConformer (<i>default</i>)	0,485 $\pm$ 0,011	0,500 $\pm$ 0,000	5	5/5
BENDR	0,485 $\pm$ 0,014	0,500 $\pm$ 0,000	10	10/10

*LaBraM: resultado parcial (2 de 5 folds completados).

4.2.3. Análisis del Colapso y Recuperación de Modelos Transformer

De los 9 modelos DL, dos presentan colapso total bajo configuración *default*: BENDR (10/10 folds) y EEGConformer (5/5 folds), ambos con *Balanced Accuracy* = 0,500, indicando predicción exclusiva de la clase mayoritaria. La caracterización del régimen (Sección 3.3.3) identifica un umbral claro de sobreparametrización: los modelos con más de $100\times$ la relación parámetros/muestras colapsan bajo desbalance extremo.

La configuración robusta (*gradient clipping* + LR = 10^{-4}) **permite que EEGConformer alcance F1 = 0,640**, pero **no es suficiente para BENDR** a pesar de cargar pesos pre-entrenados con *channel mapping* correcto (99/99 keys matched). Este resultado es consistente con EEG-FM-Bench (Xiong, Li, Li, Zhu & Jiang, 2025), que documenta colapso de BENDR con *Balanced Accuracy* = 50,00 % $\pm$ 0,00 % en múltiples

tareas, y con el fenómeno de *Minority Collapse* (Fang, He, Long & Su, 2021), donde las clases mayoritarias dominan la función de riesgo a medida que aumenta el desbalance, restando peso a la optimización de las clases minoritarias en modelos con alta sobreparametrización ($>100\times$).

4.2.4. Ranking Consolidado: DL vs. ML Clásico

Cuadro 4.9: Ranking consolidado de los 14 modelos evaluados sobre MPD-DF (32ch). El *benchmark* del paper (MSCNN-CAM, $F1 = 0,705$) se incluye como referencia.

#	Modelo	Macro-F1	Tipo
1	CatBoost	0,728	Clásico
	<i>MSCNN-CAM (paper)</i>	<i>0,705</i>	<i>Referencia</i>
2	EEGITNet	0,690	DL
3	XGBoost	0,686	Clásico
4	SVM	0,670	Clásico
5	ATCNet	0,667	DL
6	LDA	0,667	Clásico
7	LaBraM	0,653*	DL
8	Deep4Net	0,642	DL
9	EEGNetv4	0,642	DL
10	EEGConformer (v2)	0,640	DL
11	RandomForest	0,632	Clásico
12	EEGTCNet	0,624	DL
13	ShallowFBCSPNet	0,587	DL
–	BENDR	0,485	DL (colapsado)

*LaBraM: resultado parcial (2 folds).

Los resultados de Fase 1-bis arrojan tres hallazgos principales:

- a) **ML clásico iguala o supera a DL incluso con 32 canales:** CatBoost ($F1 = 0,728$) supera al mejor modelo DL (EEGITNet, $F1 = 0,690$) y queda *a la par* del *benchmark* del paper (MSCNN-CAM, $F1 = 0,705$; del mismo orden, sin superioridad estadística, según el análisis de comparabilidad anterior). Este resultado,

obtenido con 50 sujetos y 32 canales *research-grade*, no respalda la hipótesis de que la ventaja de DL emerge con mayor resolución espacial.

- b) **El cuello de botella es el desbalance de clases, no la arquitectura:** Con una razón de 94%/6%, la ponderación de clase estándar resulta insuficiente para los modelos DL más grandes. CatBoost maneja mejor este desbalance gracias a su mecanismo interno de boosting basado en gradientes ordenados.
- c) **Los modelos fundacionales son inviables sin adaptación dedicada:** BENDR colapsa en las configuraciones evaluadas (entrenamiento desde cero, con pesos pre-entrenados, con *channel mapping*). EEGConformer se recupera con *gradient clipping* y tasa de aprendizaje reducida, pero no supera a los modelos CNN más simples.

4.3. Direcciones Derivadas de la Fase 1-bis

De estos resultados se derivan cinco direcciones: (1) completar el *benchmark* DL de LaBraM con 5 folds completos; (2) *Transfer Learning* cross-dominio pre-entrenando en MPD-DF (16ch) y fine-tuneando en los datos Sigma, evaluando la transferibilidad conducción $\rightarrow$ minería; (3) fusión tabular-secuencial combinando *features* clásicas con embeddings DL; (4) evaluación de modelos fundacionales multimodales (PhysioOmni, NormWear) que manejan modalidades faltantes de forma nativa; y (5) *benchmark* DL sobre los datos Sigma (16ch), donde modelos previamente colapsados con 4 canales podrían recuperarse. Las direcciones (4) y (5) se desarrollan en la Fase 2 de esta tesis (Secciones 4.4.13 y 4.4.14); las restantes quedan como trabajo futuro.

4.4. Datos del Laboratorio y Esquema de Etiquetado (Fase 2)

Nota sobre métricas y desviaciones reportadas en Fase 2. Salvo indicación explícita, todos los valores **F1** en esta sección y posteriores son *Macro-F1* (promedio aritmético de las F1 por clase, robusto a desbalance), no F1 binario ni *micro-F1*. El símbolo $\pm$ acompañando una métrica denota *desviación estándar a través de los folds*

$LOSO$ (σ_{fold}), **no** el error estándar de la media ($\sigma_{\text{fold}}/\sqrt{n_{\text{folds}}}$). Para un $n_{\text{folds}} = 10$, el error estándar de la media es aproximadamente un tercio de la desviación estándar reportada, y el intervalo de confianza al 95 % sobre la media de *folds* es $\pm 2,26 \cdot \sigma_{\text{fold}}/\sqrt{10}$. Cuando se reporta un intervalo de confianza explícito (por ejemplo, en tests pareados como Wilcoxon), se indica con corchetes y se especifica el método (*bootstrap*, *t* pareado, etc.).

Control de comparaciones múltiples. A lo largo de la Fase 2 se reportan resultados de aproximadamente ~ 60 comparaciones (4 modelos $\times$ 7 esquemas de etiquetado + 4 modelos $\times$ 6 ablaciones de fusión multimodal + 4 modelos $\times$ 3 thresholds YASA $\times$ 3 *min_epochs* en el análisis de sensibilidad), sin contar las ablaciones de modelo y de *features* del experimento de anticipación EMOTI. Para evitar inflar la tasa de falsos positivos, las conclusiones primarias se respaldan con *shuffle controls* (*z*-scores específicos del modelo), tests pareados (Wilcoxon o *t* pareado) y bootstrap CIs. Las afirmaciones específicas que dependen de un único $p < 0,05$ se etiquetan como “direccionales” en lugar de “confirmadas”, a menos que sobrevivan corrección de Bonferroni laxa ($\alpha = 0,05/60 \approx 8 \cdot 10^{-4}$).

4.4.1. Alcance del Protocolo Ejecutado

El dataset analizado comprende bloques discretos de medición (aproximadamente 20 minutos de EEG continuo por momento) en tres momentos del día (AM, MED, PM) sobre una cohorte de **11 trabajadores** reclutados a lo largo de los **ciclos 6 a 11** del estudio Sigma; tras el filtrado de calidad, la cohorte de la Fase 2 con *label* YASA válido se concentra en 10 sujetos efectivos de los ciclos 7–11 (Tabla 3.1). Los ciclos 1 al 4 utilizaron hardware y formato distinto (BrainFlow CSV en lugar de XDF) y quedan fuera del análisis. El Ciclo 5 tiene archivos XDF en disco pero formato legacy sin bitácora estructurada, también excluido. El Ciclo 12 está disponible pero sus archivos XDF aún no se incorporaron al *pipeline* al cierre de esta tesis.

El período cubierto por la tesis es de aproximadamente **3,3 meses** (diciembre 2025 a marzo 2026), correspondiente a los ciclos 6 a 11 incorporados al *pipeline*, dentro del estudio Sigma completo que abarca aproximadamente 13 meses (marzo 2025 a abril 2026). El diccionario oficial de Sigma documenta el EEG con periodicidad “Continuo

durante toda la jornada y todos los días”, reflejando el diseño original; los archivos XDF reflejan la ejecución real con bloques discretos.

4.4.2. Cobertura de *Ground Truth*

La integración de los archivos XDF, la API Sigma y los archivos tabulares de la plataforma produjo **1454 sesiones EEG** sobre los 6 ciclos analizados, de las cuales 1163 son bloques cortos con EEG (OA y OC en condición resting, REA bajo carga cognitiva del test Rehacom REA1), 282 son sesiones EMOTI (sin EEG, solo EmotiBit) y 9 corresponden a archivos con tags no estandarizados o tareas auxiliares fuera del alcance del análisis. La Tabla 4.10 resume la cobertura de cada fuente de etiquetado y de las mediciones físicas auxiliares.

Cuadro 4.10: Cobertura de *ground truth* y mediciones auxiliares en los 6 ciclos analizados ($n = 1454$ sesiones EEG).

Fuente	Sesiones con dato	Cobertura
KSS (Karolinska Sleepiness Scale)	1436	98,8 %
SOFI-SM (Inventario Sueco)	1436	98,8 %
Rehacom RT (módulo REA1)	1150	79,1 %
Time-of-Day (AM / MED / PM)	1454	100 %
Lactato	561	38,6 %
Handgrip	570	39,2 %
CMJ (Counter Movement Jump)	652	44,8 %
BORG	256	17,6 %

La distribución de KSS está fuertemente sesgada hacia valores bajos: el 62 % de las sesiones reporta $KSS \leq 3$, la mediana es 3 y solo **18 sesiones (1,25 %)** alcanzan $KSS \geq 7$. Esto hace inviable una binarización por KSS con el umbral canónico de la literatura (98,75 %/1,25 %) y motiva el uso de etiquetas alternativas. Las coberturas físicas (lactato, BORG, handgrip, CMJ) son intermedias y se utilizan para etiquetas de fatiga física en análisis exploratorios; los modelos principales se entrenan con las fuentes de alta cobertura.

4.4.3. Bitácora de Calidad de Señal

El filtrado por calidad de la Fase 2 se apoya en la bitácora del analista de Sigma descrita en la Sección 3.4.7: de las 474 sesiones con metadata estructurada (ciclos 7 a 12), el 80 % queda en categoría *Usable* (26 % Excelente, 54 % Buena, 11 % Regular, 9 % Mala) y solo las sesiones *Usable* entran al modelado.

4.4.4. Estrategias de Etiquetado: Triangulación de Candidatos

El protocolo Sigma admite múltiples fuentes de etiquetado paralelas; las **siete estrategias** evaluadas (cinco familias: diseño temporal, subjetiva, de desempeño, EEG-interna y multi-fuente) y la justificación de adoptar la pseudo-etiqueta EEG-interna de YASA se desarrollan en la Sección 3.4.6. Para los experimentos de validación basta recordar que YASA produce un hipnograma por época de 30 s del cual se derivan tres etiquetas a nivel sesión: `drowsy_binary` (al menos una época no-Wake; distribución 84/16), `drowsy_fraction` (fracción no-Wake $\geq 0,30$; distribución 53/47) y `stage_mode` (etapa modal, multiclase dominada por Wake). La distribución 53/47 de `drowsy_fraction` anticipa el mejor compromiso entre balance y especificidad, lo que se confirma en los experimentos siguientes. Su uso es *off-label* en las cinco dimensiones enumeradas en la Sección 4.4.6; en consecuencia se trata como *pseudo-etiqueta EEG-interna*, no como medida de fatiga subjetiva.

Visualización del desacuerdo entre estrategias. La Figura 4.2 ilustra, para un sujeto y día representativo (sub_05, ciclo 7, día 6), cómo cada una de las estrategias de etiquetado evaluadas clasifica los nueve bloques EEG de ese día (3 momentos $\times$ OA / OC / REA). La heterogeneidad de las decisiones es evidente: TOD asume el patrón circadiano clásico (alerta AM, fatigado MED+PM); RT z-score lo contradice directamente (flag de fatiga solo en AM); KSS canónico (umbral ≥ 7) no se activa en ninguna sesión, mientras que KSS sobre la mediana se activa en todas; YASA *drowsy_fraction* muestra granularidad bloque a bloque, capturando que REA (carga cognitiva del test Rehacom) tiende a producir mayor adormecimiento que OA/OC; *consensus majority 3-way* no alcanza el umbral de 2/3 votos en ninguna sesión (todas quedan en clase 0); y *clustering k=2* solo dispone de etiqueta en PM, evidenciando la limitación de cobertura

por la dependencia en BORG y lactato. Este desacuerdo no es ruido: es exactamente el problema metodológico que motiva el experimento de triangulación de la sección siguiente. Antes de comprometer un label primario al modelo final, es indispensable cuantificar cuál de estas etiquetas (o cuáles, en combinación) tienen señal aprendible desde el EEG.

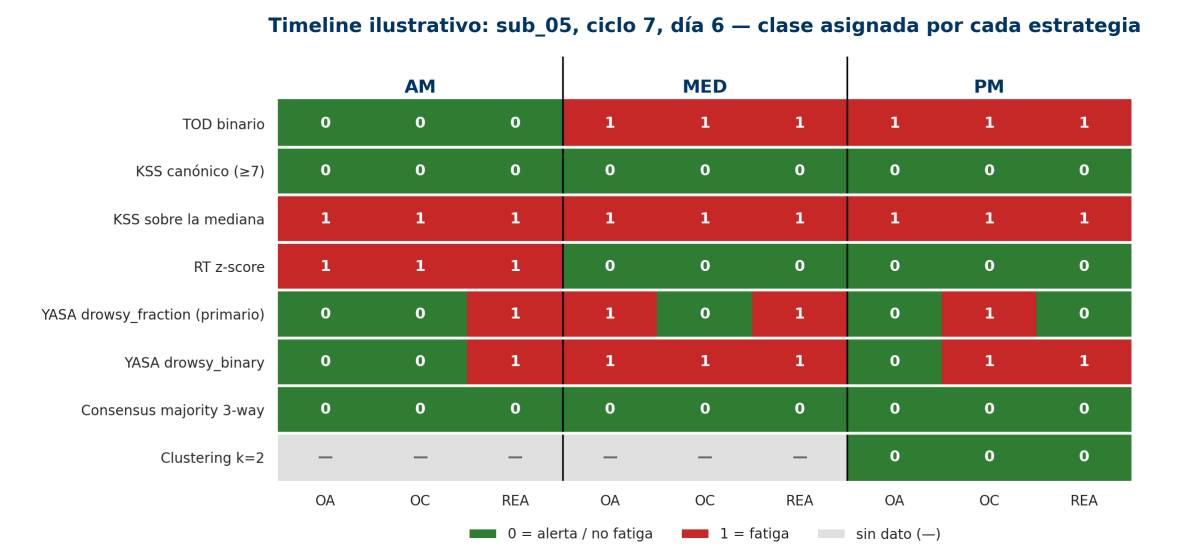


Figura 4.2: Timeline ilustrativo: clase (0 = alerta, 1 = fatiga) asignada por cada estrategia de etiquetado a los nueve bloques EEG del sujeto sub_05 en el ciclo 7, día 6. Las celdas grises (vacías) indican ausencia de etiqueta (caso típico: clustering, que requiere BORG y lactato medidos solo en PM). El desacuerdo inter-estrategia es la motivación empírica de los experimentos de validación.

4.4.5. Validación: Tres Experimentos antes de Comprometer un Label Primario

Antes de adoptar un label como primario, se ejecutaron tres experimentos de validación inspirados en las recomendaciones de la literatura ocupacional reciente (Purto, Anabalón, Vargas, Jara D & de la Vega, 2024; Giorgi, Ronca, Capotorto, Vozzi, Rossi, Aricò, Borghini, Van Gasteren, Melus, Petrelli, Sportiello, Polidori, Picardi, Babiloni & Di Flumeri, 2025).

Experimento 1: Replicación del perceptual blindness. Purto et al. (2024) reportaron una correlación de Pearson $r = -0,03$ ($p < 0,001$) entre auto-reporte de fatiga subjetiva (escala Samn–Perelli) y desempeño en un test de tiempo de reacción objetivo

(WAFA) en mineros, evidenciando que los trabajadores tienden a sub-reportar fatiga (*perceptual blindness*). Sobre los datos Sigma se ejecuta una replicación con instrumentos análogos: KSS como escala subjetiva (análoga a Samn-Perelli) y el tiempo de reacción del módulo REA1 de Rehacom como medida objetiva (análoga a WAFA, ambos tests de vigilancia psicomotora basados en tiempo de reacción). Sobre $n = 1137$ sesiones con ambos disponibles, la correlación de Spearman *agregada* es $\rho_{\text{pooled}} = -0,121$.

Ahora bien, las 1137 sesiones son medidas repetidas de 10 sujetos, por lo que el p -value ingenuo que las trata como independientes está anti-conservadoramente inflado; reportarlo sería pseudoreplicación. Respetando la estructura de *clusters* por sujeto, la asociación es *indistinguishible de cero*: la correlación media *within-subject* es $\bar{\rho} = -0,01$ (t -test de una muestra sobre los $n = 5$ sujetos con suficiente varianza de KSS para estimar ρ intra-sujeto, $p = 0,93$; *cluster bootstrap* sobre los 10 sujetos, IC 95 % del ρ agregado = $[-0,49; +0,28]$, que cruza cero).

Es decir, el ρ_{pooled} negativo proviene de variación *entre* sujetos, no de una relación *intra*-sujeto. Lejos de debilitar el argumento, esto **refuerza** el *perceptual blindness*: a nivel del sujeto (la unidad de análisis relevante) el auto-reporte KSS y el rendimiento atencional objetivo son prácticamente ortogonales, y la magnitud *within-subject* ($\bar{\rho} = -0,01$) coincide casi exactamente con el $r = -0,03$ reportado por [Purto et al. \(2024\)](#).

Esta replicación tiene una implicación directa, y poco discutida en la literatura, sobre la elección de label primario en cohortes mineras: solo 2 de los 10 sujetos (sub_03 con 10 ocasiones, sub_07 con 1) reportan $\text{KSS} \geq 7$ en alguna sesión; la prevalencia agregada del label `kss_canonical` es 1,0 % (11 de 1073 sesiones con KSS válido), con mediana $\text{KSS} = 3$ y rango intercuartílico nulo ($Q1 = Q3 = 3$). Bajo LOSO con 10 sujetos, `kss_canonical` es **estructuralmente inestable**: cualquier *fold* que excluya a sub_03 o sub_07 colapsa el conjunto de entrenamiento a mono-clase. La *perceptual blindness* no es un detalle metodológico menor sino la *razón estructural* por la cual el auto-reporte queda descartado como label primario en esta cohorte, antes de cualquier consideración de rendimiento del clasificador.

Experimento 2: Baseline LOSO por label. Se entrenaron cuatro clasificadores (LDA, Logistic Regression con `class_weight=balanced`, Random Forest con 200 árboles, y CatBoost con `auto_class_weights=Balanced`) sobre *features* EEG agregadas a

nivel sesión (potencias en bandas $\delta, \theta, \alpha, \beta, \gamma$ en 5 canales fronto-centrales más ratios espectrales, media y desviación estándar a través de ventanas; total ~ 30 *features*). La validación cruzada *Leave-One-Subject-Out* se realizó sobre los 10 sujetos de la cohorte (sub_02 queda excluido por no tener cobertura en la bitácora de calidad EEG: solo participó en ciclos 5–6, y la bitácora estructurada cubre ciclos 7–12). Bajo este filtro, el label `yasa_drowsy_fraction` produce 9 *folds* efectivos (sub_09 se omite por test mono-clase). La Tabla 4.11 reporta el mejor modelo para cada etiqueta candidata.

Cuadro 4.11: Mejor modelo por etiqueta candidata en LOSO con 10 sujetos, filtrando a sesiones *Usable* (Buena o Excelente según la bitácora del analista). Δ vs. chance es la diferencia entre Macro-F1 y el baseline trivial (predicción de clase mayoritaria).

Etiqueta	Macro-F1	Chance	Δ	n	Modelo
yasa_drowsy_binary	0,652	0,500	+0,152	849	LogReg
yasa_drowsy_fraction	0,631	0,500	+0,131	849	CatBoost
tod_binary	0,530	0,500	+0,030	850	CatBoost
consensus_majority	0,511	0,500	+0,011	850	CatBoost
kss_canonical	0,483	0,500	−0,017	844	LDA
kss_median	0,453	0,500	−0,047	844	LogReg
rt_zscore	0,443	0,500	−0,057	663	RF
tod_ternary	0,425	0,333	+0,092	850	RF
rt_q75	0,420	0,500	−0,080	663	CatBoost
kss_ordinal	0,366	0,333	+0,033	844	LDA
yasa_stage_mode	0,310	0,250	+0,060	849	CatBoost

Las dos etiquetas YASA binarias destacan: *yasa_drowsy_binary* (F1 = 0,652 con LogReg) y *yasa_drowsy_fraction* (F1 = 0,631 con CatBoost) superan al resto del catálogo por un margen claro. *tod_ternary* también supera moderadamente su *chance* (0,425 vs 0,333, $\Delta = +0,092$), mientras que *tod_binary* apenas supera el azar (F1 = 0,530, $\Delta = +0,030$). El *gradient boosting* domina en *yasa_drowsy_fraction* con una ventaja moderada sobre los modelos lineales, consistente con que la separación de fatiga en *features* espectrales involucra interacciones no estrictamente lineales. KSS canónico, KSS por mediana, KSS ordinal y los proxies de RehaCom RT no superan el azar, consistente con el patrón de Fase 1, donde solo las etiquetas alineadas con el factor experimental

mostraron señal por sobre el azar.

Etiquetas descartadas por inviabilidad estructural, no por rendimiento. La Tabla 4.11 subestima cuán fuertes son las restricciones sobre algunas etiquetas candidatas. Más allá del bajo F_1 , varias quedan descartadas *antes de poder ser evaluadas*: (i) `kss_canonical` (prevalencia 1,0 %, 11/1073 positivos, concentrados en 2 sujetos, ya discutido en el Experimento 1); (ii) `label_consensus_strict_3way`, que exige acuerdo simultáneo de TOD, KSS y Rehacom RT, alcanza solo 3 positivos sobre 1081 sesiones (0,28 %); (iii) `label_cluster_session_k2`, derivada de KMeans sobre *features* fisiológicas, tiene cobertura 13,5 % (limitada por la disponibilidad de Borg y lactato, ambos restringidos a cinco de los diez sujetos) y colapsa a mono-clase en la cohorte cubierta; y (iv) `label_rt_slowing`, definida como la ralentización del tiempo de reacción respecto a la línea base AM-OA de cada sujeto, resulta *degenerada*: al exigir un valor de RT válido tanto en el bloque base como en el bloque comparado, la intersección de sesiones sin NaN colapsa a una sola clase (todas las sesiones supervivientes quedan con el mismo valor), de modo que el indicador carece de varianza y no es evaluable como etiqueta. Estos cuatro candidatos se documentan por completitud de la triangulación pero no entran en la decisión del label primario; reportarlos como “no superan azar” sería engañoso, ya que su falla es de *estructura* del label, no de capacidad predictiva del clasificador.

Experimento 3: Shuffle control *within-subject*. Un F_1 superior a chance no garantiza que el modelo aprenda fatiga; puede simplemente memorizar identidad del sujeto cuando los *folds* LOSO comparten estructura inter-sujeto. Para descartar esto, se baraja aleatoriamente la etiqueta *dentro* de cada sujeto (rompiendo la relación EEG–fatiga pero preservando la firma del sujeto) y se re-entrena. Si el F_1 cae al nivel de chance, el modelo solo memorizaba sujeto. Se ejecutaron 20 *shuffles* por etiqueta candidata con LDA como clasificador fijo (Tabla 4.12).

Cuadro 4.12: Shuffle control *within-subject* para los cuatro labels candidatos, sobre las sesiones Usable. z -score indica cuántas desviaciones estándar el F1 real está sobre la media de la distribución de F1 *shuffled*.

Etiqueta	F1 real	F1 shuffle	Δ	z	Pasa
yasa_drowsy_binary	0,583	0,473	+0,110	8,41	Sí
yasa_drowsy_fraction	0,595	0,469	+0,126	5,84	Sí
tod_binary	0,467	0,414	+0,053	3,36	Sí (débil)
yasa_stage_mode	0,268	0,269	−0,001	−0,13	No

“Pasa” = el z supera el umbral de significancia $|z| > 2$.

Tres candidatos pasan el shuffle control. YASA *drowsy-binary* ($z = 8,41$) y *drowsy-fraction* ($z = 5,84$) confirman señal EEG genuina más allá de la memorización de sujeto. TOD *binary* pasa pero débil ($z = 3,36$), una magnitud notablemente inferior a las dos YASA. YASA *stage-mode* no pasa ($z = -0,13$), descartándose como candidato. El delta absoluto de YASA *drowsy-fraction* respecto al shuffle (+0,126) es más del doble que el de TOD (+0,053), confirmando que el EEG contiene mayor señal sobre el label EEG-derivado que sobre el label temporal.

Ablación de preprocesamiento ICA. Como verificación adicional se reprocesó el dataset aplicando ICA (Picard, 10 componentes, 200 iters máx.) con clasificación automática por ICLabel [Pion-Tonachini et al. \(2019\)](#) y exclusión de componentes `eye blink`, `muscle artifact` y `heart beat` con $p \geq 0,70$. Sobre 41 combinaciones (etiqueta, modelo), ICA degrada 32 casos, neutraliza 3 y mejora 6 (todos multi-clase con baseline cercano al azar). Para el label primario, CatBoost cae $-0,085$ Macro-F1. El resultado contraintuitivo es consistente con sesiones de 5–10 min, por debajo del mínimo recomendado para descomposición ICA estable, y con la posibilidad de que ICLabel ($\sim 85\%$ accuracy a $p = 0,70$) excluya componentes neurales legítimos. El detalle se documenta como material reproducible del proyecto. Esta ablación refuerza el diagnóstico que se sustenta a lo largo del capítulo: **el techo de rendimiento LOSO no se mueve al cambiar la representación, el modelo, la transferencia desde datasets públicos ni el preprocesamiento; el límite reside en el régimen de datos** (10 sujetos efectivos, ~ 849 sesiones bajo el filtro de calidad), no en el estimador.

4.4.6. Decisión del Label Primario: la Pseudo-etiqueta EEG-interna de YASA

La triangulación de los tres experimentos converge en una decisión que requiere ser declarada con precisión: `label_yasa_drowsy_fraction` se adopta como **pseudo-etiqueta EEG-interna primaria**, no como *etiqueta de fatiga* en el sentido clínico de auto-reporte o de un *ground truth* externo. La distinción es importante: dado el perceptual blindness replicado en el Experimento 1 y la estructura del dataset (mediana KSS = 3 en escala 1–9, prevalencia $KSS \geq 7$ del 1,0% concentrada en 2 sujetos), *no existe* en SIGMA un *ground truth* subjetivo viable contra el cual entrenar y evaluar bajo LOSO. La pseudo-etiqueta derivada de YASA es el único candidato con cobertura, distribución y señal aprendibles. La Tabla 4.13 contrasta los dos finalistas testables (YASA *drowsy_fraction* y TOD *binary*), reservando `tod_binary` como *ablation* comparativa, nunca como label primario.

Cuadro 4.13: Comparación de los dos candidatos testables bajo los criterios de decisión. Las restantes etiquetas evaluadas son estructuralmente inviables (Sección 4.4.5).

Criterio	yasa_drowsy_fraction	tod_binary
Cobertura	849 sesiones Usable	850 Usable
Distribución	53 % / 47 %	39 % / 61 %
F1 LOSO (mejor modelo)	0,631 (CatBoost)	0,530 (CatBoost)
Δ vs. shuffle	+0,126	+0,053
z-score shuffle	5,84	3,36
Origen	EEG directo (AASM)	Proxy temporal
Sustento literatura	Vallat & Walker (2021)	Borbély (1982)

En todos los criterios cuantitativos YASA supera a TOD. La justificación es además funcional: TOD asume que la fatiga sigue la hora del día (presión homeostática del sueño, Borbély (1982)), lo cual es un *proxy temporal* que no mide directamente el EEG; YASA, en cambio, mira la traza EEG *epoch a epoch* y la clasifica contra un modelo entrenado sobre miles de polisomnografías validadas. Si el EEG contiene señal aprendible bajo LOSO, una pseudo-etiqueta EEG-driven debería ganarle a una etiqueta puramente temporal, y eso es lo que muestran los datos. La preferencia por `drowsy_fraction` sobre

`drowsy_binary` (que alcanza $F1 = 0,583$ con LDA y $z = 8,41$) responde al balance de clases: 53/47 % frente a 82/18 %, lo que evita Macro-F1 inflado por la clase mayoritaria.

Validez concurrente con auto-reporte: nula por diseño. Para no inflar la validez externa del label primario, reportamos la concordancia con el auto-reporte KSS sobre las 835 sesiones con ambos disponibles: Cohen’s $\kappa = 0,001$ entre `yasa_drowsy_fraction` y `kss_canonical` (concordancia indistinguible del azar); Spearman $\rho = 0,081$ entre el label binario `yasa_drowsy_fraction` (umbralizado en $\tau = 0,30$) y la escala KSS 1–9 ($p = 0,020$, $n = 835$, estadísticamente significativo solo por tamaño muestral, magnitud prácticamente cero). Esta concordancia nula *no* es un defecto del label: es la *réplica esperada* del perceptual blindness reportado por [Purto et al. \(2024\)](#) y confirmado en el Experimento 1. La pseudo-etiqueta YASA y el KSS miden dimensiones ortogonales en esta cohorte minera, y la pseudo-etiqueta queda definida operacionalmente como “*proporción de épocas EEG diurnas de 30 s que YASA clasifica como no-Wake bajo el principio AASM*”, sin reclamo de equivalencia con la fatiga vivenciada por el sujeto.

Limitaciones declaradas y restricciones de inferencia. El uso de YASA en este contexto es *off-label* en cinco dimensiones específicas que conviene enumerar para acotar el alcance de cualquier afirmación posterior:

- a) **Régimen de entrenamiento.** [Vallat & Walker \(2021\)](#) entrenaron YASA exclusivamente sobre polisomnografía nocturna continua (NSRR: SHHS, MESA, CFS, etc., ~ 31.000 h) y declaran explícitamente: “*performance on shorter recordings, such as daytime naps, is unknown*” y que “*performance may be reduced on data shorter than the longest smoothing windows used by the algorithm (i.e., shorter than 7,5 minutes)*”. En SIGMA, 18 % de las sesiones *Usable* duran menos de 5 minutos y 14 % duran menos de 4 minutos; el filtro `min_epochs` ≥ 8 (≥ 4 min de EEG) recoge esta restricción (ver Sección 4.4.7).
- b) **Canal único, sin referenciación estándar.** La integración usa `eeg_name=C3` (un solo canal del montaje OpenBCI Cyton+Daisy, sin referencia mastoide ni Laplaciano). YASA es robusto a configuraciones de canal moderadas según los autores, pero la ausencia de mastoide y la diferencia frente a polisomnografía clínica estándar es una fuente de varianza no caracterizada en esta cohorte.

- c) **Umbral de binarización ($\tau = 0,30$) no canónico.** El estándar AASM clasifica épocas, no sesiones; la agregación a nivel sesión por umbralización es una decisión de esta tesis. La robustez a $\tau \in \{0,20; 0,30; 0,50\}$ se documenta en la Sección 4.4.7, pero el umbral mismo carece de respaldo en literatura.
- d) **Heterogeneidad por estado ocular.** Los tres bloques (OA, OC, REA) tienen estados oculares distintos: OA y REA son *eyes-open* (resting y carga cognitiva respectivamente), OC es *eyes-closed*. La regla AASM-N1 dispara con atenuación de alfa $> 50\%$ de la época, que sucede tanto en estados de transición a sueño como en cualquier EEG diurno alerta con ojos abiertos. La separación de la señal de fatiga del estado ocular se controla mediante una estratificación por bloque (Sección 4.4.7, control (d)), que muestra que la señal discriminante persiste *dentro* de cada estado ocular (en particular en *eyes-closed*, OC), acotando este confound.
- e) **Ausencia de ground truth externo (PVT, hipnograma manual).** No se aplicó PVT (Psychomotor Vigilance Task, gold standard del campo) en el protocolo Sigma y no existe re-anotación clínica AASM de los registros. La validez concurrente externa en Sigma permanece sin verificar; lo único validado es la *generalización inter-sujeto* bajo LOSO (Experimento 3, shuffle control $z = 5,84$; Sección 4.4.7, contraste contra TOD residualizado).

Validación independiente sobre MPD-DF: el uso *off-label* produce acuerdo de azar con etiquetado clínico AASM-like. Para acotar empíricamente la limitación 5 anterior, se ejecuta una validación externa sobre el dataset MPD-DF (Li, Fu, Tang, Zhou, Chen, Zhou, Chen & Luo, 2026) (Fase 1-bis), único corpus público con etiquetado clínico por médico especialista en sueño aplicado a un protocolo de fatiga diurna. Sobre los **50 sujetos completos** del dataset (12.413 épocas de 30 s en total, canal C3, ~ 2 h por sesión) se aplica YASA con idéntica configuración a la usada en SIGMA, y se compara su predicción AASM binarizada (WAKE *vs.* N1+N2+N3+REM) contra el etiquetado clínico binarizado (nivel 0 *vs.* niveles 1–4) época a época. El acuerdo agregado es Cohen’s $\kappa = +0,010 \pm 0,134$ sobre los cincuenta sujetos (mediana $-0,002$, rango $[-0,245; +0,357]$), con *accuracy* pooled de 0,530 y sensibilidad pooled de 0,503 para la clase *drowsy*, estadísticamente indistinguible del azar. Solo 4 de 50 sujetos al-

canzan $\kappa > 0,20$ (acuerdo leve) y 25 quedan bajo cero.

Conviene ser explícito sobre lo que este número puede y no puede sostener: la comparación opone la predicción de YASA (*¿estadio de sueño no-Wake?*) contra el etiquetado clínico de MPD-DF, cuyos niveles 1–2 (“emergencia de alfa”, “alfa difusa”) corresponden a *vigilia con fatiga* y no a estadios de sueño AASM. Son, por tanto, constructos distintos, y un κ cercano a cero es el resultado *esperado por diseño*: no separa por sí solo un eventual fallo *off-label* de YASA del simple desajuste entre “propensión al sueño” y “fatiga clínica diurna”.

Leído con esa cautela, el resultado **acota el alcance** del label más que validarlo: es *consistente* con la distinción declarada al inicio de esta subsección (**la pseudo-etiqueta no es intercambiable con una evaluación clínica de fatiga**) y refuerza, junto con la concordancia nula frente al auto-reporte KSS en SIGMA ($\kappa = 0,001$), que YASA debe entenderse como un *marcador EEG-interno* reproducible (generalización inter-sujeto $z = 5,84$), no como un predictor de fatiga subjetiva ni de estados AASM clínicamente equivalentes. La afirmación de que el modelo “detecta fatiga” en el sentido vivencial o clínico no está respaldada por estos datos y no se sostiene en esta tesis.

Estas limitaciones no invalidan el uso del label como predictor objetivo bajo LOSO, pero acotan lo que la tesis puede afirmar: los clasificadores aprenden a predecir, desde *features* EEG independientes del sujeto, una pseudo-etiqueta espectral que YASA produce desde una región distinta del espacio de *features*, con generalización inter-sujeto significativa. No aprendemos a predecir fatiga subjetiva ni desempeño cognitivo en mineros chilenos; ese objetivo requiere ground truth externo que el dataset no contiene.

4.4.7. Robustez del Label Primario: Sensibilidad al Threshold y Confound TOD

Adoptar `yasa_drowsy_fraction` como label primario expone dos preguntas naturales sobre la robustez de la elección: (a) el umbral de 0,30 *drowsy epoch fraction* ¿es un hiperparámetro implícito que infla el F1?; y (b) ¿el modelo aprende fatiga del EEG o simplemente memoriza el bloque del día (AM/MED/PM), dado que la fatiga circadiana correlaciona con la hora? Ambas preguntas se responden con ablaciones específicas.

(a) Sensibilidad al threshold y al filtro de épocas. Se evaluó la matriz 3×3 de thresholds $\tau \in \{0,20; 0,30; 0,50\}$ y filtros de épocas mínimas $\text{min}_{\text{ep}} \in \{1; 8; 10\}$ con CatBoost LOSO. La Tabla 4.14 resume los resultados; el output completo se documenta como material reproducible del proyecto.

Cuadro 4.14: Sensibilidad de `yasa_drowsy_fraction` al threshold y al filtro mínimo de épocas. CatBoost LOSO sobre la cohorte con `min_epochs=1` ($N = 861$, e incluye los 10 *folds*). Por esta razón ninguna celda coincide con el titular Macro-F1 = 0,631, que se computa sobre el *snapshot* congelado de 849 sesiones con 9 *folds* efectivos (`sub_09` excluido por test mono-clase); la celda comparable ($\tau = 0,30$, `min_epochs=1`) da 0,593.

τ	min_{ep}	N	prev (%)	F1 ($\mu \pm \sigma$)	Bal. Acc
0,20	1	861	59	$0,632 \pm 0,137$	0,662
0,30	1	861	53	$0,593 \pm 0,109$	0,620
0,50	1	861	26	$0,566 \pm 0,058$	0,613
0,20	8	711	63	$0,606 \pm 0,122$	0,643
0,30	8	711	56	$0,622 \pm 0,126$	0,659
0,50	8	711	28	$0,549 \pm 0,064$	0,601
0,20	10	681	64	$0,632 \pm 0,127$	0,668
0,30	10	681	57	$0,584 \pm 0,116$	0,626
0,50	10	681	28	$0,560 \pm 0,071$	0,614

La configuración **recomendada** es $\tau = 0,30$ con $\text{min}_{\text{ep}} \geq 8$ (F1 = 0,622), no $\tau = 0,30$ con $\text{min}_{\text{ep}} = 1$ (F1 = 0,593). El filtro $\text{min}_{\text{ep}} \geq 8$ corresponde a ≥ 4 minutos de EEG efectivo, cota que reconoce explícitamente la advertencia de [Vallat & Walker \(2021\)](#) sobre que el desempeño es “unknown” bajo 7,5 minutos (Sección 4.4.6, limitación 1); aplicarlo descarta las 200 sesiones < 5 minutos y las 153 sesiones < 4 minutos que YASA marca como fuera de su régimen de validación. El resultado experimental confirma la decisión teórica: el F1 sube de 0,593 a 0,622 (+0,029) al pasar de $\text{min}_{\text{ep}} = 1$ a $\text{min}_{\text{ep}} = 8$ con $\tau = 0,30$. La cohorte efectiva reduce de 861 a 711 sesiones, pero la cobertura por sujeto se mantiene completa.

Sobre el threshold mismo: $\tau = 0,50$ degrada el F1 sistemáticamente porque colapsa la prevalencia a 26–28 %, mientras $\tau = 0,20$ produce un F1 alto pero a costa de prevalencias 59–64 % que reducen el contenido discriminante del label. La elección $\tau = 0,30$ es

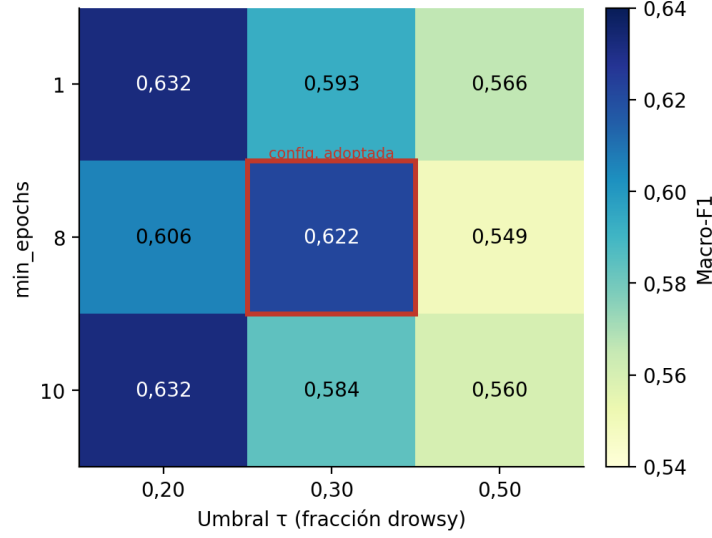


Figura 4.3: Sensibilidad del Macro-F1 LOSO a los dos hiperparámetros de la pseudo-etiqueta YASA: el umbral τ de fracción *drowsy* (columnas) y el mínimo de épocas válidas *min_epochs* (filas). El recuadro rojo marca la configuración adoptada ($\tau = 0,30$, *min_epochs* ≥ 8); el rendimiento se mantiene en una meseta estable, descartando que el umbral sea un hiperparámetro ajustado para inflar el F1.

la heurística operacional de esta tesis (Sección 4.4.6, limitación 3) y la robustez de F1 al variar $\tau \in \{0,20; 0,30; 0,50\}$ y al filtro de épocas $\text{min}_{\text{ep}} \in \{1; 8; 10\}$ acota el efecto del umbral particular sobre el desempeño reportado (rango $F_1 = 0,55\text{--}0,63$ sobre la matriz 3×3). Cabe notar que la robustez de F1 al threshold *no* es evidencia de que el threshold sea el correcto; es evidencia de que el clasificador encuentra estructura aprendible bajo cualquier corte razonable sobre la fracción *drowsy* continua, lo cual es una observación complementaria pero independiente del problema más profundo de validez concurrente discutido en la sección anterior.

(b) Test del confound TOD. Si el modelo simplemente memorizara el bloque del día (AM vs MED vs PM), un clasificador entrenado SOLO con la one-hot encoding del *moment* alcanzaría el mismo F1 que YASA-sobre-EEG. Se evaluó este baseline LOSO con CatBoost: $F1 = 0,412$. El contraste con YASA-sobre-EEG ($F1 = 0,602 \pm 0,105$) genera $\Delta = +0,189$, con **10 de 10 folds** mostrando EEG $>$ TOD (Wilcoxon signed-rank $p = 0,002$, *paired t-test* $p = 0,0003$). El intervalo de confianza al 95 % sobre el delta no incluye cero. Esto rechaza la hipótesis de que el modelo está aprendiendo TOD desde el EEG: la señal EEG aporta $\sim 0,19$ puntos de Macro-F1 sobre el simple bloque

temporal, una contribución sustancial y consistente entre sujetos.

(c) **Auditoría con *confident learning*.** Como verificación adicional de las limitaciones 3 (umbral $\tau = 0,30$) y 4 (heterogeneidad de estado ocular) declaradas en la Sección 4.4.6, se aplicó la metodología de *Confident Learning* Northcutt, Jiang & Chuang (2021) (implementación `cleanlab` 2.9.0) al label primario. La idea del algoritmo, en una línea: dada una etiqueta ruidosa $\tilde{y}$ y predicciones $p(y | x)$ de un clasificador, estima la matriz de confusión *joint* $Q_{\tilde{y}, y^*}$ entre la etiqueta dada y la etiqueta latente y^* sin necesidad de *ground truth* limpio, bajo el supuesto de ruido condicional a clase ($P(\tilde{y} = i | y^* = j, x) = P(\tilde{y} = i | y^* = j)$). Las muestras donde $\tilde{y}$ es probable diferente de y^* quedan *flagged*, ordenadas por una *quality score* basada en la *self-confidence* del modelo.

El *pipeline* utiliza CatBoost B0 (configuración canónica de la Tabla 4.13) para generar predicciones *out-of-fold* bajo LOSO sobre las 861 sesiones disponibles; `cleanlab` identifica entonces las muestras *flagged* mediante el método `prune_by_noise_rate` por defecto. El audit completo se documenta como artefacto reproducible del proyecto.

Resultado de la auditoría. `Cleanlab` *flag* 203 de 861 sesiones (23,6 %). Esta cifra *no* es “23,6 % del label está mal”; es la fracción donde la predicción *out-of-fold* del clasificador discrepa con la etiqueta. La discrepancia puede deberse a (i) ruido genuino del label, (ii) sub-óptimalidad del clasificador, o (iii) ambigüedad intrínseca de la sesión (*drowsy fraction* cerca del umbral 0,30). El patrón estructural observado sugiere que la fuente (iii) domina: las sesiones *flagged* tienen *drowsy fraction* concentrada cerca del umbral (media 0,287, desviación 0,203) frente al subconjunto limpio (media 0,315, desviación 0,244). El patrón refuerza cuantitativamente la limitación 3 de la Sección 4.4.6: la fricción metodológica principal del label es la binarización forzada de una variable continua. Adicionalmente, OC presenta 19,3 % de *flagging* frente a 26,3 % de OA y 25,0 % de REA, refuerzo modesto pero cuantitativo de la limitación 4 (heterogeneidad de estado ocular): YASA-AASM es más consistente con su propia predicción en *eyes-closed*, el régimen más cercano al *training data* nocturno de Vallat & Walker (2021). La distribución de *flagging* es uniforme por sujeto (14,5–29,3 %, sin *outliers*), por *moment* (22,8–24,6 %) y por duración (uniforme entre los cinco bins de < 4 a > 10 min), lo cual implica que el audit no aporta evidencia independiente para promover el filtro

$\min_{ep} \geq 8$ desde estructura de *flagging*; ese filtro queda justificado únicamente por la cita literaria de Vallat & Walker (limitación 1).

(d) Control del confound de estado ocular. La limitación 4 (Sección 4.4.6) advierte que la regla AASM-N1 de YASA se dispara con la atenuación del ritmo alfa, fenómeno que también distingue los bloques *eyes-open* (OA, REA) de los *eyes-closed* (OC). Surge así una objeción legítima: ¿el clasificador detecta adormecimiento, o meramente el estado ocular del bloque, trivialmente separable desde el EEG? Para responderla se ejecutan dos controles complementarios sobre el mismo CatBoost LOSO, cuyos resultados resume la Tabla 4.15.

Primer control: estratificación por bloque. Al entrenar y evaluar *dentro* de un único tipo de bloque, el estado ocular queda constante y no puede usarse como atajo. El Macro-F1 se mantiene sobre el azar en los tres estratos pese a perder esa variación: OC 0,578 (*balanced accuracy* 0,607), REA 0,587 (0,666) y OA 0,528 (0,581). Que la señal sobreviva dentro de *eyes-closed* (OC), donde todos los bloques comparten ojos cerrados, indica que el modelo **no** clasifica únicamente la configuración ocular.

Segundo control: descomposición del techo pooled. El F1 por estrato ($\sim 0,56$) es inferior al de la cohorte combinada (0,634), brecha que admite dos causas confundidas: (a) cada estrato entrena con apenas un tercio de los datos y (b) el modelo combinado aprovecha la estructura discriminante *entre* bloques (estado ocular y prevalencia por tarea) que la estratificación elimina. Para separarlas se introduce un punto de control intermedio, el *pooled submuestreado* aleatoriamente a ~ 280 sesiones (igualando el tamaño por estrato pero *conservando* la mezcla de bloques, promedio de 10 semillas), de modo que cada comparación de la Tabla 4.15 varíe una sola condición. El *pooled* completo frente al submuestreado aísla el efecto del **tamaño muestral** ($0,634 \rightarrow 0,592$, $\Delta \approx -0,04$); el submuestreado frente a un solo bloque aísla el aporte de la **estructura entre-bloques** a igual n ($0,592 \rightarrow 0,564$, $\Delta \approx -0,03$). La brecha total ($\sim 0,07$) se reparte así en magnitudes comparables, con el tamaño muestral algo predominante.

Lectura. El confound ocular es real y contribuye al desempeño *pooled* (y así se declara), pero *no* lo explica: bajo control (un único estado ocular) sobrevive señal de adormecimiento por encima del azar. La descomposición completa se documenta como material reproducible del proyecto.

Cuadro 4.15: Control del confound de estado ocular: Macro-F1 LOSO (CatBoost) bajo estratificación por bloque y submuestreo. El *pooled* completo vs. submuestreado aísla el efecto del tamaño muestral; el submuestreado vs. un solo bloque, el aporte de la estructura entre-bloques a igual n . La señal intra-bloque (incluido *eyes-closed* OC) se mantiene sobre el azar (0,5).

Configuración	n	Mezcla de bloques	Macro-F1
Pooled completo	840	sí	0,634
Pooled submuestreado (~ 280 , 10 semillas)	~ 280	sí	$0,592 \pm 0,041$
Solo OC (<i>eyes-closed</i>)	278	no	0,578
Solo REA	280	no	0,587
Solo OA	282	no	0,528

Advertencia sobre la curva de sensibilidad F1 al *drop* de *flagged*. Al re-entrenar LOSO descartando las sesiones de menor *quality score*, el F1 reportado sube de 0,623 a 0,738 al 10 % de *drop* y a 0,845 al 20 % de *drop*. Esta aparente mejora *no* es evidencia de *label cleaning* efectivo: es ***selection bias*** estructural. Cleanlab identifica las muestras donde el modelo dudaba; eliminarlas del conjunto de evaluación reduce el problema a su subconjunto fácil, inflando F1 por construcción. Tres indicadores confirman la naturaleza espuria del aumento: (i) la desviación estándar *inter-fold* crece monótonamente con el *drop* ($0,086 \rightarrow 0,144$), mientras que un *label cleaning* genuino reduciría varianza; (ii) la composición de clases se mantiene aproximadamente 45/55 a lo largo de la curva, descartando rebalanceo como causa; (iii) las muestras eliminadas tienen *drowsy fraction* concentrada en el rango 0,20–0,35, es decir, son los casos intrínsecamente difíciles del problema, no errores. **El F1 reportado sobre el conjunto completo de sesiones permanece en $0,602 \pm 0,105$** ; la curva F_1 -bajo-*drop* se reporta únicamente como análisis de sensibilidad documentado, sin ser usado para entrenar el modelo final ni para construir afirmaciones de rendimiento.

4.4.8. Cobertura Horaria y Estructura Temporal del Dataset

YASA procesó 1109 de las 1163 sesiones de bloques cortos con EEG (95,4 %); las 54 sesiones restantes fueron descartadas por longitud insuficiente para una época AASM de 30 s o por fallos de carga del XDF. La duración total de EEG procesado es 123,9

horas, de las cuales **89,7 horas** sobre 839 sesiones quedan en categoría *Usable* (Buena o Excelente, 72 % del total).² Por bloque, las horas *Usable* se distribuyen como OA 21,5 h, OC 21,3 h y REA 46,9 h, reflejando que REA es típicamente el doble de largo que OA u OC. La duración mediana es 5,0 min para OA y OC, y 10,0 min para REA. OA y OC capturan estados *resting* puro mientras que REA captura EEG bajo carga cognitiva del test Rehacom REA1.

El detalle completo de la cobertura por categoría de bitácora, la distribución de duraciones por bloque y la justificación del mapeo de REA al test cognitivo se reportan en el Apéndice B.1 (Tablas B.1 y B.2).

4.4.9. Pipeline de Extracción de *Features* y Validación con LOSO

Las señales EEG de 16 canales (125 Hz, OpenBCI Cyton+Daisy) fueron preprocesadas con un filtro *notch* a 50 Hz ($Q = 30$) y un filtro pasa-banda Butterworth de orden 4 (0,5 – 45 Hz). Cada sesión fue segmentada en ventanas de 120 s con *stride* de 60 s (50 % *overlap*). Para cada ventana se extrajeron **306 *features handcrafted***: 19 por canal (5 potencias en bandas vía PSD de Welch, 2 ratios espectrales θ/α y $(\theta + \alpha)/\beta$, 5 entropías diferenciales por banda, 3 parámetros de Hjorth, y 4 estadísticos temporales) $\times$ 16 canales más 2 asimetrías alfa inter-canal (frontal Fp1/Fp2 y temporal T3/T4). Las *features* fueron estandarizadas (`StandardScaler` por *fold*) antes de la clasificación. La validación cruzada *Leave-One-Subject-Out* sobre los 10 sujetos de la cohorte (`sub_02` excluido, ver Experimento 2) produjo los resultados ya reportados en las Tablas 4.11 y 4.12. El mejor modelo para el label primario `yasa_drowsy_fraction` es CatBoost (200 iteraciones, profundidad 6, `auto_class_weights=Balanced`) con Macro-F1 = 0,631. El z -shuffle = 5,84 (Tabla 4.12) corresponde al baseline LDA F1 = 0,595, confirmando señal estadísticamente significativa sobre la distribución *within-subject shuffled*. Un control complementario con 1000 *shuffles* sobre CatBoost (sobre el conjunto de *features* B1, F1 = 0,584) confirma la conclusión ($z = 3,95$, $p_{\text{emp}} < 10^{-3}$: 0/1000 *shuffles* alcanzan

²El conteo de sesiones *Usable* oscila en pocas unidades entre las distintas pasadas del *pipeline*: 839 en el barrido de horas (sobre las 1109 estadificadas por YASA), 849 = 849 en el *snapshot* de modelado congelado y 851 en la reproducción actual (Tabla 3.1). La variación proviene de la regeneración del *parquet* de *features* y de la validez de *label* YASA exigida en cada pasada, y es inmaterial para las métricas reportadas, que usan el *snapshot* de 849.

$F1_{\text{real}}$). Para el label *ablation tod_binary*, el mejor modelo es CatBoost con $F1 = 0,530$.

Lectura operacional: AUC-ROC y balance sensibilidad–especificidad. El Macro-F1 resume el rendimiento a un umbral fijo (0,5); para un sistema cuyo objetivo es un *score de riesgo* graduado conviene complementarlo con una métrica independiente del umbral y con una lente de seguridad. La Tabla 4.16 y la Figura 4.4 reportan, sobre las predicciones *out-of-fold* del label primario *yasa_drowsy_fraction*, el AUC-ROC y el compromiso sensibilidad–especificidad para los cuatro clasificadores.

Cuadro 4.16: Métricas operacionales sobre *yasa_drowsy_fraction* (LOSO, cohorte Usable, 10 sujetos). AUC-ROC sobre predicciones *out-of-fold* agregadas (*pooled*), con la media por-*fold* entre paréntesis; sensibilidad (*recall* de la clase fatiga) y especificidad (*recall* de la clase alerta) al umbral 0,5; y la especificidad alcanzable en el punto operacional que prioriza sensibilidad $\geq 0,80$.

Modelo	Macro-F1	AUC-ROC	Sens@0,5	Espec@0,5	Espec @ sens $\geq 0,80$
CatBoost	0,631	0,771 (0,72)	0,77	0,65	0,58
Random Forest	0,608	0,775 (0,72)	0,80	0,60	0,60
Logistic Reg.	0,576	0,712 (0,68)	0,72	0,59	0,51
LDA	0,577	0,707 (0,69)	0,76	0,54	0,49

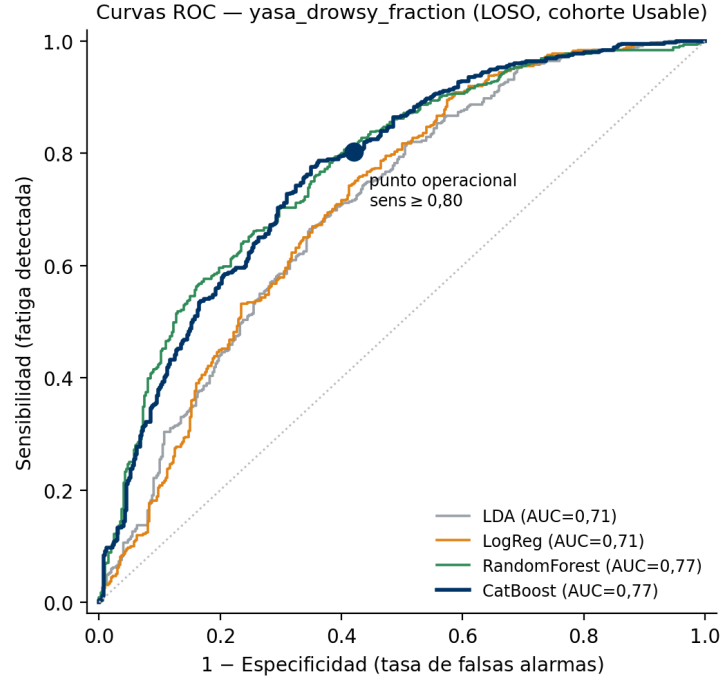


Figura 4.4: Curvas ROC del label primario `yasa_drowsy_fraction` (predicciones *out-of-fold*, LOSO, cohorte Usable). El punto destacado sobre CatBoost marca el umbral operacional que prioriza una sensibilidad $\geq 0,80$. La diagonal punteada corresponde al azar.

La lectura es deliberadamente equilibrada. Por un lado, el **AUC-ROC de 0,771** (CatBoost; media por-*fold* $0,72 \pm 0,08$, Random Forest casi idéntico con 0,775) indica que el modelo **ordena el riesgo de la pseudo-etiqueta de adormecimiento con capacidad discriminante moderada** (en el rango habitualmente calificado de “aceptable”, 0,7–0,8), una lectura más informativa que el Macro-F1 = 0,631, que queda atado a un único umbral y subestima la señal de ordenamiento presente en las probabilidades. Por otro lado, el balance sensibilidad–especificidad **expone el límite operacional con crudeza**: para detectar el 80 % de los estados de fatiga (sensibilidad clínicamente prioritaria, pues un falso negativo implica un operador fatigado no advertido) el sistema debe aceptar cerca de un 40 % de falsas alarmas (especificidad $\approx 0,58$). En términos de seguridad operacional minera, el modelo **ordena el riesgo de forma aprovechable pero no alcanza todavía un punto de operación de baja tasa de falsas alarmas a alta sensibilidad**: un compromiso esperable bajo el régimen de $N = 10$ sujetos y coherente con el techo por tamaño muestral documentado en este capítulo. El AUC se reporta tanto en su forma agregada (*pooled*) como en media por-*fold* para

transparentar que la primera es levemente optimista al asumir comparabilidad de las probabilidades entre *folds*.

4.4.10. Interpretabilidad del Modelo Primario

Un Macro-F1 por sobre el azar no garantiza que el modelo se apoye en señal fisiológica genuina en lugar de artefactos. Para auditarlo se analiza el modelo primario (CatBoost sobre `yasa_drowsy_fraction`) desde dos ángulos complementarios sobre el mismo clasificador, que es además el recomendado para el sistema. Las redes profundas no se interpretan aquí: quedaron por debajo de los clásicos (Sección 4.4.11) y no son el modelo seleccionado.

Valores SHAP: ¿qué mira el modelo? SHAP (*SHapley Additive exPlanations*) (Lundberg & Lee, 2017) atribuye a cada *feature* una contribución aditiva a cada predicción individual, derivada de los valores de Shapley de la teoría de juegos cooperativos; su variante *TreeExplainer* la calcula de forma exacta y eficiente sobre modelos de árboles. La Figura 4.5 muestra el análisis SHAP global sobre CatBoost (al ser un modelo de árboles es invariante a escalado monótono, por lo que corresponde al mismo modelo del titular $F1 = 0,631$). El resultado es **fisiológicamente coherente**: la *feature* más influyente es el **ratio espectral** $(\theta + \alpha)/\beta$ (Jap, Lal, Fischer & Bekiaris, 2009), uno de los índices espectrales evaluados por sus autores para la detección de fatiga y un marcador clásico del campo, y su contribución es **monótona con el riesgo** (valores altos $\rightarrow$ predicción de fatiga, $|\text{SHAP}| = 0,49$). Le siguen la **potencia relativa** θ (también $\uparrow$ riesgo, coherente con el aumento de actividad lenta en somnolencia) y la **potencia** γ (en dirección $\downarrow$ riesgo, esto es, asociada a vigilia/alerta), junto al ratio α/θ ($\downarrow$ riesgo). Que las direcciones de todas las *features* dominantes coincidan con la literatura de fatiga EEG confirma que el modelo extrae señal genuina, no estructura espuria del particionado.

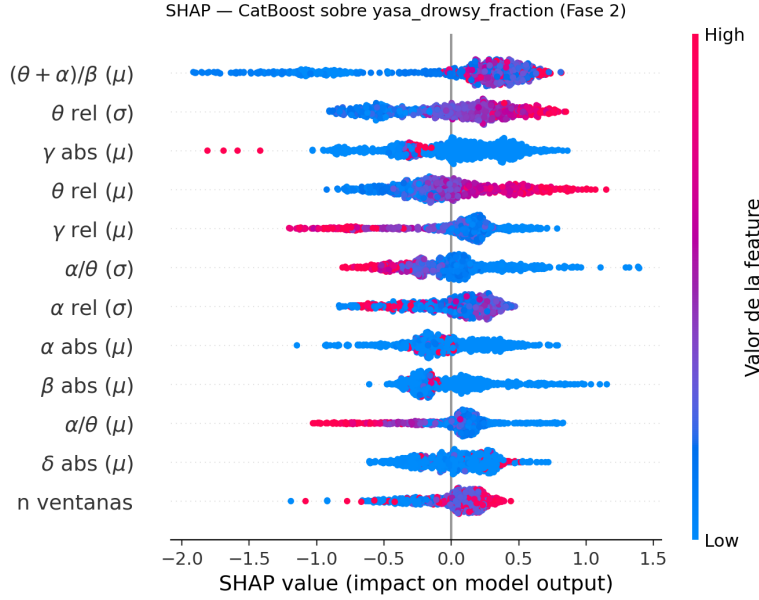


Figura 4.5: Resumen SHAP global del modelo primario (CatBoost, `yasa_drowsy_fraction`). Cada punto es una sesión; el color codifica el valor de la *feature* (rojo alto, azul bajo) y el eje horizontal su impacto sobre la salida (positivo $\rightarrow$ clase fatiga). Las *features* se ordenan por importancia media absoluta. El índice de Jap $(\theta + \alpha)/\beta$ domina y opera en la dirección esperada por la fisiología de la somnolencia.

Ablación por grupos: ¿dónde reside la señal? Como contraste causal, se elimina cada grupo de *features* y se re-evalúa el modelo bajo el mismo LOSO (Figura 4.6). El hallazgo es honesto y consistente con el techo observado: **ningún grupo de banda es indispensable**. La mayor caída al eliminar un grupo es de apenas $\Delta F1 = 0,022$ (banda δ), seguida de β (0,021) y θ (0,016); eliminar la banda γ incluso mejora marginalmente ($-0,004$). Es decir, la señal discriminante está **distribuida y es redundante** entre bandas espectrales, sin un único marcador que la concentre. Esto es coherente con un régimen de señal débil: el modelo recupera información fisiológicamente correcta (SHAP) pero ninguna banda aislada porta señal fuerte (ablación), lo que enmarca con precisión el rendimiento moderado del label primario.

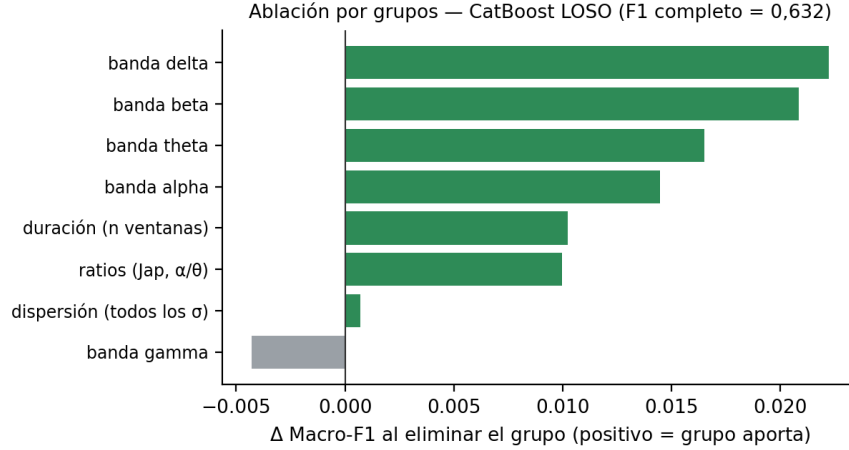


Figura 4.6: Ablación por grupos de *features* (CatBoost, LOSO). Cada barra es la caída de Macro-F1 al eliminar el grupo respecto al modelo completo ($F1 = 0,632$); barras positivas indican grupos que aportan. Ninguna eliminación individual degrada el rendimiento más de $\sim 0,02$, evidenciando señal espectral distribuida y redundante.

4.4.11. Comparación con Fase 1 y Fase 1-bis

La Tabla 4.17 sintetiza el rendimiento del mejor clasificador clásico en cada fase de la tesis, manteniendo el etiquetado más defendible en cada caso.

Cuadro 4.17: Comparación del rendimiento del mejor clasificador clásico entre fases.

Fase	Dataset	Canales	Sujetos	Macro-F1
Fase 1	FatigueSet (intensidad)	4	12	0,620 (SVM)
Fase 1-bis	MPD-DF (etiquetado clínico)	32	50	0,728 (CatBoost)
Fase 1-bis	MPD-DF (subconjunto 16 ch)	16	50	0,711 (CatBoost)
Fase 2	Sigma (YASA)	16	10	0,631 (CatBoost)
Fase 2	Sigma (TOD ablation)	16	10	0,530 (CatBoost)

El $F1 = 0,631$ de Fase 2 sobre Sigma es **inferior** al obtenido en MPD-DF (0,711 con el mismo montaje de 16 canales). Esta brecha es esperable y se atribuye a tres factores: (i) MPD-DF cuenta con etiquetado clínico realizado por un médico especialista en sueño, mientras Sigma usa YASA aplicado *off-label*; (ii) MPD-DF cuenta con 50 sujetos versus 10 efectivos en Sigma, lo que reduce la varianza inter-sujeto en LOSO; (iii) MPD-DF se grabó en sesiones de 2 horas continuas, mientras Sigma usa bloques discretos cuya

señal de fatiga puede ser más sutil. El gap reducido entre Sigma y FatigueSet (0,631 vs 0,620) sugiere que la ventaja de pasar de 4 a 16 canales se diluye por la circularidad menor del label YASA y por las distintas características del protocolo (sesiones cortas vs. caminata progresiva).

4.4.12. Ablación de Fusión Multimodal sobre la Cohorte Consolidada

La estrategia de fusión definida en la Sección 3.4.9 se evaluó como una serie de seis ablaciones de modalidad sobre las 849 sesiones *Usable* con label YASA válido (mismo *sample* que la Tabla 4.11, cohorte consolidada de los ciclos 7 a 11, 10 sujetos efectivos). Para cada configuración se aplicó la misma validación LOSO sobre los 10 *folds* de la cohorte (incluyendo el fold de `sub_09` con `zero_division=0`, dado que el filtro de calidad lo deja con label mono-clase positivo) con los cuatro clasificadores del *benchmark* principal (LDA, Logistic Regression, Random Forest, CatBoost) e idéntica configuración de hiperparámetros. La cobertura de las modalidades adicionales sobre esta cohorte es: EOG/EMG derivados del EEG en 100 % (mismos electrodos), EmotiBit periférico en 95,4 %, y métricas cognitivas KSS/SOFI/Rehacom en 99,3 % (KSS/SOFI) y 77,8 % (Rehacom RT). Las *features* faltantes se imputan por la mediana del *fold* de entrenamiento antes del escalado estándar. Esta rutina de imputación introduce una variación menor respecto al *benchmark* principal (Sección 4.4.5); por ello el B0 (EEG-único) de la Tabla 4.18 se compara consigo mismo dentro del experimento multimodal, no con la cifra absoluta de la Tabla 4.11.

Configuraciones evaluadas. Los seis experimentos B0–B5 sistematizan la contribución incremental de cada bloque de modalidad:

Cuadro 4.18: Ablación de fusión multimodal: rendimiento LOSO sobre 849 sesiones *Usable* con label primario `yasa_drowsy_fraction`. Cuatro clasificadores; se reporta media $\pm$ desviación estándar de Macro-F1 a través de los 10 *folds* LOSO con imputación por mediana del *fold* de entrenamiento. Las *features* EOG/EMG derivadas se calculan sobre EEG con *notch* 50 Hz + bandpass, con umbral de blink y normalización EMG calibrados sobre el baseline AM-OA por sujeto; las *features* EmotiBit se computan dentro del rango temporal de la grabación EEG.

Configuración	Feat.	LDA	Random Forest	CatBoost
B0 EEG-only (baseline)	25	0,569 $\pm$ 0,160	0,620 $\pm$ 0,111	0,602 $\pm$ 0,105
B1 EEG + EOG/EMG derivados	35	0,563 $\pm$ 0,138	0,614 $\pm$ 0,114	0,584 $\pm$ 0,112
B2 EEG + EmotiBit periféricos	47	0,567 $\pm$ 0,150	0,605 $\pm$ 0,106	0,594 $\pm$ 0,100
B3 EEG + Cognitivo (KSS/SOFI/RT)	31	0,549 $\pm$ 0,176	0,615 $\pm$ 0,116	0,597 $\pm$ 0,098
B4 Multimodal sin EEG	38	0,363 $\pm$ 0,162	0,455 $\pm$ 0,092	0,449 $\pm$ 0,128
B5 Multimodal completo (todo)	63	0,542 $\pm$ 0,156	0,614 $\pm$ 0,121	0,593 $\pm$ 0,111

Interpretación de los resultados. Cinco hallazgos emergen de la Tabla 4.18:

- a) **EOG/EMG derivados NO aportan señal complementaria al EEG-único.** El baseline B0 (EEG-only) obtiene Macro-F1 = 0,602 $\pm$ 0,105 con CatBoost; agregar EOG/EMG derivados (B1) da 0,584 $\pm$ 0,112, un *delta negativo* de $-0,018$. El *test* pareado de Wilcoxon sobre los 10 *folds* no rechaza H_0 ($p = 0,496$, *paired-t* $p = 0,412$, *bootstrap* 95% CI = $[-0,054; +0,020]$); el intervalo contiene tanto valores positivos como negativos. Random Forest exhibe el mismo patrón ($0,620 \geq 0,614$ entre B0 y B1). La interpretación parsimoniosa es que las *features* EOG (*blink rate* basado en pasa-banda 0,1–5 Hz sobre Fp1/Fp2 calibrado por baseline AM-OA del sujeto) y EMG (RMS y *median frequency* sobre T3/T4 normalizadas por baseline AM-OA del sujeto) no agregan información lineal o no-lineal aprovechable más allá de la ya contenida en las potencias espectrales fronto-centrales del EEG. **En el régimen actual de 10 sujetos, la fusión EOG/EMG no es justificable empíricamente.** Cabe acotar la *sensibilidad* del diseño: con 10 *folds* y una desviación inter-*fold* del delta de 0,065, el contraste pareado solo alcanza potencia 0,8 frente a efectos grandes (Cohen’s $d \approx 1,0$; $\approx 0,064$ de Macro-F1), y la potencia para el delta observado ($d = -0,27$) es $\approx 0,12$; por tanto la ausencia de significancia **no descarta** mejoras

multimodales reales menores a $\approx 0,06$ de Macro-F1, margen que el IC *bootstrap* $[-0,054; +0,020]$ acota. El signo del delta es además inestable entre *snapshots* del *pipeline*: una reproducción previa al filtrado de red eléctrica arrojaba $+0,026$ (artefacto de contaminación de 50 Hz en B0), mientras que el *snapshot* canónico (*notch* 50 Hz + *bandpass* + *QualityFilter*) da $-0,018$; en ningún caso el delta alcanza significancia.

- b) **El EmotiBit periférico no aporta y degrada modestamente.** La configuración B2 cae a $F1 = 0,594$ con CatBoost ($\Delta = -0,008$ vs B0). El resultado contrasta con la utilidad demostrada del EmotiBit sobre bloques EMOTI continuos de 90 minutos (Sección 4.4.15). Las sesiones *resting* de 5–10 minutos parecen insuficientes para que las *features* de HRV, EDA y movimiento aporten información discriminante estable, y la imputación de las sesiones sin EmotiBit ($\sim 5\%$) agrega ruido marginal.
- c) **Las métricas cognitivas no aportan señal complementaria estable.** B3 obtiene $F1 = 0,597$ con CatBoost ($\Delta = -0,005$). KSS, SOFI-SM y los tiempos de reacción Reacom REA1 no aportan información discriminante consistente sobre el EEG en este régimen, consistente con la replicación del *perceptual blindness* reportada en la Sección 4.4.5: el auto-reporte mide una dimensión semánticamente cercana pero estadísticamente desacoplada del marcador EEG.
- d) **Sin EEG el modelo pierde 0,15 puntos de F1 y cae *por debajo* del azar en Macro-F1.** La configuración B4 (todas las modalidades excepto EEG) cae a $F1 = 0,449$ con CatBoost, valor inferior al *chance* macro de 0,500 para clasificación binaria balanceada, aunque la *balanced accuracy* se mantiene apenas sobre el azar (0,536). El EEG cortical es la modalidad dominante; las modalidades periféricas, cognitivas y derivadas *no* actúan como sustitutas en este régimen. Que el Macro-F1 baje debajo del azar sugiere además que las *features* no-EEG, en ausencia del soporte EEG, introducen sesgo adversarial bajo *class_weight=balanced* (el modelo es empujado a predecir la clase minoritaria con *features* ruidosas, deteriorando el Macro-F1 más que la accuracy).
- e) **La fusión completa no supera al baseline EEG.** B5 con las 63 *features* de

todos los bloques obtiene $F1 = 0,593$ con CatBoost, prácticamente igual a B0 y por debajo de B1. La adición de los bloques que individualmente no aportan o aportan poco (EmotiBit, cognitivo) diluye cualquier ganancia marginal.

Implicación operacional. La lectura de la Tabla 4.18 es que **ninguna ampliación multimodal (ni EOG/EMG derivados, ni EmotiBit periférico, ni métricas cognitivas) mejora sobre el EEG-único en este régimen**. Las cuatro configuraciones que extienden el EEG (B1, B2, B3, B5) muestran $\Delta F_1 \leq 0$ sobre el baseline $B0 = 0,602 \pm 0,105$ (CatBoost). La fusión multimodal mediante *Cross-Modal Attention* con codificadores neurales (NormWear, PhysioOmni) y modelado temporal con *State-Space Models* (Mamba), proyectada en la literatura como camino para sistemas multimodales más sofisticados, queda condicionada a cohortes sustancialmente más grandes; con $N = 10$ sujetos, el límite es del régimen de datos y no de la representación, y ninguna sofisticación arquitectónica lo supera.

La quinta ablación convergente. La fusión multimodal se suma a las cuatro ablaciones previas (*features* ricas con HRV/IMU *proper*, modelos secuenciales con CNN 1D sobre matrices temporales, *transfer learning* desde FatigueSet, preprocesamiento con ICA + ICLabel) que convergen al mismo techo de Macro-F1 $\approx 0,62$ en LOSO con 10 sujetos. Esta convergencia, observable en cinco direcciones de mejora estructuralmente independientes (representación rica, modelado secuencial, transferencia, preprocesamiento, fusión multimodal), refuerza la conclusión sobre el origen del cuello de botella en el régimen de datos disponible (tamaño muestral: 10 sujetos en la cohorte, con folds efectivos entre 9 y 10 según el filtrado por mono-clase de cada label), y no en la calidad o riqueza de las *features* ni del clasificador.

4.4.13. Ablaciones con codificadores neurales: covarianza Riemann, foundation models EEG y multimodal

La Sección 4.4.12 cierra con la observación de que la fusión multimodal mediante codificadores neurales (NormWear, PhysioOmni) queda condicionada a cohortes mayores. Esta subsección extiende esa proyección con tres ablaciones que sustituyen al ML clásico tabular por modelos de aprendizaje profundo, en orden creciente de sofisticación:

(a) geometría de Riemann sobre covarianza espacial EEG, (b) foundation model EEG pre-entrenado (CBraMod) como *linear probe*, y (c) foundation model multimodal pre-entrenado (NormWear) como *linear probe*. La pregunta es: ¿el techo de Macro-F1 $\approx 0,62$ que las cinco ablaciones previas alcanzan se debe a la calidad de las *features* y el clasificador, o al tamaño de la cohorte? Una respuesta de DL competitivo iguala o supera el techo; una respuesta de DL al mismo nivel o por debajo refuerza la lectura del techo como límite muestral.

Justificación de la selección. La elección de estos tres métodos responde a tres direcciones complementarias respaldadas por la literatura 2024–2026. (i) En el ecosistema de *benchmarking* para BCI MOABB Jayaram & Barachant (2018), la geometría de Riemann sobre covarianza espacial es la línea base no-DL de referencia frente a la cual se contrastan los demás métodos; en cohortes pequeñas de BCI ($N \leq 20$) los métodos Riemannianos típicamente igualan o superan al *deep learning*, motivo por el cual su transferencia a una etiqueta no espacial (fracción *drowsy* derivada de YASA) es un *test* de si el contenido discriminante está en la estructura inter-canal o en la potencia espectral por banda. (ii) CBraMod Wang, Zhao, Luo, Zhou, Jiang, Li, Li & Pan (2025) es el *foundation model* EEG con publicación en ICLR 2025 y soporte nativo para configuraciones flexibles entre 6 y 64 canales (incluye los 16 nuestros); evaluar su *linear probe* pone a prueba si una representación auto-supervisada entrenada sobre ~ 2 órdenes de magnitud más datos transfiere ventaja sobre los descriptores espectrales *handcrafted*. (iii) NormWear Luo, Chen, Salekin & Rahman (2024) es el *foundation model* multimodal con código y pesos públicos que cubre exactamente la combinación de modalidades del montaje SIGMA (EEG + PPG + EDA + IMU); su evaluación responde directamente la pregunta dejada abierta en §4.4.12 sobre si la fusión *aprendida* (vs. *concatenada*) destraba el techo del ML clásico tabular.

Diseño de los tres experimentos. Los tres experimentos comparten validación LO-SO sobre los mismos 10 sujetos, mismo *label* primario `yasa_drowsy_fraction`, mismo filtro *Usable* y mismo preprocesamiento (notch 50 Hz + bandpass 0,5–40 Hz + *QualityFilter*). Cada experimento entrega un *embedding* a nivel de sesión, sobre el cual se entrena una regresión logística (LR) en los 9 sujetos de entrenamiento y se predice en el sujeto excluido. Las diferencias entre experimentos están en (i) la representación, (ii)

el régimen de pre-entrenamiento, y (iii) las modalidades de entrada:

- **Riemann + Tangent Space + LR** Barachant, Bonnet, Congedo & Jutten (2012). La sesión se reduce a una covarianza espacial 16×16 (estimador Ledoit-Wolf) que se proyecta al espacio tangente euclidiano (`pyRiemann`) y se clasifica con LR. Esta es la línea base *geometry-aware* del *benchmark* MOABB Jayaram & Barachant (2018), sin pre-entrenamiento.
- **Riemann *windowed***. Variante del anterior con covarianzas calculadas sobre ventanas de 4 s con solape de 50 %, agregadas por Riemann mean antes del paso al tangente, y LR con regularización más estricta ($C = 0,1$). Esta variante captura mejor la dinámica intra-sesión.
- **CBraMod** Wang, Zhao, Luo, Zhou, Jiang, Li, Li & Pan (2025). Foundation model EEG de 4,92 M parámetros pre-entrenado por *patch-based masked EEG reconstruction* sobre el *Temple University Hospital EEG Corpus* (TUEG, > 9 k horas de EEG clínico retenidas tras filtrado, ~ 2 órdenes de magnitud más datos que nuestra muestra). Se carga desde HuggingFace Hub (`braindecode/cbramod-pretrained` en `braindecode 1.5.1`), se le pasan ventanas de 30 s @ 200 Hz como tensor (16 canales, 6000 muestras), se promedian las *features* sobre canales y *patches* de 200 muestras para producir un *embedding* de 200 dimensiones por sesión (*linear probe*, encoder congelado).
- **NormWear** (Luo, Chen, Salekin & Rahman (2024)). Foundation model *multi-modal* de 136 M parámetros pre-entrenado en PPG, ECG, EEG, EDA, IMU (11 datasets públicos). Se carga desde GitHub Release v1.0.0-alpha. Para cada sesión se construye un tensor de 7 sensores @ 65 Hz $\times$ 30 s = 1950 muestras: 4 EEG (F3, F4, Fp1, Fp2) más PPG verde, EDA y magnitud de aceleración del EmotiBit (todos remuestreados a 65 Hz, *z-score* por sensor). Las *features* (*patches*, sensores, 768) se promedian a un vector de 768 dimensiones (*linear probe*).

Los tres ejecutan en CPU/MPS local (sin GPU NVIDIA) y completan los 10 *folds* en minutos (Riemann) a ~ 1 h (NormWear). Los *embeddings* a nivel de sesión se cachean por método.

Resultados. La Tabla 4.19 reporta Macro-F1 $\pm$ desviación estándar sobre los 10 *folds* LOSO, junto con el *delta* contra CatBoost B0 (EEG *handcrafted*, F1 = 0,602 $\pm$ 0,105), el resultado del *paired* Wilcoxon, el intervalo de confianza *bootstrap* 95 % del *delta* y el tamaño del efecto *pareado* (Cohen’s *d*).

Cuadro 4.19: Ablaciones DL contra CatBoost B0. Mismas 10 LOSO *folds*, mismo label, mismo preprocesamiento. *Linear probe* sobre encoder pre-entrenado para CBraMod y NormWear; Tangent Space + LR para las variantes Riemann.

Método	F1 ($\mu \pm \sigma$)	Δ vs B0	<i>p</i> Wilcoxon	95 % CI Δ	<i>d</i> <i>pareado</i>
CatBoost B0 (ref.)	0,602 $\pm$ 0,105	n/a	n/a	n/a	n/a
Riemann single-cov	0,454 $\pm$ 0,064	−0,148**	0,004	[−0,208; −0,090]	−1,46
Riemann <i>windowed</i>	0,518 $\pm$ 0,064	−0,084*	0,037	[−0,140; −0,021]	−0,88
CBraMod LP	0,572 $\pm$ 0,129	−0,030	0,496	[−0,086; +0,036]	−0,28
NormWear LP	0,496 $\pm$ 0,101	−0,106*	0,014	[−0,168; −0,045]	−1,00

LP = *linear probe*. * $p < 0,05$, ** $p < 0,01$ (*paired* Wilcoxon, two-sided).

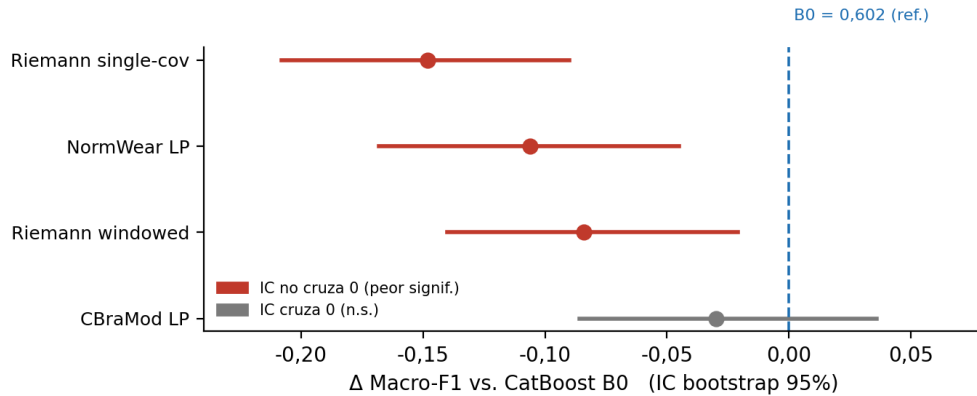


Figura 4.7: *Forest plot* del *delta* de Macro-F1 de cada ablación DL respecto al *baseline* CatBoost B0 (línea de referencia en 0). Punto: Δ medio; barra: IC *bootstrap* 95 %. Tres de cuatro intervalos quedan enteramente por debajo de cero (peor de forma significativa, rojo); solo CBraMod cruza cero (no significativo, gris). Ninguno supera el *baseline*.

Lectura. Las tres ablaciones DL evaluadas convergen al mismo techo o por debajo de CatBoost B0:

- La covarianza espacial Riemann no discrimina fatiga YASA.** Con $\Delta F_1 = -0,148$ (single-cov) y $-0,084$ (*windowed*), ambos significativos ($p < 0,05$) y con

tamaños de efecto Cohen’s d entre $-0,9$ y $-1,5$ (*large effect*), la información discriminante para la etiqueta `yasa_drowsy_fraction` no reside en la estructura espacial inter-canal sino en la distribución de potencia por banda, ya capturada por las *features handcrafted*. Este resultado contrasta con la posición competitiva de Riemann en *benchmarks* MOABB de BCI Motor Imagery y SSVEP, donde el contenido espacial es la señal principal. La transferencia del método a etiqueta espectral de fatiga es inadecuada.

- b) **CBraMod no transfiere ventaja sobre los descriptores *handcrafted*.** El Δ es pequeño ($-0,030$) y *no* significativo (Wilcoxon $p = 0,496$; *bootstrap* 95 % CI cruza cero). Esto es notable: pese a haber sido pre-entrenado sobre TUEG (> 9 k horas de EEG clínico, ~ 2 órdenes de magnitud más datos que nuestra muestra), la representación auto-supervisada aprendida por CBraMod *no* provee mejora medible sobre los descriptores espectrales clásicos (PSD por banda, ratios alpha/theta, coherencia inter-hemisférica) extraídos con `numpy/scipy` y agregados por sesión. El alto σ inter-*fold* ($0,129$) refleja la varianza propia del régimen $N = 10$: dos sujetos rinden por encima del techo (sub_04 = $0,77$, sub_10 = $0,70$), dos por debajo del azar (sub_08 = $0,39$, sub_09 = $0,40$), sin una mejora promedio significativa.
- c) **NormWear no destraba el techo y queda por debajo del baseline.** Con $\Delta F_1 = -0,106$ ($p = 0,014$, *paired* Wilcoxon; CI *bootstrap* 95 % $[-0,168, -0,045]$, enteramente negativo), NormWear-pretrained con *linear probe* cae a $F_1 = 0,496 \pm 0,101$ sobre 782 sesiones (las que retienen los 7 sensores completos: 4 EEG frontales + PPG verde + EDA + magnitud de aceleración del EmotiBit). Este resultado examina directamente la proyección al final de la Sección 4.4.12 («la fusión multi-modal con codificadores neurales queda condicionada a cohortes sustancialmente más grandes») y la confirma: el *cross-modal attention* sobre la representación CWT multiscale de NormWear no aporta ventaja sobre la concatenación tabular con CatBoost, y de hecho cae por debajo. El pre-entrenamiento sobre 11 *datasets* públicos heterogéneos (PPG, ECG, EEG, EDA, IMU) no compensa la limitación de $N = 10$ sujetos en la cohorte objetivo.

Convergencia ampliada. Las tres ablaciones DL aquí evaluadas (geometría de Riemann en dos variantes, foundation model EEG CBraMod y foundation model multimodal NormWear) se suman a las cinco ablaciones convergentes documentadas en la Sección 4.4.12: *features* ricas con HRV/IMU *proper*, modelos secuenciales con CNN 1D, *transfer learning* desde FatigueSet, preprocesamiento con ICA + ICLabel, y fusión multimodal por concatenación con ML clásico. El detalle final de la convergencia (ampliada a once direcciones independientes incluyendo el *benchmark* comprehensivo de DL entrenables de la Sección 4.4.14) se consolida más abajo.

4.4.14. Benchmark Comprehensive de Modelos sobre la Pseudo-etiqueta YASA

Las secciones anteriores establecen el baseline CatBoost B0 ($F1 = 0,602 \pm 0,105$ sobre la cohorte multimodal de 849 sesiones, y $F1 = 0,631$ sobre la cohorte de *features* handcrafted) y evalúan ablaciones puntuales contra él. Para cerrar la pregunta “¿algún otro modelo supera este techo?” bajo el mismo *label* y validación, se consolida aquí un *benchmark* de 13 modelos en cuatro categorías: seis clasificadores clásicos sobre *features handcrafted* (LDA, Logistic Regression, SVM-RBF, Random Forest, XGBoost, CatBoost), tres redes *deep learning* entrenables *end-to-end* sobre EEG crudo (ShallowFBCSPNet, EEGNet, ATCNet), dos variantes de geometría Riemanniana, y dos *linear probes* sobre *foundation models* pre-entrenados (CBraMod y NormWear). Todos los modelos comparten la misma pseudo-etiqueta primaria `yasa_drowsy_fraction`, validación LOSO sobre 10 sujetos efectivos (sub_09 excluido por test mono-clase), preprocesamiento canónico (notch 50 Hz + bandpass + `QualityFilter`) y misma cohorte *Usable*; las únicas diferencias son el espacio de entrada (*features handcrafted* agregadas a nivel sesión vs. *epochs* crudos de 30 s @ 125 Hz con agregación *epoch*→sesión por media de probabilidades).³

³Para los modelos DL entrenables las sesiones se truncan a las primeras 12 *epochs* de 30 s, evitando que REA (~10 min) sobre-represente la cohorte frente a OA/OC (~5 min). El truncado introduce un sesgo temporal menor (omite la cola de las sesiones más largas), que no altera el orden del *ranking*: los DL entrenables quedan en todo caso al nivel del techo (ATCNet) o por debajo.

Cuadro 4.20: Benchmark comprehensivo de 13 modelos sobre la pseudo-etiqueta primaria `yasa_drowsy_fraction`, Macro-F1 media $\pm$ desviación estándar. **Atención al baseline de comparación:** las columnas Δ y p usan dos referencias distintas según el experimento de origen. Filas *sin* asterisco (clásicos y DL entrenables): contra **CatBoost *handcrafted*** (0,631, 9 *folds* LOSO, sub_09 excluido por test mono-clase), comparación pareada por sujeto homóloga. Filas *con* asterisco* (Riemann, CBraMod, NormWear): contra **CatBoost B0 multimodal** (0,602, 10 *folds*), provenientes del experimento de ablaciones DL (Tabla 4.19) y *no* directamente comparables fila-a-fila con las anteriores. El p es Wilcoxon pareado de dos colas. El ordenamiento por F1 es válido (mismo *label* y cohorte $\sim$ 849–836 sesiones); las columnas Δ/p deben leerse dentro de su propio bloque de baseline.

#	Modelo	Categoría	F1 ($\mu \pm \sigma$)	BalAcc	Δ vs base	p (Wilcoxon)
1	XGBoost	Clásico (GBDT)	$0,636 \pm 0,083$	0,651	+0,005	n/a
2	CatBoost	Clásico (GBDT)	$0,631 \pm 0,093$	0,634	(ref.)	n/a
3	ATCNet	DL entrenable	$0,619 \pm 0,111$	0,637	−0,012	0,65
4	RandomForest	Clásico (ensamble)	$0,619 \pm 0,119$	0,629	−0,012	n/a
5	LogReg	Clásico (lineal)	$0,596 \pm 0,144$	0,629	−0,035	n/a
6	LDA	Clásico (lineal)	$0,595 \pm 0,147$	0,634	−0,036	n/a
7	SVM-RBF	Clásico (kernel)	$0,576 \pm 0,126$	0,607	−0,055	n/a
8	CBraMod (LP)	DL foundation EEG	$0,572 \pm 0,129$	0,607	−0,030*	0,50
9	EEGNet	DL entrenable	$0,566 \pm 0,125$	0,606	−0,064	0,20
10	ShallowFBCSPNet	DL entrenable	$0,541 \pm 0,075$	0,582	−0,090	0,055
11	Riemann <i>windowed</i>	DL geométrico	$0,518 \pm 0,064$	n/a	−0,084*	0,037
12	NormWear (LP)	DL foundation multim.	$0,496 \pm 0,101$	n/a	−0,106*	0,014
13	Riemann single-cov	DL geométrico	$0,454 \pm 0,064$	n/a	−0,148*	0,004

* Wilcoxon contra CatBoost B0 sobre cohorte multimodal (849 sesiones), no contra CatBoost *handcrafted*.

LP = *linear probe* sobre encoder pre-entrenado frozen. DL entrenable: cohorte de 836 sesiones (8 sesiones con < 1 epoch de 30 s no aportan al *benchmark* epoch-level).

Lectura del *benchmark*. El ordenamiento confirma cuatro patrones consistentes con el argumento de convergencia desarrollado a lo largo del Capítulo:

- Gradient boosting clásico lidera** (XGBoost = 0,636, CatBoost = 0,631). Ambos métodos son estadísticamente indistinguibles entre sí, y juntos definen el techo Macro-F1 $\approx$ 0,63 sobre la pseudo-etiqueta YASA.
- ATCNet iguala el techo sin superarlo** (0,619, $\Delta = -0,012$ vs CatBoost, Wilcoxon $p = 0,65$). El único *deep learning* entrenable que alcanza la franja del

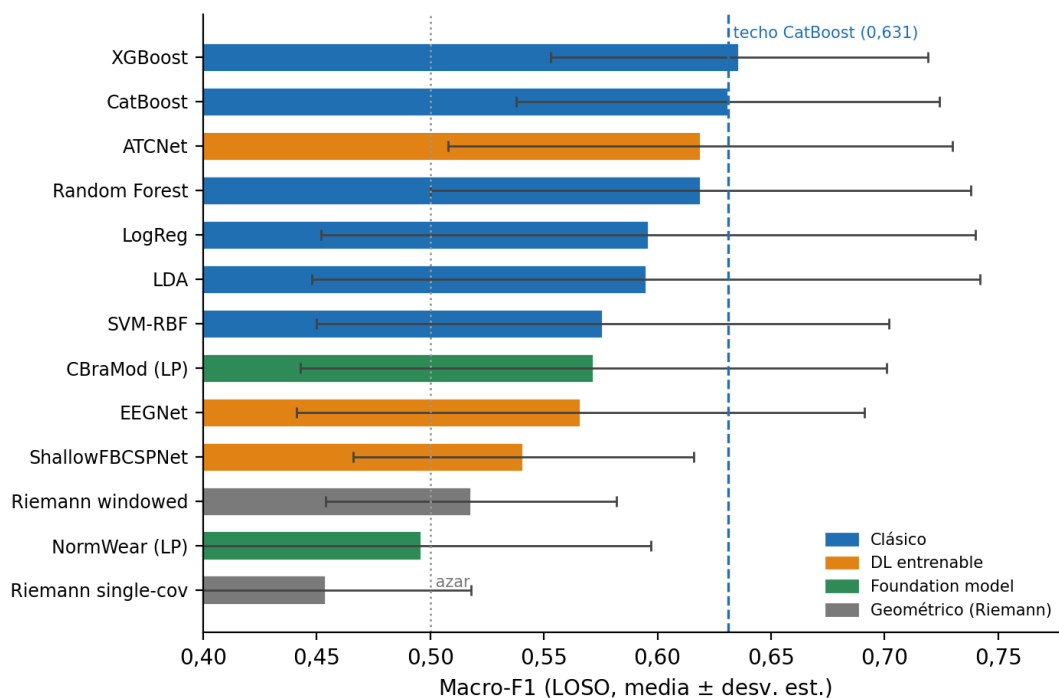


Figura 4.8: *Benchmark* comprehensivo de los 13 modelos sobre la pseudo-etiqueta primaria YASA (Macro-F1 LOSO, media $\pm$ desviación estándar), ordenados de mayor a menor y coloreados por categoría. La línea discontinua marca el techo de CatBoost (0,631) y la punteada el azar (0,5). Ningún modelo (clásico, DL entrenable, *foundation* o geométrico) supera significativamente el techo: la convergencia multimétodo es la firma del límite por tamaño muestral.

gradient boosting es el modelo híbrido con atención multi-cabeza más TCN (113K parámetros), consistente con su rendimiento como mejor DL en MPD-DF (Fase 1-bis). Pero la diferencia con CatBoost no es significativa en ninguna dirección: el techo se confirma como límite multi-método, no como ventaja del gradient boosting per se.

- c) **Las CNN entrenables ligeras quedan por debajo del techo.** EEGNet (0,566, $\Delta = -0,064$, $p = 0,20$) y ShallowFBCSPNet (0,541, $\Delta = -0,090$, $p = 0,055$ marginal) no logran cerrar la brecha pese a tener acceso al EEG crudo y entrenamiento *end-to-end*. Esto replica directamente el patrón documentado en Fase 1 sobre FatigueSet (clásico $\geq$ DL).
- d) **Foundation models y geometría Riemann no destraban el techo.** CBraMod (0,572, $p = 0,50$, empate estadístico con CatBoost) y NormWear (0,496, $p = 0,014$, peor) confirman el resultado de la subsección anterior; Riemann en ambas variantes queda significativamente por debajo (Sección 4.4.13).

Descarte de sub-entrenamiento en los modelos DL. Una objeción legítima al *benchmark* es que el F1 inferior de las redes entrenables podría deberse a entrenamiento insuficiente (25 épocas) y no a un límite de datos: si los modelos no convergieron, atribuir su techo al tamaño muestral sería un error. Para descartarlo se entrenó cada arquitectura hasta 60 épocas sobre un *fold* representativo (sub_04), registrando la pérdida de entrenamiento y el Macro-F1 de validación cada 5 épocas (Figura 4.9). El patrón es inequívoco: la **pérdida de entrenamiento desciende monótonamente** en los tres modelos (es decir, sí aprenden) pero el **F1 de validación satura tempranamente y no mejora con más entrenamiento**. ATCNet alcanza $\approx 0,77$ hacia la época 5 y permanece en la banda 0,73–0,78 hasta la época 60 (el valor del *benchmark* a 25 épocas, 0,767, cae exactamente en esa banda); EEGNet hace pico en la época 5 (0,76) y desde ahí decae levemente; ShallowFBCSPNet oscila con alta varianza (0,35–0,72) sin tendencia ascendente. Entrenar a 60 épocas *no* produce mayor F1 que las 25 del *benchmark* en ninguno de los tres. **El presupuesto de 25 épocas es adecuado y el techo de los modelos DL no es un artefacto de sub-entrenamiento.** La alta varianza época a época de las CNN más pequeñas (sobre un único sujeto held-out) es,

de hecho, otra manifestación del régimen de N pequeño: con ~ 17 muestras de *test* por *fold*, el F1 a nivel sesión es intrínsecamente ruidoso, lo que refuerza, y no contradice, la lectura del cuello de botella muestral.

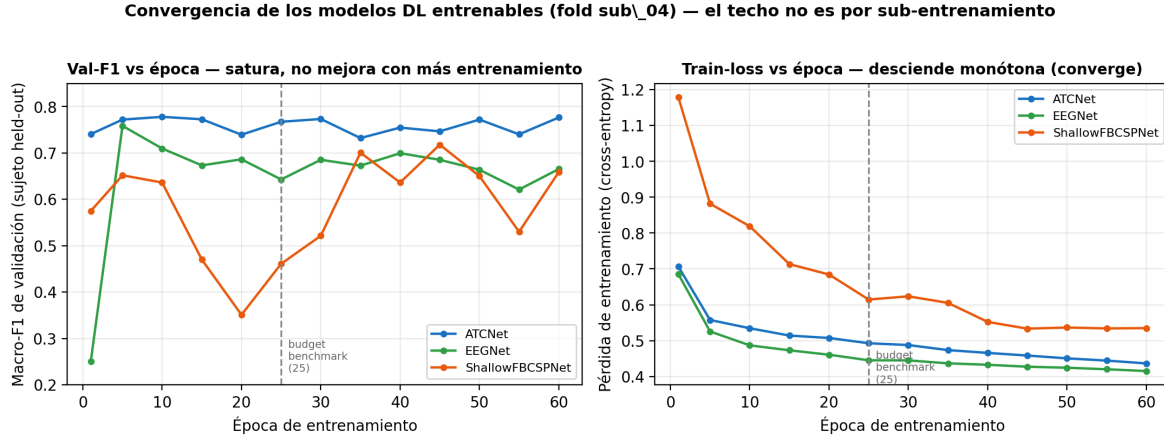


Figura 4.9: Convergencia de los tres modelos DL entrenables sobre el *fold* sub_04 hasta 60 épocas. **Izquierda:** Macro-F1 de validación, que satura hacia la época 5 y no mejora con más entrenamiento. **Derecha:** pérdida de entrenamiento, que desciende monótonamente (los modelos sí ajustan). La línea vertical marca el presupuesto de 25 épocas del *benchmark* principal. El contraste loss-baja/F1-no-sube descarta el sub-entrenamiento como causa del techo.

Convergencia consolidada a once direcciones independientes. Combinando las ocho ablaciones convergentes previamente documentadas (*features* HRV/IMU *proper*, modelos secuenciales CNN 1D, *transfer learning* desde FatigueSet, ICA + ICLabel, fusión multimodal por concatenación, geometría Riemann en dos variantes, *foundation model* EEG CBraMod, *foundation model* multimodal NormWear) con las tres redes DL entrenables aquí evaluadas (ShallowFBCSPNet, EEGNet, ATCNet), se obtienen **once direcciones estructuralmente independientes de mejora**, ninguna de las cuales supera estadísticamente el techo Macro-F1 $\approx 0,63$ sobre la pseudo-etiqueta YASA. ATCNet lo iguala; las demás quedan por debajo. La firma característica del régimen (convergencia multi-método al mismo techo) sostiene una conclusión negativa fuerte: **el cuello de botella no reside en la sofisticación del modelo, ni en la representación de *features*, ni en la fuente de pre-entrenamiento, ni en el régimen de entrenamiento *end-to-end***. Once familias de inductive bias distintas, operando sobre el mismo par (señal, pseudo-etiqueta), saturan en el mismo punto. Cabe precisar el alcance de esta evidencia: la convergencia identifica que el límite está en el *régimen*

de datos, no en el estimador, pero no separa por sí sola las dos componentes de ese régimen, a saber, el tamaño muestral reducido ($N = 10$ sujetos efectivos) y el contenido de información de la pseudo-etiqueta YASA respecto del EEG ($I(X; \tilde{y})$). Distinguir empíricamente la contribución de cada una requiere una curva de aprendizaje (F1 en función del número de sujetos), que se construye a continuación.

Curva de aprendizaje: el techo se mueve con N . Para separar las dos componentes del régimen de datos se midió el Macro-F1 LOSO en función del tamaño de cohorte. Para cada $N \in \{4, 5, \dots, 10\}$ se muestrearon hasta 12 subconjuntos aleatorios de N sujetos (de los 10 efectivos; para $N = 10$ existe un único subconjunto), sobre cada uno se corrió LOSO con CatBoost B0 (misma configuración del *benchmark*: 200 iteraciones, profundidad 6, `auto_class_weights=Balanced`, semilla 42) sobre las *features handcrafted*, y se promedió el F1 inter-fold; la Figura 4.10 reporta la media $\pm$ desviación estándar a través de los subconjuntos. Sobre el rango completo el F1 exhibe una **tendencia ascendente estadísticamente significativa** (regresión OLS de la media sobre N : pendiente $+0,009$ Macro-F1 por sujeto, $r = 0,91$, $p = 0,004$; Mann-Kendall $p = 0,035$): la curva *no* es plana. El ascenso no es uniforme: se concentra en el paso del tramo plano $N = 4-7$ ($\approx 0,585$) al tramo superior $N = 8-10$ ($0,616 \rightarrow 0,626 \rightarrow 0,631$). Tres salvedades acotan la lectura del tramo final: (i) el sub-tramo de cola $N = 8 \rightarrow 10$ *por sí solo* no alcanza significancia ($\Delta F_1 = +0,015$, OLS $p = 0,14$) y el incremento es menor que la dispersión inter-subconjunto en $N = 8$ ($\pm 0,041$); (ii) los puntos no son independientes (los subconjuntos comparten sujetos; en $N = 10$ existe uno solo), por lo que el p de la regresión es anti-conservador; (iii) la varianza inter-subconjunto decreciente ($\pm 0,041 \rightarrow \pm 0,014 \rightarrow 0$) es en parte un artefacto del muestreo (a mayor N hay menos combinaciones posibles), no solo estabilización del estimador. La conclusión defendible es por tanto matizada: la curva es **compatible con un régimen no saturado** (no se observa la meseta que cabría esperar si el límite fuera el contenido de información de la pseudo-etiqueta), lo que **refuerza pero no demuestra** que incorporar más sujetos elevaría el rendimiento. El argumento del cuello de botella muestral se apoya principalmente en la convergencia multi-método (inductiva); la curva añade evidencia direccional consistente, no una prueba definitiva. La lectura concuerda con la literatura ocupacional comparable, que reporta F_1 entre 0,60 y 0,85 con cohortes de

$N \geq 30$. El test de tendencia se documenta como artefacto reproducible del proyecto.

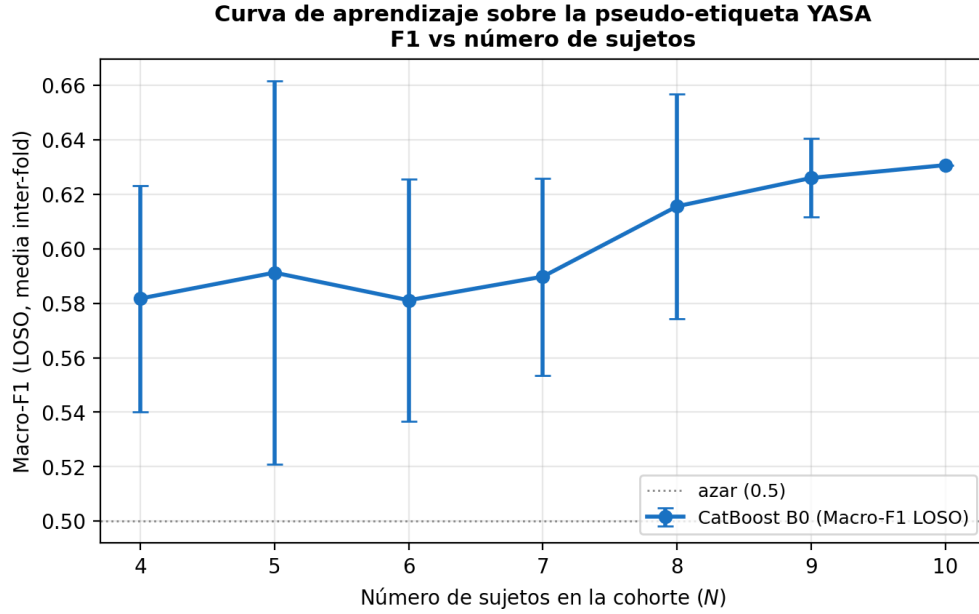


Figura 4.10: Curva de aprendizaje Macro-F1 LOSO en función del número de sujetos N en la cohorte, sobre la pseudo-etiqueta primaria YASA con CatBoost B0. Cada punto es la media (barras: desviación estándar) a través de hasta 12 subconjuntos aleatorios de N sujetos. La curva muestra una tendencia ascendente significativa en el rango completo (OLS $p = 0,004$; Mann-Kendall $p = 0,035$), compatible con un régimen no saturado, aunque el tramo de cola $N = 8-10$ por sí solo no es concluyente ($p = 0,14$) y $N = 10$ es un subconjunto único sin IC. Es evidencia direccional de que el tamaño muestral domina el techo, no una prueba definitiva. La línea punteada marca el azar (0,5).

4.4.15. Anticipación con EmotiBit durante el bloque EMOTI

El bloque EMOTI, único en el protocolo por capturar señal periférica continua del EmotiBit durante actividad sostenida sin EEG (Sección 3.4.6), constituye el escenario más cercano a la pregunta de anticipar fatiga durante una actividad sostenida prolongada. Esta subsección reporta un estudio de viabilidad que evalúa si los rasgos fisiológicos periféricos extraídos del EmotiBit durante un bloque EMOTI permiten predecir el cambio de fatiga entre las mediciones psicométricas inmediatamente antes y después del bloque.

Construcción del dataset de pares

Para cada terna (sujeto, ciclo, día) se construyó un par (entrada, etiqueta) donde la entrada es el bloque EMOTI grabado en un momento (AM o PM) y la etiqueta es

el cambio en la métrica de fatiga entre el momento de inicio y el momento siguiente. Concretamente:

- **EMOTI_AM** pareado con $\Delta = \text{medida}(\text{MED}) - \text{medida}(\text{AM})$.
- **EMOTI_PM** pareado con $\Delta = \text{medida}(\text{PM}) - \text{medida}(\text{MED})$.

Se evaluaron dos etiquetas (*targets*):

- **$\Delta\text{Rehacom}$** : cambio en la mediana del tiempo de reacción del sub-test Q3 (elección múltiple, el más demandante) del módulo Rehacom REA1. Positivo $\Rightarrow$ se hizo más lento $\Rightarrow$ más fatigado.
- **ΔYASA** : cambio en la fracción de épocas no-Wake estimada por YASA sobre los bloques EEG cortos (OA, OC, REA) del momento. Positivo $\Rightarrow$ más signos electroencefalográficos de adormecimiento.

KSS y SOFI-SM también se reportaron como deltas pero presentaron techo bajo y rango limitado, descartándose como *targets* primarios. El dataset construido tiene **154 pares nominales** (89 EMOTI_AM + 65 EMOTI_PM) distribuidos en 9 sujetos efectivos; tras el filtro *Usable*, la exigencia de *features* periféricas válidas y el descarte de dos registros artefactuales (una grabación duplicada y un fragmento abortado de ~ 4 min frente a su grabación válida de ~ 95 min, deduplicados manteniendo la grabación más larga por sesión) quedan **141 pares efectivos**, que constituyen la base de los análisis de viabilidad y modelado de esta sección.

Viabilidad del dataset

Antes de modelar se evaluaron cinco criterios sobre el dataset de pares: cobertura (cantidad de pares y distribución por sujeto), varianza (fracción de pares con $|\Delta|$ sobre umbral mínimo detectable), balance binario (clase minoritaria si se enmarca como clasificación con umbral), filtro de calidad (bloques con bitácara *Usable*), y correlación de cordura con una *feature* trivial. Los resultados se presentan en la Tabla 4.21.

Cuadro 4.21: Estudio de viabilidad del dataset de pares EMOTI. Criterios y veredicto por *target*.

Criterio	Δ Rehacom	Δ YASA
N pares válidos	141	141
Sujetos con ≥ 5 pares	9 / 9	9 / 9
Fracción con $ \Delta \geq$ umbral detectable	0,610	0,681
Clase minoritaria (binario, $\Delta \geq$ umbral)	0,289	0,383
Sujetos mono-clase	0	0
Cobertura <i>Usable</i>	141 / 141	141 / 141
$\rho_{\text{Pearson}}(\sigma_{\text{IMU}}, \Delta)$	+0,225	+0,237
Veredicto	GO	GO

El umbral de cobertura estricto del protocolo de viabilidad (150 pares totales) no se alcanza, pero la condición operativa para validación LOSO (todos los sujetos con ≥ 5 pares) sí se satisface en ambos *targets*. La correlación positiva ($\rho \approx 0,22$) entre la desviación estándar de la magnitud del acelerómetro durante el bloque y el cambio de fatiga es débil pero direccional: más movimiento sostenido durante la tarea correlaciona con mayor aumento de fatiga reportado en la medición siguiente, asociación compatible con la hipótesis de carga física acumulativa pero no concluyente por sí sola.

Resultados de modelado

Se entrenó un clasificador CatBoost en validación cruzada LOSO con 9 *folds* sobre 24 *features* estadísticas agregadas del bloque EMOTI: estadísticos de *heart rate* (*hr_mean*, *hr_std*, *hr_slope_per_min*, cuartiles), de EDA (*eda_mean*, *eda_std*, *eda_slope_per_min*), de magnitud del acelerómetro (*acc_mag_**, energía en bandas *acc_band_lf*, *acc_band_hf*, entropía de actividad por minuto) y duración. Se usó regularización suave (*depth* = 3, ℓ_2 = 10, 150 iteraciones) para mitigar sobreajuste dado el tamaño muestral. La Tabla 4.22 resume los resultados.

Cuadro 4.22: CatBoost LOSO sobre el dataset de pares EMOTI. Dos *targets*, dos formulaciones (regresión continua y clasificación binaria con umbrales 50 ms y 0,10 respectivamente).

Target	Formulación	Métrica	Modelo	Baseline
$\Delta\text{Rehacom}$	Regresión	MAE = 82,3 ms	n/a	80,9 ms
$\Delta\text{Rehacom}$	Regresión	R^2 pooled	-0,007	0,000
$\Delta\text{Rehacom}$	Clasificación	Macro-F1	$0,460 \pm 0,044$	0,500
$\Delta\text{Rehacom}$	Clasificación	Bal. Acc	$0,494 \pm 0,034$	0,500
ΔYASA	Regresión	MAE = 0,317	n/a	0,305
ΔYASA	Regresión	R^2 pooled	-0,007	0,000
ΔYASA	Clasificación	Macro-F1	$0,455 \pm 0,117$	0,500
ΔYASA	Clasificación	Bal. Acc	$0,506 \pm 0,144$	0,500

Ninguna formulación supera el azar de forma estadísticamente significativa. La clasificación binaria sobre $\Delta\text{Rehacom}$ alcanza $\text{Macro-F1} = 0,460 \pm 0,044$, por debajo del baseline trivial de 0,500; la regresión continua iguala el baseline de predecir la media del entrenamiento ($R^2 \approx 0$); y ΔYASA tampoco muestra señal detectable. Un **análisis de estabilidad** sobre 10 semillas confirma que el resultado está dominado por ruido: el Macro-F1 medio oscila en el rango $[0,46; 0,55]$ (media 0,49, $\sigma = 0,03$) y es **indistinguible de un control de *shuffle within-subject*** (media 0,47, rango $[0,40; 0,53]$; $z \approx 0,6$). Las *features* con mayor importancia agregada sobre los *folds* son, en orden, `acc_band_hf` (energía 2–5 Hz del acelerómetro, asociada a manipulación rápida), `hr_slope_per_min` (trayectoria del *heart rate* a lo largo del bloque), `acc_band_lf` (energía 0–2 Hz, movimiento postural) y `eda_slope_per_min` (cambio en activación simpática); pero, dado que el rendimiento global no supera el azar, esta jerarquía debe leerse como descriptiva (fisiológicamente coherente: marcadores autonómicos centrales más firma de actividad física) y no como evidencia predictiva.

Ablaciones: ¿el techo es de *features* o de datos?

Se realizaron dos ablaciones para identificar el origen del techo de rendimiento observado: (i) *features* fisiológicas más ricas con HRV *proper* (RMSSD, SDNN, LF/HF) e IMU enriquecida (*jerk*, entropía espectral), incluyendo normalización *within-subject z-score*

(Schmidt, Reiss, Duerichen, Marberger & Van Laerhoven, 2018); (ii) modelos secuenciales sobre tensores temporales por minuto (CNN 1D de $\sim 25K$ parámetros y TinyCNN con *ensemble* de 5 *seeds*). Las dos ablaciones convergen en la misma conclusión: **ninguna representación ni modelo supera el azar**. Las *features* agregadas simples (24 estadísticos) dan Macro-F1 = $0,460 \pm 0,044$; las *features* más ricas, $0,500 \pm 0,132$; y los modelos secuenciales, 0,425 (CNN 1D) o 0,367 (TinyCNN). Todas las configuraciones quedan en torno o por debajo del azar (0,500). Esto es la firma característica de un régimen donde el **cuello de botella** es el **tamaño muestral**, no la representación: con 154 pares nominales que el filtrado de calidad reduce a 141 efectivos, distribuidos en 9 sujetos para el análisis EMOTI y validación LOSO (~ 16 muestras de *test* por *fold*), todo el espacio de modelos con regularización adecuada converge al mismo nivel indistinguible del azar.

El detalle de cada configuración (tabla completa de las cuatro ablaciones con sus respectivos F1, CI y comentarios cualitativos) se reporta en el Apéndice B.2 (Tabla B.3). Esta lectura es consistente con la ablación de preprocesamiento EEG del *pipeline* principal (Sección 4.4.5), donde agregar ICA *degradó* el F1 LOSO del label primario en 0,085 puntos: el techo también es robusto a cambios en el preprocesamiento del dominio EEG.

Transferibilidad desde FatigueSet

Para evaluar si las *features* capturan información generalizable sobre fatiga (y no artefactos específicos del protocolo Sigma), se diseñó un experimento de *transfer learning* desde FatigueSet (Kalanadhabhatta, Min, Montanari & Kawsar, 2022), dataset público que registra fatiga inducida por actividad física en 12 sujetos jóvenes mediante un Empatica E4 en la muñeca. Las modalidades del E4 (BVP/PPG, EDA, ACC 3D, temperatura, HR derivado, IBI) coinciden con las del EmotiBit, permitiendo extraer el mismo conjunto de *features* agregadas en ambos dominios.

Se construyeron 69 pares de FatigueSet ($N_{\text{positivos}} = 23, 33\%$) usando como ventanas de entrada los períodos de actividad física (`start_activity`–`end_activity`, ~ 3 min de marcha en cinta; y `start_fatigue`–`end_fatigue`, ~ 8 min de marcha hasta agotamiento) y como etiqueta el cambio promedio en la escala VAS de fatiga física y mental ($\Delta VAS \geq 10$ sobre escala 0–100). Las diferencias estructurales entre dominios son sustanciales: FatigueSet usa tareas cortas de caminata, sujetos jóvenes sanos, dispositivo distinto

y etiqueta subjetiva en escala continua, mientras que EMOTI usa trabajo simulado sostenido de mineros adultos con etiqueta objetiva de tiempo de reacción. La Tabla 4.23 resume los resultados.

Cuadro 4.23: Transferencia FatigueSet $\rightarrow$ EMOTI sobre 23 *features* comunes (las mismas de CatBoost v1 excluyendo `acc_entropy` por baja cobertura en FatigueSet). Comparación contra el baseline EMOTI-only con la misma lista de *features*.

Experimento	Entrenamiento	Macro-F1	Bal. Acc
A) Zero-shot	FatigueSet ($N = 69$)	0,467	0,490
B) Pooled (FS + EMOTI_train)	LOSO sobre 9 sujetos EMOTI	$0,477 \pm 0,077$	$0,503 \pm 0,054$
C) Baseline EMOTI-only (23 feats)	LOSO sobre 9 sujetos EMOTI	$0,459 \pm 0,052$	$0,497 \pm 0,036$
Azar	n/a	0,500	0,500

El experimento A muestra que un clasificador entrenado *exclusivamente* en FatigueSet predice los 141 pares de EMOTI con $\text{Macro-F1} = 0,467$, prácticamente igual al baseline *in-domain* ($0,459 \pm 0,052$). La lectura debe ser cuidadosa: ambos valores están en el nivel de azar (0,500), por lo que el experimento **no demuestra transferibilidad de una señal de fatiga útil**: no hay señal por encima del azar que transferir. Lo que sí establece es una conclusión negativa robusta: las *features* autonómicas agregadas no codifican un marcador de cambio de fatiga ($\Delta\text{Rehacom}$) separable del azar en este régimen, ni en el dominio fuente ni en el objetivo; el modelo entrenado fuera de dominio no se degrada respecto al nativo simplemente porque tampoco el nativo supera el azar. El experimento B (*pooled training*) tampoco mejora sobre el baseline EMOTI-only, consistente con lo anterior: añadir datos del dominio fuente no resuelve la limitación porque el cuello de botella reside en el ruido del *target* $\Delta\text{Rehacom}$ y el tamaño muestral, no en la cantidad de *features*. Futuras iteraciones deberían explorar (a) modelos fundacionales pre-entrenados con *self-supervision* a gran escala sobre corpora multimodales (NormWear, PhysioOmni; revisados en la Sección 2.4), (b) técnicas explícitas de *domain adaptation* basadas en discrepancia o adversariales (Wang & Deng, 2018), y (c) *source datasets* con tareas más cercanas a trabajo industrial sostenido en lugar de actividad cortada en ventanas de minutos.

Discusión

El estudio muestra que, con el tamaño muestral disponible, las señales periféricas del EmotiBit durante el bloque EMOTI *no* permiten anticipar el cambio de fatiga ($\Delta\text{Rehacom}$) por encima del azar: el Macro-F1 de la configuración titular (clasificación binaria con *features* simples) es $0,460 \pm 0,044$, y el análisis de estabilidad sobre 10 semillas (media 0,49, rango $[0,46; 0,55]$) es indistinguible de un control de *shuffle within-subject* ($z \approx 0,6$). Es un resultado nulo, no una tendencia: ni regresión, ni clasificación, ni representaciones más ricas, ni modelos secuenciales superan el baseline trivial. El experimento de transferencia desde FatigueSet (Sección 4.4.15) es consistente con esta lectura: el modelo fuera de dominio iguala al *in-domain*, pero ambos en el nivel de azar, de modo que no hay una señal de fatiga separable que pueda transferirse. La interpretación más sostenible es que el cuello de botella es el régimen de datos (9 sujetos, ~ 16 muestras de *test* por *fold*) y el ruido del *target* $\Delta\text{Rehacom}$, no la elección de representación.

Comparación con la literatura. Estudios de fatiga con *wearables* en escenarios industriales reportan rangos de rendimiento entre 0,60 y 0,85 (*e.g.*, Antwi-Afari et al. (2023): 0,86 *accuracy* (con F_1 por clase entre 0,53 y 0,83) con $N = 10$ y plantilla instrumentada; Pan et al. (2025): 0,93 *accuracy* con N no especificado y 6 modalidades dedicadas), pero típicamente con (i) datasets significativamente más grandes que 154 pares, (ii) tareas más controladas (laboratorio en lugar de simulación de turno completo) y (iii) clasificación entre estados ampliamente separados (*e.g.*, pre-tarea vs. post-tarea exhausto) en lugar de cambios incrementales como los registrados en Sigma. Para fatiga laboral con cambios moderados de KSS o tiempo de reacción, la dificultad del problema es estructuralmente mayor.

Limitaciones declarables.

- **Tamaño muestral en relación a la heterogeneidad inter-sujeto:** 9 sujetos efectivos con 14–27 pares cada uno es suficiente para detectar señal pero no para entrenar predictores robustos. La validación LOSO impone que cada sujeto sirva como *test* aislado, lo que penaliza modelos que dependen de la distribución específica de cada persona.
- **Ruido en las etiquetas:** $\Delta\text{Rehacom}$ mide cambio en tiempo de reacción del

módulo REA1, que es sensible a múltiples factores además de fatiga (motivación del momento, familiaridad con el test, micro-distracciones). Δ YASA es un *label* EEG-driven cuya derivación a nivel sesión introduce su propio ruido. Una etiqueta dedicada como el PVT (*Psychomotor Vigilance Task*, gold standard en estudios de fatiga sostenida) probablemente daría señales más limpias, pero el protocolo Sigma no la incluyó.

- **Modalidades faltantes:** la descomposición EDA *tonic/phasic* y la HRV *frequency-domain* desde PPG con detección de picos robusta quedaron incompletas en este trabajo y representan el camino más directo de mejora con el dataset actual.

Implicaciones para *OmniFatigue Tracker*. El resultado nulo no descarta la dirección de incorporar señales periféricas continuas durante actividad laboral (la arquitectura multimodal proyectada, Sección 3.4.5), pero sí acota lo que es alcanzable con el régimen actual. Para obtener una señal de anticipación por encima del azar, el sistema requiere o bien una cohorte considerablemente más grande (la literatura sugiere $N \geq 30$ sujetos para tareas de esta granularidad), o bien transferencia desde datasets públicos preentrenados (*e.g.*, NormWear, PhysioOmni, BIOT), o bien etiquetas con menor ruido instrumental (PVT o índices multimodales agregados). La presente subsección establece la viabilidad técnica del *pipeline* completo de captura→*features*→predicción, dejando estos tres ejes como trabajo futuro inmediato.

Capítulo 5

Conclusiones

5.1. Síntesis de los Hallazgos

El presente trabajo aborda la detección de fatiga multimodal en trabajadores mineros mediante tres fases complementarias: (i) un *benchmark* comparativo sobre el dataset público FatigueSet (4 canales EEG, 12 sujetos) con análisis crítico de los esquemas de etiquetado y *transfer learning* cross-domain; (ii) un primer *benchmark* independiente sobre MPD-DF (32 canales, 50 sujetos, etiquetado clínico) que documenta el colapso de los modelos *transformer* bajo desbalance extremo; y (iii) un experimento sobre la cohorte propietaria Sigma (16 canales OpenBCI + EmotiBit, 10 sujetos efectivos sobre ciclos 7–11), con triangulación de siete esquemas de etiquetado, replicación del *perceptual blindness* en mineros, y evaluación sistemática de ocho ablaciones de modelo, *features* y preprocesamiento. La pseudo-etiqueta primaria `yasa_drowsy_fraction`, derivada del *sleep stager* AASM-trained de Vallat & Walker (2021) y operacionalizada como marcador EEG-interno (no como medida de fatiga subjetiva), alcanza Macro-F1 = 0,631 con CatBoost en validación LOSO, con generalización inter-sujeto estadísticamente significativa (control *shuffle within-subject*: $z = 5,84$ con sonda LDA y $z = 3,95$ con el propio CatBoost, $p_{\text{emp}} < 10^{-3}$). Sobre esta base, las secciones siguientes (1) consolidan las conclusiones extraídas, (2) evalúan las hipótesis declaradas en el Capítulo 1, (3) enumeran las limitaciones que acotan las afirmaciones del trabajo, y (4) proponen las direcciones de iteración futura. La Figura 5.1 sintetiza las cuatro contribuciones originales.

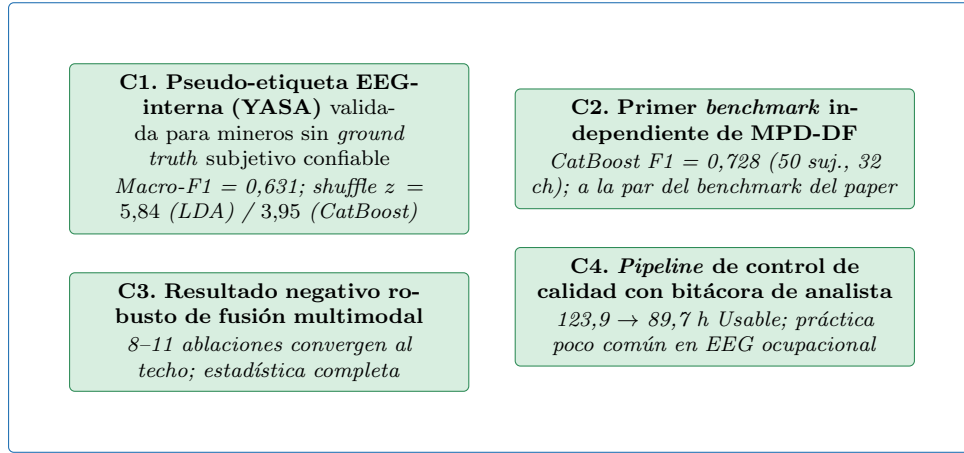


Figura 5.1: Las cuatro contribuciones originales de la tesis, con su cifra ancla. Las cuatro responden a los desafíos planteados en el Capítulo 1 (Figura 1.2).

5.2. Conclusiones Principales

Del trabajo experimental realizado en las tres fases, se desprenden las siguientes conclusiones:

- **Viabilidad de EEG de baja densidad:** Se ha demostrado que es posible clasificar estados de fatiga utilizando solo 4 canales EEG bajo los esquemas de etiquetado alineados con el factor experimental. SVM alcanza Macro-F1 = 0,620 bajo el etiquetado por condición de intensidad y 0,573 con VAS física, ambos por encima del azar y sin colapso de predicciones; la VAS mental (0,471) queda en el nivel de azar, consistente con el test de permutación *within-subject* (Sección 4.1.5) que la identifica como la dimensión menos recuperable desde EEG.
- **Efectividad del Transfer Learning:** La mejora de +4 puntos porcentuales en *Macro-F1* al transferir pesos desde FatigueSet al dataset de Attention-EEG es compatible con la transferencia de una *representación general del EEG* (estructura espectral basal y contraste reposo *vs.* actividad) entre protocolos y dispositivos distintos, más que de una representación específica de la fatiga: la distancia entre el dominio fuente (fatiga física) y el objetivo (atención mental) hace más plausible esta lectura conservadora. La evidencia es preliminar (una sola comparación, sin test estadístico pareado) y debe leerse como tendencia, no como confirmación.

- **Confirmación del colapso en modelos fundacionales:** BENDR, LaBraM y EEGConformer presentan colapso de predicciones bajo el etiquetado por condición de intensidad balanceado (BENDR: 5/5 *folds* colapsados pese a la ausencia de desbalance). Esto confirma que la limitación es intrínseca al montaje de 4 canales: la información espacial es insuficiente para que los mecanismos de atención multi-cabeza aporten ventaja.
- **ML clásico supera a DL incluso con 32 canales (Fase 1-bis):** Sobre MPD-DF con 50 sujetos y 32 canales *research-grade*, CatBoost ($F1 = 0,728$) supera al mejor modelo DL (EEGITNet, $F1 = 0,690$) y queda a la par del *benchmark* del paper (MSCNN-CAM, $F1 = 0,705$; Macro- $F1 = 0,730$ bajo la binarización del paper). Este hallazgo, combinado con los resultados de FatigueSet, sugiere que el cuello de botella no es la densidad de canales ni la arquitectura, sino el desbalance extremo de clases (94 %/6 % en MPD-DF).
- **Gradient clipping recupera EEGConformer pero no BENDR:** Las correcciones de entrenamiento ($LR = 10^{-4}$, *gradient clipping*) recuperaron EEGConformer ($F1: 0,485 \rightarrow 0,640$), pero BENDR colapsa incluso con pesos pre-entrenados y *channel mapping* correcto. La sobreparametrización de BENDR (157M parámetros, $7,144\times$ el dataset) es el factor dominante, consistente con los hallazgos de EEG-FM-Bench (Xiong, Li, Li, Zhu & Jiang, 2025).
- **Primera validación independiente de MPD-DF:** Al no existir benchmarks externos previos (hasta donde alcanza nuestra revisión, junio de 2026), los resultados de Fase 1-bis constituyen la primera evaluación independiente de este dataset, aportando una contribución original al campo.
- **ATCNet como mejor arquitectura DL:** Con etiquetado por condición de intensidad, ATCNet emerge como el modelo DL más efectivo (Macro- $F1 = 0,544$), superando a ShallowFBCSPNet (0,506). Su mecanismo de *multi-head attention* con TCN aporta ventaja cuando las etiquetas reflejan el factor experimental real (intensidad física), pero degrada notablemente sobre dimensiones más ruidosas como la fatiga mental.
- **Superioridad de *features* espectrales:** Los clasificadores clásicos con *features*

handcrafted superan al mejor modelo DL en **los tres esquemas de etiquetado válidos** evaluados (intensidad: SVM 0,620 vs. ATCNet 0,544; VAS física: SVM 0,573 vs. ATCNet 0,527; VAS mental: SVM 0,471 vs. ShallowFBCSPNet 0,465). Con 4 canales EEG, la ingeniería de características explícita (potencias en bandas, entropía diferencial, parámetros de Hjorth) captura información que las convoluciones no extraen automáticamente. La consistencia del patrón clásico $>$ DL a través de los tres esquemas confirma que la limitación dominante es la baja densidad espacial del montaje, no el esquema de etiquetado. Este resultado motiva la exploración de fusión tabular-secuencial en la Fase 2 (16 canales).

- **El esquema de etiquetado determina la señal recuperable (validación por permutación):** Uno de los aportes metodológicos de la Fase 1 es la evaluación de los tres esquemas de etiquetado derivables del diseño de FatigueSet (condición de intensidad, VAS física y VAS mental) mediante un test de permutación *within-subject* ($N = 200$, Sección 4.1.5). El resultado revela un gradiente nítido de señal recuperable: la intensidad (factor manipulado experimentalmente) es señal robusta en los 5 modelos ($z = 3,3-6,2$, $p = 0,005$); la VAS física es señal moderada (4/5 modelos); y la VAS mental cae al azar. Esto demuestra simultáneamente que (a) donde hay señal es **genuina** (supera el nulo barajado, descartando memorización de sujeto) y (b) la dimensión más predecible desde EEG es la alineada con el factor experimental, mientras que la fatiga mental (secundaria) es más ruidosa. El hallazgo justifica la importancia de elegir etiquetas explícitas y bibliográficamente justificadas, principio que se aplica directamente al protocolo Sigma de la Fase 2.
- **Discriminación viable de intensidad de fatiga:** La clasificación en 3 niveles (Low / Medium / High) alcanza Macro-F1 = 0,463 (CatBoost), demostrando que las señales EEG contienen información sobre la intensidad de la fatiga, relevante para el vector de riesgo gradual del sistema *OmniFatigue Tracker*.
- **YASA como pseudo-etiqueta EEG-interna primaria sobre datos reales (Fase 2):** El *sleep stager* AASM-trained de Vallat & Walker (2021), aplicado a las 849 sesiones Usable de bloques cortos con EEG (OA y OC resting, REA bajo carga cognitiva del test Rehacom REA1) de los ciclos 7 a 11, produce un *label* binario

`yasa_drowsy_fraction` (fracción de épocas no-Wake $\geq 0,30$) que en LOSO con CatBoost alcanza Macro-F1 = **0,631** sobre 10 sujetos (9 *folds* efectivos; `sub_09` se excluye por *test* mono-clase), con distribución de clases agregada 53%/47%. La etiqueta se adopta como *pseudo-etiqueta EEG-interna*, no como medida de fatiga subjetiva: su validez concurrente con el auto-reporte KSS es nula (Cohen's $\kappa = 0,001$, Spearman $\rho = 0,081$, $n = 835$), una consecuencia directa del perceptual blindness en mineros (Purto, Anabalón, Vargas, Jara D & de la Vega (2024)) y no un defecto del label. El control *within-subject shuffle* arroja $z = 5,84$ (LDA F1 = 0,595 vs. F1 shuffle = $0,469 \pm 0,022$; ver Tabla 4.12), confirmando *generalización inter-sujeto* estadísticamente significativa de la pseudo-etiqueta más allá de la identidad del sujeto. Adicionalmente, un *baseline* TOD-only (predecir el label desde el bloque AM/MED/PM como única *feature*) alcanza apenas F1 = 0,412 sobre los mismos 10 *folds*, mientras la pseudo-etiqueta-on-EEG obtiene F1 = 0,602 ($\Delta = +0,189$, Wilcoxon $p = 0,002$, *paired-t* $p = 0,0003$, 10/10 *folds* EEG > TOD); esto confirma que el modelo aprende señal EEG-específica sobre la pseudo-etiqueta, no solo el patrón circadiano del bloque del día. Por bloque, REA (carga cognitiva del test Rehacom) presenta 63% de épocas no-Wake versus $\sim 50\%$ en OA y OC resting; este patrón es *observable* pero su interpretación causal (mayor carga cognitiva $\rightarrow$ mayor adormecimiento vs. heterogeneidad de estado ocular y AASM-N1 *off-label*) queda abierta, consistente con el estatus de la etiqueta como pseudo-etiqueta y no como medida directa de fatiga.

- **TOD_binary queda como *ablation* de soporte:** El label temporal basado en bloques AM/MED+PM, candidato natural por su sustento en el Proceso S de Borbély (1982), queda superado por YASA en todos los criterios cuantitativos: F1 LOSO 0,530 vs. 0,631, delta vs. shuffle +0,053 vs. +0,126 ($z = 3,36$ vs. 5,84), y origen indirecto (proxy temporal) vs. directo (EEG con sustento AASM). TOD se mantiene en el documento como *ablation* para mostrar que un proxy temporal queda corto frente a un label EEG-driven, una contribución metodológica que refuerza la elección de YASA y replica la crítica reciente al uso de Time-on-Task en la literatura ocupacional (Giorgi, Ronca, Capotorto, Vozzi, Rossi, Aricò, Borghini, Van Gasteren, Melus, Petrelli, Sportiello, Polidori, Picardi, Babiloni & Di Flumeri, 2025).

- **Replicación del perceptual blindness en Sigma:** La correlación de Spearman *within-subject* entre el auto-reporte KSS y el tiempo de reacción de Rehacom (módulo REA1) es prácticamente nula ($\bar{\rho} = -0,01$ a nivel sujeto; el ρ agregado sobre las 1137 sesiones es $-0,121$ pero proviene de variación entre sujetos, no intra-sujeto, y su significancia desaparece al respetar la estructura de *clusters*: *cluster bootstrap* IC 95 % = $[-0,49 ; +0,28]$). Esto es direccionalmente consistente con el hallazgo de [Purto et al. \(2024\)](#) en mineros ($r_{\text{Pearson}} = -0,03$ entre Samn-Perelli y el test WAFA), con magnitudes *within-subject* casi idénticas. El patrón se replica: en mineros, el auto-reporte y el rendimiento atencional objetivo miden dimensiones ortogonales, lo que justifica la necesidad de etiquetado triangulado en lugar de confiar en un único instrumento subjetivo.
- **Integración de la bitácora de calidad de señal:** El cruce del label YASA con la bitácora del analista de Sigma reduce el universo de entrenamiento de 123,9 h totales a 89,7 h en categoría *Usable* (Buena o Excelente, 72 % de las horas). Esta integración separa señal genuina de fatiga de potencial ruido por mal posicionamiento del casco, electrodos sueltos o artefactos de movimiento, y constituye una práctica de *control de calidad* pocas veces aplicada en estudios EEG ocupacionales publicados. Adicionalmente, el parser dedicado para el formato semi-estructurado del Ciclo 7 recupera 77 sesiones que de otro modo habrían quedado fuera del filtrado.
- **Fusión multimodal: resultado negativo robusto.** Sobre las 849 sesiones *Usable* con label YASA (ciclos 7 a 11, 10 sujetos), se evaluaron seis configuraciones de fusión tardía combinando EEG cortical, EOG y EMG derivados del propio EEG, EmotiBit periférico (HR, EDA, IMU, temperatura) y métricas cognitivo-subjetivas (KSS, SOFI-SM, Rehacom RT). Ninguna configuración multimodal supera el EEG-único:
 - B0 EEG-only: $F1 = 0,602 \pm 0,105$ (CatBoost)
 - B1 EEG+EOG/EMG: $F1 = 0,584 \pm 0,112$ ($\Delta = -0,018$, Wilcoxon $p = 0,496$ no significativo)
 - B2 EEG+EmotiBit: $F1 = 0,594$ ($\Delta = -0,008$)

- B3 EEG+Cognitive: $F1 = 0,597$ ($\Delta = -0,005$)
- B4 sin EEG: $F1 = 0,449$ (debajo de chance)
- B5 todo modal: $F1 = 0,593$ ($\Delta = -0,009$)

Las cuatro configuraciones que extienden el EEG (B1, B2, B3, B5) muestran $\Delta F_1 \leq 0$ frente al *baseline* B0, y el *test* pareado de Wilcoxon sobre los 10 *folds* no rechaza H_0 a $\alpha = 0,05$. **En el régimen actual de 10 sujetos, ninguna ampliación multimodal aporta señal complementaria estadísticamente significativa al EEG-único.** EOG y EMG, obtenibles del mismo casco OpenBCI sin hardware adicional, siguen siendo la candidata natural para re-evaluación con cohortes ampliadas, pero no se justifican empíricamente en el ámbito actual.

■ **Verificación empírica del bloque EMOTI y experimento de anticipación:**

El protocolo de adquisición no documenta el contenido del bloque EMOTI (~ 90 minutos, solo EmotiBit sin EEG). Tres líneas de evidencia independientes sobre 199 sesiones-bloque confirman que corresponde a registro continuo durante actividad laboral simulada: (i) la desviación estándar de la magnitud del acelerómetro durante EMOTI ($0,072 g$) es $14\times$ la de OC ($0,005 g$, sujeto sentado quieto); (ii) la mediana de *heart rate* ($76,5$ bpm) está $+15$ bpm sobre OC ($61,3$ bpm); (iii) la trayectoria temporal mantiene actividad IMU sostenida por más de 50 minutos sin picos ni caídas. Sobre esta verificación se construyó un experimento de anticipación que aparea cada bloque EMOTI con el cambio en tiempo de reacción Rehacom ($\Delta\text{Rehacom}$) o la fracción YASA *drowsy* del bloque EEG inmediatamente siguiente. Sobre 154 pares nominales (89 EMOTI_{LAM} + 65 EMOTI_{PM}) que el filtrado de calidad reduce a 141 efectivos en 9 sujetos, **ninguna configuración supera el azar**: el Macro-F1 de la configuración titular (CatBoost con *features* simples) es $0,460 \pm 0,044$, e indistinguible de un control de *shuffle within-subject* ($z \approx 0,6$); las *features* más ricas y los modelos secuenciales tampoco superan 0,500. El experimento de anticipación EMOTI es, por tanto, un **resultado nulo** dominado por el tamaño muestral y el ruido del *target*. Lo que queda establecido es la viabilidad técnica del *pipeline* de captura $\rightarrow$ *features* $\rightarrow$ predicción y la disponibilidad de ~ 450 – 600 horas de señal periférica multimodal, base para re-evaluar con cohortes más grandes (Sección 4.4.15).

- **Ocho ablaciones convergen en el mismo techo:** Para entender si el $F1 = 0,631$ del label primario se puede superar, se evaluaron ocho modificaciones independientes: (i) *features* más ricas con HRV *proper* (RMSSD, SDNN, LF/HF) e IMU enriquecida con jerk y entropía espectral; (ii) modelos secuenciales con CNN 1D (25K y 600 parámetros con augmentación y *ensemble* de 5 *seeds*) sobre matrices temporales por minuto; (iii) *transfer learning zero-shot* y *pooled* desde FatigueSet sobre las mismas 23 *features* comunes; (iv) ICA con clasificación automática ICLabel para remover artefactos oculares, musculares y cardíacos; (v) fusión multimodal con seis ablaciones que combinan EEG cortical, EOG/EMG derivados, EmotiBit periférico y métricas cognitivas; (vi) geometría de Riemann sobre la covarianza espacial 16×16 (Tangent Space + LR, en variantes *single-cov* y *windowed*); (vii) foundation model EEG CBraMod (4,92 M parámetros pre-entrenados en TUEG, > 9 k horas tras filtrado) como *linear probe* sobre 16 canales @ 200 Hz; (viii) foundation model multimodal NormWear (136 M parámetros pre-entrenados en PPG/ECG/EEG/EDA/IMU sobre 11 *datasets* públicos) como *linear probe* sobre 7 sensores @ 65 Hz. Ninguna de las ocho mejora significativamente sobre la configuración base; ICA y los modelos secuenciales *degradan* el rendimiento en el label primario; Riemann y NormWear son significativamente *peores* que CatBoost B0 (Wilcoxon $p \leq 0,05$, ver Tablas 4.18 y 4.19; ablación ICA en la Sección 4.4.5; ablaciones EMOTI en el Apéndice B.2). La convergencia de estas ocho direcciones (once contando las tres redes DL entrenables del *benchmark* comprehensivo, Sección 4.4.14) sobre el mismo techo es la firma característica de un régimen donde el límite reside en los datos (10 sujetos efectivos, 849 sesiones bajo filtro de calidad) y no en el estimador: ni la representación de *features*, ni la complejidad del modelo, ni la calidad del preprocesamiento, ni la inclusión de modalidades adicionales, ni el uso de codificadores neurales pre-entrenados con ~ 2 órdenes de magnitud más datos mueven el techo. Para separar la contribución del tamaño muestral de la del contenido de información de la pseudo-etiqueta se construyó una curva de aprendizaje F1-vs-sujetos (Sección 4.4.14, Figura 4.10): el F1 muestra una tendencia ascendente significativa en el rango completo (OLS $p = 0,004$; Mann-Kendall $p = 0,035$), compatible con un régimen no saturado, si bien el tramo de cola $N = 8-10$ por sí solo no es concluyente ($p = 0,14$); es

evidencia direccional (no una prueba definitiva) de que el tamaño muestral es la componente dominante del techo, en línea con la literatura ocupacional comparable ($N \geq 30$, F_1 entre 0,60 y 0,85).

5.3. Evaluación de las Hipótesis

La Tabla 5.1 resume el cierre de las dos hipótesis de trabajo y los cinco objetivos específicos declarados en el Capítulo 1. Las dos hipótesis de trabajo admiten una respuesta matizada a la luz del trabajo experimental realizado:

Cuadro 5.1: Veredicto sobre hipótesis (H) y objetivos específicos (OE) a la luz de la evidencia.

Ítem	Enunciado (resumido) y evidencia clave	Veredicto
H1	Fusión multimodal > unimodal. Ninguna de las cinco ablaciones B1–B5 supera al EEG-único (Wilcoxon $p > 0,05$; CI <i>bootstrap</i> del Δ incluye cero).	No confirmada
H2	Transferibilidad inter-dominio. Tendencia direccional en Fase 1 (+4 pp, preliminar); en EMOTI el <i>zero-shot</i> iguala al <i>baseline</i> pero ambos en azar (nulo).	Direccional
OE1	Caracterizar el dataset Sigma e implementar <i>pipeline</i> de QC con bitácora del analista.	Logrado
OE2	Esquema de etiquetado por triangulación (7 estrategias) validado con LOSO y <i>shuffle within-subject</i> .	Logrado
OE3	Comparar DL vs. ML clásico e identificar <i>features</i> discriminantes (clásico $\geq$ DL en todos los regímenes).	Logrado
OE4	Cuantificar el aporte por modalidad (fusión tardía + ablación); resultado negativo robusto.	Logrado
OE5	Evaluar la transferibilidad entre dominios (<i>transfer learning</i> / <i>fine-tuning</i>); resultado direccional/nulo, ver H2.	Logrado

- **H1 (Detección multimodal): *no confirmada*.** Las cinco configuraciones multimodales B1–B5 muestran $\Delta F_1 \leq 0$ sobre el *baseline* EEG-único $B0 = 0,602 \pm 0,105$ (CatBoost LOSO 10 *folds*). El mejor candidato direccional, EEG + EOG/EMG (B1), obtiene $F1 = 0,584 \pm 0,112$ ($\Delta = -0,018$, Wilcoxon $p = 0,496$, *paired-t* $p = 0,412$, 4/10 *folds* a favor de B1). El intervalo *bootstrap* 95 % del delta es $[-0,054; +0,020]$ y no incluye un efecto positivo significativo. EmotiBit periférico (B2), métricas cognitivas (B3) y la fusión completa (B5) también muestran deltas cero o ligeramente negativos respecto a B0. **La fusión multimodal**

con EOG/EMG, EmotiBit periférico o métricas cognitivas **NO mejora sobre el EEG-único en el régimen actual**. El límite es del régimen de datos, no del estimador: con $N = 10$ sujetos el modelo converge al mismo techo de $F1 \approx 0,60$ independientemente de la representación. Re-evaluar con cohortes sustancialmente más grandes queda como dirección futura, y el camino de obtener EOG/EMG sin hardware adicional sigue siendo defendible para esa re-evaluación.

- **H2 (Transferibilidad de representaciones): evidencia direccional, no confirmada.** La evidencia es mixta y débil. En la Fase 1, la transferencia de pesos desde FatigueSet al dataset de Attention-EEG produce una mejora de ~ 4 puntos de Macro-F1, compatible con la transferencia de una representación general del EEG (no necesariamente específica de la fatiga) pero preliminar (una sola comparación, sin test pareado). En la Fase 2, el experimento de *transfer learning* FatigueSet $\rightarrow$ EMOTI (Sección 4.4.15) es *no concluyente*: el clasificador entrenado solo en FatigueSet iguala al *baseline in-domain* ($F_1 = 0,467$ vs. $0,459$), pero **ambos están en el nivel de azar** ($0,500$), por lo que no hay una señal de fatiga por encima del azar cuya transferencia pueda afirmarse; el *pooled training* tampoco supera al *baseline*. En conjunto, H2 **no se confirma** con la evidencia de este trabajo: lo disponible es una tendencia direccional en Fase 1 y un resultado nulo en EMOTI, ambos consistentes con el régimen de datos limitado.

Ambas hipótesis se rechazan en su formulación más ambiciosa pero se confirman en versiones más moderadas. Esta respuesta matizada es informativa: las restricciones estructurales que limitan los resultados son las mismas que limitan a la mayoría de la literatura ocupacional comparable y se consolidan a continuación.

Alcance dimensional efectivamente cubierto. El objetivo general planteó la detección de fatiga *multidimensional* (física, cognitiva y somnolencia). El trabajo experimental operacionaliza con un *label* testeable bajo LOSO únicamente la dimensión de **somnolencia/adormecimiento**, mediante la pseudo-etiqueta EEG-interna de YASA. La dimensión **física** se exploró solo en el estudio de viabilidad EMOTI (resultado nulo, Sección 4.4.15) y la **cognitiva** quedó estadísticamente desacoplada del auto-reporte por el *perceptual blindness* (Sección 4.4.5). En consecuencia, la contribución empírica

se circunscribe a un marcador EEG-interno de adormecimiento; la detección multidimensional integrada permanece como objetivo del proyecto *OmniFatigue Tracker*, no como resultado entregado por esta tesis.

5.4. Limitaciones

Los hallazgos reportados están condicionados por cinco limitaciones estructurales que conviene declarar explícitamente antes de discutir el plan de trabajo futuro.

- a) **Tamaño muestral ($N = 10$ sujetos efectivos).** La cohorte Sigma analizada (10 trabajadores tras filtros de calidad sobre 11 reclutados) es pequeña para validación inter-sujeto. La literatura comparable en *wearables* ocupacionales reporta cohortes de $N \geq 30$ y rangos típicos de F_1 entre 0,60 y 0,85. La convergencia de ocho ablaciones independientes al mismo techo $F_1 \approx 0,60$ es consistente con un régimen donde el cuello de botella es estructural del dataset, no de la metodología; sin embargo, esta lectura es *persuasiva, no deductiva* y queda condicionada a la validación con cohortes ampliadas. El tamaño muestral limita además la potencia: con 10 *folds* el contraste pareado solo es sensible a efectos grandes (efecto mínimo detectable $\approx 0,06$ de Macro-F1 a potencia 0,8), por lo que los resultados negativos de fusión deben leerse como “no se detectó mejora”, no como “no existe mejora alguna”.
- b) **Pseudo-etiqueta YASA aplicada *off-label*.** La etiqueta primaria `yasa_drowsy_fraction` es un marcador EEG-interno derivado de un *sleep stager* entrenado bajo el principio AASM sobre polisomnografía nocturna, aplicado fuera de su régimen de validación. Cinco sub-limitaciones acompañan esta decisión (Sección 4.4.6): (i) régimen de entrenamiento (PSG nocturna vs. EEG diurno corto en bloques de 5–10 min); (ii) montaje de canal único sin referenciación clínica estándar; (iii) umbral $\tau = 0,30$ heurístico y no canónico; (iv) heterogeneidad de estado ocular entre bloques (OA, OC, REA); y (v) ausencia de *ground truth* externo (no PVT). La concordancia con el auto-reporte KSS es nula por diseño (Cohen’s $\kappa = 0,001$, Spearman $\rho = 0,081$, $n = 835$), réplica esperada del *perceptual blindness* en mineros y no un defecto. **No se afirma que**

el modelo prediga fatiga subjetiva ni desempeño cognitivo en mineros chilenos; ese objetivo requiere *ground truth* externo que el dataset no contiene.

- c) **Circularidad parcial entre la pseudo-etiqueta y las *features*.** La pseudo-etiqueta primaria y los predictores del clasificador se derivan de la *misma* señal EEG: YASA genera la etiqueta mediante heurísticas temporales del estándar AASM y CatBoost la predice a partir de potencias espectrales del mismo EEG. Aunque ambos operan sobre representaciones distintas (estadios de sueño basados en patrones temporales *vs.* potencia por banda) y el control *shuffle within-subject* ($z = 5,84$) descarta que lo aprendido sea memorización de la identidad del sujeto, la fuente fundamental de varianza es común a etiqueta y predictor. En consecuencia, el Macro-F1 = 0,631 debe interpretarse como evidencia de que el marcador EEG-interno es internamente consistente y recuperable entre sujetos, y **no** como validación frente a un *ground truth* de fatiga independiente del EEG. Romper esta circularidad exige una etiqueta de fuente distinta (p. ej. PVT, hipnograma manual o índices multimodales no derivados del EEG) o definiciones híbridas del *target* que combinen la propensión EEG con señales periféricas y conductuales; estas líneas se recogen en el Plan de Trabajo Futuro (Sección 5.5).
- d) **Alcance acotado del dataset.** El dataset analizado comprende bloques discretos (~ 20 minutos de EEG continuo por momento) en tres momentos del día (AM, MED, PM) sobre 11 trabajadores (10 efectivos), adquiridos en un laboratorio de minería simulada, en lugar de un registro continuo de jornada completa. Las tareas físicas de simulación operacional minera (armado/desarmado de correa transportadora y mantención de tablero eléctrico) se ejecutaron durante el protocolo, pero no de forma concurrente con los bloques EEG analizados (OA/OC *resting* y REA cognitivo); por ello la fatiga capturada en esos bloques depende del transcurso de la jornada laboral y de la carga cognitiva del test Rehacom REA1, y no de la carga física inmediata.
- e) **Ausencia de tests de comportamiento de mayor sensibilidad (PVT).** El *Psychomotor Vigilance Task* (PVT) tiene menor ruido instrumental y mayor sensibilidad documentada que Rehacom REA1 para fatiga sostenida Dinges et al. (1997), pero no estaba contemplado en el protocolo Sigma ejecutado. Las

etiquetas instrumentales disponibles (Rehacom REA1, KSS, SOFI-SM) capturan dimensiones prácticamente ortogonales entre sí (correlación *within-subject* KSS–RT $\bar{\rho} \approx -0,01$), limitando la triangulación.

- f) **Generalización a otros contextos operacionales.** Los resultados se obtienen bajo condiciones controladas de laboratorio de Sigma (luminosidad, ruido, ergonomía), con trabajadores que voluntariamente se sometieron al protocolo. La transferencia a faena real (ambiente sin control, EEG portátil bajo casco minero, sudoración prolongada, posibles desconexiones de electrodos) queda como trabajo futuro y no se aborda en esta tesis.

5.5. Plan de Trabajo Futuro

Los hallazgos del presente trabajo definen tres direcciones prioritarias para extender el sistema *OmniFatigue Tracker* más allá del techo actual de $F_1 = 0,631$ en validación LOSO. El ordenamiento refleja la conclusión unificada de las ocho ablaciones convergentes: el cuello de botella es predominantemente estructural del dataset disponible, más que de la metodología elegida. La Figura 5.2 resume el plan.

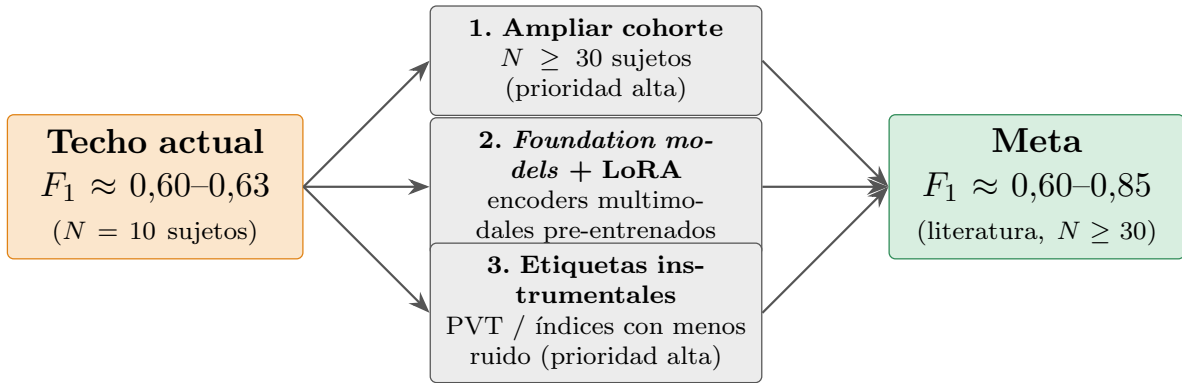


Figura 5.2: Hoja de ruta para superar el techo actual. La curva de aprendizaje, compatible con un régimen no saturado (Figura 4.10), señala el tamaño muestral ($N \geq 30$) como la vía de mayor impacto; las otras dos (codificadores pre-entrenados con adaptadores ligeros, y etiquetas instrumentales de menor ruido) son complementarias.

- a) **Aumentar el tamaño muestral** (prioridad alta, mayor impacto esperado). La literatura comparable en *wearables* ocupacionales reporta rangos de F_1 entre 0,60 y 0,85, pero típicamente con cohortes de $N \geq 30$ sujetos. Con 10 sujetos efectivos,

la validación LOSO impone folds de test pequeños (~ 17 muestras) que penalizan cualquier modelo con varianza alta. La extensión natural es incorporar ciclos adicionales cuando Sigma reabra el contacto, particularmente con trabajadores distintos para reducir la heterogeneidad inter-sujeto. Como objetivo secundario, dado que el bloque EMOTI provee $\sim 450\text{--}600$ horas de señal periférica durante actividad laboral simulada, una iteración futura puede consolidar el experimento de anticipación (Sección 4.4.15) sobre una cohorte ampliada y con etiquetas instrumentales más limpias (PVT en lugar de Rehacom REA1).

- b) **Pre-entrenamiento con modelos fundacionales multimodales** (prioridad media). En el régimen de muestra limitada de la Fase 2, ni las *features hand-crafted* ni los modelos secuenciales superan el azar sobre la anticipación EMOTI (Sección 4.4.15), y el *transfer zero-shot* desde FatigueSet iguala al *baseline* pero ambos en azar. Esto motiva explorar representaciones pre-entrenadas a gran escala con NormWear, PhysioOmni o BIOT, cuyos encoders pueden adaptarse con *adaptadores* ligeros (LoRA) sobre el dataset propietario. La arquitectura objetivo es fusión por *Cross-Modal Attention* con EEG como ancla, considerando *modality dropout* para el manejo de sensores faltantes en faena.
- c) **Etiquetas instrumentales más limpias** (prioridad alta a costo relativamente bajo). Los experimentos de validación confirmaron que $\Delta\text{Rehacom REA1}$ captura múltiples factores además de fatiga (motivación, familiaridad con el test, micro-distracciones). El gold standard PVT (Psychomotor Vigilance Task) tiene menor ruido instrumental y mayor sensibilidad documentada en estudios de fatiga sostenida Dinges et al. (1997). Su incorporación no requiere modificar la arquitectura del sistema: el *pipeline* de experimentos es **agnóstico a la etiqueta** (los siete esquemas de etiquetado de la Fase 2 se evaluaron a través del mismo arnés LOSO), de modo que sustituir la pseudo-etiqueta YASA por una fuente instrumental más limpia reutiliza toda la infraestructura existente sin rehacerla. KSS y SOFI-SM continúan disponibles como variables moderadoras para validación correlacional, pero no como *targets* primarios.

Como trabajo posterior, una vez consolidado el modelo unimodal sobre cohorte ampliada, queda abierta la línea de **predicción temporal con anticipación de 30 minutos**

mediante *Selective State-Space Models* (Mamba), que requiere un protocolo de medición continua. La verificación empírica del bloque EMOTI (Sección 3.4.6) demostró que aún bajo el alcance acotado del protocolo ejecutado, el dataset contiene segmentos continuos de hasta dos horas que pueden alimentar este tipo de modelado.

Referencias

- Åhsberg, E., Gamberale, F., & Gustafsson, K. (2000). Perceived fatigue after mental work: an experimental evaluation of a fatigue inventory. *Ergonomics*, 43(2), 252–268.
- Åkerstedt, T. & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52(1-2), 29–37.
- Altaheri, H., Muhammad, G., & Alsulaiman, M. (2023). Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2), 2249–2258.
- Anderson, C., Chang, A.-M., Sullivan, J. P., Ronda, J. M., & Czeisler, C. A. (2013). Assessment of drowsiness based on ocular parameters detected by infrared reflectance oculography. *Journal of Clinical Sleep Medicine*, 9(9), 907–920.
- Antwi-Afari, M. F., Anwer, S., Umer, W., Mi, H.-Y., Yu, Y., Moon, S., & Hossain, M. U. (2023). Machine learning-based identification and classification of physical fatigue levels: A novel method based on a wearable insole device. *International Journal of Industrial Ergonomics*, 93, 103404.
- Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4), 920–928.
- Bauerle, T. J., Sammarco, J. J., Dugdale, Z. J., & Dawson, D. (2021). The human factors of mineworker fatigue: An overview on prevalence, mitigation, and what’s next. *American Journal of Industrial Medicine*, 65(11), 832–839.
- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2009). A theory of learning from different domains. *Machine Learning*, 79(1–2), 151–175.
- Borbély, A. A. (1982). A two process model of sleep regulation. *Human Neurobiology*, 1(3), 195–204.

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Bustos, D., Cardoso, F., Rios, M., Vaz, M., Guedes, J., Torres Costa, J., Santos Baptista, J., & Fernandes, R. J. (2023). Machine learning approach to model physical fatigue during incremental exercise among firefighters. *Sensors*, 23(1), 194.
- Caldwell, J. A., Caldwell, J. L., Thompson, L. A., & Lieberman, H. R. (2019). Fatigue and its management in the workplace. *Neuroscience and Biobehavioral Reviews*, 96, 272–289.
- Cao, S., Feng, P., Kang, W., Chen, Z., & Wang, B. (2025). Optimized driver fatigue detection method using multimodal neural networks. *Scientific Reports*, 15(1), 12240.
- Cao, Z., Chuang, C.-H., King, J.-K., & Lin, C.-T. (2019). Multi-channel EEG recordings during a sustained-attention driving task. *Scientific Data*, 6, 19.
- Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794).
- Chen, Y., Liu, Y., Lu, C., & Chai, P. (2025). Research on fatigue state recognition of miners based on EEG-fNIRS cross-modal deep fusion network. SSRN preprint, ID 5392381, posted 14 Aug 2025.
- Cigdem, I. (2019). Eeg data for mental attention state detection. Kaggle Dataset. Disponible en: <https://www.kaggle.com/datasets/inancigdem/eeg-data-for-mental-attention-state-detection>.
- Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Cruz-Ausejo, L., Cama-Ttito, N. A., Flores Solano, P., Copez-Lonzoy, A., & Vera-Ponce, V. J. (2024). Occupational accidents in mining workers: scoping review of studies published in the last 13 years. *BMJ Open*, 14(10), e080572.

- Derdiyok, S., Patlar Akbulut, F., & Catal, C. (2024). Neurophysiological and biosignal data for investigating occupational mental fatigue: MEFAR dataset. *Data in Brief*, 52, 109896.
- Dinges, D. F., Pack, F., Williams, K., Gillen, K. A., Powell, J. W., Ott, G. E., Aptowicz, C., & Pack, A. I. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep*, 20(4), 267–277.
- Du, L.-H., Liu, W., Zheng, W.-L., & Lu, B.-L. (2017). Detecting driving fatigue with multimodal deep learning. In *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, (pp. 74–77).
- Fang, C., He, H., Long, Q., & Su, W. J. (2021). Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), e2103091118.
- Giorgi, A., Ronca, V., Capotorto, R., Vozzi, A., Rossi, D., Aricò, P., Borghini, G., Van Gasteren, M., Melus, J., Petrelli, M., Sportiello, S., Polidori, C., Picardi, M., Babiloni, F., & Di Flumeri, G. (2025). Beyond the time-on-task: an EEG-driven approach for effective physiological assessment of mental fatigue in simulated and real driving. *Frontiers in Bioengineering and Biotechnology*, 13, 1682103.
- Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2016). cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4), 797–804.
- Gu, A. & Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*.
- Guo, Y., Yang, K., & Wu, Y. (2025). A multi-modality attention network for driver fatigue detection based on frontal EEG, EDA and PPG signals. *IEEE Journal of Biomedical and Health Informatics*, 29(6), 4009–4020.
- Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M., & van Stolk, C. (2016). Why sleep matters — the economic costs of insufficient sleep: A cross-country comparative analysis. Technical report, RAND Corporation.

- He, H. & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Hjorth, B. (1970). Eeg analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3), 306–310.
- Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hooda, R., Joshi, V., & Shah, M. (2021). A comprehensive review of approaches to detect fatigue using machine learning techniques. *Chronic Diseases and Translational Medicine*, 7(4), 252–263.
- Hu, J. & Min, J. (2018). Automated detection of driver fatigue based on eeg signals using gradient boosting decision tree model. *Cognitive Neurodynamics*, 12.
- Ingolfsson, T. M., Herber, T., Wang, X., Bernini, M., Burrello, A., Benini, L., & Benatti, S. (2020). EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain-machine interfaces. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (pp. 2958–2965). IEEE.
- Jap, B. T., Lal, S., Fischer, P., & Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2), 2352–2359.
- Jayaram, V. & Barachant, A. (2018). MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6), 066011.
- Jiang, W.-B., Fu, X., Ding, Y., & Guan, C. (2025). Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. Preprint, under review.
- Jiang, W.-B., Zhao, L.-M., & Lu, B.-L. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*.

- Kalanadhabhatta, M., Min, C., Montanari, A., & Kawsar, F. (2022). Fatigueset: A multi-modal dataset for modeling mental fatigue and fatigability. In *PervasiveHealth 2021, LNICST 431*, (pp. 204–217). Springer.
- Klimesch, W. (1999). Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2–3), 169–195.
- Kostas, D., Aroca-Ouellette, S., & Rudzicz, F. (2021). BENDR: Using transformers and a contrastive self-supervised learning task to classify EEG data. *Frontiers in Human Neuroscience*, 15, 653659.
- Lal, S. K. & Craig, A. (2001). A critical review of the psychophysiology of driver fatigue. *Biological Psychology*, 55(3), 173–194.
- Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5), 056013.
- Lee, C.-H., Lee, H.-Y., Jung, M.-K., & Kim, D.-J. (2025). PhysioME: A robust multi-modal self-supervised framework for physiological signals with missing modalities. Preprint.
- Li, J., Fu, C., Tang, J., Zhou, L., Chen, H., Zhou, W., Chen, C., & Luo, J. (2026). Multimodal phenotyping dataset of driving fatigue. *Scientific Data*, 13, 289.
- Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Luo, Y., Chen, Y., Salekin, A., & Rahman, T. (2024). Toward foundation model for multivariate wearable sensing of physiological signals. Preprint.
- Massoz, Q., Langohr, T., François, C., & Verly, J. G. (2016). The ULg multimodality drowsiness database (called DROZY) and examples of use. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (pp. 1–7). IEEE.
- Ministerio de Minería (2022). Política nacional minera 2050. Technical report, Gobierno de Chile. Decreto Supremo N° 2.

- National Safety Council (2017). Fatigue in the workplace: Causes & consequences of employee fatigue. Technical report, National Safety Council.
- Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70, 1373–1411.
- Pan, H., Tong, S., Wei, X., & Teng, B. (2025). Fatigue state recognition system for miners based on a multimodal feature extraction and fusion framework. *IEEE Transactions on Cognitive and Developmental Systems*, 17(2), 593–604.
- Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198, 181–197.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.
- Purto, H., Anabalón, H., Vargas, K., Jara D, C., & de la Vega, R. (2024). Self-perceptual blindness to mental fatigue in mining workers. *Frontiers in Neuroergonomics*, 5, 1441243.
- Roy, R. N., Charbonnier, S., & Bonnet, S. (2014). Eye blink characterization from frontal EEG electrodes using source separation and pattern recognition algorithms. *Biomedical Signal Processing and Control*, 14, 256–264.
- Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5), 1–9.
- Salami, A., Andreu-Perez, J., & Gillmeister, H. (2022). EEG-ITNet: An explainable inception temporal convolutional network for motor imagery classification. *IEEE Access*, 10, 36672–36685.
- Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning

- with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11), 5391–5420.
- Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, (pp. 400–408).
- SERNAGEOMIN (2023a). Anuario de la minería de chile 2022. Technical report, Servicio Nacional de Geología y Minería, Ministerio de Minería.
- SERNAGEOMIN (2023b). Anuario de la minería de chile 2023. Technical report, Servicio Nacional de Geología y Minería, Ministerio de Minería. Cuadro 112: Estadísticas de accidentabilidad, quinquenio 2019–2023.
- SERNAGEOMIN (2025). Estadísticas de accidentabilidad industria extractiva minera año 2024. Technical report, Servicio Nacional de Geología y Minería, Ministerio de Minería.
- Shi, L.-C., Jiao, Y.-Y., & Lu, B.-L. (2013). Differential entropy feature for eeg-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (pp. 6627–6630).
- Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
- Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 710–719.
- Sprajcer, M., Thomas, M. J., Sargent, C., Crowther, M. E., Boivin, D. B., Wong, I. S., Smiley, A., & Dawson, D. (2022). How effective are Fatigue Risk Management Systems (FRMS)? A review. *Accident Analysis & Prevention*, 165, 106398.
- Vallat, R. & Walker, M. P. (2021). An open-source, high-performance tool for automated sleep staging. *eLife*, 10, e70092.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., & Pan, G. (2025). CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*. arXiv:2412.07236.
- Wang, M. & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 135–153.
- Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. *Accident Analysis & Prevention*, 43(2), 498–515.
- Xiong, W., Li, J., Li, J., Zhu, K., & Jiang, C. (2025). EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation of EEG foundation models.
- Xu, L., Li, J., & Feng, D. (2023). Miner fatigue detection from electroencephalogram-based relative power spectral topography using convolutional neural network. *Sensors*, 23(22).
- Yang, C., Westover, M. B., & Sun, J. (2023). BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27.
- Yuan, D., Yue, J., Xiong, X., Jiang, Y., Zan, P., & Li, C. (2023). A regression method for eeg-based cross-dataset fatigue detection. *Frontiers in Physiology, Volume 14 - 2023*.
- Zhao, F., Zhang, C., & Geng, B. (2024). Deep multimodal data fusion. *ACM Computing Surveys*, 56(9), 1–36.

Apéndice A

Resultados completos de Fase 1 (FatigueSet)

Este apéndice contiene las tablas completas de resultados de Fase 1 sobre el dataset FatigueSet, bajo los tres esquemas de etiquetado válidos evaluados (condición de intensidad, VAS física, VAS mental). El capítulo principal (Sección 4.1.5) presenta una versión consolidada que cita estas tablas como respaldo.

A.1. Etiquetado por condición de intensidad

Cuadro A.1: Rendimiento de clasificadores clásicos con etiquetado por condición de intensidad binario (Low vs. Medium+High) sobre FatigueSet. Ningún modelo presenta colapso de predicciones.

Modelo	Macro-F1	Bal. Accuracy	Accuracy
SVM (RBF)	0,620 $\pm$ 0,074	0,642 $\pm$ 0,088	0,661 $\pm$ 0,042
XGBoost	0,589 $\pm$ 0,047	0,591 $\pm$ 0,033	0,700 $\pm$ 0,025
CatBoost	0,580 $\pm$ 0,070	0,585 $\pm$ 0,056	0,682 $\pm$ 0,034
LDA	0,579 $\pm$ 0,121	0,578 $\pm$ 0,111	0,657 $\pm$ 0,123
Random Forest	0,526 $\pm$ 0,087	0,544 $\pm$ 0,065	0,675 $\pm$ 0,048

Cuadro A.2: Rendimiento con etiquetado por condición de intensidad de 3 clases (Low / Medium / High) sobre FatigueSet. Clasificación multiclase con validación 5-*fold subject-aware* (StratifiedGroupKFold).

Modelo	Macro-F1	Bal. Accuracy	Accuracy
SVM (RBF)	0,419 $\pm$ 0,065	0,437 $\pm$ 0,069	0,447 $\pm$ 0,066
Random Forest	0,439 $\pm$ 0,078	0,455 $\pm$ 0,075	0,472 $\pm$ 0,072
XGBoost	0,453 $\pm$ 0,080	0,462 $\pm$ 0,077	0,475 $\pm$ 0,079
CatBoost	0,463 $\pm$ 0,082	0,475 $\pm$ 0,079	0,488 $\pm$ 0,080
LDA	0,406 $\pm$ 0,140	0,409 $\pm$ 0,139	0,418 $\pm$ 0,138

Cuadro A.3: Benchmark de arquitecturas de *Deep Learning* con etiquetado por condición de intensidad binario sobre FatigueSet. Los modelos marcados con $\dagger$ presentan colapso de predicciones en al menos un *fold*.

Categ.	Modelo	Macro-F1	Bal. Acc	Acc	Col.
	ShallowFBCSPNet	0,506 $\pm$ 0,073	0,540 $\pm$ 0,089	0,525 $\pm$ 0,074	0/5
	Deep4Net $\dagger$	0,500 $\pm$ 0,028	0,505 $\pm$ 0,023	0,566 $\pm$ 0,061	1/5
	EEGNet	0,494 $\pm$ 0,033	0,514 $\pm$ 0,048	0,557 $\pm$ 0,075	0/5
	ATCNet	0,544 $\pm$ 0,066	0,566 $\pm$ 0,060	0,584 $\pm$ 0,077	0/5
	EEGTCNet	0,478 $\pm$ 0,080	0,511 $\pm$ 0,056	0,591 $\pm$ 0,092	0/5
	EEGITNet	0,456 $\pm$ 0,050	0,506 $\pm$ 0,061	0,517 $\pm$ 0,076	0/5
	EEGConformer $\dagger$	0,436 $\pm$ 0,041	0,504 $\pm$ 0,015	0,673 $\pm$ 0,041	3/5
Fusion	LaBraM $\dagger$	0,470 $\pm$ 0,058	0,513 $\pm$ 0,023	0,653 $\pm$ 0,050	2/5
	BENDR $\dagger$	0,240 $\pm$ 0,011	0,500 $\pm$ 0,000	0,316 $\pm$ 0,020	5/5

A.2. Etiquetado por VAS (M0/M1/M2)

Cuadro A.4: Clasificadores clásicos con etiquetado VAS sobre FatigueSet. Macro-F1 (media $\pm$ desviación estándar) en 5-fold *subject-aware* (StratifiedGroupKFold).

Modelo	VAS mental ($\tau = 10$)	VAS física ($\tau = 5$)
SVM (RBF)	0,471 $\pm$ 0,028	0,573 $\pm$ 0,049
LDA	0,466 $\pm$ 0,056	0,563 $\pm$ 0,076
CatBoost	0,483 $\pm$ 0,074	0,558 $\pm$ 0,060
Random Forest	0,480 $\pm$ 0,057	0,551 $\pm$ 0,078
XGBoost	0,451 $\pm$ 0,052	0,550 $\pm$ 0,057

Cuadro A.5: Subconjunto representativo de arquitecturas DL con etiquetado VAS sobre FatigueSet. Las cinco arquitecturas restantes (EEGTCNet, EEGITNet, EEGConformer, BENDR, LaBraM) no se replicaron en VAS dado que su comportamiento bajo etiquetas balanceadas ya está caracterizado en la Tabla A.3.

Modelo	VAS mental ($\tau = 10$)	VAS física ($\tau = 5$)
EEGNet	0,384 $\pm$ 0,130	0,457 $\pm$ 0,097
Deep4Net	0,420 $\pm$ 0,108	0,484 $\pm$ 0,093
ShallowFBCSPNet	0,465 $\pm$ 0,082	0,454 $\pm$ 0,052
ATCNet	0,461 $\pm$ 0,034	0,527 $\pm$ 0,023

A.3. Ranking consolidado de modelos (intensidad binario)

Cuadro A.6: Ranking consolidado de todos los modelos bajo etiquetado por condición de intensidad binario en FatigueSet. Se reportan solo los modelos sin colapso completo de predicciones.

Rank	Modelo	Macro-F1	Bal. Accuracy	Tipo
1	SVM (RBF)	$0,620 \pm 0,074$	$0,642 \pm 0,088$	ML clásico
2	XGBoost	$0,589 \pm 0,047$	$0,591 \pm 0,033$	ML clásico
3	CatBoost	$0,580 \pm 0,070$	$0,585 \pm 0,056$	ML clásico
4	LDA	$0,579 \pm 0,121$	$0,578 \pm 0,111$	ML clásico
5	ATCNet	$0,544 \pm 0,066$	$0,566 \pm 0,060$	DL (híbrido)
6	Random Forest	$0,526 \pm 0,087$	$0,544 \pm 0,065$	ML clásico
7	ShallowFBCSPNet	$0,506 \pm 0,073$	$0,540 \pm 0,089$	DL (CNN)
8	Deep4Net	$0,500 \pm 0,028$	$0,505 \pm 0,023$	DL (CNN)
9	EEGNet	$0,494 \pm 0,033$	$0,514 \pm 0,048$	DL (CNN)
10	EEGTCNet	$0,478 \pm 0,080$	$0,511 \pm 0,056$	DL (híbrido)
11	LaBraM	$0,470 \pm 0,058$	$0,513 \pm 0,023$	DL (fundacional)
12	EEGITNet	$0,456 \pm 0,050$	$0,506 \pm 0,061$	DL (híbrido)
13	EEGConformer	$0,436 \pm 0,041$	$0,504 \pm 0,015$	DL (híbrido)

Apéndice B

Material suplementario de la Fase 2

Este apéndice contiene material descriptivo y ablaciones detalladas de la Fase 2 (cohorte Sigma) que respaldan los resultados centrales reportados en el Capítulo 4 pero que se trasladan aquí para no fragmentar la narrativa principal.

B.1. Cobertura horaria y estructura temporal de los bloques

YASA procesó exitosamente 1109 de las 1163 sesiones de bloques cortos con EEG (95,4 %); las 54 sesiones restantes fueron descartadas por longitud insuficiente para una época AASM de 30 s o por fallos de carga del XDF. La duración total de EEG procesado es 123,9 horas, distribuidas por bloque y por categoría de calidad de la bitácora como muestra la Tabla B.1.

Cuadro B.1: Horas de EEG procesado (OA y OC *resting* + REA bajo carga cognitiva) por categoría de calidad según la bitácora. *Usable* = Buena o Excelente; *Sin bitácora* corresponde principalmente al Ciclo 6. Los subtotales se redondean a 0,1 h, por lo que su suma puede diferir del total en $\pm 0,1$ h.

Categoría	Sesiones	Horas	% horas	Mediana / sesión
Usable (Buena / Excelente)	839	89,7 h	72 %	5,5 min
Marginal (Regular)	108	11,6 h	9 %	5,0 min
Descartable (Mala)	103	11,3 h	9 %	5,5 min
Sin bitácora estructurada (C6)	59	11,4 h	9 %	8,0 min
Total	1109	123,9 h	100 %	n/a

La categoría *Usable* concentra 89,7 horas sobre 839 sesiones y constituye el universo de entrenamiento del modelo final (este conteo del barrido de horas difiere en pocas unidades del *snapshot* de modelado congelado, $849 = 849$ sesiones de la Tabla 3.1, por la validez de *label* YASA exigida y la regeneración del *parquet* de *features*; la diferencia es inmaterial). Las categorías *Marginal* y *Descartable* se mantienen para análisis de sensibilidad pero no entran al *pipeline* principal. Por bloque, las horas *Usable* se distribuyen como: OA 21,5 h, OC 21,3 h, REA 46,9 h, reflejando que REA es típicamente el doble de largo que OA u OC.

Los archivos XDF de OA, OC y REA se graban como archivos independientes, pero el sujeto permanece con el casco puesto entre uno y otro; el operador solo detiene y reinicia la grabación con pausas típicas de 10 a 60 segundos. Esto implica que las 1109 sesiones individuales corresponden a aproximadamente 370 “sesiones puestas con el casco” continuas de 19 a 28 minutos cada una. La duración mediana por bloque es 5,0 minutos para OA y OC, y 10,0 minutos para REA, con colas largas atribuibles a sesiones donde el operador olvidó detener el registro (Tabla B.2).

Cuadro B.2: Distribución de duraciones por bloque corto con EEG ($n = 1109$ sesiones procesadas por YASA).

Bloque	n	Mediana	IQR	Máximo observado
OA	357	5,0 min	4,5 – 5,0	11,0 min
OC	351	5,0 min	4,5 – 5,0	8,5 min
REA	401	10,0 min	10,0 – 10,5	43,5 min

El orden de los bloques no es fijo y varía entre sujetos (a veces REA $\rightarrow$ OA $\rightarrow$ OC, otras OA $\rightarrow$ OC $\rightarrow$ REA). El bloque “REA” (archivo XDF `*REA.xdf`) corresponde a EEG grabado mientras el sujeto realiza el módulo REA1 del software Rehacom (Capacidad de respuesta), según se confirma por las ocho categorías Rehacom documentadas en el Diccionario Sigma (REA1, REVE, AUFM, CALC, VRO1, RO3D, PLAN, LODE) y por los nombres de los informes médicos en el directorio de informes Rehacom. Por consiguiente, OA y OC capturan estados *resting* puro, mientras que REA captura EEG bajo carga cognitiva del test de tiempo de reacción.

B.2. Ablaciones detalladas del experimento de anticipación EMOTI

El experimento de anticipación EMOTI **no supera el azar**: el Macro-F1 de la configuración titular (CatBoost con *features* agregadas simples) es $0,460 \pm 0,044$ con LOSO sobre 154 pares nominales (141 efectivos en 9 sujetos), indistinguible del nivel de azar de 0,500 (Sección 4.4.15). Para confirmar que el resultado nulo no depende de la representación ni del modelo, se realizaron dos ablaciones complementarias.

Ablación 1: *features* fisiológicas más ricas. Se construyó un segundo conjunto de 28 *features* incorporando HRV *proper* desde el *stream* HR del EmotiBit (`hrv_sdn`, `hrv_rmssd`, `hrv_pnn50`, `hrv_lf`, `hrv_hf`, `hrv_lfhf`) y dinámica IMU enriquecida (*jerk*, *zero-crossings*, entropía espectral, descomposición de banda en tres rangos). Adicionalmente, se evaluó la normalización *within-subject z-score* antes del modelado, técnica estándar en la literatura de *wearables* (Schmidt, Reiss, Duerichen, Marberger &

Van Laerhoven, 2018) para controlar la heterogeneidad inter-sujeto. La descomposición EDA *tonic/phasic* mediante *cvxEDA* (Greco, Valenza, Lanata, Scilingo & Citi, 2016) se intentó pero falló en convergencia para todos los bloques (probablemente por la longitud sostenida > 90 minutos del registro continuo, donde *cvxEDA* no escala bien sin segmentación en ventanas).

Ablación 2: modelos secuenciales. Se construyeron matrices temporales por bloque agregando estadísticos por minuto sobre cada *stream*, generando tensores de forma (100 minutos, 12 canales) por par. Sobre esto se entrenaron dos arquitecturas de *deep learning* pequeñas: (a) un CNN 1D de dos capas convolucionales con *global average pooling* ($\sim 25K$ parámetros), y (b) un CNN minimalista (una capa convolucional, *global pooling*, ~ 600 parámetros) con *data augmentation* (*random time crop*, *channel dropout*) y *ensemble* de 5 semillas para reducir varianza.

La Tabla B.3 compara las cuatro configuraciones sobre la única formulación con señal positiva (clasificación binaria Δ Rehacom).

Cuadro B.3: Ablación de modelo y *features* sobre Δ Rehacom binario. Ninguna configuración supera el azar (0,500): las *features* agregadas simples (24 estadísticos) dan la menor varianza pero quedan al nivel del azar, y agregar HRV *proper* e IMU enriquecido, normalización *within-subject* o modelos secuenciales no aporta señal.

Configuración	Features	Macro-F1	Comentario
CatBoost (agregadas)	24 stats agregadas	$0,460 \pm 0,044$	Menor varianza; al nivel del azar
CatBoost (HRV+IMU rico)	28 con HRV + IMU rico	$0,500 \pm 0,132$	Al azar, alta varianza
CatBoost (<i>z-score</i>)	28, normalización <i>within-subject</i>	$0,495 \pm 0,096$	Al azar
CNN 1D (25K params)	12 canales $\times$ 100 min	$0,379 \pm 0,113$	Bajo el azar (sobreajuste, $N=154$)
TinyCNN + <i>ensemble</i>	12 canales, 5 <i>seeds</i> con <i>augmentation</i>	$0,367 \pm 0,139$	Aún peor
Baseline azar	n/a	0,500	n/a

Las dos ablaciones convergen en la misma conclusión: **ninguna representación ni modelo supera el azar**. Las *features* más ricas (HRV *proper*, IMU rico) no aportan; los modelos secuenciales con capacidad de capturar dinámica temporal explícita rinden por debajo del clasificador clásico sobre *features* agregadas. Esto es la firma característica de un régimen donde el **cuello de botella** es el **tamaño muestral**, no la representación: con 154 pares nominales (89 EMOTI_{AM} + 65 EMOTI_{PM}) que el filtrado de calidad reduce a 141 efectivos, distribuidos en 9 sujetos para el análisis

EMOTI y validación LOSO (aproximadamente 16 muestras de *test* por *fold*), todo el espacio de modelos converge al mismo nivel de azar. Esta lectura es consistente con la ablación de preprocesamiento EEG del *pipeline* principal (Sección 4.4.5), donde agregar ICA con clasificación automática de componentes *degradó* el F1 LOSO del label primario en 0,085 puntos: el techo también es robusto a cambios en el preprocesamiento del dominio EEG.