UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

UN MODELO SECUENCIAL PROFUNDO PARA DETECTAR EVENTOS EN EL ELECTROENCEFALOGRAMA DEL SUEÑO

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

NICOLÁS IGOR TAPIA RIVAS

PROFESOR GUÍA:
PABLO ESTÉVEZ VALENCIA

MIEMBROS DE LA COMISIÓN:
CLAUDIO PÉREZ FLORES
RODRIGO SALAS FUENTES
JOSÉ CORTÉS BRIONES

SANTIAGO DE CHILE
2022

RESUMEN DE LA TESIS PARA OPTAR AL
GRADO DE MAGÍSTER EN CIENCIAS DE LA
INGENIERÍA, MENCIÓN ELÉCTRICA
Y AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: NICOLÁS IGOR TAPIA RIVAS
FECHA: 2022
PROF. GUÍA: PABLO ESTÉVEZ VALENCIA

UN MODELO SECUENCIAL PROFUNDO PARA DETECTAR EVENTOS EN EL ELECTROENCEFALOGRAMA DEL SUEÑO

Los husos de sueño y los complejos K son eventos transitorios del electroencefalograma del sueño. Han sido asociados con diversas funciones cognitivas y podrían diagnosticar enfermedades. Sin embargo, la detección manual limita su estudio porque es lenta y varía significativamente entre expertos, motivando enfoques automáticos. Los mejores detectores están basados en aprendizaje profundo, pero ninguno procesa secuencialmente el contexto ni se ha validado extensamente. Ambas brechas se abordan al proponer y validar el detector *Recurrent Event Detector v2* (REDv2), considerando dos variantes según si su entrada es la señal en el tiempo o en un espacio tiempo-frecuencia. La evaluación en varias bases de datos anotadas muestra que REDv2 alcanza el estado del arte. Además, extensos experimentos en datos anotados, artificiales, y no anotados, muestran que REDv2 aprende a usar características conocidas, generaliza bien, requiere de poco ajuste para adaptarse a otros expertos, y genera detecciones que reproducen tendencias demográficas conocidas. El procesamiento secuencial mejora el desempeño, y REDv2 demuestra ser un detector confiable. La variante sobre un espacio tiempo-frecuencia no entrega mejor desempeño, aunque podría mejorar la interpretabilidad de mapas de relevancia en la entrada. El diseño se podría reutilizar para predecir otros tipos de eventos transitorios en el EEG.

A mis padres

Agradecimientos

Agradezco al Dr. Pablo Estévez, mi profesor guía, por haber creído en mí desde el principio al invitarme al Laboratorio de Inteligencia Computacional y entregarme gran autonomía para definir mi trabajo. Adicionalmente, le agradezco por los excelentes computadores que me facilitó a través del laboratorio. Sin ellos aún estaría esperando los resultados de mis experimentos. Quiero agradecer también a todos en el laboratorio por su acogida y por la oportunidad de reírnos juntos de las dificultades. Además, gran parte de lo que aprendí sobre la teoría y práctica de *deep learning* fue gracias a ustedes. Agradezco la ayuda del INTA, en particular del Dr. Marcelo Garrido, por su gran voluntad para acompañarnos en varias ocasiones a revisar los datos. Agradezco al Dr. José Cortés por corregir el borrador y por mostrarse siempre tan interesado y entusiasmado con lo que hacíamos. También quiero agradecer al Dr. Claudio Pérez y al Dr. Rodrigo Salas por aceptar ser parte de mi comisión y por corregir el borrador. Por último, agradezco a la Agencia Nacional de Investigación y Desarrollo (ANID) por financiar parcialmente este trabajo a través de la beca Magíster Nacional/2019 - 22191803 y a través del proyecto FONDECYT - 1171678.

En una nota más personal, quiero partir agradeciendo profundamente a mis padres por darme un hogar, por preocuparse de que no me faltara nada, y por permitirme concentrarme en mis estudios. Sé que muchas veces no entendían por lo que estaba pasando en mi carrera, pero me tuvieron paciencia y confianza. Gracias por estar ahí y regalarme un refugio que fue invaluable en los momentos más estresantes. Sin su apoyo, no sería ni el profesional ni la persona que soy ahora. Espero ser siempre motivo de su orgullo para devolverles aunque sea una parte de lo que han dado por mí.

Sofi, te agradezco por ser una gran compañera de vida. Hemos compartido gran parte de mi carrera y todo el magíster, y siempre fuiste cariñosa conmigo, me apoyaste incondicionalmente, y me animaste en mis dificultades. Gracias por tu comprensión cuando me faltaba tiempo, por tantas onces improvisadas cuando se hacía larga la jornada, y por tantos momentos felices. Gracias por ser la hermosa persona que eres y por inspirarme a ser una mejor persona cada día. Muchas gracias a mis maravillosos amigos que han sido una constante en mi vida y han sido clave en todo este proceso: Nacho, Esteban, Maxi, Coque. Ustedes hicieron de estos años universitarios unos años memorables, llenos de historias. Las cátedras y las notas se perderán en el aire, pero nuestras vivencias siempre estarán conmigo y me sacarán una sonrisa. Para terminar, no puedo dejar de agradecer al Germán por ser un gran amigo, y un gran aliado junto al Esteban durante el magíster. Gracias a ustedes dos por hacerme creer que el magíster era algo divertido.

Tabla de Contenido

Capítulo 1

Introducción

1.1. Motivación

Para diagnosticar un sujeto, o validar una hipótesis en una muestra de sujetos, es común que los expertos en medicina primero deban determinar manualmente la existencia o ausencia de patrones específicos sobre diversas señales biológicas. Este primer paso puede llegar a ser muy repetitivo y demandante en tiempo. Su costo puede impactar directamente en los resultados si se sacrifica calidad por rapidez, por ejemplo al contratar una o más personas con poca experiencia para completar la tarea. Incluso cuando la tarea es realizada por un experto, es común que se introduzcan inconsistencias debido a la variabilidad que existe entre diversos expertos o dentro de un mismo experto en diferentes instantes de tiempo. Un algoritmo automático de detección ofrece una alternativa rápida y consistente, haciendo que la diferencia entre preparar los datos de un sujeto o mil sujetos sea insignificante. Además, consigue que los expertos enfoquen sus esfuerzos en lo que hacen mejor: analizar e investigar dichos patrones.

En esta tesis, se investiga el problema de detección y localización precisa de **husos de sueño** y **complejos K**, patrones transitorios cortos de la actividad eléctrica cerebral que aparecen durante el sueño. La investigación del sueño cumple un rol central en el estudio del funcionamiento y desarrollo saludable del cerebro. Para investigarlo, se registra un conjunto de señales fisiológicas llamado polisomnograma (PSG), entre las que destaca el electroencefalograma (EEG), es decir, la medición de la actividad eléctrica cerebral. Durante el sueño, el cuerpo transita entre cinco etapas llamadas W, R, N1, N2 y N3 [1], en donde W, R, y N representan vigilia, sueño REM, y sueño no-REM (NREM), respectivamente. En promedio, la mitad del sueño ocurre en etapa N2, caracterizada por la aparición de husos de sueño y complejos K en el EEG. Ambos patrones están involucrados en diversas funciones cognitivas que incluyen la memoria, el aprendizaje, y el procesamiento de estímulos [2, 3]. Además, sus características son potenciales indicadores de enfermedades como la esquizofrenia [4] y la epilepsia [5].

La detección de estos eventos del EEG significa anotar instantes de inicio y fin en el registro de interés, y es un paso necesario para avanzar en su investigación. Por ejemplo, su anotación permite cuantificar su densidad o estadísticos de diversas características morfológicas de la señal, indicadores utilizados para profundizar en la investigación de su relación con procesos cerebrales o

enfermedades. La detección manual experta, que es el actual estándar de calidad, se desea automatizar por dos razones principales. La detección manual demanda mucho tiempo porque los expertos deben inspeccionar el PSG visualmente en segmentos de 20 s o 30 s, llamados épocas. Además, se ha reportado una significativa variabilidad entre las anotaciones de distintos expertos debido a definiciones oficiales subjetivas que dependen últimamente del criterio experto individual [6].

Los detectores automáticos tradicionales de la literatura se basan en la extracción de características de la señal diseñadas a mano, cuya principal ventaja es que son interpretables. Sin embargo, justamente debido a que la definición de los husos de sueño y complejos K es ambigua, es muy difícil determinar manualmente lo que debería hacer una máquina para emular la detección. En efecto, se ha reportado que los detectores tradicionales incurren en muchos falsos positivos [6, 3, 7]. Recientemente, métodos basados en aprendizaje profundo (*deep learning* en inglés), cuyas características son aprendidas simultáneamente con el detector, se han aplicado a la detección de husos de sueño y complejos K alcanzando el estado del arte [8, 9, 7, 10].

Los métodos de aprendizaje profundo aplicados a este problema están principalmente basados en redes neuronales convolucionales: ninguno de estos métodos procesa secuencialmente un contexto grande usando capas neuronales diseñadas específicamente para eso. Las arquitecturas convolucionales han tenido éxito en el procesamiento de imágenes, e incluso de espectrogramas de señales de voz. A pesar de ello, no es claro si son las más adecuadas para detectar con exactitud los patrones transitorios cortos en el EEG del sueño. Por ejemplo, un huso de sueño podría consistir solo de cuatro oscilaciones, haciendo desafiante no solo determinar su existencia sino que también su inicio y fin. La ausencia de procesamiento secuencial podría limitar la capacidad de explotar eficientemente el contexto temporal, una fuente central de información para los expertos. En efecto, los husos de sueño y los complejos K son descritos como patrones sobresalientes en el EEG [1], implicando un concepto de contexto, y la estructura temporal a largo plazo de la señal puede proveer características útiles, tales como actividad circundante compatible y artefactos en la señal. Por último, el comportamiento y la generalización de los métodos de aprendizaje profundo propuestos en la literatura no han sido extensamente validados para que los expertos entiendan los límites de su funcionamiento y los adopten a pesar de sus características no interpretables.

En esta tesis, se desarrolla un detector llamado *Recurrent Event Detector v2* (REDv2), basado en capas convolucionales para extraer características locales de la señal y en capas recurrentes para modelar secuencialmente la señal a largo plazo. Su diseño permite atenuar las limitaciones de otros detectores de la literatura, y los experimentos realizados ofrecen una extensa validación del detector tanto cuantitativamente (desempeño y generalización) como cualitativamente (caracterización de los eventos detectados). Además, se evalúan dos variantes de la estructura base de REDv2 con diferentes representaciones de entrada, según si se usa la señal EEG directamente en el tiempo (REDv2-Time) o en un espacio tiempo-frecuencia gracias a la Transformación de Wavelet Continua (CWT) [11] (REDv2-CWT). Resultados preliminares de esta tesis, usando un modelo preliminar llamado *Recurrent Event Detector* (RED), fueron publicados y presentados en la *2020 International Joint Conference on Neural Networks* (IJCNN) [12].

1.2. Hipótesis

Las hipótesis de esta tesis son:

H1 El desempeño de un detector de husos de sueño y complejos K mejoraría al incluir en su diseño una etapa de procesamiento secuencial del contexto a largo plazo, i.e., de una escala mayor que la duración típica del evento a detectar.

H2 El desempeño del detector mejoraría al usar como entrada la señal transformada a un espacio de tiempo-frecuencia, con respecto al caso de usar directamente la señal en el tiempo.

H3 Un detector basado en un modelo secuencial de aprendizaje profundo, luego de ser entrenado, serviría para detectar husos de sueño y complejos K con buen desempeño en bases de datos externas (no usadas para el entrenamiento).

1.3. Objetivo general

El objetivo general de la tesis es proponer, desarrollar y validar un método de detección de eventos transitorios cortos en el EEG del sueño, como husos de sueño y complejos K, basado en aprendizaje profundo y que considere un modelado secuencial de las señales a largo plazo.

1.4. Objetivos específicos

Los objetivos específicos de esta tesis son:

O1 Desarrollar e implementar una red neuronal profunda con modelado secuencial a largo plazo que resuelva el problema de detección de husos de sueño y complejos K.

O2 Comparar el modelo propuesto con el estado del arte en detección de husos de sueño y complejos K usando al menos dos bases de datos.

O3 Evaluar el comportamiento del modelo propuesto ante señales perturbadas, artificiales, y provenientes de otras bases de datos.

O4 Validar rangos estadísticos y tendencias demográficas de los husos de sueño detectados por el modelo propuesto con respecto a aquellos reportados en la literatura, usando un gran conjunto de datos sin anotaciones expertas.

1.5. Alcances

Para acotar la complejidad de esta tesis, se evaluarán eventos solo durante la etapa N2 del sueño y usando como entrada un solo canal de EEG.

1.6. Organización del documento

El resto del documento se organiza como sigue. En el Capítulo 2 se describen los antecedentes necesarios para comprender en profundidad el contexto del trabajo. Se introduce al lector el dominio en donde ocurren los patrones a detectar: el electroencefalograma (EEG) del sueño. Se describen sus estándares de medición y anotación, y sus fuentes de variación conocidas. Se sigue con una descripción teórica de las representaciones de tiempo-frecuencia que se pueden obtener de una señal, con énfasis en el método utilizado en REDv2-CWT. Después, se revisan superficialmente varios conceptos y técnicas del aprendizaje profundo que permiten comprender el método de

detección desarrollado a aquellos lectores que no estén familiarizados con el área. El capítulo cierra con una revisión de los trabajos en la literatura que son relevantes para el problema de detección de husos de sueño y complejos K. Con los antecedentes necesarios ya establecidos, en el Capítulo 3 se plantea el problema concreto que se quiere resolver, se describen los métodos y materiales usados para evaluar el desempeño, y se presenta el método de detección propuesto basado en un modelo secuencial profundo. Además, se propone un método de aumento de datos para entrenar el modelo. El capítulo cierra con varios detalles experimentales en preparación al capítulo siguiente. El Capítulo 4 comienza evaluando la efectividad del método de aumento de datos propuesto, para después profundizar en la evaluación del desempeño del detector propuesto, con sus dos variantes, y las características de sus detecciones ante diversos escenarios. En el Capítulo 5 se discuten las ventajas y limitaciones del detector propuesto, su generalización, su interpretabilidad, y sus principales implicaciones. Al final del documento, se concluye verificando el cumplimiento de los objetivos planteados, resumiendo las principales contribuciones del trabajo, y entregando recomendaciones de investigación futura.

Capítulo 2

Antecedentes

2.1. El electroencefalograma (EEG) del sueño

2.1.1. Actividad eléctrica del cerebro

En el cerebro existe una extensa red de células especializadas en el procesamiento, transmisión y almacenamiento de la información llamadas *neuronas*. Una neurona se comunica con otra a través de intercambios químicos (o, en algunos casos, eléctricos [13]) para estimular o inhibir la transmisión de pulsos de voltaje (*disparos*) a lo largo de la membrana celular de la neurona receptora, que a su vez puede estimular o inhibir la respuesta de otras células. El patrón de disparo de una neurona, dependiente de los patrones de sus entradas, es la unidad básica para la codificación neuronal de la información en el sistema nervioso [14].

En ensambles neuronales, gracias a circuitos neuronales retroalimentados, dichos patrones pueden entrar en sincronía, implicando el disparo aproximadamente simultáneo de un gran grupo de neuronas de forma periódica. En dicho escenario, la suma de los campos eléctricos generados por cada neurona da lugar a una oscilación neuronal de gran escala que oscila a una frecuencia específica, cuya amplitud depende del grado de sincronía del ensamble [14]. Se cree que las oscilaciones neuronales cumplen un rol central en el funcionamiento del cerebro, por lo que caracterizarlas y comprenderlas es un área activa de investigación en neurociencia [15].

La actividad eléctrica que es conducida al cuero cabelludo proviene principalmente de oscilaciones neuronales de gran escala de la corteza cerebral, con intensidades típicamente en la escala de 10–100 μV. Además, varias oscilaciones neuronales pueden coexistir simultáneamente, enriqueciendo la composición espectral de la señal eléctrica. En promedio, el espectro presenta potencias que decaen aproximadamente como f^{-a} con $a \in [1, 2]$ [15]. En él, se han diferenciado bandas de frecuencia que han sido consistentemente asociadas a funciones específicas, llamadas delta (0–4 Hz), theta (4–8 Hz), alfa (8–13 Hz), beta (13–30 Hz) y gamma (30–100 Hz) [15].

2.1.2. Registro de la actividad cerebral mediante EEG

La actividad eléctrica del cerebro puede ser medida por un arreglo de electrodos adheridos a la superficie del cuero cabelludo para generar un EEG. El EEG se compone de varios *canales* que

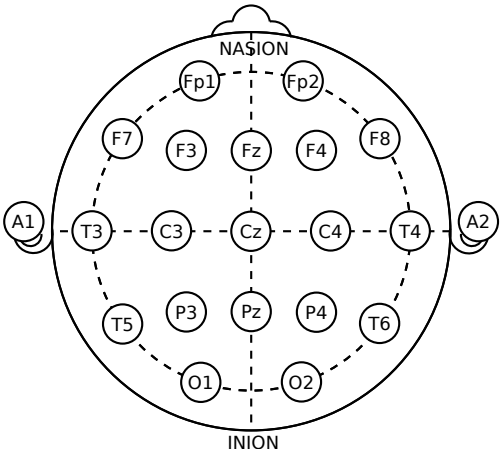

Figura 2.1: Sistema internacional 10-20 para los electrodos del EEG. El nasión se encuentra en el puente nasal. Imagen de Dominio Público.

representan la diferencia de potencial entre dos electrodos. La elección del par de electrodos se puede hacer de dos formas principales. La primera opción es utilizar una derivación unipolar, en donde todos los electrodos usan el mismo electrodo de referencia, usualmente ubicado detrás de las orejas (sobre el hueso mastoideo) o en la línea media del cráneo. La segunda opción es utilizar una derivación bipolar, en donde cada canal corresponde a la diferencia entre dos electrodos adyacentes. En general, se registra el EEG con una derivación unipolar ya que otros esquemas de referencia pueden ser fácilmente obtenidos a partir de ella [16].

La ubicación y nomenclatura de cada electrodo del EEG están estandarizadas para asegurar reproducibilidad. El sistema internacional 10-20 [17] está ampliamente adoptado y es la base para otros sistemas con arreglos de electrodos más densos (ver Figura 2.1).[1] El sistema etiqueta cada ubicación con una letra y un número. La letra representa la región y puede ser Fp (pre-frontal), F (frontal), T (temporal), P (parietal), O (occipital) y C (central). El número que complementa la etiqueta es par para el hemisferio derecho e impar para el hemisferio izquierdo. Por ejemplo, el canal C3 indica la señal medida por un electrodo ubicado en la región central izquierda. Existen también electrodos ubicados en la línea media del cráneo que son complementados por la letra Z en lugar de un número (e.g., el canal Fz). Por último, existen dos electrodos ubicados detrás de las orejas, que se pueden utilizar como referencia y que se etiquetan con la letra A o M. Ambos electrodos podrían estar físicamente conectados o bien podrían ser posteriormente promediados para formar una referencia de orejas conectadas (*linked ear (LE) reference* en inglés).

El proceso de adquisición de la señal consiste en varias etapas. En primer lugar, cada electrodo debe establecer una conexión eléctrica con el cuero cabelludo, ya sea con la ayuda de un gel conductor o en contacto directo. Luego, la señal de voltaje de cada electrodo es capturada por un circuito analógico que la amplifica y la filtra. Finalmente, un conversor analógico-digital digitaliza la señal a una frecuencia de muestreo dada. Las frecuencias de corte del filtro se fijan para remover bandas que contienen principalmente ruido y para remover aquellas componentes que superen la frecuencia de Nyquist del conversor y así evitar *aliasing* [16].

[1] Los arreglos que siguen el sistema 10-20 se componen usualmente de menos de 20 canales. Actualmente, es común realizar estudios con 64 canales. Los sistemas más densos alcanzan 256 canales.

Tabla 2.1: Artefactos comunes en el EEG.

Tipo	Características
Artefactos cardíacos	Se distinguen generalmente por su sincronía con el ritmo cardíaco, que puede monitorearse con un electrocardiograma (ECG).
Artefactos musculares	Se caracterizan por un espectro relativamente plano de alta frecuencia, y se originan por la tensión de músculos cercanos al electrodo, que puede monitorearse con un electromiograma (EMG).
Artefactos oculares	Se originan por el movimiento de los globos oculares, al ser éstos dipolos eléctricos, que puede monitorearse con un electrooculograma (EOG).
Artefactos de adquisición	Se originan por un funcionamiento incorrecto del equipo de adquisición. Por ejemplo, debido a un desplazamiento o desconexión momentánea de un electrodo producto de un movimiento brusco, o un corto-circuito entre dos electrodos (e.g., el gel conductor de ambos hacen contacto).
Artefactos ambientales	Se originan por señales externas tales como la corriente alterna de la línea eléctrica, que inyecta una oscilación de 50 Hz o 60 Hz dependiendo del lugar, o ruido proveniente de dispositivos cercanos.

2.1.3. Artefactos en el EEG

En la práctica, el EEG captura varias fuentes eléctricas que no corresponden a actividad cerebral. Las contaminaciones intensas provocan la aparición de patrones no neuronales en el EEG conocidos como *artefactos*. Los artefactos más comunes son cardíacos, musculares, oculares, de adquisición, y ambientales, descritos en la Tabla 2.1 [18].

Durante el sueño, el sujeto se encuentra mayoritariamente inmóvil y sin pestañear, y las frecuencias de interés son menores a 35 Hz. Por lo tanto, la presencia de artefactos es significativamente menor, aunque siguen existiendo. Por ejemplo, una persona podría tensar partes de su rostro, moverse bruscamente en la cama, o encontrarse en momentos del sueño caracterizados por movimientos oculares rápidos.

2.1.4. Estándares para el estudio del sueño

El sueño se estudia con la ayuda de un conjunto de señales fisiológicas, llamado polisomnograma (PSG), que se registra durante todo el tiempo en que el sujeto duerme, usualmente una noche completa. Entre las señales más importantes se encuentran el EEG, el EMG, y el EOG. El registro adquirido puede ser utilizado por expertos en sueño para detectar patrones y analizar la estructura general del sueño. Por ejemplo, pueden estudiar los cambios inducidos en la estructura del sueño a gran y pequeña escala por la realización de una actividad, el consumo de una sustancia, o el diagnóstico de una enfermedad.

Los expertos utilizan criterios estandarizados en combinación con su experiencia y su conocimiento del paciente para medir y evaluar el PSG. Actualmente, se utiliza el manual de la American Academy of Sleep Medicine (AASM) [1], publicado por primera vez en el año 2007 y actualizado periódicamente. Este manual integra y actualiza el antiguo manual para etapas del sueño de Rechtschaffen & Kales (R&K) [19]. Para evaluar la estructura del sueño, la AASM recomienda

que el experto utilice las siguientes señales: el EOG medido en el ojo izquierdo y derecho; el EEG medido en un canal occipital, central y frontal; y el EMG medido debajo del mentón.

Para el EEG, la AASM recomienda utilizar canales del hemisferio derecho, que pueden ser respaldados por sus equivalentes en el hemisferio izquierdo cuando los canales principales presentan problemas. Las derivaciones recomendadas son O2-M1, C4-M1 y F4-M1, que pueden ser respaldadas por las derivaciones O1-M2, C3-M2 y F3-M2. En caso de requerir una alternativa a la recomendación principal, se acepta usar Cz-Oz, C4-M1 y Fz-Cz. Además, se recomienda utilizar una frecuencia de muestreo de al menos 200 Hz (idealmente 500 Hz), aplicar un filtro pasa-banda con frecuencias de corte 0,3 Hz y 35 Hz, y utilizar electrodos con una impedancia no mayor a 5 kΩ.

El experto, durante su análisis, visualiza la señal en ventanas de unos cuantos segundos a la vez, llamadas *épocas* o *páginas* (aludiendo a su registro analógico en papel). El manual de la AASM recomienda épocas de 30 s, con una relación de aspecto para su visualización de 10 mm cada 1 s y 7,5–10 mm cada 50 μV. El antiguo manual de R&K [19] recomienda épocas de 20 s, por lo que se pueden encontrar PSG anotados con cualquiera de estas dos segmentaciones dependiendo de su fecha de origen.

2.1.5. Etapas del sueño

Durante el sueño, una persona transita a través de diferentes estados cerebrales que han sido agrupados en etapas del sueño. La AASM reconoce cinco etapas que abarcan la etapa de vigilia, el sueño de movimiento oculares rápidos o REM (*rapid eye movement* en inglés), y el sueño no REM o NREM, y define los patrones que deben ser observados en el EEG, EOG y EMG para anotar la ocurrencia de cada una [1]. En la Tabla 2.2 se describe la composición característica de cada etapa según la AASM, su equivalente en el antiguo manual de R&K [19], y la fracción de la noche que típicamente ocupan [20].

El experto asigna una de estas cinco etapas a una época o página del PSG según los criterios indicados en el manual de la AASM, su experiencia, y su conocimiento del paciente (e.g., la forma de sus patrones típicos). Si no es posible determinar la etapa debido a problemas en la señal, el experto puede anotar la época con una etiqueta que indique esta situación (e.g., «?»). Como resultado, se obtiene una secuencia de etapas de sueño para todas las épocas de un registro, llamada *hipnograma*.

En adultos sanos, el sueño se inicia con sueño NREM, aumentando su profundidad progresivamente hasta transitar al sueño REM. El sueño NREM y REM continúan alternándose cíclicamente a lo largo de la noche en ciclos de 90 a 110 minutos en promedio, resultando en cuatro a seis ciclos por noche. En general, los episodios de sueño REM se alargan durante el transcurso de la noche, y la etapa N3 disminuye progresivamente su presencia dentro de los episodios de sueño NREM, pudiendo incluso desaparecer hacia el final de la noche, a medida que la etapa N2 toma progresivamente mayor protagonismo [20].

2.1.6. Complejos K y husos de sueño

La mitad del sueño ocurre en la etapa N2, caracterizada por la aparición de dos eventos llamados *complejos K* y *husos de sueño* (ver Figura 2.2). Aunque ambos eventos ocurren principalmente durante la etapa N2, también pueden observarse durante la etapa N3. Su morfología distintiva los

Tabla 2.2: Etapas del sueño según la AASM.

Etapa	Etapa equivalente de R&K	Significado	Fracción de la noche	Características
W	Wake	Estado de vigilia.	$< 5\,\%$	Presenta actividad alfa en el EEG durante más de media época en el canal occipital, alto tono muscular, y parpadeos y movimientos oculares similares a los que ocurren al leer.
N1	1	Primera fase de sueño NREM.	2–5 %	La actividad alfa en el EEG se atenúa y se reemplaza por actividad de frecuencia mixta de baja amplitud durante más de media época. No ocurren husos de sueño ni complejos K (ondas transitorias en el EEG), el tono muscular permanece relativamente alto aunque menos que en la etapa W, y hay movimientos oculares suaves por la somnolencia de la persona.
N2	2	Segunda fase de sueño NREM.	45–55 %	Su inicio es marcado por la aparición de complejos K o husos de sueño en el EEG. El EEG presenta actividad de frecuencia mixta de baja amplitud, aunque sin la actividad delta lenta (0,5–2 Hz) característica de la etapa N3. El tono muscular es bajo, aunque no existe un criterio específico para el EMG y el EOG.
N3	3 y 4	Tercera fase de sueño NREM.	10–20 %	Presenta actividad delta lenta (0,5–2 Hz) de amplitud pico a pico mayor a 75 μV en el canal frontal del EEG durante más del 20 % de la época. Podrían presentarse husos de sueño y complejos K. El tono muscular es generalmente menor que en la etapa N2, aunque no existe un criterio específico para el EMG y el EOG.
R	REM	Fase de sueño REM.	20–25 %	Presenta actividad de frecuencia mixta de baja amplitud en el EEG, mínimo tono muscular, y ráfagas de movimientos oculares rápidos.

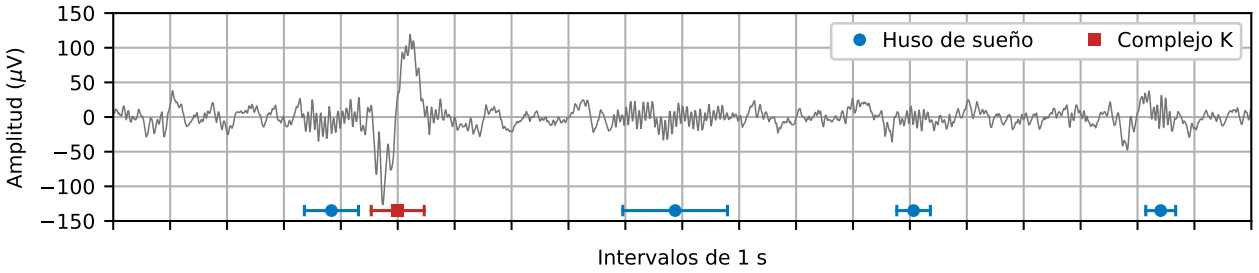

Figura 2.2: Cuatro husos de sueño (●) y un complejo K (■) en un EEG en etapa N2.

ha vuelto objetos de estudio importantes en la medicina del sueño para identificar e investigar el sueño saludable y patológico.

Los complejos K, también llamados ondas K, son ondas bifásicas de gran amplitud (son el evento cerebral con mayor amplitud) que consisten en una componente negativa abrupta seguida de una componente positiva usualmente más lenta, y tienden a preceder a un huso de sueño [1]. En general, el patrón tiene una duración en el rango 0,5–1 s [18], localizándolo principalmente en la banda delta lenta (0,5–2 Hz). Sin embargo, una duración menor a 0,5 s lo excluye como criterio para anotar etapas de sueño [1]. El patrón no tiene un criterio de amplitud estándar pero destaca de la actividad de fondo del EEG, sobre todo en los canales frontales [1]. En adultos, su amplitud pico a pico tiende a estar en el rango 100–400 μV, y tanto su amplitud como su densidad decrecen en adultos mayores [18]. La aparición de complejos K ha sido asociada con el procesamiento de estímulos, tanto internos como externos, y con la mantención del sueño [18]. Además, su actividad anormal ha sido identificada como un biomarcador de varias patologías, tales como la epilepsia [5], la apnea obstructiva del sueño [21], y el síndrome de las piernas inquietas [22].

Por otro lado, los husos de sueño, también llamados ondas sigma, son ráfagas de actividad oscilatoria localizadas en la banda *sigma* (11–16 Hz) [1]. Su envolvente tiene una amplitud menor en los extremos y mayor en su interior, dándole al patrón una forma similar a un huso para hilar. En general, el patrón tiene una duración en el rango 0,5–1,2 s [18], aunque puede ser tan corto como 0,3 s [6, 23]. Sin embargo, de forma análoga a los complejos K, una duración menor a 0,5 s lo excluye como criterio para anotar etapas de sueño [1]. El patrón tampoco tiene un criterio de amplitud estándar, requiriéndose que represente un tren de ondas predominante con respecto a la actividad de fondo del EEG, sobre todo en los canales centrales [1]. Tanto su amplitud como su densidad decrecen en adultos mayores [18]. En una caracterización reciente, la amplitud pico a pico fue de 31 ± 7 μV en adultos jóvenes y 26 ± 7 μV en adultos mayores [23]. Los husos de sueño han sido asociadas con el aprendizaje y la consolidación de la memoria [24]. Además, su actividad anormal ha sido identificada como un biomarcador de enfermedades neuropsiquiátricas como la esquizofrenia [4].

Los husos de sueño tienen un rango de frecuencias que depende fuertemente de la edad y del sujeto. En adultos, el rango es 11–16 Hz aunque suele estar concentrado en 12–14 Hz [18]. En una caracterización reciente, la frecuencia promedio en adultos fue de $13{,}1 \pm 0{,}8$ Hz [23]. En niños, en cambio, el rango de frecuencia tiende a concentrarse en el extremo inferior del rango adulto debido a una mayor actividad de las frecuencias inferiores a 12 Hz [25]. El rango de frecuencias se desplaza progresivamente con la edad hacia frecuencias mayores hasta estabilizarse durante la adolescencia [25, 24]. Adicionalmente, en adultos, la potencia en la banda sigma durante la etapa N2 tiene

un pico que existe principalmente en 12,5–14 Hz, mientras que durante la etapa N3 aparece o se intensifica un segundo pico de menor intensidad por debajo de 12,5 Hz. Ambos picos son estables a través de diferentes noches para un mismo sujeto, mientras que son característicamente diferentes entre sujetos, siendo comparado a una huella digital del sueño NREM [24].

2.1.7. Desplazamientos de los datos

Las señales y las anotaciones expertas poseen varias fuentes de variabilidad que, según su intensidad, pueden inducir un *desplazamiento* importante de la distribución de probabilidad esperada de los datos. La variabilidad puede producirse en tres niveles distintos: en la actividad cerebral, en el registro de dicha actividad, y en la anotación de dichos registros.

Desplazamiento de la población. Las características de la actividad eléctrica cerebral en el cuero cabelludo varían significativamente entre distintos sujetos. En primer lugar, el espectro de la actividad cerebral en 8–16 Hz durante el sueño NREM muestra la mayor variabilidad entre individuos [26], afectando particularmente a los picos de frecuencia de la banda sigma [24]. En segundo lugar, las características demográficas de una población en particular también inciden en el tipo de actividad cerebral que se espera observar. Por ejemplo, la edad, el sexo, y la existencia de enfermedades puede cambiar la amplitud de las ondas y la densidad de eventos cortos como husos de sueño y complejos K [20, 24].

Desplazamiento de la adquisición. El proceso utilizado para adquirir las señales por medio de EEG introduce variabilidad en los registros. Diferentes equipos de EEG pueden tener diferentes parámetros eléctricos que afectan la señal muestreada. Por ejemplo, pueden haber cambios en el tipo de electrodos (húmedos o secos), la impedancia, la ganancia, o la calidad del filtro anti-aliasing. La señal también puede adquirirse con una calidad diferente debido a cambios en el comportamiento de los sujetos (e.g., su nivel de movimiento) o en la intensidad de los artefactos. Por último, la referencia y los canales utilizados durante la adquisición pueden cambiar, modificando las señales disponibles. Por ejemplo, otro estudio puede tener uno o más canales ausentes, o puede haber usado una derivación bipolar en lugar de una unipolar.

Desplazamiento de la anotación. La anotación de etapas de sueño, husos de sueño, o complejos K en registros de EEG, al menos aquellos medidos en el cuero cabelludo, posee una variabilidad significativa debido a la imprecisión de las definiciones oficiales y la compleja diversidad de las señales. La concordancia entre expertos para anotar etapas de sueño es de 86 % [27]. Por otro lado, el *F1-score* entre un solo experto y un consenso de varios expertos para anotar husos de sueño es de 0,65–0,76 [6, 23].[2] Por último, el *F1-score* entre expertos para anotar complejos K es de 0,32–0,78 [28, 29, 30]. En estos problemas, el experto solo observa la compleja actividad cerebral a través del EEG del cuero cabelludo, mediciones inherentemente ruidosas (e.g., distorsionadas por conducción volumétrica). Las anotaciones podrían ser más precisas y consistentes con la ayuda de electrodos intracraneales. Sin embargo, los riesgos asociados a realizar una abertura en el cráneo para instalar los electrodos intracraneales hace que esta técnica sea poco atractiva para el estudio del sueño. En su ausencia, la siguiente mejor opción es un consenso de expertos sobre el EEG. Su alto costo lo excluye de la práctica estándar, limitando la anotación a uno o dos expertos en la mayoría

[2] El *F1-score* es una medida de desempeño que mide la similitud entre dos conjuntos de anotaciones, con valores entre 0 (peor) y 1 (mejor). Más detalles sobre la medición de esta métrica se dan en la Sección 3.2.1.

de los estudios. En husos de sueño, se ha realizado dicho consenso con el objetivo de evaluar el desempeño de expertos, no expertos y detectores automáticos, y generar una base de datos de alta calidad para el desarrollo de mejores detectores [23]. Incluso ante un consenso, existen diferencias en las políticas de anotación de diferentes grupos que aparecen por su modo particular de resolver ambigüedades.

2.2. Representaciones de tiempo-frecuencia

Una señal puede transitar del dominio del tiempo al dominio de la frecuencia usando la Transformada de Fourier. Sin embargo, en el EEG es común querer analizar la evolución temporal de las componentes de frecuencia. Para conseguirlo, se requiere obtener una representación de la señal que capture ambos dominios en simultáneo. Aunque existen varios métodos para ello, en todos se cumple el principio de incertidumbre: El producto entre la resolución en el tiempo (Δt) y la resolución en la frecuencia (Δf) está acotado inferiormente. Es decir, existe un compromiso entre ambas dimensiones.

Dos de los métodos más conocidos son la Transformada de Fourier de Tiempo Corto (*Short-Time Fourier Transform*, STFT) y la Transformada de Wavelet Continua (*Continuous Wavelet Transform*, CWT).[3] Tanto la STFT como la CWT se formulan como una proyección de la señal $x(t)$ sobre un banco de filtros complejos $\{\psi_f(t)\}_{f\in\mathbb{F}}$. Es decir, la transformada $\mathcal{T}[x]$ corresponde a la convolución dada por

$$\mathcal{T}[x](t,f) = \int x(\tau)\psi_f^*(t-\tau)d\tau. \tag{2.1}$$

Aunque ψ_f puede ser arbitrario, sobre todo para la CWT, en este trabajo se restringe el análisis al caso en que

$$\psi_f(t) := w_f(t)\exp(j2\pi ft), \tag{2.2}$$

en donde $w_f(t) \in \mathbb{R}$ es la *función de ventana* específica para la frecuencia f. Así, es directo interpretar a cada ψ_f como un filtro pasa-banda de frecuencia central $f \in \mathbb{F}$ y a $\mathcal{T}[x](t,f)$ como la componente de dicha frecuencia. Como la transformada es compleja, es posible obtener tanto la magnitud como la fase de cada componente de frecuencia en el tiempo. La elección de la función de ventana determina la transformada.

2.2.1. Transformada de Fourier de Tiempo Corto (STFT)

La STFT utiliza una ventana de ancho L independiente de f. Una elección sencilla es la ventana rectangular, definida por

$$w_L^{\text{rect}}(t) = \begin{cases} 1 & \text{si } t \in [-L/2,\ L/2], \\ 0 & \text{si no.} \end{cases} \tag{2.3}$$

Esta ventana introduce distorsiones en el espectro debido a la discontinuidad de los bordes. Por esta razón, es común reemplazarla por una función de ventana más suave que minimice dichos efectos de borde. Una de las funciones de ventana más utilizadas es la ventana de Hann, de forma

[3] Se describen las nociones básicas relevantes para la tesis. El lector interesado puede profundizar consultando [11].

acampanada, definida por

$$w_L^{\text{Hann}}(t) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{2\pi t}{L}\right)\right) & \text{si } t \in [-L/2,\ L/2], \\ 0 & \text{si no.} \end{cases} \tag{2.4}$$

Si bien la STFT se puede entender teóricamente como la convolución en (2.1), en la práctica se aprovecha el ancho constante de su ventana y se implementa dividiendo la señal en una serie de ventanas de ancho L y obteniendo el espectro en cada una por medio de la Transformada Rápida de Fourier (*Fast Fourier Transform*, FFT). Suponiendo una frecuencia de muestreo f_s y ventanas de L muestras (i.e., ancho L), la FFT retorna los coeficientes de las frecuencias $0 \leq f < f_s/2$ con un paso $\Delta f = f_s/L$. Además, el ancho de la ventana implica que $\Delta t = L/f_s$. En consecuencia, $\Delta t \Delta f = 1$, haciendo explícito el compromiso entre ambas resoluciones.

2.2.2. Transformada de Wavelet Continua (CWT)

La CWT es un método alternativo a la STFT en donde los filtros no tienen ancho constante, flexibilizando el compromiso $\Delta t \Delta f$ para cada frecuencia. La CWT depende de un filtro base llamado *wavelet madre*, denotado aquí por $\psi(t)$, que se escala a diferentes anchos (*escalas*) para formar el banco de filtros completo. Es decir, para cada escala s, se tiene la wavelet escalada

$$\psi_s(t) := \frac{1}{s}\psi\left(\frac{t}{s}\right), \tag{2.5}$$

implicando que el ancho de los filtros del banco varía en proporción directa con s, en lugar de mantenerse constante como en la STFT.[4]

Una de las wavelets más usadas es la wavelet Morlet compleja, que corresponde a una sinusoide con ventana Gaussiana, dada por

$$\psi(t) := \frac{2}{\sqrt{\pi\beta}} \exp\left(-\frac{t^2}{\beta}\right) \exp\left(j2\pi t\right), \tag{2.6}$$

donde β controla el ancho de la ventana Gaussiana —a mayor β, más ancha— y en consecuencia el compromiso $\Delta t \Delta f$ de referencia. Al escalar esta wavelet según (2.5), la formulación presentada en (2.6) permite hacer $s = 1/f$. Es decir, el filtro a la escala s tiene frecuencia central f. Así, se tiene que la CWT con wavelet Morlet compleja corresponde a reemplazar en (2.2) la función de ventana dada por

$$w_s^{\text{Morlet}}(t) = \frac{2}{s\sqrt{\pi\beta}} \exp\left(-\frac{t^2}{s^2\beta}\right). \tag{2.7}$$

La CWT podría localizar sus filtros en las mismas frecuencias centrales que la STFT eligiendo $s = 1/f$ con f el arreglo de frecuencias de la STFT, típicamente una progresión lineal. Sin embargo, se recomienda una progresión geométrica para la CWT debido a la manera en que varía

[4]La formulación presentada corresponde a la normalización L1, que asegura que todas las wavelets tienen la misma ganancia en su frecuencia central. Esto es conveniente para analizar EEG porque preserva las amplitudes relativas de las ondas. Otra alternativa disponible es usar una normalización L2, $\psi_s^{L2}(t) = \psi(t/s)/\sqrt{s}$, que asegura que todas las wavelets tienen la misma energía.

el compromiso $\Delta t \Delta f$ para diferentes frecuencias. En efecto, el adelgazamiento progresivo de los filtros a medida que aumenta la frecuencia (disminuye la escala) provoca un ensanchamiento progresivo de sus respuestas en frecuencia, haciendo redundante utilizar frecuencias muy cercanas en el rango de frecuencias altas. Por lo tanto, se recomienda que las escalas decaigan exponencialmente, i.e., que las frecuencias centrales aumenten exponencialmente, para disminuir la redundancia de la transformación y así usar de forma óptima el número total de escalas.

2.2.3. Comparación entre STFT y CWT

Al usar una ventana acampanada en la STFT (e.g., ventana de Hann), se tiene que la forma general de la ventana dista poco de la ventana Gaussiana utilizada en la wavelet Morlet. Por lo tanto, la principal diferencia entre la STFT y la CWT radica en la variación del ancho de cada filtro en el banco. La STFT impone el mismo ancho L, y en consecuencia los mismos Δt y Δf, a todas las frecuencias. En cambio, la CWT usa un ancho variable que permite modificar las resoluciones Δt y Δf bajo la restricción $\Delta t \Delta f = 1$ a lo largo del banco de filtros.

La Transformada de Fourier asume que la señal es estacionaria. Como las señales en el EEG no son estacionarias, es crítico ajustar el intervalo de tiempo durante el cuál se asume estacionariedad para obtener una buena aproximación. El ancho constante de la STFT implica asumir estacionariedad durante el mismo intervalo de tiempo para todas las componentes. Sin embargo, en el EEG, las frecuencias más altas tienden a mostrar dinámicas más rápidas en comparación con las frecuencias lentas, por lo que el supuesto de estacionariedad impacta de forma diferente a cada frecuencia. Convenientemente, la potencia de frecuencias más altas es extraída usando ventanas más pequeñas en la CWT, implicando una mejor resolución temporal pero a costa de una peor resolución en frecuencia. En general, esto último no representa un problema, ya que las bandas de frecuencia de interés del EEG tienden a ensancharse a medida que aumenta la frecuencia, implicando que el discernimiento entre frecuencias vecinas se hace progresivamente menos crítico. En este sentido, la CWT optimiza el compromiso entre ambas resoluciones al asignar una resolución temporal acorde a la velocidad de las variaciones esperadas, siendo más apropiada para capturar eventos transitorios en el EEG.

Por otro lado, la STFT es más rápida y puede introducir menos efectos de borde. Gracias a la FFT, la STFT tiene un bajo costo computacional cuando se compara con la CWT, que se calcula por convolución directa. Además, la CWT posee filtros notoriamente más anchos en sus frecuencias más bajas con respecto a las más altas, aumentando significativamente el costo computacional de descomponer las frecuencias bajas. Adicionalmente, el compromiso de referencia base que se escoja en la CWT (i.e., el valor de β) impacta directamente en la magnitud de los efectos de borde introducidos por sus filtros, pudiendo superar con creces aquellos introducidos por la STFT. Por ejemplo, si se diseñan ambas transformaciones para que se utilice un filtro de ancho 0,5 s en la frecuencia 10 Hz (i.e., cinco oscilaciones de dicha frecuencia), la STFT introducirá 0,25 s de distorsión en cada borde en todas las frecuencias, mientras que la CWT introducirá mayores distorsiones mientras menores sean las frecuencias: el filtro asociado a 1 Hz sería de 5 s de ancho. Por lo tanto, si se desean eliminar los efectos de borde, la CWT requerirá de mayor longitud de señal que la STFT para poder recortar el resultado y obtener el mismo largo final.

2.2.4. Una transformación intermedia

La STFT y la CWT se pueden interpretar como dos casos extremos respecto a la estrategia de variación del ancho de los filtros: independiente de la frecuencia para la STFT y en proporción inversa a la frecuencia para la CWT. En el software de código libre EEGLAB [31], uno de los más usados para estudiar señales de EEG, se le permite al usuario elegir una transformación basada en wavelets con un comportamiento intermedio gracias a un factor $q \in [0, 1]$, llamado *factor de expansión de escala*. Según este nuevo parámetro, $q = 0$ corresponde al comportamiento de la STFT y $q = 1$ corresponde al comportamiento de la CWT.

Tomando como referencia la CWT, sea s_q una escala expandida por el factor q dada por

$$s_q = qs + (1 - q)s_{\max}, \tag{2.8}$$

en donde $s_{\max} = 1/f_{\min}$ es la escala asociada a $f_{\min}$, la mínima frecuencia de la transformación. Es directo que $s_{q=0} = s_{\max}$ (i.e., constante), $s_{q=1} = s$, y $s_{q,\max} = s_{\max}$ (i.e., el ancho del filtro más lento es independiente de q). Entonces, el nuevo banco de filtros tiene una función de ventana dada por por

$$w_{s,q}(t) = \frac{2}{s_q\sqrt{\pi\beta}} \exp\left(-\frac{t^2}{s_q^2\beta}\right). \tag{2.9}$$

El caso $q = 1$ (CWT, en donde (2.9) es igual a (2.7)) implica que cada filtro muestra siempre el mismo número de oscilaciones (i.e., todas las wavelets preservan la plantilla de la wavelet madre). Exigir el mismo número de oscilaciones para todas las frecuencias podría ser inadecuado para algunas aplicaciones, en donde se podría elegir como alternativa $q < 1$ para admitir progresivamente más oscilaciones en las frecuencias más altas. Este valor se puede ajustar para acercarse tanto como se desee al otro extremo, $q = 0$, que corresponde al caso en que la ventana no varía su ancho (como la STFT) y en consecuencia el número de oscilaciones dentro de cada filtro del banco aumenta proporcionalmente a la frecuencia. En EEGLAB, se recomienda configurar q en un valor ligeramente menor a 1. De esta forma, se conservan en gran parte las ventajas de usar la CWT y se compensa su desventaja de usar filtros demasiado anchos para las frecuencias más bajas, disminuyendo el costo computacional y los efectos de borde.

2.3. Aprendizaje profundo

El aprendizaje profundo (*deep learning* en inglés) es un conjunto de técnicas dentro del área del aprendizaje de máquinas que ha ganado popularidad gracias a sus impresionantes resultados en el procesamiento de imágenes, video, audio y texto [32], y que se ha adoptado progresivamente en otras áreas como el procesamiento de EEG, alcanzando el estado del arte en diversas tareas [33]. Como toda técnica de aprendizaje de máquinas, el aprendizaje profundo ajusta los parámetros de sus modelos usando un entrenamiento sobre un conjunto de datos. Sin embargo, se diferencia en que sus modelos aprenden simultáneamente la mejor representación de los datos, permitiendo su aplicación exitosa en datos *en bruto* o mínimamente preprocesados y con alta dimensionalidad. Estos modelos se llaman *redes neuronales artificiales*, o simplemente redes neuronales, por su inspiración en el sistema nervioso. En esta sección, se describen sus nociones básicas.[5]

[5]El lector interesado puede profundizar consultando [34].

2.3.1. Redes neuronales artificiales

La neurona artificial es la unidad básica de una red neuronal artificial, dada por

$$z = \phi\left(b + \sum_{j=1}^{d} w_j x_j\right), \tag{2.10}$$

en donde $\mathbf{x} \in \mathbb{R}^d$ es la entrada; $w_j \in \mathbb{R}$ y $b \in \mathbb{R}$ son los parámetros ajustables (*entrenables*) de la neurona llamados pesos y sesgo, respectivamente; y ϕ es una función no lineal. Esta función, llamada función de activación, *activa* o *inhibe* a la neurona según la ponderación de sus entradas. Se caracteriza por transitar entre dos estados según si su argumento es negativo (inhibición) o positivo (activación). Dentro de las opciones más comunes se encuentran las funciones sigmoide, tangente hiperbólica, y unidad lineal rectificada (*Rectified Linear Unit*, ReLU) [35].

Un grupo de m neuronas pueden procesar la misma entrada en paralelo para formar una capa neuronal de ancho o dimensión m y una salida $\mathbf{z} \in \mathbb{R}^m$ dada por

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} \phi\left(b_1 + \sum_{j=1}^{d} w_{1j} x_j\right) \\ \phi\left(b_2 + \sum_{j=1}^{d} w_{2j} x_j\right) \\ \vdots \\ \phi\left(b_m + \sum_{j=1}^{d} w_{mj} x_j\right) \end{bmatrix} = \phi\left(\mathbf{b} + \mathbf{W}\mathbf{x}\right), \tag{2.11}$$

en donde ϕ se aplica componente a componente, y $\mathbf{b} \in \mathbb{R}^m$ y $\mathbf{W} \in \mathbb{R}^{m \times d}$ son el vector de sesgos y la matriz de pesos de la capa, respectivamente. Otros tipos de capas neuronales son posibles al agregar más estructura a (2.11), descritas en la Sección 2.3.5. Varias capas pueden componerse haciendo que la salida de una sea la entrada de otra, formando una red neuronal. Se ha demostrado que esta clase de modelos son aproximadores universales de funciones [36].

La configuración de las capas y su conexión entre ellas corresponde a la arquitectura de la red neuronal. La última capa (capa de salida) requiere una dimensión y activación compatible con la salida deseada. En cambio, las capas restantes (capas ocultas) no están restringidas. Esto permite una gran flexibilidad en el diseño de la arquitectura que se resuelve típicamente de forma experimental y en base a algunos principios de diseño. En una arquitectura, cada capa puede ser de distinto tipo, y se pueden componer arbitrariamente mientras formen un grafo acíclico entre la entrada y la salida de la red. Por ejemplo, pueden existir ramas paralelas que concatenan sus salidas, o conexiones laterales unidireccionales.

2.3.2. Entrenamiento por retropropagación del error

Una red neuronal representa una familia de funciones f_θ parametrizada por $\theta \in \Theta$ (pesos y sesgos). El entrenamiento aproxima un óptimo θ^* para el problema. En el aprendizaje supervisado, si $\mathbf{y}$ es la salida deseada ante una entrada $\mathbf{x}$, se quiere que $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$ esté a una mínima distancia de $\mathbf{y}$. Este ajuste se consigue optimizando la esperanza de una función de pérdida diferenciable $\ell(\mathbf{y}, \hat{\mathbf{y}})$ cuyo mínimo se alcanza cuando $\hat{\mathbf{y}} = \mathbf{y}$. En problemas de clasificación de C clases, $\mathbf{y} \in [0, 1]^C$ representa la distribución de probabilidad sobre dichas clases, y es común minimizar la entropía

cruzada entre la salida de la red y la salida deseada, dada por

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c). \tag{2.12}$$

Si $\mathcal{L}(\theta)$ es la esperanza de $\ell(\mathbf{y}, \hat{\mathbf{y}})$, y se tiene un conjunto de entrenamiento de N ejemplos $\{(\mathbf{x}_i,\ \mathbf{y}_i)\}_{i=1}^{N}$, el entrenamiento resuelve

$$\theta^* = \arg\min_{\theta\in\Theta} \mathcal{L}(\theta) = \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{y}^i, f_\theta(\mathbf{x}^i)). \tag{2.13}$$

La diferenciabilidad de f_θ y ℓ permite minimizar usando *gradiente descendente*, un algoritmo iterativo dado por

$$\theta_{k=0} = \theta_0, \qquad\qquad \theta_k = \theta_{k-1} - \alpha\nabla_\theta\mathcal{L}(\theta_{k-1}), \tag{2.14}$$

en donde θ_0 es la inicialización y α es la tasa de aprendizaje. Para amortiguar el costo computacional, el gradiente $\nabla_\theta\mathcal{L}(\theta) = \frac{1}{N}\sum_{i\leq N} \nabla_\theta\ell(\mathbf{y}^i, \hat{\mathbf{y}}^i)$ se estima de forma insesgada en cada iteración con un lote o *batch* aleatorio de $M \ll N$ ejemplos de entrenamiento. El algoritmo resultante se llama *gradiente descendente estocástico* (*stochastic gradient descent*, SGD) debido a la estocasticidad que introduce dicha estimación del gradiente. SGD converge si α es una serie decreciente en las iteraciones que satisface $\sum_k \alpha_k = \infty$ y $\sum_k \alpha_k^2 < \infty$ [37]. Los M gradientes $\nabla_\theta\ell$ requeridos en cada iteración se obtienen eficientemente con un método basado en la regla de la cadena llamado *retropropagación del error* (*error backpropagation* en inglés), con un costo computacional similar a la evaluación de la red.

2.3.3. Técnicas para mejorar el gradiente

El gradiente juega un rol central en el entrenamiento. Se han propuesto varias técnicas para asegurar su calidad y evitar la divergencia o estancamiento del entrenamiento, algunas de las cuales se describen en la Tabla 2.3.

2.3.4. Técnicas para mejorar la generalización

El objetivo principal no es aprender una función que minimice la pérdida promedio en el conjunto de entrenamiento, sino que aprender una función suficientemente general para minimizar la esperanza de la función de pérdida bajo la distribución de probabilidad desconocida de los datos. Para permitir el ajuste, la red neuronal debe ser suficientemente grande, pero un gran tamaño aumenta el riesgo de sobre-ajuste al conjunto de entrenamiento. Se han propuesto varias técnicas que regularizan el entrenamiento para disminuir el riesgo de sobre-ajuste, algunas de las cuales se describen en la Tabla 2.4.

2.3.5. Tipos de capas neuronales

Se han propuesto varios tipos de capas neuronales que están diseñadas para aprender eficientemente ante un tipo determinado de entradas. A continuación se describen las capas más comunes, en donde se omite la función de activación a la salida cuando es posible.

Tabla 2.3: Técnicas para mejorar la calidad del gradiente.

Técnica	Descripción
Función de activación	Impacta el valor del gradiente y la tasa a la que se desvanece a través de las capas. Es común usar ReLU a menos que se requieran valores acotados. ReLU no satura sus activaciones, atenuando el desvanecimiento del gradiente en redes profundas.
Inicialización	La inicialización aleatoria de los pesos impacta el gradiente al inicio del entrenamiento y la probabilidad de una buena convergencia. Las más populares son la de Xavier Glorot [38] (ideal para activaciones sigmoides) y la de Kaimin He [39] (ideal para activación ReLU). Ambas aseguran una varianza constante a través de las capas ocultas.
Regla de actualización	La tasa de aprendizaje y la dirección de descenso son los factores más influyentes en el entrenamiento, por lo que se han propuesto varios métodos para su ajuste. El optimizador Adam (*adaptive moment estimation*) [40] es uno de los más populares. A diferencia de SGD, este método usa una ponderación histórica del gradiente como dirección de descenso (inercia), escala la tasa de aprendizaje para cada parámetro de forma independiente usando una ponderación histórica de la norma del gradiente, y es invariante a la escala de la función de costo.
Recorte del gradiente	En algunas arquitecturas, particularmente las recurrentes, el gradiente podría ser demasiado grande en algunas iteraciones, implicando actualizaciones grandes que podrían poner en riesgo el éxito del entrenamiento [41]. Para evitar esto, la norma del gradiente $\|\|\nabla_\theta \mathcal{L}(\theta)\|\|$ se puede restringir a una norma máxima $g_{\max}$.
Batch normalization	El entrenamiento se puede hacer más robusto a la inicialización y se puede acelerar al introducir capas de normalización entre las capas ocultas. Una de las más utilizadas es *batch normalization*, acortado como *batchnorm* [42]. Esta capa normaliza cada componente a media cero y varianza unitaria usando las estadísticas del batch de entrenamiento actual, y luego aplica una transformación lineal entrenable. Durante la inferencia, la normalización se fija con las estadísticas históricas de todo el conjunto de entrenamiento.

Tabla 2.4: Técnicas para regularizar el entrenamiento.

Técnica	Descripción
Aumento de datos	Mientras más grande y diverso sea el conjunto de entrenamiento, más difícil es memorizarlo, mejorando la generalización. Un conjunto puede aumentar su tamaño artificialmente usando técnicas de aumento de datos que aplican transformaciones para las cuales el modelo debe ser invariante (mantener la etiqueta $\mathbf{y}$) o equivariante (modificar la etiqueta $\mathbf{y}$ con la misma transformación de $\mathbf{x}$).
Detención temprana	Durante el entrenamiento, se monitorea el desempeño del modelo en un conjunto independiente de datos, llamado conjunto de validación, para estimar la generalización. El entrenamiento se detiene cuando el desempeño de validación empeora.
Dropout	Disminuye la dependencia en neuronas específicas y aproxima ensambles de varias redes en una sola red neuronal [43]. Durante el entrenamiento, la capa de *dropout* enmascara cada componente con cero con una probabilidad ρ (típicamente $\leq 0{,}5$) y escala sus salidas en $1/(1-\rho)$ para preservar la magnitud esperada de una combinación lineal de ellas. Durante la inferencia, implementa la función identidad. Las capas de *dropout* disminuyen la capacidad de las capas que afectan, por lo que suelen usarse cuando hay alta dimensionalidad.

Capa densa. Ya descrita en (2.11). Esta capa tiene todas sus neuronas conectadas a toda la entrada.[6] Una capa densa de m unidades transforma un vector $\mathbf{z}_1 \in \mathbb{R}^d$ en un vector $\mathbf{z}_2 \in \mathbb{R}^m$ según $\mathbf{z}_2 = \mathbf{W}\mathbf{z}_1 + \mathbf{b}$, con $\mathbf{W} \in \mathbb{R}^{m \times d}$ sus pesos y $\mathbf{b} \in \mathbb{R}^m$ sus sesgos.

Capa softmax. Es una capa sin parámetros que aproxima suavemente la detección de la componente máxima. Tiene una salida que suma 1, como una distribución de probabilidad. Transforma un vector $\mathbf{z} \in \mathbb{R}^d$ en un vector $\mathbf{p} \in \mathbb{R}^d$ según $\mathbf{p} = \exp(\mathbf{z}) / \sum_{j=1}^{d} \exp(\mathbf{z})_j$, en donde la exponencial se aplica componente a componente. Se utiliza generalmente como capa de salida en problemas de clasificación, en donde $\mathbf{p}$ es la probabilidad predicha y $\mathbf{z}$ es el *logit*.

Capa convolucional. Diseñada para procesar datos equivariantes bajo traslaciones en uno o más de sus ejes. Los datos podrían tener estructura 2D, como imágenes (equivariantes en su altura y ancho), o 1D, como series de tiempo (equivariantes en el tiempo). Las neuronas se agrupan en filtros, y cada filtro contiene neuronas que comparten parámetros y que están conectadas a una vecindad de la entrada. Además, la salida de la capa preserva la estructura equivariante. De esta forma, cada filtro implementa una convolución con la entrada a lo largo de los ejes equivariantes. Sea $\mathbf{z}_1 \in \mathbb{R}^{h_1 \times w_1 \times d}$ un tensor 3D que representa una entrada de estructura 2D de alto h_1, ancho w_1, y d canales. Una capa convolucional 2D de m filtros y tamaño de filtro (k_1, k_2) transforma el tensor de entrada en otro tensor 3D $\mathbf{z}_2 \in \mathbb{R}^{h_2 \times w_2 \times m}$ por medio de la convolución $\mathbf{z}_2 = \mathbf{W} * \mathbf{z}_1 + \mathbf{b}$, en donde $\mathbf{W} \in \mathbb{R}^{k_1 \times k_2 \times d \times m}$ es la matriz de pesos que agrupa los m filtros de forma (k_1, k_2, d), y $\mathbf{b} \in \mathbb{R}^m$ es el vector de sesgos que agrupa el sesgo de cada filtro. Los bordes de $\mathbf{z}_1$ se pueden rellenar con ceros para que $h_2 = h_1$ y $w_2 = w_1$ (conocido como *zero-padding*). Además, la convolución se puede hacer con saltos, i.e., cada s_1 posiciones en la altura y cada s_2 posiciones en el ancho, en lugar de hacerse en todas las posiciones posibles (equivalente a $s_1 = s_2 = 1$), implicando $h_2 = h_1/s_1$ y $w_2 = w_1/s_2$. Por otro lado, si se tiene $\mathbf{z}_1 \in \mathbb{R}^{l_1 \times d}$, un tensor 2D que representa una entrada de estructura 1D de largo l_1 y d canales, se puede aplicar una capa convolucional 1D de m filtros, tamaño de filtro k, y salto s. Su comportamiento es análogo al caso 2D al introducir un eje equivariante adicional redundante de tamaño 1 y usar un tamaño de filtro $(k, 1)$.

Capa convolucional dilatada. En cada filtro se introduce un tasa de dilatación n_{dilation} que permite aumentar el campo receptivo sin aumentar el número de parámetros. Específicamente, un filtro de tamaño (k_1, k_2) se comporta como un filtro de tamaño $(1 + n_{\text{dilation}}(k_1 - 1), 1 + n_{\text{dilation}}(k_2 - 1))$ al intercalar $n_{\text{dilation}} - 1$ ceros entre cada parámetro a lo largo de los ejes equivariantes. La convolución ordinaria se obtiene cuando $n_{\text{dilation}} = 1$.

Capa de pooling. Disminuyen la dimensionalidad a lo largo de los ejes equivariantes. Esta capa segmenta una entrada 2D a lo largo de sus ejes equivariantes en grupos (*pools*) de tamaño (p_1, p_2) y con saltos (s_1, s_2) entre grupos. Luego, cada grupo se colapsa a un escalar independientemente para cada canal. Grupos sin intersección se consiguen cuando $s_1 \geq p_1$ y $s_2 \geq p_2$, y típicamente se elige $s_1 = p_1$ y $s_2 = p_2$. El colapso podría ser el promedio (capa de *average pooling*) o el máximo (capa de *max pooling*).

Capa recurrente vainilla. Procesa series de tiempo de forma secuencial y compartiendo los parámetros en cada instante tiempo. Sea $\mathbf{z} \in \mathbb{R}^{l \times d}$ una entrada de l muestras y d dimensiones o

[6] También se conoce como capa totalmente conectada (*fully-connected layer* en inglés) o capa lineal (*linear layer* en inglés).

canales por muestra. Una capa de dimensión m procesa la entrada al aplicar, para cada $k = 1, \ldots, l$, la recurrencia

$$\mathbf{h}_{k,\bullet} = \phi\left(\mathbf{W}\begin{bmatrix}\mathbf{z}_{k,\bullet}\\ \mathbf{h}_{k-1,\bullet}\end{bmatrix} + \mathbf{b}\right), \tag{2.15}$$

en donde $\mathbf{h} \in \mathbb{R}^{l\times m}$ es la serie de salida compuesta por los *estados* de la capa, $\mathbf{h}_{0,\bullet} = 0$ es el estado inicial, $\mathbf{W} \in \mathbb{R}^{m\times(d+m)}$ es la matriz de pesos, $\mathbf{b} \in \mathbb{R}^m$ es el vector de sesgos, y ϕ es la función de activación, normalmente una tangente hiperbólica para evitar la divergencia de la recurrencia. Esta capa recurrente básica presenta problemas para aprender dependencias temporales de largo plazo.

Capa recurrente Long Short-Term Memory (LSTM). También procesa series de tiempo pero con mayor capacidad para aprender dependencias temporales de largo plazo al complementar el estado con una memoria y un mecanismo de compuertas que permiten manipular la memoria linealmente [44]. A diferencia de una capa recurrente vainilla, una capa LSTM de dimensión m procesa la entrada al aplicar, para cada $k = 1, \ldots, l$, la recurrencia

$$\mathbf{f}_k = \text{sigmoid}\left(\mathbf{W}_f\begin{bmatrix}\mathbf{z}_{k,\bullet}\\ \mathbf{h}_{k-1,\bullet}\end{bmatrix} + \mathbf{b}_f\right), \qquad \mathbf{i}_k = \text{sigmoid}\left(\mathbf{W}_i\begin{bmatrix}\mathbf{z}_{k,\bullet}\\ \mathbf{h}_{k-1,\bullet}\end{bmatrix} + \mathbf{b}_i\right), \tag{2.16}$$

$$\mathbf{o}_k = \text{sigmoid}\left(\mathbf{W}_o\begin{bmatrix}\mathbf{z}_{k,\bullet}\\ \mathbf{h}_{k-1,\bullet}\end{bmatrix} + \mathbf{b}_o\right), \qquad \tilde{\mathbf{c}}_k = \tanh\left(\mathbf{W}_c\begin{bmatrix}\mathbf{z}_{k,\bullet}\\ \mathbf{h}_{k-1,\bullet}\end{bmatrix} + \mathbf{b}_c\right), \tag{2.17}$$

$$\mathbf{c}_k = \mathbf{f}_k \odot \mathbf{c}_{k-1} + \mathbf{i}_k \odot \tilde{\mathbf{c}}_k, \qquad \mathbf{h}_{k,\bullet} = \mathbf{o}_k \odot \tanh(\mathbf{c}_k), \tag{2.18}$$

en donde $\mathbf{c}_k \in \mathbb{R}^m$ es la memoria interna, $\mathbf{c}_0 = 0$ es la memoria inicial, $\mathbf{f}_k \in \mathbb{R}^m$ es la compuerta de olvido, $\mathbf{i}_k \in \mathbb{R}^m$ es la compuerta de entrada, $\mathbf{o}_k \in \mathbb{R}^m$ es la compuerta de salida, $\tilde{\mathbf{c}}_k \in \mathbb{R}^m$ es la memoria candidata, $\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_o, \mathbf{W}_c \in \mathbb{R}^{m\times(d+m)}$ son las matrices de pesos, y $\mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_o, \mathbf{b}_c \in \mathbb{R}^m$ son los vectores de sesgos. Para que la memoria se preserve por más pasos al inicio del entrenamiento, $\mathbf{b}_f$ se inicializa en 1 en lugar de 0 [45].

Capa recurrente bidireccional. Concatena las salidas de dos capas recurrentes independientes que procesan la misma entrada pero en direcciones temporales contrarias, integrando contexto pasado y futuro. En el caso de la capa LSTM, la versión bidireccional se abrevia normalmente como BLSTM.

2.4. Trabajos relacionados

En esta sección, se revisa la bibliografía relevante para el problema de detección de husos de sueño y complejos K. En primer lugar, se presentan los métodos y materiales utilizados en la literatura para evaluar el desempeño de un detector arbitrario. Después, se revisan los métodos de detección propuestos en la literatura, con especial énfasis en aquellos basados en aprendizaje profundo. A continuación, se identifican limitaciones y brechas en los trabajos existentes que justifican la realización de esta tesis. Por último, se describen las limitaciones fundamentales que existen para un detector arbitrario desde una perspectiva causal, las que permiten establecer expectativas realistas para el desempeño en distintos escenarios.

2.4.1. Medición del desempeño

Al detectar husos de sueño o complejos K en una señal de EEG, se define un conjunto de *eventos* detectados, que debe ser comparado con el conjunto de eventos etiquetados para calcular métricas

de desempeño [6]. En la literatura se identifican cuatro métodos para medir el desempeño: por muestra, por segmento, por evento, y por sujeto.

Evaluación por muestra. Es el método más sencillo, en donde cada muestra de la señal se asigna a la clase negativa (actividad de fondo) o positiva (interior de un evento) y se determinan aciertos o falsos para cada una independientemente. Sin embargo, este método tiene dos desventajas. En primer lugar, por la baja presencia de eventos en comparación a la actividad de fondo, la mayoría de las muestras son verdaderos negativos, lo que provoca que métricas que usan dicho tipo de acierto (e.g., especificidad) sean demasiado optimistas y de poco valor para discriminar entre detectores. En segundo lugar, todo tipo de discordancia entre las detecciones y las etiquetas se penaliza, y se penaliza durante todo el largo. No se distingue entre un traslape parcial y una ausencia de traslape, a pesar de consistir en dos tipos diferentes de error (instantes de inicio y fin versus existencia). Además, cada evento pondera en el desempeño proporcionalmente a su largo, aún cuando la intuición dice que todo evento debería ponderar lo mismo.

Evaluación por segmento. En este método, la señal se segmenta sin traslape en ventanas consecutivas de ancho constante (e.g., 1 s o 10 s), y cada ventana se asigna a la clase negativa o positiva según un criterio basado en sus muestras interiores. Por ejemplo, que la ventana pertenece a la clase mayoritaria de sus muestras, o que pertenece a la clase positiva si al menos una de sus muestras es positiva. Si bien este método compensa en algún grado las limitaciones de la evaluación por muestra, requiere la elección de un ancho de ventana que es arbitrario. Si es muy corta, cada evento puede estar en múltiples ventanas, tendiendo a los problemas de la evaluación por muestra. Si es muy larga, se pierde la asociación entre etiqueta y detección.

Evaluación por evento. En este método, se hace un apareamiento explícito entre el conjunto de eventos y de detecciones por máximo traslape. Una vez que se encuentran los apareamientos, se asignan verdaderos positivos cuando se cumple algún criterio de concordancia entre la etiqueta y la detección. Cuando no se cumple dicho criterio, se asignan falsos. Este método tiene la ventaja de que cada evento pondera lo mismo y que los errores se penalizan una sola vez. Sin embargo, tiene la desventaja de que es necesario definir un criterio de concordancia. Por otro lado, no están definidos los verdaderos negativos, por lo que no se pueden ocupar métricas que usen dicho tipo de aciertos. Este es el método recomendado en [6] y seguido por varias publicaciones recientes.

Evaluación por sujeto. En este método no se calculan las métricas usuales de desempeño, sino que se mide la similitud que existe entre parámetros calculados en la señal completa usando el conjunto de etiquetas y el conjunto de detecciones. Por ejemplo, se puede comparar la densidad de eventos o su duración promedio. La similitud se puede medir en base a la calidad del ajuste lineal entre las etiquetas y las detecciones, por ejemplo usando el coeficiente R^2. En [6], se recomienda complementar la evaluación por evento usando la evaluación por sujeto.

Como se mencionó anteriormente, el método recomendado para calcular métricas de detección es por evento. En la literatura se identifican varios criterios de concordancia para utilizar este método. Entre los principales destacan: intersección no nula, intersección suficiente, razón de intersección suficiente, tiempos de inicio cercanos, y centros cercanos. En la intersección no nula, cualquier intersección entre el evento etiquetado y la detección determina un verdadero positivo. En la intersección suficiente, se requiere que la intersección dure al menos un intervalo de tiempo.

En la razón de intersección suficiente, se requiere que la razón entre el intervalo de intersección y el intervalo que abarca la unión de ambos eventos (*Intersection over Union*, IoU), supere un cierto umbral (e.g., 0,2). En los tiempos de inicio cercanos, se requiere que la distancia entre el inicio de ambos eventos sea menor a un cierto intervalo de tiempo (e.g., 0,2 s). Por último, en los centros cercanos se usa un criterio similar al de los tiempos de inicio cercanos pero usando los centros de cada evento.

En [6] se recomienda la razón de intersección (IoU) suficiente. Con respecto a los demás, este criterio tiene la ventaja de ser independiente de la duración de los eventos, y es sensible a todo tipo de desalineación. Aunque el IoU tiene un máximo de 1 (coincidencia perfecta), dicho valor casi nunca se alcanza en la práctica, lo que justifica tolerar un error. A modo de referencia, expertos individuales que anotan husos de sueño en adultos jóvenes alcanzan una mediana de 0,81 en IoU al compararse con un consenso de expertos, con rango intercuartil 0,68–0,90 [23]. Usando una evaluación por evento con criterio de IoU, se puede obtener el número de falsos positivos, falsos negativos, y verdaderos positivos para calcular métricas. En [6], se recomienda utilizar *recall* (proporción de etiquetas detectadas) y *precision* (proporción de detecciones verdaderas), ambas independientes del número de verdaderos negativos. Estas dos métricas se pueden combinar en el *F1-score*, su promedio armónico, para resumir el desempeño.

Si bien las recomendaciones anteriores se han seguido en varias de las publicaciones recientes, aún existe una significativa heterogeneidad que dificulta comparar la literatura existente. Esto es particularmente grave en publicaciones más viejas, previas a la publicación de las recomendaciones actuales. Además, existen otras dos dimensiones que limitan la comparación con la literatura: la definición del conjunto válido de eventos, y los datos.

Conjunto válido de eventos. Para efectos de anotar etapas de sueño, los husos de sueño deben durar al menos 0,5 s [1]. Sin embargo, los husos podrían ser tan cortos como 0,3 s [6]. Por lo tanto, el conjunto válido más extenso para evaluar el desempeño contempla una duración mínima de 0,3 s. En la literatura, el desempeño se ha medido con diversas duraciones mínimas, siendo la más común 0,5 s. Esto da lugar a comparaciones injustas con detectores que admiten una duración mínima menor, como 0,3 s, ya que las duraciones más cortas son más propensas a falsos (tanto positivos como negativos).

Datos. Los detectores propuestos hace casi una década o más, fueron desarrollados y evaluados usando diferentes bases de datos privadas. Dada la variabilidad entre expertos, es difícil establecer cuál detector es mejor que otro según dichas evaluaciones. Además, dichas bases de datos son en su gran mayoría pequeñas (segmentos de señal de 1 o 2 sujetos), introduciendo más varianza al desempeño reportado. El problema de las bases privadas se atenuó con la publicación de las bases de datos públicas DREAMS, tanto para husos de sueño [46] como para complejos K [29], las que fueron rápidamente utilizadas para validar nuevos detectores. A pesar de ello, las bases DREAMS siguen siendo pequeñas, consistentes en segmentos de 30 min de 8–10 sujetos. No fue hasta la publicación de la base de datos pública MASS [47], en particular su subconjunto 2 (MASS-SS2) con anotaciones de husos de sueño y complejos K, que se pudo contar con una base de datos grande: 19 registros de noche completa con toda la etapa N2 anotada. Desde la publicación de MASS-SS2, las bases DREAMS han caído en desuso. Aún cuando MASS-SS2 ha representado un gran paso para estandarizar el desarrollo de detectores, solo contiene anotaciones realizadas por un experto: dos expertos que anotan independientemente husos de sueños, y un experto que anota

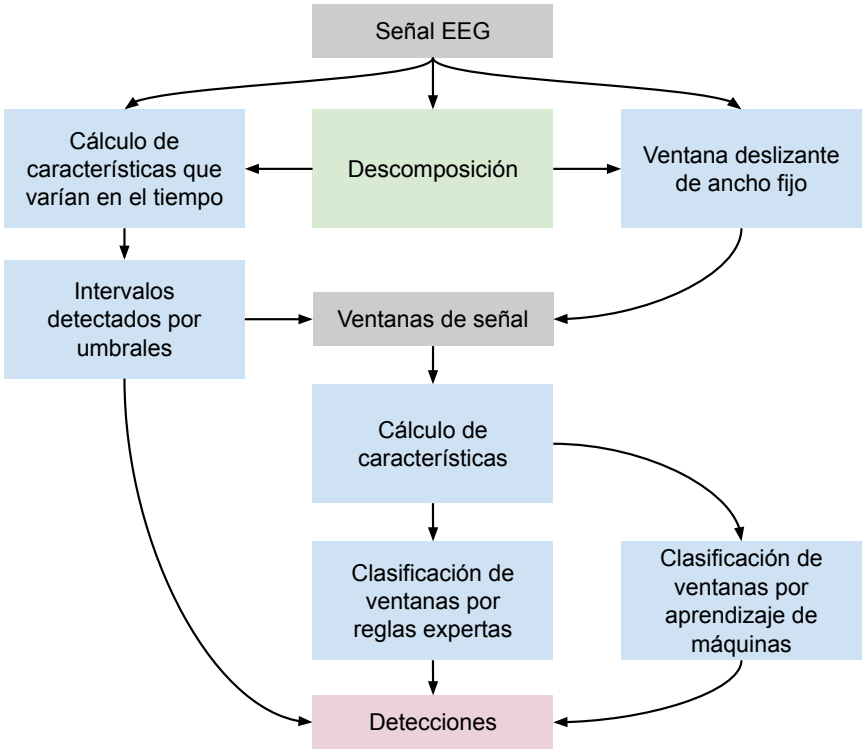

Figura 2.3: Resumen de los métodos de detección tradicionales basados en características.

complejos K. En husos de sueño, la diferencia entre los dos expertos hace inútil un consenso. Por lo tanto, aún está presente el problema de la variabilidad entre expertos. Además, al analizar los espectros, se observa que la muestra de 19 registros es capaz de representar bien la variabilidad espectral en la banda delta (donde existen los complejos K), pero no así aquella variabilidad en la banda sigma (donde existen los husos de sueño). Esta observación es consistente con los reportes de que la banda delta y sigma tienen baja y alta variabilidad individual, respectivamente [26, 24]. Por lo tanto, el problema de detección de husos de sueño requiere más sujetos. Recientemente, se ha publicado MASS-MODA [23], una base de datos pública basada en MASS con anotaciones de husos de sueño basadas en un consenso de varios expertos y sobre segmentos de señal de 180 sujetos. La base MASS-MODA se protege de la variabilidad entre expertos gracias al consenso, y cubre una gran variabilidad de señales gracias al número de sujetos, al costo de consistir en un menor volumen de datos (en horas) en comparación a MASS-SS2.[7]

2.4.2. Métodos de detección de husos de sueño y complejos K

Varios métodos han sido propuestos para la detección de husos de sueño y complejos K en el EEG del sueño. Para efectos de este trabajo, se distinguen dos grandes grupos: aquellos basados en características —temporales o espectrales— diseñadas a mano (métodos tradicionales), y aquellos basados en aprendizaje profundo. Entre los métodos tradicionales revisados, se pueden identificar varias estrategias generales de detección, resumidas en la Figura 2.3.

Algunos métodos calculan una o más características que varían en cada instante de tiempo para aplicar umbrales. Luego de aplicar umbrales, se determinan intervalos de tiempo que pueden con-

[7]Más detalles de MASS-SS2 y MASS-MODA se entregan en la Sección 3.2.2

siderarse directamente como detecciones [48] (también los métodos evaluados en [6] y el método usado en [49]). Entre los métodos de este subgrupo se encuentran los detectores clásicos que usan la amplitud instantánea de alguna banda de frecuencia para determinar la existencia de eventos (banda sigma en husos de sueño, banda delta en complejos K). Otros métodos usan los intervalos detectados como ventanas candidatas. La señal en su interior se representa por un conjunto de características adicionales, y con ellas se clasifica cada ventana como negativa (actividad de fondo) o positiva (evento de interés). Algunos métodos clasifican las ventanas candidatas usando reglas expertas diseñadas a mano [50, 29, 51, 52, 53, 54, 55], mientras que otros métodos usan algoritmos de aprendizaje de máquinas tradicionales, como *Support Vector Machine* o *Random Forest*, para aprender la clasificación de ventanas [56, 57]. Otro grupo de métodos no detecta ventanas candidatas, sino que desliza una ventana de ancho fijo, en cuyo interior se calculan características y se clasifican ya sea usando reglas expertas [58, 59] o aprendizaje de máquinas [60, 61, 62]. Luego de deslizar la ventana sobre toda la señal, etiquetas positivas consecutivas constituyen una detección. Por ejemplo, el detector A7 [59] usa ventanas de 0,3 s con un paso de 0,1 s para calcular cuatro características (potencia sigma, potencia sigma relativa, y covarianza y correlación entre banda sigma y señal original) que son pasadas por un umbral y se combinan con reglas sencillas para detectar husos de sueño. Por último, algunos métodos primero descomponen la señal EEG para después usar una o varias componentes de alguna de las formas descritas anteriormente, bajo la hipótesis de que la descomposición aísla los eventos de interés y facilita su detección [63, 64, 65, 66]. Por ejemplo, el detector Spinky [65] descompone la señal en una componente oscilatoria y una componente transitoria, que usa después para detectar husos de sueño y complejos K, respectivamente.

Los métodos tradicionales tienen un desempeño limitado, particularmente con los falsos positivos, debido a la variabilidad biológica que ocurre entre sujetos y al proceso subjetivo de anotación que hace que la transferencia de conocimiento desde los expertos del sueño a un algoritmo sea una tarea difícil. Adicionalmente, las características diseñadas a mano son específicas para cada evento y normalmente no generalizan a otras clases de eventos. Como una alternativa, los métodos basados en aprendizaje profundo pueden aprender características automáticamente a partir de los datos que pueden generalizar a registros nuevos. Estos métodos alcanzan el estado del arte en detección de husos de sueño y complejos K. En esta tesis, se desarrolla un detector basado en aprendizaje profundo, por lo que estos métodos son los más relevantes para el trabajo actual.

Uno de los métodos propuestos es DOSED [8], desarrollado para eventos cortos en general, basado en una red neuronal convolucional que procesa la señal EEG en el tiempo en segmentos de 20 s para detectar husos de sueño y complejos K. Luego de la extracción convolucional de características, se aplica una capa *fully-connected* con varias salidas sobre la representación de todo el segmento de 20 s para tomar en cuenta el contexto temporal. A lo largo del segmento se ubican anclas equiespaciadas con una duración por defecto (para husos de sueño y complejos K, las anclas tienen duración de 1 s y separación de 0,25 s), y DOSED predice tres salidas para cada ancla: la existencia de un evento, la desviación que se debe aplicar al centro por defecto, y la desviación que se debe aplicar a la duración por defecto.

En segundo lugar se tiene SpindleNet [9], desarrollado para husos de sueño, basado en redes neuronales aplicadas sobre una ventana deslizante de 0,25 s. En cada ventana independientemente, se aplican dos redes neuronales de igual arquitectura, compuestas por capas convolucionales seguidas de capas recurrentes (LSTM). La primera red neuronal procesa la señal original, mientras que la segunda procesa la envolvente de la banda sigma. Sus salidas se concatenan con características

espectrales de la misma ventana para predecir la clase de la ventana usando capas *fully-connected*.

En tercer lugar se tiene el método propuesto en [7], abreviado aquí como DKL-KC, desarrollado para complejos K. En este método, aquellos picos negativos que superen un umbral se usan para determinar el centro de ventanas candidatas de 6 s. En dichas ventanas, se calcula la Transformada de Wavelet Discreta (*Discrete Wavelet Transform*, DWT) y se aplica una red neuronal *fully-connected*, en cuya salida se ajusta un proceso Gaussiano (combinación que recibe el nombre de *Deep Kernel Learning*, DKL). Finalmente, el método entrega la probabilidad de clase de la ventana de 6 s, y en consecuencia si el pico negativo candidato es parte de un complejo K, pero no predice una duración más precisa para cada complejo K detectado.

En cuarto lugar, se tiene SpindleU-Net [10], desarrollado para husos de sueño (aunque podría aplicarse también a complejos K), basado en una red neuronal convolucional que procesa la señal EEG en el tiempo en segmentos de 20 s. La red neuronal es una versión 1D de la arquitectura U-Net [67], con un decodificador modificado que modula las conexiones saltadas del codificador antes de ser usadas, por medio de compuertas de atención de grilla (no confundir con atención secuencial, como la de [68]). Este método predice una segmentación densa del segmento de 20 s, i.e., asigna una probabilidad de clase a cada muestra para ser binarizada con un umbral.

Por último, se tiene el método propuesto en [69], desarrollado para husos de sueño, basado en una red neuronal convolucional aplicada sobre ventanas deslizantes. Específicamente, se deslizan ventanas con tres duraciones distintas: 0,5 s, 0,75 s, y 1 s. En cada ventana se aplica la misma red neuronal convolucional con *spatial pyramid pooling*, i.e., se usa el promedio temporal de los filtros en varios niveles de resolución como entrada para capas *fully-connected* que predicen la clase de la ventana. Así, en cada muestra de la señal se tienen tres probabilidades, una por cada ancho de ventana, y se mantiene la máxima. Finalmente, la probabilidad de cada muestra se binariza con un umbral. A diferencia de los otros cuatro método descritos, este método no fue evaluado en MASS-SS2, sino que solo en DREAMS, por lo que es difícil establecer cómo se compara con ellos.

2.4.3. Limitaciones de los detectores existentes

En la mayoría de los detectores propuestos se han identificado riesgos de sobre-ajuste, lo que dificulta la interpretación de sus resultados. Los riesgos consisten en no reportar claramente el proceso de partición de los datos, no asegurar la independencia del diseño con el desempeño (i.e., el conjunto de validación se mezcla con el conjunto de prueba), o bien no reportar claramente el origen de las decisiones de diseño. Gran parte de este problema se puede atribuir a los pocos datos históricamente disponibles, lo que a su vez aumenta la varianza de los resultados. Si bien dicha escasez se ha ido atenuando con MASS-SS2, las publicaciones recientes que usan MASS-SS2 no están libres de los riesgos de sobre-ajuste porque diseñan (u omiten cómo diseñan) y evalúan el desempeño usando todos los datos de MASS-SS2 en una validación cruzada de tres subconjuntos (entrenamiento, validación y prueba). En cada partición de dicho esquema, el desempeño del conjunto de validación y del conjunto de prueba son en efecto independientes, pero no así el desempeño *promedio* de validación y el desempeño *promedio* de prueba, debido a la alternancia de todos los datos. Por esta razón, elegir la arquitectura o los hiperparámetros en base al desempeño promedio de validación (bajo dicho esquema) y fijar el mismo diseño para todas las particiones es un riesgo de sobre-ajuste.

Los detectores tradicionales (basados en características) tienen la ventaja de la interpretabilidad ya que sus predicciones pueden ser explicadas al inspeccionar sus características. Es más, aquellos que ni siquiera usan aprendizaje de máquinas para las reglas de clasificación son más transparentes aún y brindan un control más sencillo al usuario para hacer reajustes. Sin embargo, la literatura evidencia que dichos detectores son poco robustos. En primer lugar, los detectores que usan características básicas (y en consecuencia intuitivas) incurren en muchos falsos positivos [3]. Está bien establecido, por ejemplo, que en husos de sueño no basta utilizar la amplitud sigma, y es necesario usar características complementarias tales como la potencia relativa entre la banda sigma y otras bandas. Por otro lado, el enfoque de clasificación de ventanas, ya sean candidatas o deslizantes de ancho fijo, induce una limitación importante en la resolución y el contexto. Por un lado, el contexto está limitado al interior de cada ventana, lo que no permite poner en perspectiva la forma de onda observada si se desea que el tamaño de la ventana represente la duración del evento (en ventanas candidatas) o permita indicar su ubicación con buena resolución (en ventanas deslizantes). Además, una ventana deslizante pequeña permite en principio localizar mejor el inicio y el fin de los eventos, pero deja muy poca señal para calcular características, y solo algunas clases de eventos admiten ventanas más pequeñas que su duración típica: se podría hacer en husos de sueño por ser una oscilación, pero no en complejos K donde hay que observar el evento completo para identificarlo. Una ventana mediana, del orden de la duración del evento, evitaría cortar el evento o lo haría poco (ya que el evento es de duración variable), y podría abarcar señal suficiente para calcular buenas características, pero no brinda contexto circundante. Una ventana significativamente mayor que la duración del evento entregaría mayor información y contexto circundante, pero es difícil localizar con exactitud el evento por medio de la clasificación de dichas ventanas.

En general, el uso de características induce en sí mismo una limitación porque es complejo desarrollar buenas características. Por ejemplo, luego de varios años de literatura acumulada al respecto, se ha desarrollado el detector A7 [59], un detector de husos de sueño competitivo con *F1-score* consistentemente por encima de 70 % y que está completamente basado en características y reglas expertas. Esto es difícil de replicar en otras clases de eventos menos investigados, y el esfuerzo invertido en el desarrollo de características para una clase de eventos no se puede generalizar o transferir a los demás. Es decir, no se puede *re-ajustar* el detector para detectar otra clase de eventos sin requerir un rediseño de sus características, en lugar de solo modificar el valor de sus parámetros. Esto hace que el desarrollo de detectores basados en características sea lento y poco sistemático, sobre todo porque es difícil traducir manualmente el conocimiento experto de este problema. El uso de aprendizaje de máquinas sobre las características diseñadas no resuelve esta dificultad, ya que solo reemplaza el diseño de reglas expertas. Estas limitaciones motivan la exploración de detectores basados en aprendizaje profundo que, manteniendo el mismo diseño, pueden aprender a partir de los datos las características necesarias para detectar una variedad de eventos.

Los detectores de husos de sueño y complejos K basados en aprendizaje profundo alcanzan actualmente el estado del arte, pero no están libres de limitaciones. En estos detectores se han identificado los mismos riesgos de sobre-ajuste mencionados anteriormente, que afectan la arquitectura diseñada o los hiperparámetros seleccionados. Además, se han identificado dos brechas generales que son específicas para estos métodos: **ausencia de un procesamiento secuencial del contexto** y **ausencia de una validación extensa**.

En primer lugar, el contexto es una fuente importante de información para la detección. En el manual oficial de la AASM, los husos de sueño y los complejos K no solo son descritos en base a su

morfología, sino que también en base a su apariencia en comparación con la actividad circundante [1]. Por ejemplo, en ninguno de los dos eventos se tiene un criterio de amplitud absoluta. En su lugar, se requiere que sean *sobresalientes* en el EEG. Además, la dinámica de la señal circundante puede ayudar a discriminar falsos positivos al indicar la presencia de actividad incompatible o de artefactos. A pesar de ello, los métodos propuestos en [9] y [69] usan una ventana deslizante de 0,25 s y 0,5–1 s, respectivamente, e ignoran toda la señal fuera de la ventana para la clasificación. Por lo tanto, no usan el contexto que rodea al evento para la predicción, al contrario de lo realizado por expertos. Además, el enfoque de ventana deslizante es ineficiente debido a cálculos redundantes. Los métodos restantes, es decir, los propuestos en [8], [7] y [10], usan un contexto mayor para sus predicciones, de 20 s, 6 s, y 20 s respectivamente. Sin embargo, ninguno de estos métodos procesa secuencialmente dicho contexto usando capas neuronales diseñadas específicamente para eso, como las capas de auto-atención [68] o las capas recurrentes. El contexto se agrega convolucionalmente en [8] y [10], y con capas *fully-connected* en [7]. La ausencia de procesamiento secuencial podría limitar la capacidad de explotar el contexto para analizar de forma precisa la dinámica temporal del evento y su entorno (mejorando su discriminación), sobre todo la dinámica de la vecindad inmediata para determinar los instantes en que este inicia y termina (mejorando su localización).

En segundo lugar, una desventaja de usar métodos basados en aprendizaje profundo en lugar de métodos basados en características es la ausencia de una explicación detallada de sus predicciones. Esto es resultado de su comportamiento de caja negra: si bien el cálculo de las características aprendidas es transparente, no son interpretables en términos útiles para los expertos. Poder justificar predicciones ha sido ampliamente reconocido como un factor importante en áreas como la medicina [70]. Por lo tanto, la ausencia de interpretabilidad puede provocar que estos métodos, aún cuando son el estado del arte, no sean utilizados. Para evitar este problema, se puede llevar a cabo una validación extensa del detector desarrollado para evaluar si su comportamiento se alinea con las expectativas que tiene el experto de un buen detector, en diversos escenarios y dimensiones. Como resultado, el experto ganaría una intuición del funcionamiento del detector, de sus rangos de operación, y de la calidad de sus detecciones sobre diversas señales. Con esto, se podría compensar su falta de comprensión de los mecanismos que gobiernan las predicciones, aumentando la probabilidad de su uso en la práctica.

2.4.4. Causalidad: Limitaciones fundamentales

Según la revisión bibliográfica, aún hay oportunidades de mejora. Sin embargo, existen limitaciones fundamentales que surgen tras un análisis causal del problema. Como se argumenta en [71], existen tareas causales ($\mathbf{x} \Rightarrow \mathbf{y}$) y anticausales ($\mathbf{y} \Rightarrow \mathbf{x}$), y la categoría determina el efecto de algunos métodos de entrenamiento y de los desplazamientos del conjunto de evaluación con respecto al conjunto de entrenamiento. El problema de esta tesis es una tarea causal porque el experto determina la existencia y los instantes de inicio y fin ($\mathbf{y}$) en base a los patrones que se observen en la señal recolectada ($\mathbf{x}$).[8]

Los detectores de husos de sueño y complejos K enfrentan el problema de bases de datos pequeñas, motivando la exploración de métodos que compensen la falta de datos como entrenamientos semi-supervisados y técnicas de aumento de datos. Típicamente, el entrenamiento semi-supervisado aprovecha datos no etiquetados (i.e., información proveniente solo de $p(\mathbf{x})$) para mejorar el modelo bajo el supuesto de que entradas similares se etiquetan de forma similar. En una

[8]En general, según se argumenta en [71], las tareas de segmentación son causales.

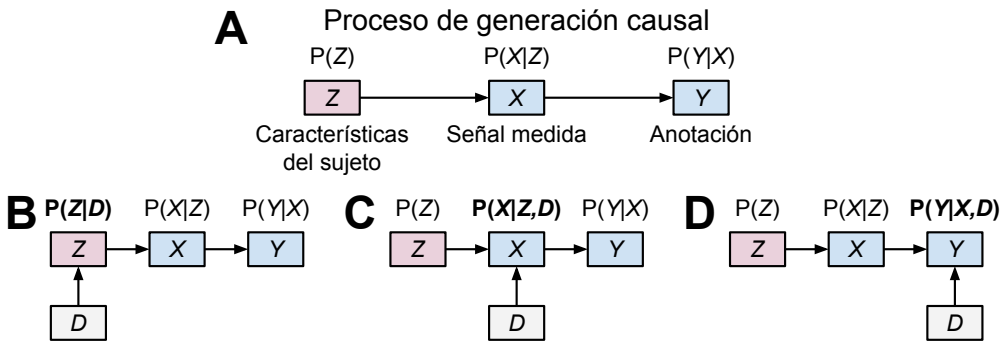

Figura 2.4: Desplazamientos de los datos. (**A**) En el proceso de generación de los datos para la detección de husos de sueño o complejos K, las señales medidas (X) causan las anotaciones (Y), es decir, es una tarea causal. Además, las señales medidas dependen de variables no observadas en los datos (Z), propias de un individuo o de una demografía. Sobre este proceso, pueden ocurrir desplazamientos de la población (**B**), de la adquisición de las señales (**C**), o de las anotaciones (**D**). En cada desplazamiento, la distribución de la variable afectada también depende del conjunto de datos D (de entrenamiento o de evaluación) que se esté utilizando. Adaptado de [71].

tarea causal, $p(\mathbf{x})$ —la causa— y $p(\mathbf{y}|\mathbf{x})$ —el mecanismo— son independientes, implicando que $p(\mathbf{x})$ no entrega más información de $p(\mathbf{y}|\mathbf{x})$. En otras palabras, la semi-supervisión no beneficia generalmente el desempeño en tareas causales. Por otro lado, las técnicas de aumento de datos generan más muestras posibles a partir de muestras conocidas, entregando más información de la distribución conjunta $p(\mathbf{x}, \mathbf{y})$. Por lo tanto, estas técnicas son adecuadas para mejorar el desempeño en tareas causales y anticausales.

Como se adelantó en la Sección 2.1.7, un detector de husos de sueño y complejos K puede ser afectado por desplazamientos en la población, en la adquisición, y en la anotación. Cada uno de estos desplazamientos afecta la tarea causal de interés en momentos diferentes del proceso de generación de los datos (ver Figura 2.4). A continuación se describe el efecto de cada desplazamiento en un modelo entrenado solo en el conjunto de entrenamiento, es decir, sin acceso a muestras etiquetadas del conjunto de evaluación y en consecuencia sin la posibilidad de re-entrenamiento.

Efecto de un desplazamiento de población. Es posible aprender un detector que se puede aplicar ante desplazamientos en la población en la medida en que la distribución de entrenamiento cubra el soporte de la distribución de evaluación (algo en lo que puede ayudar el aumento de datos en bases de datos pequeñas), ya que no es posible conocer el desempeño del modelo al extrapolar más allá del soporte de la distribución de entrenamiento. Sin embargo, las imperfecciones del entrenamiento podrían introducir dependencias falsas que ligan el modelo de inferencia estimado con la población (i.e., se desplaza la probabilidad condicional estimada en dirección a la distribución empírica observada en la población de entrenamiento), por lo que el desempeño podría empeorar en la nueva población. En este caso, el desplazamiento se puede compensar durante el entrenamiento con una reponderación de los ejemplos igual a $p_{\text{test}}(\mathbf{x})/p_{\text{train}}(\mathbf{x})$ [72]. Esto es posible solo cuando se tiene acceso a muestras de $p_{\text{test}}(\mathbf{x})$ y se puede estimar fidedignamente la razón en cuestión a partir de las muestras, lo que en datos de alta dimensionalidad como el EEG es un problema en sí mismo.

Efecto de un desplazamiento de adquisición. Análogamente al caso anterior, el principal problema es un comportamiento impredecible cuando el conjunto de evaluación no tiene un soporte contenido en el del conjunto de entrenamiento. En primer lugar, se puede atenuar su efecto a través

de la normalización de los datos, i.e., transformar las señales a una representación estándar, para maximizar el traslape de las distribuciones. Además, si las nuevas condiciones de adquisición son conocidas, se puede utilizar un aumento de datos para intentar simular dichas condiciones durante el entrenamiento y así expandir el soporte de la distribución de entrenamiento a la región de interés.

Efecto de un desplazamiento de anotación. A pesar de que el detector se puede hacer más robusto a los dos desplazamientos descritos anteriormente, que afectan la estimación del modelo principalmente a través de $p(\mathbf{x})$, los desplazamiento de anotación, que modifican directamente la distribución condicional $p(\mathbf{y}|\mathbf{x})$ a estimar, presentan un desafío mayor y no se conoce una solución a menos que se hagan supuesto fuertes acerca de los mecanismos de dicho desplazamiento. En consecuencia, para compensar el desplazamiento se requiere una recalibración manual de las salidas o la disponibilidad de anotaciones provenientes de la nueva distribución. En la práctica, esto implica que es difícil aprender un detector que generalice correctamente a otra base de datos sin un reajuste si el criterio experto de anotación cambia a través de un mecanismo desconocido, un fenómeno usual al cambiar de anotador o grupo de anotadores como se encontró en la revisión bibliográfica.

Capítulo 3

Metodología

3.1. Vista general del problema

Datos. Sea $S = \{s_i\}_{i=1}^{N}$ una colección de N sujetos. Durante una noche completa del sujeto s_i, se registra una señal de EEG $\mathbf{x}^i \in \mathbb{R}^{T_i \times C}$ de T_i muestras y C canales con frecuencia de muestreo f_s. Además, existe una máscara binaria de anotación $\mathbf{m}^i \in \mathbb{R}^{T_i}$ que indica qué intervalos de la señal deben ser anotados. Un experto o un grupo de expertos, al analizar la señal de EEG del sujeto s_i, determina una colección de M_i eventos $Y^i = \{(t_j^{\text{start}}, t_j^{\text{end}})\}_{j=1}^{M_i}$, donde el evento j-ésimo inicia en t_j^{start} y termina en t_j^{end}. Dichos eventos consisten en husos de sueño o en complejos K según la aplicación.

Problema. Se desea diseñar un modelo f_θ de parámetros θ tal que $Y^i \approx f_\theta(\mathbf{x}^i, \mathbf{m}^i)$. Es decir, se desea que el modelo f_θ (detector) extraiga automáticamente (detecte) una colección de eventos (detecciones) $\hat{Y}^i = f_\theta(\mathbf{x}^i, \mathbf{m}^i)$, a partir de la señal EEG $\mathbf{x}^i$ —proveniente de un sujeto **no visto** durante su diseño y ajuste— que aproxime la colección experta Y^i.

Solución. El enfoque utilizado para concretar este sistema consiste en predecir segmentaciones densas a partir de ventanas de señal de largo fijo. A partir de una señal $\mathbf{x}$ de largo T, se extraen consecutivamente ventanas de largo T_w. Sea $\mathbf{x}_{k:k+T_w}$ la ventana de señal que comienza en la muestra k. El modelo f_θ procesa dicha ventana y entrega $\mathbf{b}_{k:k+T_w} = f_\theta(\mathbf{x}_{k:k+T_w})$, una salida binaria de largo T_w en donde 1 indica que la muestra se encuentra al interior de un evento mientras que 0 indica que no. Luego de aplicar el modelo a todas las ventanas, se obtiene la salida binaria $\mathbf{b}$ para la señal completa, que se transforma a una colección de eventos $\{(t_j^{\text{start}}, t_j^{\text{end}})\}_{j=1}^{M}$ al extraer los intervalos de salidas contiguas iguales a 1. Dicha colección se post-procesa de acuerdo a criterios específicos del tipo de evento de interés, y, de ser necesario, se remueven elementos para asegurar que $\forall j,\ \exists k \in [t_j^{\text{start}}, t_j^{\text{end}}],\ \mathbf{m}_k = 1$. La colección final obtenida es la detección $\hat{Y}$.

Alcances. Por simplicidad, se usa un solo canal ($C = 1$) con frecuencia de muestreo $f_s = 200$ Hz, y ventanas de predicción de 20 s ($T_w = 4000$). Además, solo se consideran anotaciones durante la etapa N2. Por lo tanto, $\mathbf{m}$ indica intervalos en etapa N2, o bien intervalos anotados si es que solo una porción de la etapa N2 fue anotada por expertos.

3.2. Evaluación del modelo

3.2.1. Métricas de desempeño

Significancia estadística. La significancia estadística de la diferencia entre las medias de dos muestras independientes de resultados (e.g., entre los desempeños promedios de dos modelos) se evalúa usando el t-test de varianzas desiguales de Welch [73]. A menos que se indique lo contrario, se usa un nivel de significancia de $P < 0{,}05$.

Análisis por eventos basado en IoU. Se sigue el análisis por eventos de [6], en donde dos eventos se comparan usando la razón entre la intersecciόm y la unión (*intersection over union*, IoU). Dado un evento real $A = (t_A^{\text{start}}, t_A^{\text{end}})$ y una detección $B = (t_B^{\text{start}}, t_B^{\text{end}})$, el IoU entre A y B está dado por

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \text{máx}\left(0,\ \frac{\text{mín}(t_A^{\text{end}}, t_B^{\text{end}}) - \text{máx}(t_A^{\text{start}}, t_B^{\text{start}})}{\text{máx}(t_A^{\text{end}}, t_B^{\text{end}}) - \text{mín}(t_A^{\text{start}}, t_B^{\text{start}})}\right), \tag{3.1}$$

en donde $\text{IoU}(A, B) \in [0, 1]$ con un valor ideal de 1 alcanzado solo cuando $A = B$. Con esta medida se encuentra un apareamiento de máximo IoU entre una colección de eventos reales y una colección de detecciones, en donde se recorre en orden creciente cada evento real y se aparea con la detección con la que tiene máximo IoU sujeto a un IoU de apareamiento mayor a cero (i.e., que exista un traslape) y a que la detección no se encuentre apareada con otro evento real. Así se asegura que cada evento es detectado a lo más por una detección, y que cada detección detecta a lo más a un evento real. Al completar el apareamiento, se tiene un verdadero positivo (*true positive*, TP) sí y solo sí el IoU de apareamiento es igual o mayor a un umbral $\tau_{\text{IoU}} \in (0, 1)$. De lo contrario, los eventos reales constituyen un falso negativo (FN) y las detecciones constituyen un falso positivo (FP). Eventos reales y detecciones sin pareja siempre son FN y FP (i.e., se tratan con un IoU de apareamiento igual a cero por definición). En consecuencia, la matriz de confusión (i.e., los valores TP, FN y FP) es una función del umbral τ_{IoU}, así como todas las métricas derivadas de ella.[1]

Métricas para una colección de eventos. Usando este marco, se calcula

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3.3}$$

$$\text{F1-score} = 2\left(\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right)^{-1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{3.4}$$

Al igual que en [6], se reportan estas métricas usando $\tau_{\text{IoU}} = 0{,}2$. Adicionalmente, para independizar la métrica del umbral específico, se propone una versión *average* definida como el área bajo la curva formada por la métrica en función de $\tau_{\text{IoU}} \in (0, 1)$. Es decir,

$$\text{Average Metric} = \int_0^1 \text{Metric}(\tau_{\text{IoU}}) \mathrm{d}\tau_{\text{IoU}}, \quad \text{Average Metric} \in [0, 1]. \tag{3.5}$$

En particular, la métrica *Average F1-score* (AF1) se propone como métrica central para guiar el diseño y ajuste del modelo. Al integrar el *F1-score* en todos los umbrales, un AF1 alto requiere

[1] No se usan verdaderos negativos por no estar bien definidos en este problema, por lo que tampoco se usan métricas que dependan de dicha cantidad como la especificidad.

dos aspectos de un buen detector: detecciones con localizaciones correctas (medido con un bajo τ_{IoU}) y con duraciones correctas (medido con un alto τ_{IoU}). Por último, sea $\mathcal{M}(Y, \hat{Y})$ el conjunto de todos los apareamientos (A, B) generados (i.e., $\mathcal{M}$ no contiene a los eventos reales y detecciones que quedan sin un par), y sea $\mathrm{IoU}(\mathcal{M})$ el conjunto de los valores de IoU de los pares en $\mathcal{M}$. Para analizar directamente la exactitud con la que se predicen los instantes de inicio y fin de los eventos, se reporta el histograma de $\mathrm{IoU}(\mathcal{M})$, y se propone reportar junto a las métricas de detección (*recall*, *precision*, y *F1-score*) una cuarta métrica que corresponde al promedio $\mathbb{E}(\mathrm{IoU}(\mathcal{M}))$, abreviada aquí como **mIoU**.

Métricas para una colección de sujetos. El análisis por eventos planteado en [6] se extiende aquí para introducir más control en el análisis de colecciones de sujetos. Al nivel de la colección de sujetos, una métrica puede ser obtenida por un *macro-promedio* o por un *micro-promedio*. Además, la dispersión del desempeño puede ser analizada usando una *dispersión entre subconjuntos* y una *dispersión entre sujetos*. A continuación se definen estos conceptos.

Macro-promedio y micro-promedio. Sea $S = \{s_i\}_{i=1}^{N}$ una colección de N sujetos, cada uno con un EEG y anotaciones. Se hace un apareamiento en cada $s_i \in S$ independientemente y luego, dado un τ_{IoU}, se calcula la matriz de confusión TP_i, FP_i, y FN_i. Para una métrica cualquiera que sea función de la matriz de confusión, la métrica de la colección S se puede obtener por un macro-promedio o por un micro-promedio. La métrica por macro-promedio está dada por

$$\mathrm{Metric}^{\mathrm{macro}}(S) := \frac{1}{N} \sum_{i=1}^{N} \mathrm{Metric}\left(\mathrm{TP}_i, \mathrm{FP}_i, \mathrm{FN}_i\right). \tag{3.6}$$

Es decir, la métrica se calcula en cada sujeto y se promedian, consiguiendo que cada sujeto pondere lo mismo. Por otro lado, la métrica por micro-promedio está dada por

$$\mathrm{Metric}^{\mathrm{micro}}(S) := \mathrm{Metric}\left(\sum_{i=1}^{N} \mathrm{TP}_i, \sum_{i=1}^{N} \mathrm{FP}_i, \sum_{i=1}^{N} \mathrm{FN}_i\right). \tag{3.7}$$

Es decir, se combinan todas las matrices de confusión y se calcula una única métrica, consiguiendo que cada evento, en lugar de cada sujeto, pondere lo mismo. Los conceptos de macro-promedio y micro-promedio no son exclusivos de las métricas de detección, sino que se pueden extender a cualquier función de una colección de eventos que se defina como la sumatoria de una magnitud por evento, como puede el mIoU, la densidad de eventos, la amplitud media, entre otras. En general, el micro-promedio corresponde a realizar una única agregación global, mientras que el macro-promedio corresponde a una agregación por sujeto seguida de una agregación de estos resultados intermedios. En general, se opta por un macro-promedio a menos que la base de datos o el experimento no permita asegurar un número suficiente de ejemplos en todos los sujetos para obtener un resultado confiable.

Dispersión entre subconjuntos y entre sujetos. Para la misma colección S, se define la dispersión entre subconjuntos y la dispersión entre sujetos como dos maneras de analizar la dispersión del desempeño. La dispersión entre subconjuntos, generados por ejemplo como particiones de una validación cruzada, surge naturalmente al calcular la métrica por macro-promedio o por micro-promedio en cada subconjunto independientemente. Si se tienen N_F subconjuntos, cada uno con

una fracción del total N de sujetos, se obtienen N_F valores de la métrica, de donde se puede calcular la desviación estándar. Es decir, corresponde a la estadística típicamente esperada de una validación cruzada. Este tipo de evaluación muestra la variación del desempeño dentro de una colección al variar la submuestra utilizada o al ejecutar varios entrenamientos sobre los mismos subconjuntos. En otras palabras, mide la robustez del desempeño promedio. Por otro lado, la dispersión entre sujetos requiere calcular una métrica independientemente para cada sujeto. En consecuencia, este análisis es práctico solo cuando es posible o confiable obtener una métrica por sujeto. Dado que algunos sujetos podrían aparecer repetidos en la evaluación (e.g., por varias repeticiones de una validación cruzada), a cada sujeto se le asigna el desempeño promedio obtenido en todas sus apariciones, de modo que cada sujeto es representado por un solo escalar. De esta forma, se tienen N valores de la métrica, de donde se puede calcular la desviación estándar o bien graficar la distribución de valores.[2] Este tipo de evaluación muestra la robustez del desempeño ante nuevos sujetos e identifica si existen sujetos muy distantes con respecto al grupo general (i.e., anómalos).

3.2.2. Bases de datos

Se usan varios tipos de bases de datos para evaluar y analizar el modelo propuesto desde varias perspectivas. En primer lugar, se usan *bases de datos etiquetadas por expertos* para evaluar el desempeño del modelo y compararlo con el desempeño de otros detectores de la literatura. Dentro de esta categoría se encuentran las bases de datos MASS-SS2, MASS-MODA e INTA-UCH. Una fracción de MASS-SS2, especificada en la Sección 3.2.3, determina el diseño del modelo propuesto (i.e., arquitectura e hiperparámetros). En segundo lugar, se propone usar *bases de datos artificiales* para analizar el comportamiento del modelo en una situación artificial asociada a un comportamiento esperado. Dentro de esta categoría se encuentran las bases de datos con etiquetas artificiales (CAP-S1, CAP-S2 y CAP-A7) y la base simulada PINK. Por último, se usan grandes *bases de datos no etiquetadas*, en donde solo se conocen las etapas de sueño, para analizar cualitativamente las estadísticas de los parámetros de los eventos detectados durante la etapa N2 por un modelo ya entrenado. Dentro de esta categoría se encuentran las seis bases de datos del repositorio *National Sleep Research Resource* (NSRR) utilizadas en [49].

A continuación se describe cada una de las bases de datos mencionadas. Como se describe en la Sección 3.3.5, las señales son siempre divididas en páginas de 20 s para entrenar y se requiere seleccionar solo aquellas páginas *válidas* (e.g., en etapa N2). Sin embargo, algunas bases de datos usadas para entrenar usan otros esquemas de segmentación (e.g., páginas de 30 s para indicar etapas del sueño, o bloques de señal para ser anotados). Cuando sea el caso, se describe también la estrategia usada para permitir la división en páginas de 20 s. En todos los casos, cada sujeto posee un solo PSG de noche completa con etapas de sueño anotadas. Los datos con etiquetas expertas se resumen en la Tabla 3.1, los datos con etiquetas artificiales se resumen en la Tabla 3.2, y los datos sin etiquetas se resumen en la Tabla 3.3. La duración media y la densidad media (medida en eventos por minuto, epm) se calculan por macro-promedio, a excepción de MASS-MODA en donde se calculan por micro-promedio debido a sus registros individuales demasiado cortos.

[2]Si el desempeño de la base de datos se calcula usando un micro-promedio, o si el número de repeticiones de los sujetos no es el mismo para todos (esto último no ocurre en esta tesis), el promedio de estos N escalares no coincide con el promedio entre subconjuntos, que es el promedio reportado para una validación cruzada y en consecuencia corresponde al desempeño promedio reportado del modelo.

Tabla 3.1: Descripción de los datos con etiquetas expertas. «SS» es huso de sueño y «KC» es complejo K. E1 y E2 corresponden a los dos expertos que hicieron las marcas en MASS-SS2. Las siglas usadas para las bases de datos se definen en el texto principal.

	MASS-SS2-E1SS	-E2SS	-KC	MASS-MODA	INTA-UCH
Sujetos	15			180	10
Edad (años)	18–33			18–76	10
f_s^{original} (Hz)	256			256	200
Canal	C3-LE			C3-A2 o C3-LE	F4-C4
Evento	SS	SS	KC	SS	SS
Anotación	1 exp.			Consenso 31–42 exp.	1 exp.
Anotado	Etapa N2			Bloques de 115 s	Etapa N2
Tamaño útil (h)	60,01			24,97	32,94
Total eventos	9.990	21.846	8.781	5.272	12.237
Densidad (epm)	2,72	6,02	2,49	3,52	6,31
Duración (s)	0,83	1,20	0,73	0,84	1,24

Tabla 3.2: Descripción de los datos con etiquetas artificiales (i.e., detecciones de otros métodos) de husos de sueño. Las siglas usadas para las bases de datos se definen en el texto principal.

	CAP-S1	CAP-S2	CAP-A7
Sujetos	80		
Edad (años)	14–77		
f_s^{original} (Hz)	100–512		
Canal	C4-A1 o C3-A2		
Anotación	Detector S1	Detector S2	Detector A7
Anotado	Etapa N2		
Tamaño útil (h)	251,64		
Total eventos	75.237	58.417	51.597
Densidad (epm)	4,82	3,76	3,35
Duración (s)	1,02	1,02	1,00

Tabla 3.3: Descripción de los datos sin etiquetas del *National Sleep Research Resource* (NSRR). Las siglas usadas para las bases de datos se definen en el texto principal.

	CHAT	CCSHS	CFS	SHHS	MrOS	SOF	Combinación
Sujetos	1.053	513	711	5.648	2.862	437	11.224
Sexo (% mujeres)	52,7 %	49,5 %	55,0 %	52,3 %	0,0 %	100,0 %	40,9 %
Edad (años)	4–10	16–19	6–88	39–90	67–90	75–90	4–90
f_s^{original} (Hz)	200–512	128	128–256	125–250	256	128	125–512
Canal	C3-A2	C3-A2	C3-A2	C3-A2 o C4-A1	C3-A2	C3-A2	C3-A2 o C4-A1
Tamaño útil (h)	2.271,2	1.734,7	2.271,0	18.755,6	10.158,8	1.356,7	36.548,1

MASS-SS2. La base de datos *Montreal Archive of Sleep Studies* (MASS) [47], disponible públicamente, contiene un PSG de noche completa de 200 sujetos adultos con frecuencia de muestreo 256 Hz, y está dividida en cinco subconjuntos. El segundo subconjunto (MASS-SS2) contiene 19 adultos jóvenes (18–33 años) con etapas de sueño anotadas en páginas de 20 s según el antiguo estándar R&K. Además, se tienen anotaciones de husos de sueño por dos expertos (E1 y E2) y de complejos K por un experto (E1) durante la etapa N2 del canal C3-LE. De los 19 sujetos, solo 15 fueron anotados por E2. Además, E1 usó procedimientos estándar de anotación mientras que E2 tuvo acceso a la señal EEG filtrada en la banda sigma (11–16 Hz) y no usó una duración mínima, alejándose del procedimiento estándar. En general, esto provoca que el acuerdo entre E1 y E2 para husos de sueño sea muy bajo, con E2 marcando cerca del doble de husos que E1 y con duraciones más largas, haciendo de E2 un anotador más sensible. Por ejemplo, si E2 se trata como un detector de E1, se obtiene por macro-promedio: *F1-score* 56,4 %, *recall* 95,5 %, *precision* 41,2 %, y mIoU 60,8 %. Dado que E1 $\cup$ E2 $\approx$ E2 y E1 $\cap$ E2 $\approx$ E1, se decidió no experimentar con consensos de ambos expertos. De los 19 registros disponibles, solo se utilizan los 15 registros que fueron anotados por ambos expertos para los experimentos. Para facilitar la comparación con desempeños reportados en la literatura, en un apartado de los resultados se reporta el desempeño en los 19 registros (Tabla 4.2). Debido a que E2 no utilizó una duración mínima para los husos de sueño, las anotaciones más cortas que 0,3 s fueron eliminadas como en [6]. Para efectos de evaluación, MASS-SS2 contribuye con tres bases etiquetadas pero a partir de las mismas señales de EEG: husos de sueño según E1 (MASS-SS2-E1SS), husos de sueño según E2 (MASS-SS2-E2SS), y complejos K según E1 (MASS-SS2-KC).

MASS-MODA. En [23] se seleccionaron 180 sujetos de MASS y se anotaron husos de sueño en extractos de señal usando un consenso de varios expertos a través de la plataforma *Massive Online Data Annotation* (MODA). Estas anotaciones están disponibles públicamente y al ser combinadas con las señales de MASS forman la base de datos MASS-MODA. La demografía está dividida en dos *fases*: la *fase 1* consiste en 100 adultos jóvenes (edad promedio 24,1 años) y la *fase 2* consiste en 80 adultos mayores (edad promedio 62,0 años). Para la anotación, se extrajeron bloques de señal de 115 s en etapa N2 sin artefactos. En la fase 1, en 1 sujeto se extrajeron 2 bloques, en 84 sujetos se extrajeron 3 bloques, y en 15 sujetos se extrajeron 10 bloques. Por otro lado, en la fase 2, en 65 sujetos se extrajeron 3 bloques y en 15 sujetos se extrajeron 10 bloques. Considerando ambas fases, hay 30 sujetos con 10 bloques de señal anotada (19,2 min) y 150 sujetos con 2 o 3 bloques de señal anotada (3,8–5,8 min). El canal anotado es C3-A2, o bien C3-LE cuando el canal A2-LE no está disponible para usarse como referencia. Durante la anotación, se usaron 42 expertos en la fase 1 y 31 expertos en la fase 2. Más del 95 % de la señal disponible para anotar fue vista por al menos 3 expertos, resultando en una anotación por consenso de expertos de alta calidad. Para efectos de evaluación, la máscara de anotación válida para esta base de datos está dada por los bloques de señal anotados en lugar de las etapas N2 de los registros. Además, por simplicidad, se consideraron 2,5 s de señal adicional en cada borde de los bloques para formar intervalos de 120 s, compatible con 6 páginas consecutivas de 20 s.

INTA-UCH. La base de datos del Instituto de Nutrición y Tecnología de los Alimentos (INTA) de la Universidad de Chile, abreviada por INTA-UCH, es una base privada recolectada entre abril del 2002 y septiembre del 2005 con un equipo Cadwell Easy II. Consiste en un PSG de noche completa muestreado a 200 Hz de 10 niños (10 años), con etapas de sueño anotadas en páginas de

30 s según el antiguo estándar R&K.[3] Se anotaron husos de sueño en el canal F4-C4 con un criterio de duración mínima de 0,5 s y separación mínima de 0,5 s. Cada anotación fue realizada por un solo experto, pero varios expertos participaron anotando diferentes porciones de la base de datos. Las páginas de 30 s fueron transformadas a páginas de 20 s que fueron asignadas a la etapa N2 (i.e., tratadas como *válidas*) si intersectan al menos parcialmente una anotación de etapa N2 de la segmentación original.

CAP-S1, CAP-S2 y CAP-A7. La base de datos de patrones alternantes cíclicos (*cyclic alternating pattern*, CAP) del sueño [74] consiste en un PSG de noche completa de 108 sujetos sanos y enfermos, con etapas de sueño anotadas en páginas de 30 s según el antiguo estándar R&K, y sin anotaciones de husos de sueño o complejos K. La frecuencia de muestreo varía por registro en el rango 100–512 Hz. De estos 108 sujetos, se seleccionan 80 (14–77 años) en base a la calidad de la señal, y se ignoran las páginas con amplitudes mayores a 300 μV.[4] Estos registros se usan para crear bases de datos anotadas artificialmente. Para generar anotaciones artificiales determinísticas de husos de sueño, se propone usar detectores basados en aplicar umbrales y reglas a características interpretables comúnmente utilizadas en la literatura, i.e., sin aprendizaje de máquinas. De esta forma, las anotaciones artificiales se pueden considerar aproximaciones del conocimiento experto del problema. Se consideran tres detectores: dos de implementación propia basados en reglas clásicas (detectores S1 y S2), y el detector A7 [59]. En los tres casos, las detecciones separadas por menos de 0,5 s son combinadas, y luego aquellas con una duración menor a 0,5 s o mayor a 3,0 s son removidas.[5] En breve, cada detector consiste en lo siguiente.

- El detector S1 es un algoritmo simple que consiste en calcular la amplitud instantánea de la banda sigma y aplicar sobre ella dos umbrales, τ_H y $\tau_L = \lambda_L \tau_H$ con $\lambda_L \in (0, 1)$. Los eventos se detectan si superan τ_H por al menos 0,3 s, en cuyo caso se fija su duración usando τ_L. Sus parámetros se ajustan para maximizar el AF1 en todo MASS-MODA, resultando en $\tau_H = 10$ μV y $\lambda_L = 0{,}86$. Con estos parámetros, S1 en MASS-MODA alcanza por sobre-ajuste: *F1-score* 63,8 %, *recall* 69,5 %, *precision* 58,9 %, y mIoU 69,1 %.
- El detector S2 es un algoritmo simple similar a S1 con la diferencia de que $\tau_H = \lambda_H \bar{\sigma}$, con $\bar{\sigma}$ la mediana de la amplitud instantánea de la banda sigma considerando todas las páginas en N2 de la noche. Esta estrategia es similar a la empleada por el detector basado en wavelet usado en [49] para analizar NSRR. Al igual que S1, sus parámetros se ajustan para maximizar el AF1 en todo MASS-MODA, resultando en $\lambda_H = 2{,}9$ y $\lambda_L = 0{,}8$. Con estos parámetros, S2 en MASS-MODA alcanza por sobre-ajuste: *F1-score* 65,0 %, *recall* 67,5 %, *precision* 62,7 %, y mIoU 70,7 %.
- El detector A7, propuesto para detectar husos de sueño en [59], reporta el mejor desempeño entre los detectores tradicionales según [59] y [23]. Está basado en cuatro características provenientes de estadísticos calculados en la banda sigma y entre la banda sigma y la señal original, en ventanas de 0,3 s con un paso de 0,1 s. Si bien las ventanas no permiten la extracción de características con dependencias de largo plazo, algunas características son normalizadas considerando sus estadísticas en un segmento mayor, de 30 s, permitiendo un

[3] La base de datos consiste en 11 archivos de PSG, ya que un sujeto tiene su noche separada en dos archivos. Para esta tesis se tuvo cuidado en unir ambas partes, resultando en 10 señales.

[4] Se descartaron sujetos con señales con alta interferencia, contaminaciones rítmicas extrañas, o ausencia de un pico significativo en la banda sigma durante la etapa N2.

[5] Se decide usar como duración mínima 0,5 s en lugar de 0,3 s debido a que los detectores clásicos entregan muchos falsos positivos en eventos muy cortos.

uso limitado del contexto. Se usan los parámetros del detector reportados en [59].

Las detecciones son generadas para el canal C4-A1, o el canal C3-A2 en su ausencia, durante toda la etapa N2. Se utiliza la misma transformación a páginas de 20 s que la descrita para INTA-UCH. Al revisar las distribuciones de duración, amplitud y frecuencia de las anotaciones artificiales, se encuentra que cumplen con los rangos esperados para husos de sueño (mencionados en la Sección 2.1.6). Además, las anotaciones de S1 muestran una distribución de amplitud homogénea entre sujetos, mientras que las de S2 y A7 muestran diferencias notorias entre sujetos, con un nivel de heterogeneidad similar entre los dos detectores.

PINK. Una variable aleatoria llamada *ruido rosado* tiene un espectro que decae como una ley de potencia ($f^{-a},\ a \geq 0$), una tendencia que también se observa en el espectro promedio de la actividad cerebral. Se propone generar una base de datos de señales artificiales compuesta únicamente por ruido rosado para analizar la respuesta de los modelos ante señales con estadísticas similares a las cerebrales pero que, por construcción, no poseen patrones estables y recurrentes a lo largo de la noche como husos de sueño o complejos K. Las señales se generan de tal forma que imitan el espectro promedio de MASS-SS2 durante la etapa N2 sin considerar el pico de la banda sigma.[6] El método de generación se resume en la Figura 3.1, y los detalles de su implementación se pueden consultar en el Anexo A. Con este método, se simulan 25 señales muestreadas a 200 Hz de 1 h de duración (i.e., 180 páginas de 20 s por señal).[7]

NSRR. Para analizar la calidad de las detecciones de husos de sueño obtenidas al aplicar el detector en datos externos, se usa una combinación de seis bases de datos del repositorio *National Sleep Research Resource* (NSRR) [75], abreviadas por CHAT [76], CCSHS [77], CFS [78], SHHS [79], MrOS [80] y SOF [81]. Cada una contiene un PSG de noche completa de diversas demografías, con etapas de sueño anotadas en páginas de 30 s según el antiguo estándar R&K, y sin anotaciones de husos de sueño o complejos K. Algunas bases contienen un segundo PSG para sus sujetos, pero aquí solo se considera la primera medición. En total, hay 11.630 sujetos disponibles, cuya frecuencia de muestreo varía por registro en el rango 125–512 Hz. Siguiendo el estándar para estudiar husos de sueño, solo se consideran páginas en etapa N2 del canal C3-A2, o del canal C4-A1 cuando el canal anterior está contaminado. A diferencia de [49], no se transforman las señales para corregir artefactos. En cambio, para disminuir el efecto de artefactos, se miden varias características de la señal en cada página y se quitan del análisis aquellas que escapan de los rangos observados en MASS. Específicamente, se usan los siguientes criterios para determinar la validez de una página:

- La amplitud debe permanecer en $[-200,\ 200]\ \mu$V (basado en el percentil 98 de los complejos K en MASS-SS2-KC).
- La desviación estándar debe estar en $[5{,}0895,\ 37{,}4640]\ \mu$V (rango observado en MASS-MODA).
- Se calcula el espectro $p(f)$ como el promedio de la FFT en ventanas de 2 s, y se ajusta en dicho espectro una ley de potencia $\hat{p}(f) = af^b$ por regresión lineal, para $f \in [2, 10) \cup$

[6] El pico de la banda sigma (11–16 Hz) se remueve porque no corresponde a un ritmo de fondo del EEG sino que a patrones discretos (husos de sueño). Si se mantuviera al generar ruido, se generaría una señal con alta actividad sigma de forma uniforme a lo largo del tiempo.

[7] A las 180 páginas mencionadas se agrega una página extra de señal en cada borde para entregar contexto en caso de ser requerido, dando un total de 182 páginas de señal en donde solo 180 son usadas como segmentos de predicción.

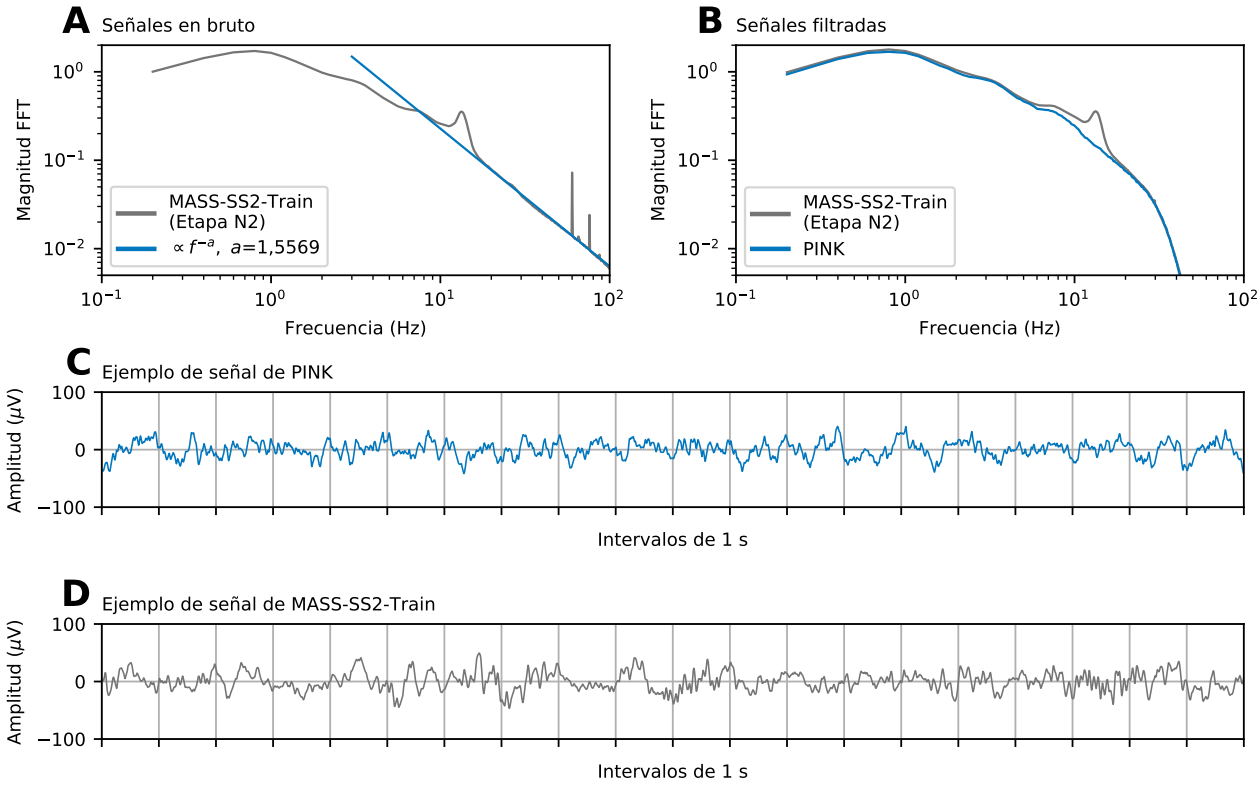

Figura 3.1: Simulación de ruido rosado para base de datos PINK. (**A**) Se calcula el espectro promedio de las señales en bruto de MASS-SS2-Train (porción de MASS-SS2 usada para diseño) durante la etapa N2, y el pico en 11–16 Hz se remueve usando un ajuste de ley de potencia. (**B**) Las señales artificiales de PINK son ruido blanco transformado para copiar el espectro modificado de MASS-SS2. Luego del preprocesamiento, ambas bases de datos tienen un espectro promedio prácticamente igual fuera de la banda 11–16 Hz. (**C** y **D**) A modo de comparación, se muestra un extracto de 20 s de PINK y de MASS-SS2-Train.

$(17, 30]$ Hz. Luego se impone que las características de este ajuste se encuentren dentro del rango observado en MASS-MODA:

- La escala a debe estar en $[1{,}2120,\ 34{,}8881]$.
- El exponente b debe estar en $[-1{,}8910,\ -0{,}6305]$.
- La desviación $p(f)/\hat{p}(f)$ para $f \in [2, 30]$ Hz debe ser a lo más 8,0416.
- El coeficiente de ajuste R^2 de la regresión para $f \in [2, 10) \cup (17, 30]$ Hz debe ser al menos 0,6996.

Finalmente, se descartan aquellos sujetos con menos de 60 minutos de etapa N2 válida, o que no poseen información de edad o sexo. Al final del proceso de selección, se preserva el 96,5 % de los sujetos y casi el 90 % de las páginas N2. Específicamente, se tienen 11.224 sujetos que abarcan el rango etario 4–90 años, de los cuales un 40,9 % son mujeres. Por lo tanto, la base de datos NSRR provee de un gran volumen de datos para obtener estadísticas robustas. En total, consiste en 36.548,1 h de señal, que corresponde a 145 veces CAP, 609 veces MASS-SS2, y 1.465 veces MASS-MODA. Para NSRR, no es necesario convertir las páginas de 30 s a páginas de 20 s porque no se usa para entrenar.

3.2.3. Partición de los datos para la evaluación

Todas las particiones de los datos en conjuntos de entrenamiento, validación y prueba se hacen asegurando que no se comparten sujetos entre los conjuntos. En general, se usan esquemas de validación cruzada de k particiones en donde se alternan los tres conjuntos. Específicamente, se obtiene una permutación aleatoria de los N sujetos (según una semilla fijada de antemano), y en ella se forman secuencialmente k grupos de $\lceil N/k \rceil$ sujetos. Si $k\lceil N/k \rceil > N$, se completa el último grupo usando los primeros sujetos de la permutación obtenida por la siguiente semilla. En cada iteración del esquema, un grupo actúa como conjunto de prueba, y el conjunto de validación corresponde al conjunto de prueba de la siguiente iteración del esquema de forma circular, i.e., el conjunto de validación de la k-ésima iteración es el conjunto de prueba de la primera iteración. De esta forma, todos los sujetos participan alguna vez como validación o prueba. Se decide usar $k = 5$. Para el caso de MASS-MODA, como cada sujeto puede pertenecer a una de cuatro categorías (con o sin 10 bloques; fase 1 o fase 2), la validación cruzada es estratificada en dichas categorías.

Solo en MASS-SS2, por ser la base de datos utilizada para **diseñar el detector**, se utiliza un esquema diferente. En primer lugar, de los $N = 15$ sujetos disponibles, se selecciona y aparta un conjunto de prueba representativo de 4 sujetos (MASS-SS2-Test) (ver Anexo B). El conjunto de los 11 sujetos restantes (MASS-SS2-Train) es el **único conjunto que determina el diseño del detector**, disminuyendo el riesgo de sobre-ajuste. Así, en MASS-SS2 se usan tres esquemas:

- Para experimentos de diseño, solo se usa MASS-SS2-Train, para así obtener métricas y tomar decisiones independientes del conjunto MASS-SS2-Test. Se usa un esquema de validación cruzada de 5 particiones. Por simplicidad, el sujeto 1 siempre está en el conjunto de entrenamiento y los 10 restantes se distribuyen según el esquema de validación cruzada $\mathrm{train} : \mathrm{val} : \mathrm{test} = 6 : 2 : 2$.
- Para evaluar el desempeño final en MASS-SS2, MASS-SS2-Train se utiliza en un esquema de validación cruzada de 4 particiones con $\mathrm{train} : \mathrm{val} = 8 : 3$, y en cada partición se evalúa el modelo en el conjunto de prueba independiente dado por MASS-SS2-Test.
- Por completitud y para dar más opciones de comparación con la literatura, también se evalúa el desempeño haciendo participar a todos los sujetos de MASS-SS2, usando tanto $N = 15$ como $N = 19$ (i.e., incluyendo registros no anotados por E2), en el mismo esquema de validación cruzada de 5 particiones usado en las otras bases de datos. Sin embargo, ya que hay sujetos de MASS-SS2-Train entre los sujetos de prueba, **el desempeño obtenido no es independiente** de las decisiones de diseño.

En MASS-SS2 con $N = 15$, MASS-MODA, e INTA-UCH, se generan tres ciclos de validación cruzada para obtener 15 particiones distintas, en donde cada sujeto se usa tres veces como sujeto de prueba y validación. En las tres bases de datos basadas en CAP, por su mayor tamaño, solo se usa un ciclo para obtener 5 particiones distintas, en donde cada sujeto se usa una sola vez como sujeto de prueba y validación. En MASS-SS2 con $N = 19$, se generan cuatro ciclos pero solo se usan las primeras 19 particiones, en donde cada sujeto se usa cuatro veces como sujeto de prueba y validación. En el caso de la evaluación en MASS-SS2-Test, se generan tres ciclos pero solo se usan las primeras 11 particiones, para que así cada sujeto de MASS-SS2-Train sea utilizado tres veces como sujeto de validación. En todos los casos, las semillas usadas son los números naturales partiendo del 0. Finalmente, en todas las bases de datos se calculan las métricas de las particiones por macro-promedio, a excepción de MASS-MODA en donde se calculan por micro-promedio

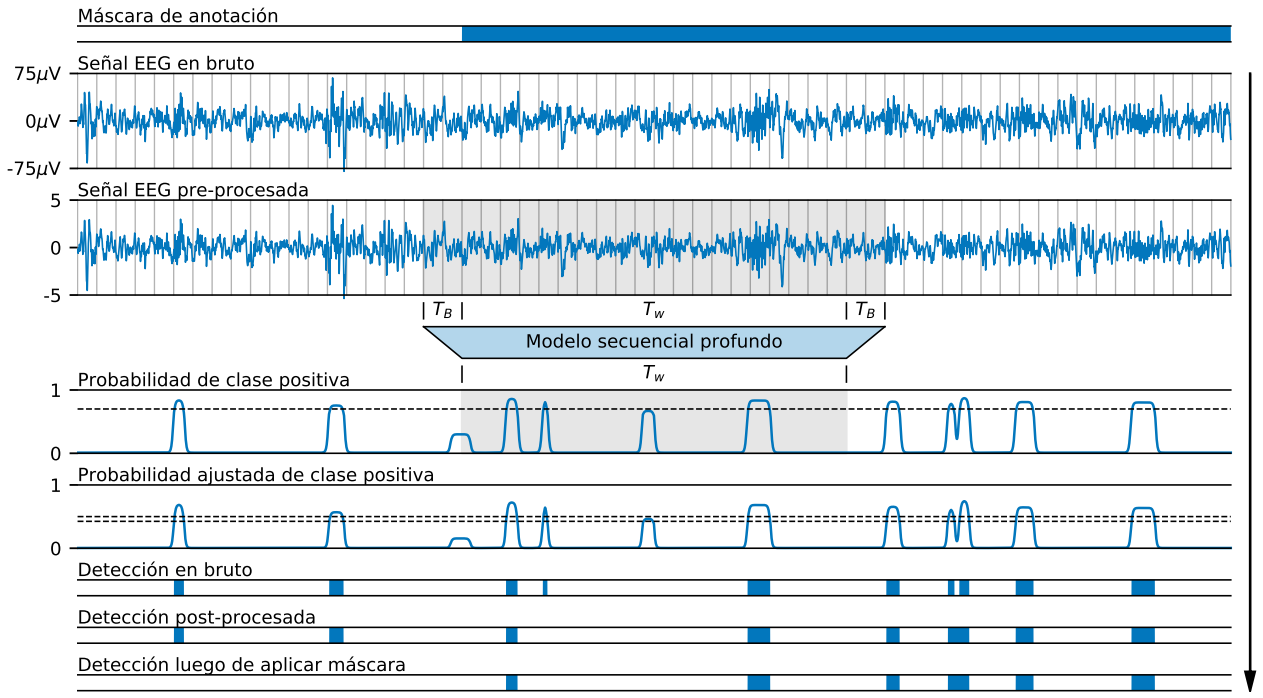

Figura 3.2: Vista global del método de detección propuesto llamado REDv2. El EEG se divide en ventanas de T_w muestras, y en cada una se aplica el modelo secuencial profundo para generar una segmentación densa. Para evitar efectos de borde, el modelo utiliza T_B muestras adicionales a cada lado. En la señal se muestran marcadores cada 1 s. El valor de T_B depende la arquitectura de la red neuronal.

debido a sus registros individuales demasiado cortos.

3.3. Detector propuesto basado en aprendizaje profundo

3.3.1. Vista global del método

La Figura 3.2 entrega una vista global del método propuesto para detectar eventos como husos de sueño y complejos K en el EEG del sueño. El registro completo de EEG se pre-procesa por medio de un filtrado, remuestreo a 200 Hz, y normalización. Después, se aplica un modelo secuencial profundo a través de todo el EEG en ventanas independientes de $T_w = 4000$ muestras (20 s). La salida del modelo es la probabilidad de pertenecer a la clase negativa (fondo) o positiva (evento) de cada muestra de la ventana. Al concatenar las predicciones de todas las ventanas, se genera una segmentación densa para todo el registro. Para evitar efectos de borde en la predicción, en cada ventana se agrega un borde de T_B muestras que son descartadas a la salida del modelo.

Por la arquitectura de su red neuronal, y por representar una actualización respecto al detector preliminar presentado en [12], el detector se llama *Recurrent Event Detector v2* (REDv2). Este detector considera una etapa convolucional para la extracción de características locales, seguida de una etapa recurrente para la integración del contexto temporal de la ventana completa. Además, se presentan dos variantes de la estructura base de REDv2, cada una con una representación de entrada distinta: REDv2-Time, que usa la señal EEG directamente en el tiempo, y REDv2-CWT, que primero transforma la señal a un espacio tiempo-frecuencia gracias a la Transformada de Wavelet Continua (CWT).

Luego de obtener la probabilidad para cada muestra de todo el registro, se usa un umbral τ_p para definir la existencia de un evento, y un umbral $\tau_L < \tau_p$ para definir su duración. Primero, la probabilidad de la clase positiva se ajusta para que superar τ_p en la probabilidad original sea equivalente a superar 0,5 en la probabilidad ajustada. Sobre la probabilidad ajustada, se aplica τ_L para determinar detecciones, descartando aquellas que no superan nunca 0,5. Finalmente, se aplica un post-procesamiento específico del tipo de evento buscado, y se aplica una máscara de anotación para descartar aquellas detecciones generadas en momentos inválidos del registro.

3.3.2. Preprocesamiento de señales

Filtrado. Se aplica un filtro pasa-banda con frecuencias de corte en 0,1 Hz y 35 Hz sin distorsión de fase. Específicamente, se trata de un filtro pasa-banda tipo Butterworth de orden 3, aplicado hacia adelante y hacia atrás para obtener un filtrado sin distorsión de fase. En el caso particular de MASS-MODA, se aplica el mismo filtro pasa-banda con frecuencias de corte en 0,3 Hz y 30 Hz sin distorsión de fase utilizado en su publicación original [23]. Específicamente, se trata de un filtro pasa-alto tipo Butterworth de orden 10 con frecuencia de corte en 0,3 Hz en formato SOS (*series second-order sections*) aplicado hacia adelante y hacia atrás, seguido de un filtro pasa-bajo tipo Butterworth de orden 10 con frecuencia de corte en 30 Hz en formato SOS aplicado hacia adelante y hacia atrás. Debido a que MASS-MODA se compone de segmentos cortos en lugar de señales de noche completa, se agregaron 10 s de señal en cada borde que se cortaron después de filtrar para evitar efectos de borde.

Remuestreo. De ser necesario, las señales filtradas son remuestreadas a 200 Hz usando un método de filtrado polifásico (*polyphase filtering* en inglés). Si $g = \mathrm{GCD}(f_s^{\mathrm{old}},\ f_s^{\mathrm{new}})$ es el máximo común divisor entre la frecuencia de muestreo original f_s^{old} y la deseada f_s^{new}, la señal se sobremuestrea f_s^{new}/g veces intercalando ceros, se filtra con un pasa-bajos de respuesta finita (i.e., una función sinc con ventana) sin distorsión de fase, y luego se submuestrea f_s^{old}/g veces saltando muestras.

Normalización. Las señales resultantes, cuya media ya es cero por el filtrado, se normalizan dividiendo por la desviación estándar calculada usando la combinación de los conjuntos de entrenamiento y validación de la partición en uso (i.e., se calcula en cada entrenamiento).[8] Para su cálculo, se recolectan todas las muestras de las etapas R, N1, N2, y N3 cuyo valor absoluto no es mayor al percentil 99 de su registro para limitar la influencia de magnitudes anómalas. Para evitar la duplicación en memoria de la base de datos provocada por un cálculo directo de la desviación estándar (algo problemático en bases grandes), se utiliza un cálculo alternativo. Si se tiene un total de N señales, para cada señal $\mathbf{x}_i$ de T_i muestras, $i = 1, \ldots, N$, se retornan los resultados parciales $a_i = \sum_{k=1}^{T_i} x_k^2$ junto al valor de T_i. Con estos resultados parciales se calcula la desviación estándar σ haciendo

$$\sigma^2 = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = \mathbb{E}[x^2] = \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} T_i}. \tag{3.8}$$

[8] Se encontró que es mejor usar una desviación estándar global en lugar de una desviación estándar independiente para cada señal (i.e., normalizar cada señal con su propia estadística), probablemente debido a que las anotaciones, al menos en MASS-SS2-Train, se basaron en parte en la magnitud absoluta de la señal, y dicha relación absoluta se pierde con una normalización individualizada.

Una vez normalizadas, los valores por debajo de -10 y por encima de 10 son recortados a dicho límite. En MASS-SS2-Train, dicho límite está en el orden de 160–170 μV.

3.3.3. Transformación de señales con wavelets

Una representación de entrada alternativa se obtiene al aplicar una transformada de wavelets sobre la señal EEG, usando la wavelet Morlet compleja (i.e., exponencial compleja con ventana Gaussiana) con un factor de expansión de escala $q = 0{,}9$, y con un ancho inicializado en β_0 y configurado como un parámetro entrenable (ver Sección 2.2). Además, se utilizan $N_s = 32$ escalas entre $f_{\min} = 0{,}5$ Hz y $f_{\max} = 30$ Hz usando una progresión geométrica $\{s_i\}_{i=1}^{N_s}$ dada por

$$s_i = \frac{1}{f_{\max}} \left(\frac{f_{\max}}{f_{\min}} \right)^{\frac{i-1}{N_s - 1}}. \tag{3.9}$$

Teóricamente, la ventana Gaussiana de la wavelet Morlet tiene un soporte infinito, pero en la práctica se usa un soporte finito. Para la función $\exp(-x^2/(2\sigma^2))$, una heurística común es truncar el soporte a alguna vecindad proporcional a σ, i.e., a $x \in [-k_\sigma \sigma,\ k_\sigma \sigma]$ con $k_\sigma > 0$. En la formulación de la wavelet, esta heurística corresponde a truncar el soporte a $[-k_\sigma s_q \sqrt{\beta/2},\ k_\sigma s_q \sqrt{\beta/2}]$. Una elección común es $k_\sigma = 3$, concentrando el 99,7 % del área bajo la curva. Siguiendo esta elección, cada wavelet se trunca al intervalo $[-\eta s_q \sqrt{4{,}5\beta},\ \eta s_q \sqrt{4{,}5\beta}]$, en donde $\eta \geq 1$ es una expansión opcional. Debido a que β se puede modificar durante el entrenamiento pero el tamaño de las wavelets se fija en la inicialización (usando $\beta = \beta_0$), se fija $\eta = 1{,}5$ para permitir un incremento de β.

La wavelet asociada a la frecuencia central $f_{\min}$ es la más ancha, con ancho temporal dado por $2\eta\sqrt{4{,}5\beta}/f_{\min} \approx 12{,}73\sqrt{\beta_0}$ s. Por lo tanto, para evitar efectos de borde y así asegurar una representación precisa, la transformada se calcula con T_B^{CWT} muestras adicionales de señal EEG en cada borde que son recortadas luego de obtener el resultado. El borde T_B^{CWT} debe ser al menos igual a la mitad del ancho de la wavelet más grande, es decir, al menos igual a

$$T_B^{CWT} = f_s \eta \sqrt{4{,}5\beta}/f_{\min} \approx 6{,}3640 f_s \sqrt{\beta_0}, \tag{3.10}$$

que corresponde a $\lceil 1272{,}8\sqrt{\beta_0} \rceil$ muestras a $f_s = 200$ Hz. Con la configuración especificada, al aplicar la transformada a una señal EEG de largo T_{input} usando un salto de $s \geq 1$ muestras, se obtiene una salida de largo temporal $T_{\text{output}} = (T_{\text{input}} - 2T_B)/s$, con N_s escalas, y dos canales (parte real y parte imaginaria). En otras palabras, la salida es un tensor $\mathbf{z} \in \mathbb{R}^{T_{\text{output}} \times N_s \times 2}$.

El parámetro β (o su inicialización β_0) se puede seleccionar o buscar de forma más conveniente al relacionarlo con el número de oscilaciones observadas en la wavelet. Si se cuenta el número de oscilaciones N_o de la wavelet asociada a la escala s que ocurren dentro del soporte truncado $[-k_\sigma s_q \sqrt{\beta/2},\ k_\sigma s_q \sqrt{\beta/2}]$, se obtiene

$$N_o = k_\sigma \sqrt{2\beta} \left(\frac{s_q}{s} \right) \Leftrightarrow \beta = \frac{N_o^2}{2k_\sigma^2} \left(\frac{s}{s_q} \right)^2. \tag{3.11}$$

Esta relación permite elegir β a partir de una elección de N_o, una variable más fácil de interpretar.[9] Para contar N_o, $k_\sigma = 3$ podría ser muy grande porque en las colas del soporte las oscilaciones

[9]Notar que si $q = 1$ (correspondiente a la CWT original), se tiene que $s_q = s$, implicando que N_o es constante para todas las escalas tal como se espera.

Tabla 3.4: Posibles valores del ancho de la wavelet según (3.11) cuando $q = 0{,}9$ y $k_\sigma = 2{,}2214$. El mínimo T_B^{CWT} para evitar efectos de borde se calcula según (3.10).

β	N_o (13 Hz)	N_o (1 Hz)	Mínimo T_B^{CWT}
0,0331	2	0,63	232 (1,16 s)
0,0744	3	0,94	348 (1,74 s)
0,1323	4	1,26	463 (2,32 s)
0,2068	5	1,57	579 (2,90 s)
0,2978	6	1,89	695 (3,48 s)
0,4053	7	2,20	811 (4,06 s)

son prácticamente imperceptibles, implicando una sobre-estimación de N_o. Un mejor valor de k_σ para efectos de conteo se puede obtener al calcular el ancho L de la ventana de Hann que mejor aproxima cerca del origen a la ventana Gaussiana. Usando una expansión de Taylor de primer orden en torno a $x = 0$ en la función exponencial para la ventana Gaussiana y en la función coseno para la ventana de Hann, se obtiene

$$w_L^{\text{Hann}}(x) = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi x}{L}\right)\right) = 1 - \frac{\pi^2 x^2}{L^2} + O(x^4) \tag{3.12}$$

$$w_\sigma^{\text{Gauss}}(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right) = 1 - \frac{x^2}{2\sigma^2} + O(x^4) \tag{3.13}$$

Descartando el residuo e igualando ambas expansiones, se obtiene un ancho $L = \pi\sqrt{2}\sigma$. Por otro lado, el ancho también es $L = 2k_\sigma\sigma$, implicando $k_\sigma = \pi/\sqrt{2} \approx 2{,}2214$. Usando este valor de k_σ en (3.11), se puede determinar la inicialización β_0 en base a un N_o deseado a una escala s específica. Tomando como referencia la frecuencia central de la banda sigma (i.e., 13 Hz), se determinan posibles valores de β para N_o entre 2 y 7 en la Tabla 3.4, en donde también se indica, para cada β, el N_o a la frecuencia 1 Hz y el mínimo T_B^{CWT} requerido.

3.3.4. Arquitectura del modelo secuencial profundo

Esquema general de la arquitectura. La Figura 3.3A ilustra, a un nivel funcional, los bloques que conforman la arquitectura de la red neuronal del detector REDv2. Su entrada es un segmento de EEG de largo $T_w + 2T_B$, con T_w fijo en 4000 (20 s) y T_B un borde por definir, y su salida es una segmentación densa con una predicción cada 8 muestras de entrada, i.e., de largo $T_w/8$. Se distinguen tres etapas: codificación local, contextualización, y clasificación muestra a muestra. La codificación local es un bloque convolucional cuyo objetivo es transformar la señal EEG a una serie de tiempo multivariada, en donde en cada instante de tiempo se tienen características extraídas de una vecindad acotada (corto plazo). A la salida de este bloque, la serie de tiempo ya está submuestreada en un factor de 8. Luego de la codificación local, se tiene la etapa de contextualización, que es un bloque secuencial cuyo objetivo es integrar la información temporal de toda la ventana de entrada. A la salida de este bloque, se tiene una serie de tiempo a la misma resolución temporal pero en donde en cada muestra se han integrado características de su contexto pasado y futuro de largo plazo. Finalmente, la clasificación muestra a muestra se realiza con una capa convolucional 1D de kernel unitario y dos canales, seguida de una capa softmax. Esta etapa asigna a cada instante de tiempo la probabilidad de pertenecer al fondo (clase 0 o negativa) o al interior de un evento de interés (clase 1 o positiva). Como ya se mencionó, la salida final tiene una resolución temporal 8 veces menor que la señal EEG. Para $f_s = 200$ Hz, esto significa que las etiquetas son predichas

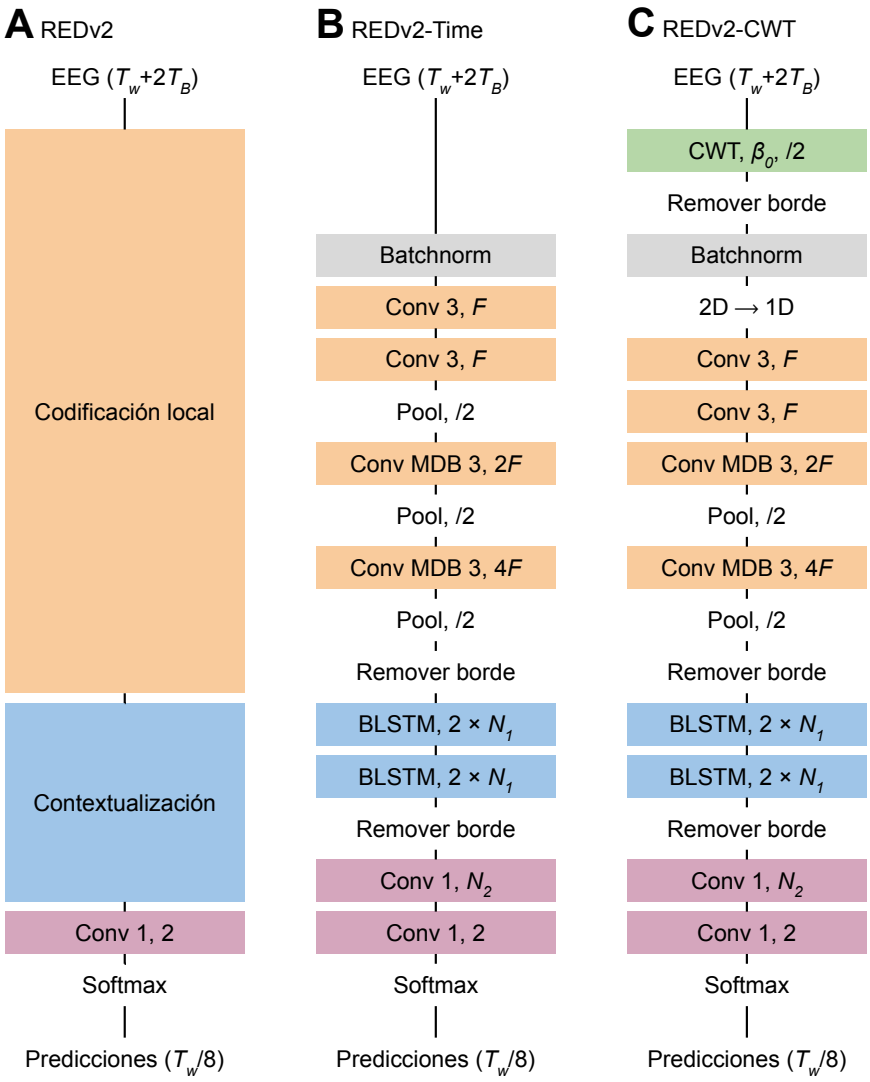

Figura 3.3: Arquitectura del modelo propuesto. (**A**) Diagrama de bloques funcional de REDv2. (**B**) Arquitectura del modelo REDv2-Time. (**C**) Arquitectura del modelo REDv2-CWT.

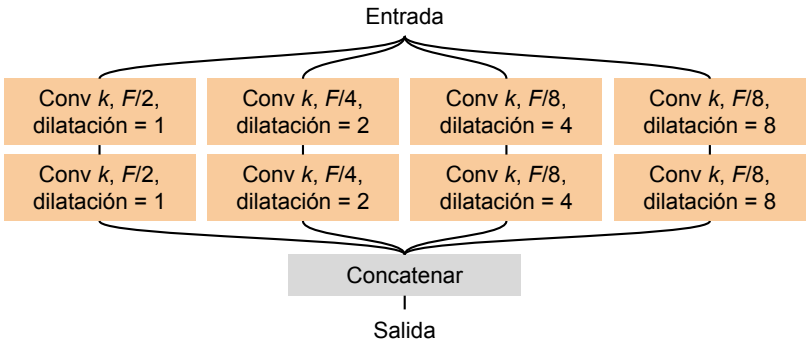

Figura 3.4: Definición del bloque llamado Conv MDB k, F. Posee la misma cantidad de parámetros que una secuencia de dos capas Conv k, F, pero con mayor campo receptivo.

cada 40 ms en lugar de cada 5 ms. Esta predicción sigue siendo suficientemente densa para eventos como husos de sueño y complejos K que duran al menos 300 ms. Para el entrenamiento, se usa directamente esta resolución menor, mientras que para la inferencia se retorna a la resolución original por interpolación lineal.

Exploración de la etapa de codificación local. Se exploraron varias alternativas para implementar esta etapa convolucional. Para simplificar la búsqueda, solo se exploraron arquitecturas 1D que utilizan la señal EEG directamente en el tiempo, y se fijó la etapa de contextualización en una red BLSTM. Entre las técnicas exploradas se tienen: capas convolucionales normales; capas convolucionales dilatadas; conexiones residuales; ramas paralelas; pre-cálculo de bandas de frecuencia; pre-cálculo de la amplitud sigma; y pre-cálculo de razones de potencia. En general, no se obtuvo un beneficio al pre-calcular características. Durante la exploración, se determinó un bloque convolucional 1D con múltiples dilataciones en paralelo que permiten aumentar el campo receptivo total sin comprometer el número de parámetros ni la capacidad de extraer características en vecindades pequeñas de forma eficiente. El bloque, ilustrado en la Figura 3.4, se llama *Convolutional Multi-dilated Block* (Conv MDB), con kernel de ancho k y un total de F canales a la salida.

Exploración de la etapa de contextualización. Una vez determinado el diseño de la codificación local, se exploraron varias alternativas para implementar esta etapa de integración de todo el contexto disponible. Entre las técnicas exploradas se tienen: capas LSTM y BLSTM; capas GRU [82] y BGRU; conexiones residuales; capas de auto-atención [68]; redes convolucionales temporales [83]; y redes convolucionales de codificación-decodificación como [67]. En general, se encontró que las alternativas convolucionales alcanzan un menor mIoU, y que las alternativas de auto-atención requieren pre-entrenamiento para acercarse al desempeño de las alternativas recurrentes, sin entregar ganancias.

Modelo REDv2-Time. Como resultado de la exploración realizada, se determina una arquitectura base aplicada directamente sobre la señal EEG. Esta arquitectura se llama REDv2-Time y se ilustra en la Figura 3.3B. Al principio, se aplica una capa de *batchnorm* para permitir corregir la normalización e introducir pequeñas variaciones en la amplitud durante el entrenamiento a modo de regularización. La codificación local se compone de capas convolucionales 1D con un número inicial de filtros F, kernel de tamaño 3, y *zero-padding*. Cada capa convolucional es seguida por una capa de *batchnorm* y activación ReLU. El submuestreo se hace progresivamente con capas de *pooling* de tamaño 2, y cada vez que se aplica se duplica el número de filtros. La contextualización se compone de dos capas BLSTM, cada una con N_1 neuronas por dirección, seguidas de una capa convolucional de kernel unitario, N_2 canales, y activación ReLU. Se aplica *dropout* a la entrada de cada capa con probabilidad ρ_1, ρ_2 y ρ_2, respectivamente. El campo receptivo total de la codificación local es 204 (1,02 s), por lo que a su salida se remueve un borde de 0,6 s para eliminar efectos de borde. Para asegurar un mínimo contexto bidireccional en la contextualización, se remueve un borde de 2 s a la salida de la última capa BLSTM. Por lo tanto, en REDv2-Time se requiere un borde de $T_B = T_B^{Conv} + T_B^{LSTM}$ con $T_B^{Conv} = 120$ (0,6 s) y $T_B^{LSTM} = 400$ (2 s). Los hiperparámetros se fijan experimentalmente en $F = 64$, $N_1 = 256$, $N_2 = 128$, $\rho_1 = 0{,}2$, $\rho_2 = 0{,}5$, y submuestreo por *average pooling*.

Exploración del uso de la CWT. Una vez determinada la arquitectura REDv2-Time, se exploraron varias alternativas para usar la representación tiempo-frecuencia de la CWT como entrada

en lugar de la señal en el tiempo. Para ello, se fijaron los parámetros de la transformación según lo descrito en la Sección 3.3.3. Se eligió el ancho inicial de la wavelet β_0 a partir del espacio de búsqueda indicado en la Tabla 3.4. Además, se exploraron diversas formas de aprovechar esta transformada en la red neuronal: se evaluó pre-calcular la magnitud y la fase de la CWT; y se evaluó si tratar la transformada como una entrada 1D (preservando las convoluciones de REDv2-Time) o 2D (cambiando las convoluciones 1D a 2D pero usando dilataciones solo en el eje temporal).

Modelo REDv2-CWT. Como resultado de la exploración realizada, se determina una arquitectura alternativa aplicada sobre la CWT de la señal EEG. Esta arquitectura se llama REDv2-CWT y se ilustra en la Figura 3.3C. La CWT usa los parámetros descritos en la Sección 3.3.3, además de un salto de 2 muestras para disminuir el costo computacional. Luego de la transformación, se remueve un borde de T_B^{CWT}, cuyo valor depende del ancho inicial de la wavelet β_0, para evitar efectos de borde. Después, se aplica una capa de *batchnorm* a través del eje de las frecuencias para permitir normalizar cada componente de forma independiente. Para conectar con la secuencia de convoluciones 1D mostradas en REDv2-Time, el tensor 2D $\mathbf{z} \in \mathbb{R}^{T_{\text{output}} \times N_s \times 2}$ cambia a un tensor 1D $\mathbf{z}' \in \mathbb{R}^{T_{\text{output}} \times 2N_s}$. El resto de la arquitectura de REDv2-CWT es igual a la de REDv2-Time a excepción de que, debido al submuestreo inicial de la CWT, se quita la primera capa de *average pooling*. Se elige inicializar el ancho de la wavelet en $\beta_0 = 0{,}1323$. Por lo tanto, en REDv2-CWT se requiere un borde de $T_B = T_B^{CWT} + T_B^{Conv} + T_B^{LSTM}$ con $T_B^{CWT} = 462$ (2,31 s), $T_B^{Conv} = 120$ (0.6 s) y $T_B^{LSTM} = 400$ (2 s).[10]

Exploración del tamaño del contexto. La elección de una ventana de 20 s está basada en las páginas de 20 s del hipnograma de MASS-SS2. Se exploró usar otros tamaños de ventana como entrada, y se encontró que el desempeño se satura con una ventana de 15–23 s. Por simplicidad, se mantuvo la ventana de 20 s. También se exploró complementar la arquitectura base (con entrada de 20 s) con una función simple de un contexto mayor, del orden de 1 a 2 minutos. Específicamente, se exploró aplicar sesgos o modulaciones dinámicas en diferentes capas de la arquitectura base. Se exploraron dos formas de calcular sesgos o modulaciones: por medio de una red *fully-connected* aplicada sobre el espectro del contexto grande; y por medio de una red convolucional 1D aplicada directamente sobre la señal del contexto grande, copiando la arquitectura de DOSED [8] a excepción de su capa de salida. No se encontraron beneficios en estas alternativas.

Entrada multicanal. Si bien la descripción de las arquitecturas se enfoca en usar un solo canal de EEG, es posible usar un número arbitrario de canales en la entrada de REDv2. En el caso de REDv2-Time, se deben concatenar los canales antes de la primera convolución. En el caso de REDv2-CWT, primero se debe calcular la CWT en cada canal por separado, y luego se deben concatenar las representaciones antes de la primera convolución.

3.3.5. Entrenamiento del modelo

Inicialización. Debido a las funciones de activación, todos los pesos de las capas convolucionales se inicializan usando la inicialización de Kaimin He, mientras que los pesos de las capas recurrentes se inicializan usando la inicialización de Xavier Glorot. Todos los sesgos se inicializan en cero a excepción de dos casos: el sesgo de la compuerta de olvido de la capa LSTM que se inicializa en

[10] El valor de T_B^{CWT} fue elegido como el número par de muestras más cercano al mínimo valor requerido, para ser compatible con el cálculo de la CWT con un salto de 2 muestras.

uno; y el sesgo de la salida de la clase 1 que se inicializa en $\log(p_1/(1-p_1))$ con p_1 un valor pequeño (aquí fijado en $p_1 = 0{,}1$) para que el modelo prediga aproximadamente $\mathbb{P}(y=1) = p_1$ en la inicialización, y así evitar que la pérdida provocada por la clase frecuente domine el inicio del entrenamiento y provoque inestabilidad, según lo recomendado en [84].

Extracción de ejemplos de entrenamiento. Cada señal EEG del conjunto de entrenamiento tiene un largo arbitrario y se divide de forma consecutiva y disjunta en páginas de 20 s, análogo al antiguo estándar de R&K (ver Sección 2.1.4). Además de la señal EEG, se tienen dos series de tiempo del mismo largo que se dividen de la misma forma: las etiquetas y las máscaras de anotación. La serie de las etiquetas es una secuencia binaria de etiquetas expertas que representan la existencia de un evento en cada instante de tiempo (i.e., es la vectorización de la colección de eventos anotados por el experto), y en consecuencia representan la salida ideal del modelo. La serie de las máscaras de anotación es la definida en la Sección 3.1. Para estas tres series, las páginas asignadas a la etapa N2 conforman el banco de páginas de entrenamiento. De estas páginas se deben extraer ejemplos de entrenamiento: segmentos de EEG de largo $T_{\text{input}} = T_w + 2T_B$ ($T_w = 4000$ [20 s], $T_B = 520$ [2,60 s] para REDv2-Time, $T_B = 982$ [4,91 s] para REDv2-CWT), segmentos de etiquetas de largo T_w, y segmentos de máscaras de largo T_w. Para aumentar el número de segmentos disponibles y mejorar la robustez a traslaciones temporales, los segmentos son recortados de forma aleatoria durante el entrenamiento. Específicamente, se eligen aleatoriamente M páginas del banco de páginas de entrenamiento, y al interior de cada página se elige aleatoriamente un centro que se usa como el centro del segmento a extraer.

Generación de batches. En la práctica, eventos como los husos de sueño y complejos K son raros en comparación con la actividad cerebral de fondo, pudiendo existir páginas con uno o ningún evento. Para generar batches balanceados, se cuenta el número de muestras pertenecientes a un evento al interior de cada página del banco de entrenamiento y se calcula la mediana de dichas cantidades. Usando esta mediana, el banco de entrenamiento se divide en dos bancos: uno compuesto de las n_1 páginas que están por debajo de la mediana, y otro compuesto de las n_2 páginas que están por encima. En cada iteración, se genera un batch de tamaño M extrayendo $M/2$ ejemplos de cada banco. Dado que hay dos fuentes de datos para formar cada batch, se define una época de entrenamiento como el número máximo de iteraciones con páginas únicas de ambos bancos. En consecuencia, una época de entrenamiento se compone de $\lfloor 2\min(n_1, n_2)/M \rfloor$ iteraciones.

Función de pérdida. Sea $\mathbf{y} \in \{0,1\}^{T_w}$ el vector binario de etiquetas de un cierto segmento de entrenamiento, $\mathbf{m} \in \{0,1\}^{T_w}$ la máscara binaria de anotación asociada, y $\mathbf{p} \in \mathbb{R}^{(T_w/8)\times 2}$ las probabilidades predichas para cada clase por el modelo propuesto. Como el modelo entrega salidas con una frecuencia de muestreo 8 veces menor a la original, las etiquetas ($\mathbf{y}$) y las máscaras ($\mathbf{m}$) de cada ejemplo se submuestrean usando una capa de *average pooling* de tamaño 8 y se redondean al entero más cercano. Las redes neuronales que predicen segmentaciones densas como el modelo propuesto se pueden entrenar para minimizar la entropía cruzada ponderada del segmento, dada por

$$\ell(\mathbf{y}, \mathbf{p}) = -\frac{1}{\sum_{k=1}^{T_w/8} w_{\text{class}}(y_k)} \sum_{k=1}^{T_w/8} w_{\text{class}}(y_k) \log p_k(y_k), \tag{3.14}$$

en donde $w_{\text{class}}(y)$ es el peso de la clase y. En esta tesis, se usa una función de pérdida de entropía cruzada *enmascarada*, dada por

$$\ell(\mathbf{y}, \mathbf{p}) = -\frac{1}{\sum_{k=1}^{T_w/8} m_k w_{\text{class}}(y_k)} \sum_{k=1}^{T_w/8} m_k w_{\text{class}}(y_k) \log p_k(y_k), \tag{3.15}$$

en donde la suma de las pérdidas por muestra está normalizada en cada segmento de forma independiente para asegurar que cada uno contribuye a la pérdida total del batch de forma comparable. En esta función de pérdida se incluye la máscara como un ponderador, permitiendo ignorar aquellas muestras que se encuentren fuera de los intervalos anotados por el experto que pueden aparecer a causa de la extracción aleatoria de segmentos durante el entrenamiento. Dado que no se encontraron beneficios al utilizar pesos diferentes para cada clase, se fija $w_{\text{class}}(y = 0) = w_{\text{class}}(y = 1) = 1$.

Exploración de la función de pérdida. Se exploraron otras funciones de pérdida y esquemas de ponderación, sobre todo inspirados en el fuerte desbalance de las etiquetas provocado por la rareza de la clase positiva. Sin embargo, en ningún caso se encontró una ganancia en desempeño con respecto a la entropía cruzada (función de pérdida seleccionada). Los esquemas de ponderación explorados son: pesos por clase constantes; pesos por clase proporcionales a la frecuencia dentro del segmento; pesos en el borde de los eventos que decaen exponencialmente (como en [67]); y selección aleatoria de muestras de la clase negativa para acotar el desbalance (como en [8]). Las funciones de pérdida exploradas son: entropía cruzada; entropía cruzada con penalización de entropía; entropía cruzada con suavización de etiquetas; entropía cruzada con recorte rectangular en las predicciones suficientemente buenas; entropía cruzada con recorte suave en las predicciones suficientemente buenas; pérdida de Dice; pérdida de *hinge*; pérdida focal [84]; y pérdida focal interpolada con la entropía cruzada.

Optimización. El entrenamiento minimiza (3.15). Se utiliza el optimizador Adam [40] con tasa de aprendizaje $\alpha = 10^{-4}$, decaimientos exponenciales $\beta_1 = 0{,}9$ y $\beta_2 = 0{,}999$, $\varepsilon = 10^{-7}$, y tamaño de batch 32. Además, se realiza una detención temprana utilizando como criterio el AF1 de las detecciones inferidas en el conjunto de validación con umbral de probabilidad $\tau_p = 0{,}5$ (ver Sección 3.3.6 para conocer el proceso de inferencia) calculado al final de cada época de entrenamiento. Si el criterio no mejora durante 5 épocas consecutivas, la tasa de aprendizaje se divide por la mitad y se reinicia el conteo de épocas sin mejora. Si se alcanzan 200 épocas de entrenamiento o no existe mejora del criterio luego del cuarto decaimiento de la tasa de aprendizaje, se detiene el entrenamiento. Al finalizar el entrenamiento, se retorna el mejor modelo encontrado según el criterio. Como se recomienda recortar el gradiente en arquitecturas recurrentes [41], se restringe la norma global del gradiente a 1.

Optimización de ajuste fino. Si el modelo fue pre-entrenado, la optimización mantiene el esquema descrito anteriormente pero con una tasa de aprendizaje inicial de $\alpha = 5 \cdot 10^{-5}$ y con un máximo de 3 decaimientos de la tasa de aprendizaje en lugar de 4. Así, el pre-entrenamiento se trata como si reemplazara la fase inicial de un entrenamiento normal (i.e., antes del primer decaimiento de la tasa de aprendizaje).

3.3.6. Inferencia

Vector de probabilidad. El modelo predice una señal completa usando segmentos de largo $T_w = 4000$ (20 s) con un paso $T_w/2$. Cuando se completa la inferencia, se conserva la mitad central de cada segmento predicho (atenuando así efectos de borde) para formar el vector de probabilidad de la señal completa.

Colección de detecciones. Para generar las detecciones (i.e., los instantes de inicio y fin), el vector de probabilidad obtenido se sobremuestrea linealmente en un factor de 8 para alcanzar la frecuencia de muestreo original. Inspirado en detectores tradicionales, se utilizan dos umbrales: τ_p y $\tau_L < \tau_p$. Sea $p \in (0, 1)$ la probabilidad de salida para la clase 1 (presencia de un evento) y $\tau_p \in (0, 1)$ el umbral de detección. La **probabilidad ajustada** $\tilde{p} \in (0, 1)$ tiene umbral de detección 0,5, recuperando la interpretación usual de predecir la clase más probable, y se define como

$$\tilde{z} = \log\left(\frac{p}{1-p}\right) - \log\left(\frac{\tau_p}{1-\tau_p}\right), \tag{3.16}$$

$$\tilde{p} = \frac{1}{1+\exp(-\tilde{z})}. \tag{3.17}$$

Luego, una detección se define como un intervalo en el cual las probabilidades ajustadas de la clase 1 son al menos τ_L (se decide la duración), y en al menos una muestra tiene un valor de al menos 0,5 (se decide la existencia). De existir una estrategia de postprocesamiento (descrita en la Sección 3.3.7 para husos de sueño y complejos K), esta se aplica a la colección de detecciones obtenida. Finalmente, solo se retornan las detecciones que están al menos parcialmente contenidas en la máscara binaria de anotación (i.e., intervalos de señal válidos).[11] El umbral τ_p se ajusta al final del entrenamiento con una búsqueda de grilla entre 0 y 1 con un paso de 0,02 para maximizar el AF1 de la combinación de los conjuntos de entrenamiento y validación de la partición en uso.[12] Por simplicidad, se fija $\tau_L = 0{,}85 \cdot 0{,}5 = 0{,}425$. Con estos parámetros, se encontró que usar dos umbrales, uno para detectar (τ_p) y otro para la duración (τ_L), mejora el desempeño con respecto a usar solo uno, sobre todo en el *precision*. Reconociendo el logit en el primer término de (3.16), se observa que la transformación a probabilidad ajustada puede incorporarse directamente en la red neuronal al restar $\log(\tau_p/(1-\tau_p))$ al sesgo de salida de la clase 1.

Probabilidad de un evento. El modelo entrega probabilidades para cada muestra, y un evento es un intervalo de varias muestras. Para el análisis se define como probabilidad de un evento el percentil 75 de las probabilidades ajustadas de sus muestras interiores. Se decide usar el percentil 75 como escalar representativo en lugar de la mediana o el promedio para mayor robustez frente a los bordes del evento, típicamente de menor probabilidad.

3.3.7. Postprocesamiento de detecciones

Detecciones de husos de sueño. Basado en procedimientos estándar [1], se combinan las predicciones más cercanas que Δ_{sep} y luego se eliminan las predicciones más cortas que $\Delta_{\min}$. Si bien los

[11] En la práctica, se trata de los segmentos en etapa N2, con la excepción de MASS-MODA en donde se trata de los segmentos vistos por los expertos.

[12] Se obtienen mejores resultados que usar solo el conjunto de validación, probablemente debido a que el tamaño pequeño del conjunto de validación de MASS-SS2-Train introduce mucha varianza.

husos de sueño no tienen una duración máxima, en la práctica la distribución de duración de dichos eventos está acotada [49], por lo que se eliminan las predicciones más largas que $2\Delta_{\max}$ y aquellas predicciones con una duración entre $\Delta_{\max}$ y $2\Delta_{\max}$ son recortadas para mantener un centro de duración $\Delta_{\max}$. Basado en [6], se usa $\Delta_{\text{sep}} = 0{,}3$ s, $\Delta_{\min} = 0{,}3$ s y $\Delta_{\max} = 3$ s para las predicciones de sujetos adultos. Para sujetos infantes, se usa $\Delta_{\text{sep}} = 0{,}5$ s, $\Delta_{\min} = 0{,}5$ s y $\Delta_{\max} = 5$ s por recomendación de los expertos del INTA.

Detecciones de complejos K. Estos eventos deberían durar al menos 0,5 s [1], por lo que se eliminan las predicciones más cortas que 0,3 s para dar holgura a la predicción de eventos cortos. Múltiples complejos K podrían ser predichos como un solo evento por la ausencia de una separación mínima, y en MASS-SS2-KC dichos eventos se consideran disjuntos. Para separar detecciones combinadas, se propone un nuevo post-procesamiento, inspirado en la detección de picos negativos de detectores tradicionales como el propuesto en [65]. En experimentos se encontró que sube el desempeño gracias a que aumenta el IoU y reduce falsos negativos. El post-procesamiento de separación consiste en lo siguiente. Dentro de cada predicción, se usa la señal filtrada por un pasa-bajos (tipo Butterworth de orden 3, aplicado hacia adelante y hacia atrás) con frecuencia de corte en 4 Hz para detectar picos negativos. Para evitar efectos de borde, se ignoran los picos más cercanos que 0,05 s al instante inicial de la detección o más cercanos que 0,2 s al instante final. Luego, los picos sin un cruce por cero entre ellos se agrupan en su punto medio para que cada candidato a complejo K sea representado por un único pico negativo. Si queda más de un pico al final de estos pasos, la predicción se separa en el instante medio entre cada par de picos consecutivos. Luego de esta separación, nuevamente se eliminan las predicciones más cortas que 0,3 s.

3.4. Filtrado de señales con filtro finito

Cuando se requiere filtrar una banda de frecuencia angosta (como en el método de aumento de datos descrito en la Sección 3.5.2), o cuando se requiere implementar el filtro usando convoluciones (e.g., para ejecutarlo en una GPU), un filtro con respuesta al impulso finita (*finite impulse response*, FIR) es más conveniente que uno de respuesta infinita (e.g., los filtros tipo Butterworth). Un filtro FIR pasa-bajos común se obtiene al aplicar una función de ventana a una función tipo sinc.[13] Específicamente, si $\mathbf{h}^{\text{LP}} \in \mathbb{R}^{2n+1}$ es la respuesta al impulso de un filtro FIR pasa-bajos de largo (orden) $2n+1$, el filtro (digital) está dado por

$$h_k^{\text{LP}} = \frac{1}{Z} w_{2n+1}^{\text{Hann}}(k) \frac{\operatorname{sen}(2\pi f_{\text{cutoff}} k / f_s)}{k}, \quad k \in \{-n, \ldots, n\}, \tag{3.18}$$

en donde f_s es la frecuencia de muestreo, f_{cutoff} es la frecuencia de corte, w_{2n+1}^{Hann} es una ventana de Hann de largo $2n+1$, y $Z = \sum_k h_k^{\text{LP}}$ es una constante de normalización para asegurar ganancia unitaria en la banda de paso. La elección de un número impar de muestras permite evitar desfases en la señal filtrada al convolucionar. Por otro lado, un filtro FIR pasa-altos es equivalente a restar a la señal original la señal filtrada por un pasa-bajos con la misma frecuencia de corte. Por lo tanto,

[13] La función sinc, de soporte infinito, es un filtro pasa-bajos *ideal* porque tiene transformada de Fourier igual a una ventana rectangular en el origen. Sin embargo, el soporte de la función sinc se hace finito a través de una función de ventana para hacer factible un filtro FIR. Esto provoca que el corte en el dominio de la frecuencia sea suave en lugar de abrupto (introduciendo lo que conoce como un *ancho de banda de transición*) y que aparezcan ondulaciones que pueden distorsionar la señal. La respuesta del filtro FIR se controla a través de la función de ventana y el orden (tamaño) del filtro.

si $\mathbf{h}^{\mathrm{HP}} \in \mathbb{R}^{2n+1}$ es la respuesta al impulso de un filtro FIR pasa-altos de largo (orden) $2n+1$, el filtro (digital) está dado por

$$h_k^{\mathrm{HP}} = \begin{cases} 1 - h_k^{\mathrm{LP}} & \text{si } k = 0 \\ -h_k^{\mathrm{LP}} & \text{si } k > 0 \end{cases}, \quad k \in \{-n, \ldots, n\}, \tag{3.19}$$

en donde $\mathbf{h}^{\mathrm{LP}} \in \mathbb{R}^{2n+1}$ tiene la misma frecuencia de corte (f_{cutoff}) que $\mathbf{h}^{\mathrm{HP}}$. Por último, un filtro pasa-banda con frecuencias de corte $f_{\mathrm{cutoff}}^{\mathrm{low}}$ y $f_{\mathrm{cutoff}}^{\mathrm{high}}$, con $f_{\mathrm{cutoff}}^{\mathrm{low}} < f_{\mathrm{cutoff}}^{\mathrm{high}}$, se obtiene al aplicar un filtro pasa-altos con frecuencia de corte $f_{\mathrm{cutoff}}^{\mathrm{low}}$ seguido de un filtro pasa-bajos con frecuencia de corte $f_{\mathrm{cutoff}}^{\mathrm{high}}$.

De forma similar a la wavelet Morlet, el filtro definido en (3.18) preserva su *forma* (i.e., la relación entre el período y el decaimiento de la envolvente) si su orden $L = 2n+1$ es inversamente proporcional a f_{cutoff}. Con dicha estrategia, el compromiso $\Delta t \Delta f$ sigue el mismo comportamiento que en la CWT (ver Sección 2.2), que en este caso consiste en que el acortamiento del filtro hacia las frecuencias altas permite localizar las dinámicas temporales a costa de aumentar el ancho de la banda de transición.[14] Como un punto intermedio, se decide hacer L inversamente proporcional a $\sqrt{f_{\mathrm{cutoff}}}$ según

$$L(f_{\mathrm{cutoff}}) = \frac{L_{\mathrm{ref}}}{\sqrt{f_{\mathrm{cutoff}}}}, \tag{3.20}$$

en donde L_{ref} es el ancho del filtro con corte en 1 Hz.[15] Se fija en $L_{\mathrm{ref}} = 6$ s para limitar la banda de transición al intervalo $(f_{\mathrm{cutoff}} - 1,\ f_{\mathrm{cutoff}} + 1)$ para $f_{\mathrm{cutoff}} \leq 16$ Hz. El valor de L sugerido por (3.20) se reemplaza por $2\lfloor L/2 \rfloor + 1$ para asegurar un valor impar.

3.5. Aumento de datos

Se aplican transformaciones aleatorias en la entrada que preservan las anotaciones en cada iteración del entrenamiento. Las primeras dos transformaciones son inyecciones sencillas de ruido, independientes de la entrada, cuyo objetivo es asegurar que el modelo sea robusto a variaciones insignificantes. En cambio, las últimas dos transformaciones son una propuesta más agresiva que consiste en la adición de señales aleatorias controladas, llamadas *ondas* y *anti-ondas* (descritas en la Sección 3.5.2), a la señal de entrada que modifican su composición espectral. La propuesta está motivada en el hecho de que dicha composición presenta una variabilidad significativa entre sujetos que es difícil de representar en un conjunto acotado de sujetos de entrenamiento. Sin embargo, en base al conocimiento experto del problema, algunas fuentes de variabilidad son irrelevantes para la detección, hecho que se puede explotar para aumentar la robustez del detector. Debido a que se actúa modificando la composición espectral, se deben tomar supuestos respecto al tipo de modificaciones que no afectan las anotaciones, dependiendo de si se trata del interior de un evento o no

[14] Si bien la banda de transición se acerca al caso ideal (transición abrupta) con un filtro FIR grande, una respuesta al impulso grande puede provocar una elongación innecesaria de las dinámicas temporales que ocurren en la banda de frecuencia de interés a través de un decaimiento lento de oscilaciones transitorias (e.g., un huso de sueño).

[15] Esta estrategia para expandir el ancho del filtro FIR con respecto al ancho dado por una proporción inversa (tipo CWT) se diseñó antes de conocer la expansión interpolada dada por el factor de expansión de escala q usado por EEGLAB (ver Sección 2.2.4). La expansión dada por (3.20) implica $L(f_{\mathrm{cutoff}}) \to 0$ cuando $f_{\mathrm{cutoff}} \to \infty$, a diferencia del caso de usar q que implicaría $L(f_{\mathrm{cutoff}}) \to C(1-q)s_{\max} > 0$ con C una constante. Es decir, la expansión a través de q asegura en el límite una fracción fija del ancho más grande. Como (3.20) se usó solo para $f_{\mathrm{cutoff}} \leq 16$ Hz, no se espera una diferencia notable en la práctica.

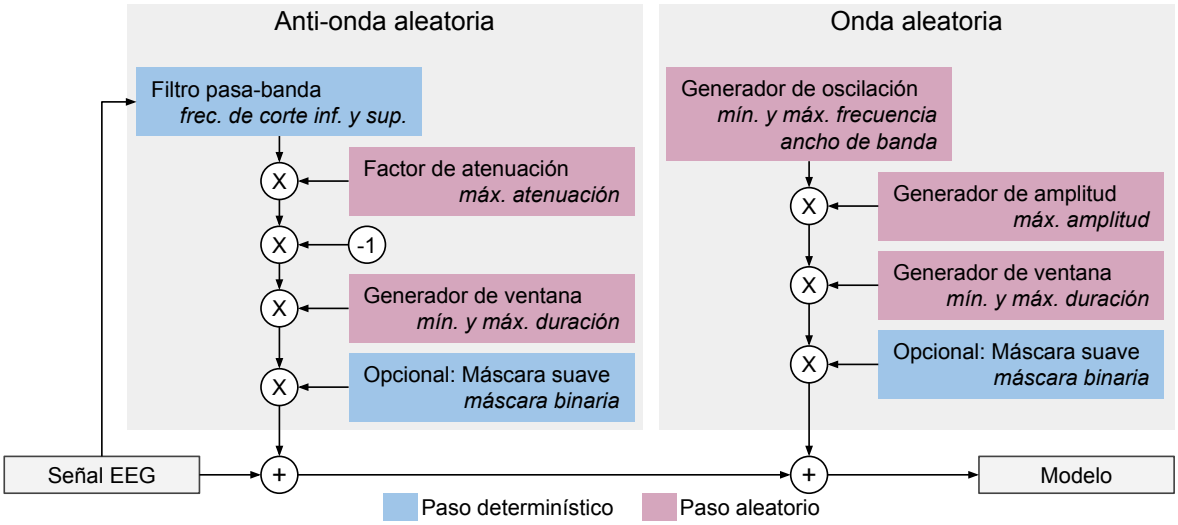

Figura 3.5: Esquema general de la adición de ondas y anti-ondas durante el entrenamiento. Los parámetros de cada bloque se indican en cursiva. Se muestra una sola anti-onda y una sola onda, pero se pueden aplicar varias con diferentes parámetros.

y del tipo del mismo. Al evaluar el modelo, todas las transformaciones aleatorias se desactivan. A continuación se describen estas transformaciones.

3.5.1. Ruido independiente

En primer lugar, cada muestra x de la señal de entrada se reemplaza por $x + u$ en donde u es un ruido muestreado uniformemente de $[-A_{\text{add}},\ A_{\text{add}}]$ y de forma independiente para cada muestra. La amplitud del ruido se fija arbitrariamente en $A_{\text{add}} = 1\ \mu$V, amplitud mucho menor a la de los complejos K y los husos de sueño. Además, para el caso de REDv2-CWT, cada escala s de la CWT se reemplaza por $s/(1 + v)$ en donde v es un ruido muestreado uniformemente de $[-A_{\text{scale}},\ A_{\text{scale}}]$ y de forma independiente para cada escala. Esta última transformación tiene el efecto de cambiar la frecuencia central de cada wavelet de f a $f+vf$, provocando que la wavelet desplace aleatoriamente su pico de extracción de potencia dentro de la banda $[f - A_{\text{scale}}f,\ f + A_{\text{scale}}f]$. Se fija $A_{\text{scale}} = 0{,}02$ para asegurar que el ancho de la banda de desplazamiento dentro de la banda sigma (11–16 Hz) sea del orden de 0,5 Hz.

3.5.2. Adición de ondas y anti-ondas

Luego de aplicar el ruido independiente a la señal de entrada, se aplican transformaciones consistentes en la adición de señales llamadas *ondas aleatorias* y *anti-ondas aleatorias*, cuyo esquema general se muestra en la Figura 3.5. Antes de describir la implementación de dichas transformaciones, se requiere describir el método de generación de tres tipos de señales: una señal aleatoria suave, una ventana aleatoria suave, y una máscara suave basada en las etiquetas. La señal aleatoria suave es útil para el generador de oscilación y el generador de amplitud de las *ondas aleatorias*, y se genera a partir del Algoritmo 1, en donde una señal de ruido uniforme se suaviza con una ventana de Hann de ancho L (fijado en $L = 0{,}5$ s) y se normaliza para tomar valores en el rango $[a_{\min},\ a_{\max}]$.[16] Por otro lado, la ventana aleatoria suave (salida del generador de ventana) es útil

[16]La ventana de Hann aplicada como un filtro, o en general cualquier ventana de promedio móvil, constituye un

Algoritmo 1 Generación de señal aleatoria suave.

Entrada: Número de muestras (T), valor mínimo ($a_{\min}$), valor máximo ($a_{\max}$), largo de ventana móvil (L).

Salida: Señal de T muestras acotada en $[a_{\min},\ a_{\max}]$ ($\mathbf{x}$).

$\mathbf{h}^{\mathrm{LP}} \leftarrow w_L^{\mathrm{Hann}} / \mathrm{Sum}(w_L^{\mathrm{Hann}})$
$\mathbf{x} \leftarrow \mathrm{RandomUniform}(\mathrm{size} = T,\ \mathrm{min} = -1,\ \mathrm{max} = 1)$
$\mathbf{x} \leftarrow \mathbf{x} * \mathbf{h}^{\mathrm{LP}}$
$\mathbf{x} \leftarrow (\mathbf{x} - \mathrm{Min}(\mathbf{x}))/(\mathrm{Max}(\mathbf{x}) - \mathrm{Min}(\mathbf{x}))$
$\mathbf{x} \leftarrow a_{\min} + (a_{\max} - a_{\min})\mathbf{x}$

Algoritmo 2 Generación de ventana con duración y localización aleatoria y transiciones suaves.

Entrada: Número de muestras (T), duración mínima ($d_{\min}$), duración máxima ($d_{\max}$).

Salida: Señal de ventana de T muestras con duración d e inicio t_0 ($\mathbf{w}$).

$d \leftarrow \mathrm{RandomUniform}(\mathrm{size} = 1,\ \mathrm{min} = d_{\min},\ \mathrm{max} = d_{\max})$
$t_0 \leftarrow \mathrm{RandomUniform}(\mathrm{size} = 1,\ \mathrm{min} = 0,\ \mathrm{max} = T - d)$
$\Delta \leftarrow 0{,}1d$
$a_1 \leftarrow t_0 + \Delta$
$a_2 \leftarrow t_0 + d - \Delta$
$\tau \leftarrow \Delta/4$
$\mathbf{w}^{\mathrm{onset}} \leftarrow \mathrm{Zeros}(T)$
$\mathbf{w}^{\mathrm{offset}} \leftarrow \mathrm{Zeros}(T)$
for $k = 0$ to $T - 1$ **do**
  $w_k^{\mathrm{onset}} \leftarrow \mathrm{Sigmoid}\left((k - a_1)/\tau\right)$
  $w_k^{\mathrm{offset}} \leftarrow \mathrm{Sigmoid}\left((k - a_2)/\tau\right)$
end for
$\mathbf{w} \leftarrow \mathbf{w}^{\mathrm{onset}} - \mathbf{w}^{\mathrm{offset}}$

Algoritmo 3 Generación de máscara suave basada en las etiquetas.

Entrada: Señal binaria de etiquetas ($\mathbf{y}$), largo de ventana móvil (L), región válida (r_{mask}).

Salida: Máscara suave del mismo largo que $\mathbf{y}$ ($\mathbf{m}$).

$\mathbf{h}^{\mathrm{Enlarge}} \leftarrow \mathrm{Ones}(L)$
$\mathbf{y}^{\mathrm{Enlarged}} \leftarrow \mathbf{y} * \mathbf{h}^{\mathrm{Enlarge}}$
$\mathbf{y}^{\mathrm{Enlarged}} \leftarrow \mathrm{Clip}(\mathbf{y}^{\mathrm{Enlarged}},\ \mathrm{min} = 0,\ \mathrm{max} = 1)$
$\mathbf{h}^{\mathrm{LP}} \leftarrow w_L^{\mathrm{Hann}} / \mathrm{Sum}(w_L^{\mathrm{Hann}})$
$\mathbf{m} \leftarrow \mathbf{y}^{\mathrm{Enlarged}} * \mathbf{h}^{\mathrm{LP}}$
if $r_{\mathrm{mask}} = \mathrm{fondo}$ **then**
  $\mathbf{m} \leftarrow 1 - \mathbf{m}$
end if

para localizar las perturbaciones en un intervalo específico de la señal de entrada, y se genera a partir del Algoritmo 2, en donde se selecciona un ancho y un instante de inicio aleatorios y, en lugar de una ventana rectangular, se genera una ventana que posee un comienzo y un final en forma de sigmoide. La transición completa en forma de sigmoide ocupa aproximadamente el 20 % del ancho de la ventana en cada borde. Por último, la máscara suave (aplicada opcionalmente al final) es útil para aquellos casos en los que se requiere restringir la perturbación al interior o al exterior de los eventos de interés, y se genera a partir del Algoritmo 3, en donde la señal binaria de las etiquetas se suaviza usando una ventana de Hann de ancho L (fijado en $L = 0{,}2$ s) para introducir una transición suave de ancho $L/2$ en cada borde de los eventos. Antes de suavizar las etiquetas, se expanden hacia el exterior ambos bordes de cada evento en $L/2$ para asegurar que en el interior de los eventos no existe ningún valor suavizado (i.e., el evento está completamente protegido). Si se desea restringir la perturbación al interior de los eventos ($r_{\text{mask}} = \text{evento}$), el resultado anterior es directamente la máscara. En cambio, si lo que se desea es restringir la perturbación al exterior ($r_{\text{mask}} = \text{fondo}$), se invierte el valor del resultado (i.e., se reemplaza m por $1 - m$) para formar la máscara.

Una vez descritos estos tres tipos de señales auxiliares, se procede a describir la implementación de las perturbaciones propuestas. En primer lugar, una *anti-onda aleatoria* tiene como objetivo atenuar una cierta banda de frecuencia (ver Figura 3.5). Se comienza filtrando la señal de entrada $\mathbf{x} \in \mathbb{R}^T$ con un filtro pasa-bandas FIR (descrito en la Sección 3.4) entre las frecuencias de corte $f_{\text{cutoff}}^{\text{low}}$ y $f_{\text{cutoff}}^{\text{high}}$ para extraer la oscilación $\mathbf{z} \in \mathbb{R}^T$ que permite atenuar la actividad de la banda $f_{\text{cutoff}}^{\text{low}}$–$f_{\text{cutoff}}^{\text{high}}$. Además, se muestrea uniformemente un factor de atenuación $a \in [0,\ A_{\text{max}}^{\text{atten}}]$, con $A_{\text{max}}^{\text{atten}} \leq 1$, y se multiplica con la oscilación $\mathbf{z}$ usando un signo negativo para que, al sumar la *anti-onda* a la señal de entrada, se tenga como resultado una atenuación de la banda en un factor a. Luego, se genera una función de ventana $\mathbf{w} \in \mathbb{R}^T$ con una duración muestreada uniformemente en el rango $[d_{\text{min}},\ d_{\text{max}}]$ (ver Algoritmo 2). Opcionalmente, se genera una máscara $\mathbf{m} \in [0, 1]^T$ para restringir la perturbación al interior o al exterior de los eventos según el indicador r_{mask} (ver Algoritmo 3). Así, la *anti-onda aleatoria* $\mathbf{u} \in \mathbb{R}^T$ está dada por

$$\mathbf{u}(f_{\text{cutoff}}^{\text{low}}, f_{\text{cutoff}}^{\text{high}}, A_{\text{max}}^{\text{atten}}, d_{\text{min}}, d_{\text{max}}, r_{\text{mask}}) = \mathbf{m} \odot \mathbf{w} \odot (-a\mathbf{z}), \tag{3.21}$$

en donde $\odot$ es el producto muestra a muestra. Se fija $A_{\text{max}}^{\text{atten}} = 1$ a menos que se indique lo contrario. En la Figura 3.6B se muestra un ejemplo de este método.

En segundo lugar, una *onda aleatoria* tiene como objetivo inyectar actividad a una cierta banda de frecuencia, la cual es independiente de la señal de entrada (ver Figura 3.5). Para una banda de frecuencia objetivo f_{min}–f_{max}, se comienza muestreando una frecuencia inferior $f_a \in [f_{\text{min}},\ f_{\text{max}} - \Delta f]$ de forma uniforme y se utiliza para generar una señal aleatoria $\mathbf{f} \in [f_a,\ f_a + \Delta f]^T$ (ver Algoritmo 1) que representa la frecuencia instantánea de la oscilación a generar con ancho de banda Δf. Esta frecuencia instantánea determina la oscilación $\mathbf{z} \in \mathbb{R}^T$ dada por

$$z_k = \cos\left(2\pi \sum_{i=0}^{k} f_k / f_s\right), \quad k \in \{0, \ldots, T-1\}, \tag{3.22}$$

en donde f_s es la frecuencia de muestreo. Para determinar la amplitud de esta oscilación, se muestrea una amplitud superior $a_{\text{sup}} \in [0,\ A_{\text{max}}]$ de forma uniforme y una amplitud inferior

filtro pasa-bajos cuya frecuencia de corte se controla con L: es menor a medida que L crece.

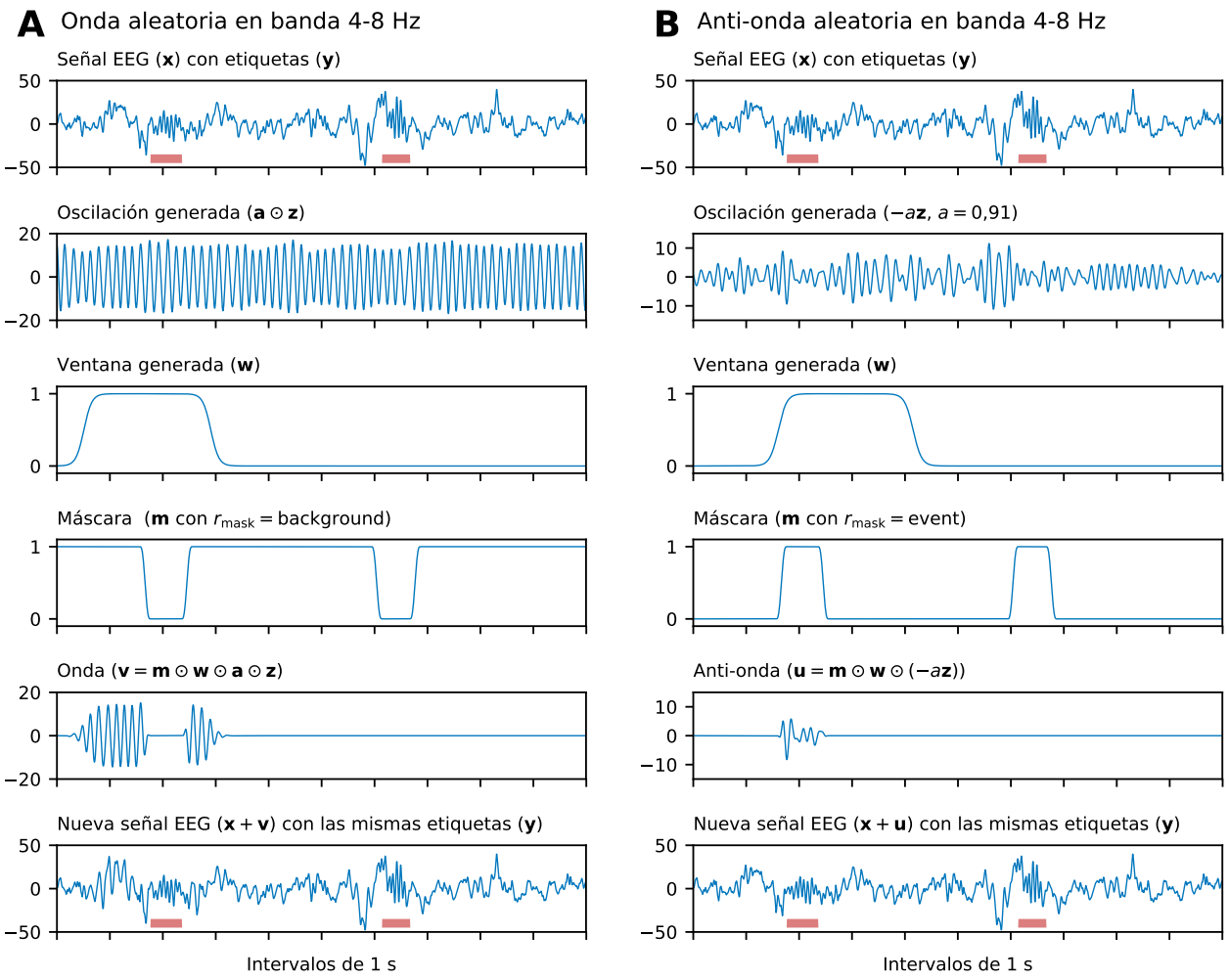

Figura 3.6: Ejemplo de los métodos propuestos de aumento de datos llamados *ondas aleatorias* (**A**) y *anti-ondas aleatorias* (**B**) ante etiquetas de husos de sueño. Los parámetros usados son los indicados para la banda 4–8 Hz en las Tablas 3.5 y 3.6.

$a_{\text{inf}} \in [0,\ a_{\text{sup}}]$ de forma uniforme. Ambos valores se utilizan para generar una señal aleatoria $\mathbf{a} \in [a_{\text{inf}},\ a_{\text{sup}}]^T$ (ver Algoritmo 1) que representa la amplitud instantánea de la oscilación. Después, al igual que en el caso de la *anti-onda aleatoria*, se genera una función de ventana $\mathbf{w} \in \mathbb{R}^T$ (ver Algoritmo 2) y, opcionalmente, una máscara $\mathbf{m} \in [0,1]^T$ (ver Algoritmo 3). Así, la *onda aleatoria* $\mathbf{v} \in \mathbb{R}^T$ está dada por

$$\mathbf{v}(f_{\text{min}}, f_{\text{max}}, \Delta f, A_{\text{max}}, d_{\text{min}}, d_{\text{max}}, r_{\text{mask}}) = \mathbf{m} \odot \mathbf{w} \odot \mathbf{a} \odot \mathbf{z}. \tag{3.23}$$

En la Figura 3.6A se muestra un ejemplo de este método.

Para una señal de entrada $\mathbf{x} \in \mathbb{R}^T$ se pueden generar y sumar N_u *anti-ondas aleatorias* y N_v *ondas aleatorias*, cada una perturbando una componente distinta, es decir,

$$\mathbf{x} \leftarrow \mathbf{x} + \sum_{i=1}^{N_u} \mathbf{u}_i + \sum_{i=1}^{N_v} \mathbf{v}_i. \tag{3.24}$$

Debido a que las *anti-ondas* dependen de la señal a perturbar (por el filtro pasa-banda), se generan y aplican primero las *anti-ondas* de forma secuencial: la *anti-onda* $\mathbf{u}_i$ se genera en base a la señal $\mathbf{x}+$

Tabla 3.5: Parámetros para el método propuesto de aumento de datos llamado *ondas aleatorias*.

Evento	ID	$f_{\min}$ (Hz)	$f_{\max}$ (Hz)	Δf	$A_{\max}$ (μV)	$d_{\min}$ (s)	$d_{\max}$ (s)	r_{mask}
Huso de sueño	1	0,5	2	1	18	3	5	todo
	2	4	8	2	20	1	5	fondo
	3	7	10	2	12	1	5	fondo
Complejo K	1	11	16	2	10	1	5	todo

Tabla 3.6: Parámetros para el método propuesto de aumento de datos llamado *anti-ondas aleatorias*.

Evento	ID	$f_{\text{cutoff}}^{\text{low}}$ (Hz)	$f_{\text{cutoff}}^{\text{high}}$ (Hz)	$A_{\max}^{\text{atten}}$	$d_{\min}$ (s)	$d_{\max}$ (s)	r_{mask}
Huso de sueño	1	-	2	0,5	3	5	todo
	2	4	8	1	1	5	evento
	3	7	10	1	1	5	evento
Complejo K	1	11	16	1	1	5	todo

$\sum_{j=1}^{i-1} \mathbf{u}_j$, para cada $i = 1, \ldots, N_u$ y con $\mathbf{u}_0 = 0$. Se decide aplicar siempre todas las perturbaciones durante el entrenamiento, en lugar de decidir una probabilidad de ocurrencia, ya que la amplitud de las perturbaciones admite valores insignificantes (i.e., un factor de atenuación cercano a cero en el caso de una *anti-onda*, o una amplitud máxima cercana a cero en el caso de una *onda*).

Los parámetros de las *ondas* y las *anti-ondas* a utilizar se muestran en la Tabla 3.5 y en la Tabla 3.6, respectivamente. La configuración de las perturbaciones es específica al tipo de evento (huso de sueño o complejo K) debido a los supuestos que se deben tomar para preservar las etiquetas. Si bien se deben fijar varios parámetros, las configuraciones escogidas son fijadas en base al conocimiento del problema y las estadísticas de amplitud de las señales en MASS-SS2-Train, es decir, sin ajuste experimental, evitando el alto costo computacional de una búsqueda de grilla.

La elección de las bandas de frecuencia se basa en conocimiento experto. Para el caso de husos de sueño, se toma como punto de partida la razón de potencia utilizada en [59], igual a la potencia en la banda sigma (11–16 Hz) dividida por la potencia de la señal sin la banda delta (4,5–30 Hz). En base a esto, se asume que las anotaciones se preservan si dicha razón de potencia es igual o mayor **dentro del evento**, o si es igual o menor **fuera del evento**. En consecuencia, se elige modificar las bandas delta, theta, y alfa de acuerdo a dicho principio, recordando que una *anti-onda* disminuye la potencia en la banda mientras que una *onda* la aumenta. La razón de potencia descrita sugiere que la banda delta es irrelevante para definir la existencia de un huso de sueño, por lo que puede aumentar o disminuir en cualquier región ($r_{\text{mask}} = \text{todo}$, es decir, no se utiliza una máscara). Sin embargo, experimentos preliminares sugieren que la actividad en la banda 2–4 Hz (delta rápida) tiene una leve influencia en los resultados de la detección, al contrario de la banda 0,5–2 Hz (delta lenta) en donde no se encontraron efectos.[17] Por lo tanto, solo se perturba la banda delta lenta, y por precaución se restringe la atenuación provocada por la *anti-onda* en la banda delta lenta a $A_{\max}^{\text{atten}} = 0{,}5$. Por último, la banda alfa se reemplaza por la banda 7–10 Hz para evitar contaminar la banda sigma. Para el caso de complejos K, se toma como punto de partida que constituyen ondas bifásicas *sobresalientes* y que pueden ser co-ocurrentes con husos de sueño. Para no afectar su

[17]Dada la observación de que la banda delta podría ser irrelevante, se evaluó removerla en el preprocesamiento. Se encontró que eliminar la banda 0,5–2 Hz no mejora ni empeora el desempeño, mientras que eliminar la banda 0,5–4 Hz introdujo una leve caída en el *F1-score* y en el mIoU.

forma ni su amplitud relativa al resto de la señal (afectando su cualidad de sobresaliente), se decide solo perturbar la banda sigma.

La elección de la amplitud máxima $A_{\max}$ de las *ondas* se basa en las estadísticas de amplitud de las señales en MASS-SS2-Train, usando las anotaciones del experto E1. Sea $f_{\min}$–$f_{\max}$ una banda de frecuencia a perturbar usando la máscara determinada por r_{mask}. Para cada sujeto en MASS-SS2-Train, se filtra la señal en la banda $f_{\min}$–$f_{\max}$ usando un filtro pasa-banda FIR (descrito en la Sección 3.4) y se calcula la amplitud de la señal analítica dada por la transformada de Hilbert. Luego, para cada página de 20 s en etapa N2, se extrae la amplitud máxima considerando solo muestras ubicadas en el exterior o en el interior de eventos dependiendo de si $r_{\text{mask}} = \text{fondo}$ o $r_{\text{mask}} = \text{evento}$, respectivamente. En cada sujeto de forma independiente, se calcula el histograma de dichas amplitudes con un paso de 1 μV. Después, todos los histogramas se promedian y se elige $A_{\max}$ como la mediana de la distribución resultante. Si $r_{\text{mask}} = \text{todo}$, se calcula la amplitud máxima utilizando $r_{\text{mask}} = \text{evento}$. Estas estadísticas se describen detalladamente en el Anexo C.

3.6. Medición de parámetros de husos de sueño y complejos K

Para analizar la calidad de las detecciones, se miden algunos parámetros clásicos: duración, amplitud pico a pico (PP), y frecuencia central. En husos de sueño, la señal EEG primero se filtra con un pasa-banda FIR en 9,5–16,5 Hz (ver Sección 3.4). Sobre esta señal filtrada se calcula la amplitud PP y la frecuencia central de forma similar a [23]. La amplitud PP del huso se obtiene como la máxima diferencia entre picos consecutivos, i.e., entre pares de mínimos y máximos vecinos. Por otro lado, la frecuencia central del huso se obtiene por FFT. Específicamente, se aísla la señal dentro del huso, se rellenan sus bordes con ceros hasta alcanzar 10 s de duración, se obtiene su FFT, y se extrae la frecuencia de máxima potencia. En complejos K, la señal EEG primero se filtra con un pasa-bajos FIR en 8 Hz. La amplitud PP se obtiene simplemente como la diferencia entre el máximo y el mínimo dentro del complejo K.

3.7. Comparación con la literatura

Se ajustan y entrenan algunos detectores de la literatura con implementación disponible que representen un desempeño de referencia relevante, usando el mismo esquema de evaluación y el mismo post-procesamiento descrito en esta tesis. Específicamente, se seleccionan los detectores llamados DOSED [8], A7 [59] y Spinky [65]. El detector DOSED representa una referencia muy relevante por ser una red convolucional que procesa segmentos de EEG de 20 s para predecir la existencia y la duración de eventos cortos, como pueden ser husos de sueño o complejos K. Para la evaluación, se utilizan los hiperparámetros reportados en su publicación y se ajusta el umbral de probabilidad al final del entrenamiento usando el mismo procedimiento usado en el modelo propuesto porque produce mejores resultados. Por otro lado, los detectores A7 y Spinky representan buenas referencias que no están basadas en aprendizaje de máquinas, sino exclusivamente en técnicas de procesamiento de señales y umbrales. El detector A7 es específico para husos de sueño, y sus cuatro umbrales se ajustan con una búsqueda de grilla para maximizar el AF1 del conjunto de entrenamiento. En cambio, el detector Spinky detecta tanto husos de sueño como complejos K, pero solo se usan sus resultados en complejos K porque su desempeño en la detección de husos de sueño es peor que DOSED y A7. Su umbral también se ajusta con una búsqueda de grilla para maximizar el AF1 del conjunto de entrenamiento. Adicionalmente, el detector Spinky solo entrega

el instante de tiempo del pico negativo del complejo K, por lo que se asume que el evento comienza 0,1 s antes y termina 1,3 s después de dicho pico como en su publicación.

Los otros cuatro detectores de la literatura que están basados en redes neuronales (ver Sección 2.4) quedan fuera de este marco común por no brindar una implementación con la funcionalidad de entrenar. Sin embargo, dado que tres de ellos fueron evaluados en MASS-SS2 usando validación cruzada, se agregan sus desempeños reportados en la configuración experimental más cercana para la comparación. Para el caso de SpindleNet, propuesto en [9], se compara su resultado para la unión de expertos en MASS-SS2 con la validación cruzada de MASS-SS2-E2SS con $N = 15$. Para el caso del detector propuesto en [7], abreviado aquí como DKL-KC, se compara su resultado para MASS-SS2-KC con la validación cruzada de MASS-SS2-KC con $N = 19$. A diferencia de los demás detectores basados en redes neuronales, DKL-KC solo permite conocer el instante de tiempo del pico negativo del complejo K (de forma análoga a Spinky). Por último, para el caso de SpindleU-Net, propuesto en [10], se compara su resultado para MASS-SS2-E1SS con la validación cruzada de MASS-SS2-E1SS con $N = 19$, y su resultado para MASS-SS2-E2SS con la validación cruzada de MASS-SS2-E2SS con $N = 15$. En [10] solo reportan la dispersión entre sujetos. Sin embargo, también reportan el desempeño en cada sujeto por separado. Por lo tanto, para asegurar una comparación justa de la dispersión, se usa dicha tabla para simular el esquema de validación cruzada usado en este trabajo y estimar la dispersión entre subconjuntos. Tanto SpindleNet como SpindleU-Net fueron evaluados usando un post-procesamiento de duración diferente al utilizado aquí. Mientras que REDv2, DOSED y A7 son evaluados usando una duración mínima de 0,3 s, SpindleNet fue evaluado con una duración mínima de 0,4 s y SpindleU-Net con 0,5 s.

3.8. Código

Este trabajo está implementado en el lenguaje de programación Python. Se usan varias librerías disponibles en este lenguaje. Se usa extensivamente el ecosistema Scipy de computación científica, particularmente NumPy (manipulación de arreglos), Matplotlib (gráficos), Pandas (manipulación de tablas), y el módulo signal de Scipy (procesamiento de señales). Los archivos EDF de las señales se leen con pyEDFlib o MNE. La implementación y el entrenamiento de las redes neuronales se hace con TensorFlow. Con la ayuda de capas convolucionales, también se usa TensorFlow para implementar filtros FIR y la CWT. Para acelerar la evaluación del desempeño, algunos ciclos se paralelizan usando joblib. Por último, las proyecciones (e.g., Kernel PCA) y las regresiones lineales se consiguen usando scikit-learn.

Capítulo 4

Resultados

4.1. Efecto del aumento de datos

Para observar el efecto del método de ondas aleatorias y anti-ondas aleatorias, se evalúa el desempeño de REDv2-Time cuando se activan (probabilidad de ocurrencia igual a 1) o desactivan (probabilidad de ocurrencia igual a 0) dichas transformaciones, manteniendo todos los demás hiperparámetros constantes. La evaluación se realiza solo usando MASS-SS2-Train, respetando el principio de no tomar decisiones de diseño fuera de dicho conjunto. Se realizan 6 repeticiones de una validación cruzada de 5 particiones, cada una con una semilla aleatoria diferente (30 particiones en total). Esto entrega 6 evaluaciones por sujeto. Por brevedad, solo se evalúa usando las anotaciones del experto E1.

La Tabla 4.1 resume las métricas de desempeño del experimento, en donde se reporta la dispersión tanto entre particiones como entre sujetos (ver Sección 3.2.1). Además, la Figura 4.1 muestra cada uno de los puntos (particiones o sujetos) que son representados en las estadísticas de la Tabla 4.1. Aunque la media del *F1-score* aumenta en 0,4 % en E1SS y 0,1 % en KC al utilizar aumento de datos, dichos cambios no son estadísticamente significativos ($P > 0{,}05$). Además, la dispersión entre sujetos del *F1-score* decrece en 0,7 % en E1SS y se mantiene similar en KC al utilizar aumento de datos. El cambio en la dispersión en E1SS es más visible en las Figuras 4.1A y 4.1B. Las particiones distantes de la diagonal en la Figura 4.1A tienden a estar más cerca de ella al utilizar aumento de datos. Este efecto es más pronunciado en el sector de bajo *recall*. En la Figura 4.1B, los dos sujetos de menor *recall* desplazan su punto de operación de forma más notoria que el resto

Tabla 4.1: Desempeño de REDv2-Time en MASS-SS2-Train, con y sin aumento de datos de ondas aleatorias y anti-ondas aleatorias. Se reporta el promedio y la desviación estándar de las particiones, y en paréntesis la desviación estándar entre sujetos.

Evento	Aumento	F1-score (%)	Recall (%)	Precision (%)	mIoU (%)
E1SS	No	$79{,}0 \pm 3{,}4$ $(4{,}7)$	$82{,}3 \pm 9{,}0$ $(11{,}2)$	$78{,}4 \pm 6{,}9$ $(9{,}9)$	$83{,}8 \pm 2{,}1$ $(3{,}0)$
	Sí	$79{,}4 \pm 2{,}9$ $(4{,}0)$	$83{,}4 \pm 7{,}8$ $(9{,}8)$	$77{,}8 \pm 6{,}8$ $(9{,}5)$	$84{,}0 \pm 1{,}9$ $(2{,}8)$
KC	No	$83{,}7 \pm 2{,}3$ $(3{,}2)$	$85{,}4 \pm 3{,}6$ $(5{,}4)$	$82{,}6 \pm 3{,}4$ $(5{,}6)$	$90{,}3 \pm 0{,}7$ $(0{,}9)$
	Sí	$83{,}8 \pm 2{,}4$ $(3{,}2)$	$85{,}4 \pm 3{,}8$ $(5{,}5)$	$82{,}8 \pm 4{,}0$ $(6{,}0)$	$90{,}1 \pm 0{,}9$ $(1{,}1)$

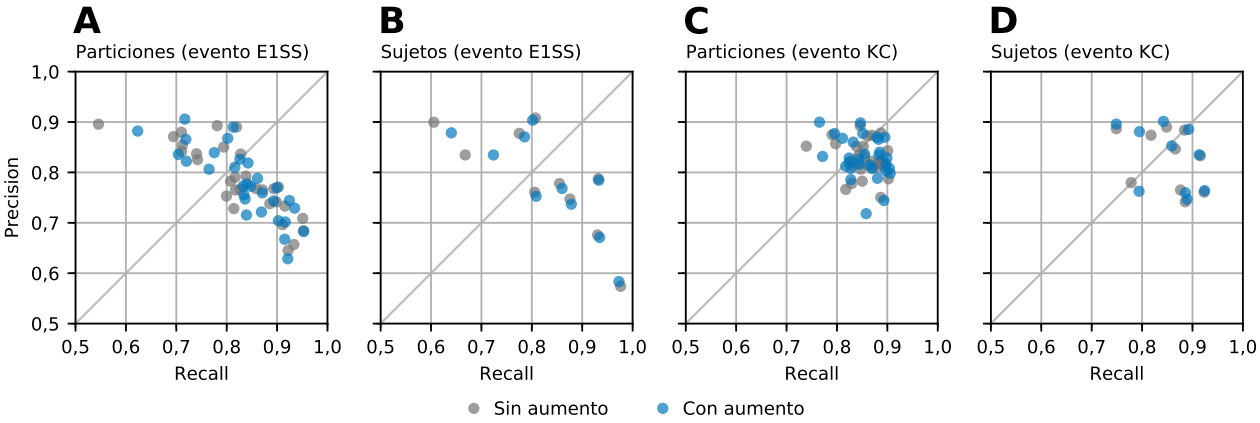

Figura 4.1: Desempeño de REDv2-Time en MASS-SS2-Train, con (en azul) y sin (en gris) aumento de datos de ondas aleatorias y anti-ondas aleatorias. El *recall* y el *precision* en husos de sueño (MASS-SS2-E1SS; **A** y **B**) y en complejos K (MASS-SS2-KC; **C** y **D**) se muestra de dos formas: (**A** y **C**) particiones individuales ($N = 30$), y (**B** y **D**) sujetos individuales ($N = 10$).

de los sujetos, y lo hacen hacia la diagonal. Por el contrario, el sujeto con el menor *precision* (el único con un promedio por debajo de 60 %) permanece similar con o sin aumento de datos. Las Figuras 4.1C y 4.1D muestran que la dispersión en KC permanece similar. En comparación al caso de E1SS, la dispersión inicial de KC (sin aumento de datos) es notoriamente menor y no está tan afectada por sujetos con diferencias grandes entre *recall* y *precision*, que es donde el aumento de datos muestra su mayor efecto en E1SS. En todos los experimentos siguientes se utiliza siempre el aumento de datos.

4.2. Desempeño comparado con la literatura

En esta sección, se evalúa el desempeño de los modelos propuestos y se compara con el obtenido por los detectores seleccionados de la literatura. Además del desempeño típico, se evalúa la calidad de las detecciones comparando sus parámetros con los de las anotaciones, tanto al nivel del evento como del sujeto. También se analiza el desempeño para distintos subconjuntos de los eventos de acuerdo a intervalos de parámetros. Por último, se analiza la generalización al transferir los detectores directamente de una base de datos a otra. Ejemplos de falsos negativos, falsos positivos, y verdaderos positivos, tanto de husos de sueño como de complejos K, se ilustran en el Anexo D.

4.2.1. Desempeño general

La Tabla 4.2 reporta las métricas obtenidas para todos los modelos y bases de datos con etiquetas expertas usando el esquema de particiones descrito en la Sección 3.2.3. Se reporta el promedio y la desviación estándar entre las particiones evaluadas. Además, la desviación estándar entre sujetos se reporta en la Figura 4.2. Como se indica en la Sección 3.2.3, el desempeño en las particiones reportado en la Tabla 4.2 se obtiene por macro-promedio, excepto en MASS-MODA en donde se obtiene por micro-promedio. Por otro lado, las dispersiones reportadas en la Figura 4.2 para MASS-SS2 corresponden al esquema de validación cruzada completa con $N = 15$. Con el fin de obtener métricas y parámetros fiables en MASS-MODA, la dispersión entre sujetos, así como todos los análisis siguientes que requieren mediciones por sujeto, se calcula considerando solo los sujetos con 10 bloques de señal que además tienen al menos 10 anotaciones de husos de sueño ($N = 28$).

Tabla 4.2: Desempeño de la detección, obtenido con el esquema de partición de datos descrito en la Sección 3.2.3. Se reporta el promedio y la desviación estándar entre las particiones evaluadas.

Datos	Detector	F1-score (%)	Recall (%)	Precision (%)	mIoU (%)
MASS-SS2-E1SS	**REDv2-Time**	81,0 ± 0,5	82,6 ± 1,0	80,3 ± 1,2	85,1 ± 0,2
(independiente)[a]	**REDv2-CWT**	80,9 ± 0,4 [ns]	82,4 ± 1,3	80,3 ± 1,7	84,7 ± 0,4 [s]
	DOSED [8]	78,0 ± 0,5 [s]	77,7 ± 2,4	79,8 ± 2,0	75,3 ± 1,3 [s]
	A7 [59]	69,7 ± 0,4 [s]	82,7 ± 1,9	61,2 ± 1,5	74,9 ± 0,2 [s]
MASS-SS2-E2SS	**REDv2-Time**	85,1 ± 0,5	83,9 ± 0,9	87,4 ± 1,6	78,0 ± 0,2
(independiente)	**REDv2-CWT**	85,1 ± 0,5 [ns]	84,7 ± 0,8	86,4 ± 1,4	78,0 ± 0,2 [ns]
	DOSED [8]	81,8 ± 0,7 [s]	79,7 ± 1,4	85,0 ± 1,4	73,8 ± 0,5 [s]
	A7 [59]	73,4 ± 0,1 [s]	82,8 ± 0,0	66,4 ± 0,2	74,8 ± 0,0 [s]
MASS-SS2-KC	**REDv2-Time**	83,3 ± 0,4	82,4 ± 1,4	85,1 ± 0,9	90,5 ± 0,2
(independiente)	**REDv2-CWT**	83,5 ± 0,4 [ns]	82,1 ± 0,7	85,7 ± 0,7	90,3 ± 0,3 [s]
	DOSED [8]	78,0 ± 0,9 [s]	76,5 ± 1,9	80,5 ± 1,9	72,2 ± 1,3 [s]
	Spinky [65]	65,7 ± 0,2 [s]	65,0 ± 1,8	67,7 ± 2,2	42,3 ± 0,1 [s]
MASS-MODA	**REDv2-Time**	81,8 ± 1,4	83,1 ± 2,7	80,6 ± 2,5	83,4 ± 0,5
(CV completo)[b]	**REDv2-CWT**	81,5 ± 1,3 [ns]	82,4 ± 2,4	80,8 ± 2,7	83,2 ± 0,5 [ns]
	DOSED [8]	77,5 ± 1,7 [s]	76,4 ± 2,8	78,9 ± 3,0	71,4 ± 1,1 [s]
	A7 [59]	73,3 ± 1,9 [s]	74,1 ± 2,1	72,8 ± 3,6	71,0 ± 0,9 [s]
INTA-UCH	**REDv2-Time**	83,9 ± 4,0	84,8 ± 5,5	83,8 ± 5,7	75,6 ± 2,5
(CV completo)	**REDv2-CWT**	83,5 ± 4,3 [ns]	84,7 ± 5,8	83,3 ± 5,8	75,5 ± 2,4 [ns]
	DOSED [8]	78,1 ± 6,2 [s]	78,6 ± 11,8	80,0 ± 6,6	68,4 ± 4,3 [s]
	A7 [59]	78,4 ± 4,6 [s]	77,7 ± 6,3	80,4 ± 7,4	69,6 ± 2,8 [s]
MASS-SS2-E1SS	**REDv2-Time**	80,8 ± 2,1	83,6 ± 4,1	79,7 ± 5,3	84,8 ± 1,2
(CV completo)	**REDv2-CWT**	80,8 ± 2,0 [ns]	83,8 ± 4,4	79,6 ± 5,5	84,5 ± 1,2 [ns]
($N = 15$)	DOSED [8]	76,8 ± 2,9 [s]	79,7 ± 5,9	77,5 ± 7,8	74,7 ± 2,1 [s]
	A7 [59]	73,0 ± 3,4 [s]	80,1 ± 4,0	68,1 ± 5,5	73,9 ± 1,0 [s]
MASS-SS2-E1SS	**REDv2-Time**	80,5 ± 2,1	83,6 ± 4,2	78,8 ± 3,3	84,7 ± 1,0
(CV completo)	**REDv2-CWT**	79,9 ± 2,5 [ns]	81,7 ± 5,3	79,8 ± 3,3	84,5 ± 1,1 [ns]
($N = 19$)	SpindleU-Net [10] [c]	80,3 ± 2,3 [ns]	83,0 ± 3,3	78,8 ± 3,4	73,5
MASS-SS2-E2SS	**REDv2-Time**	86,1 ± 2,0	86,3 ± 3,9	86,8 ± 3,6	78,7 ± 1,1
(CV completo)	**REDv2-CWT**	86,1 ± 2,1 [ns]	86,9 ± 4,1	86,1 ± 3,8	78,8 ± 1,0 [ns]
($N = 15$)	DOSED [8]	82,5 ± 2,5 [s]	84,0 ± 5,0	82,5 ± 4,9	73,1 ± 1,1 [s]
	A7 [59]	74,9 ± 2,8 [s]	81,5 ± 3,1	70,0 ± 4,3	74,7 ± 1,1 [s]
	SpindleU-Net [10] [c]	85,4 ± 2,7 [ns]	86,2 ± 4,5	86,4 ± 4,9	Desconocido
	SpindleNet [9] [c]	83,0 ± 2,0 [s]	85.2	81,0 ± 3,2	Desconocido
MASS-SS2-KC	**REDv2-Time**	83,7 ± 1,5	85,1 ± 3,2	83,0 ± 3,2	90,6 ± 0,6
(CV completo)	**REDv2-CWT**	83,8 ± 1,4 [ns]	84,9 ± 2,5	83,3 ± 2,5	90,4 ± 0,5 [ns]
($N = 15$)	DOSED [8]	78,1 ± 2,2 [s]	79,2 ± 3,7	77,7 ± 3,3	72,3 ± 1,4 [s]
	Spinky [65]	63,1 ± 3,8 [s]	61,6 ± 3,5	65,6 ± 6,3	41,2 ± 1,6 [s]
MASS-SS2-KC	**REDv2-Time**	83,6 ± 1,7	85,0 ± 3,6	82,9 ± 2,3	90,4 ± 0,4
(CV completo)	**REDv2-CWT**	83,8 ± 1,7 [ns]	84,7 ± 3,6	83,5 ± 2,1	90,3 ± 0,5 [ns]
($N = 19$)	DKL-KC [7] [c]	78,0 ± 2,0 [s]	80,0 ± 5,0	77,0 ± 5,0	Desconocido

[a] Conjunto de prueba restringido a MASS-SS2-Test (ver Sección 3.2.3).
[b] Todos los datos participan en la validación cruzada (ver Sección 3.2.3).
[c] Se indica y adapta el desempeño reportado en la publicación original.
[ns] Diferencia no significativa con REDv2-Time ($P > 0{,}05$).
[s] Diferencia significativa con REDv2-Time ($P < 0{,}05$).

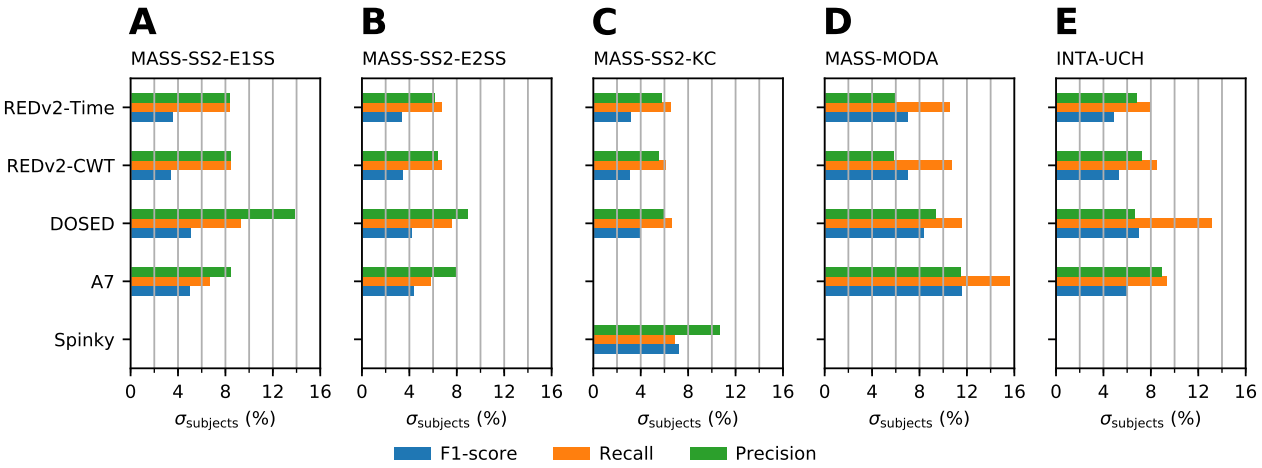

Figura 4.2: Dispersión (desviación estándar) entre sujetos del desempeño de la detección. En MASS-MODA solo se consideran los sujetos con suficientes datos.

En general, REDv2-Time y REDv2-CWT no muestran diferencias significativas entre ellos en *F1-score* y en la mayoría de las mediciones de mIoU, aunque en algunos casos REDv2-Time presenta una pequeña pero significativa ventaja en mIoU frente a REDv2-CWT. Así como en este caso, en varios de los resultados que siguen no existen diferencias relevantes entre las variantes REDv2-Time y REDv2-CWT. Por simplicidad, cada vez que sea irrelevante diferenciar entre ambos modelos, se usa la designación REDv2.

Para ambos modelos, el desempeño en MASS-SS2-Test (evaluación independiente al diseño) se mantiene cercano al desempeño obtenido cuando se usan todos los sujetos de MASS-SS2 (evaluación con riesgo de sobre-ajuste). Por lo tanto, el sobre-ajuste a los sujetos de MASS-SS2 se puede considerar despreciable. A continuación, se restringe el análisis a los resultados obtenidos con el esquema de validación cruzada completa en todas las bases de datos por brevedad. Además, esto permite considerar desempeños reportados en la literatura.

En los escenarios evaluados en la Tabla 4.2, REDv2 consigue un *F1-score* más alto respecto a DOSED, A7, Spinky, SpindleNet, y DKL-KC. En cambio, REDv2 no se diferencia significativamente del *F1-score* reportado para SpindleU-Net en MASS-SS2-E1SS ($N = 19$) y MASS-SS2-E2SS. Sin embargo, SpindleU-Net reporta un menor mIoU que REDv2 en MASS-SS2-E1SS ($\approx$11 % de diferencia). La ausencia de la desviación estándar no permite hacer el test estadístico, pero un análisis de sensibilidad muestra que para que dicha diferencia no sea significativa, la desviación estándar debe ser mayor a 9. Debido a que en la Tabla 4.2 la desviación estándar del mIoU tiende a ser menor que la del F1-score, y dado que en la publicación original se reporta que SpindleU-Net tiene dispersiones similares a DOSED, es muy probable que esta diferencia sea significativa. Lamentablemente, no se reporta el mIoU de SpindleU-Net para MASS-SS2-E2SS. En general, REDv2 también consigue una menor dispersión entre sujetos respecto a los demás detectores. A pesar de ello, dicha dispersión, en particular para *recall* y *precision*, sigue siendo perfectible. Entre las bases de datos, las menores dispersiones se consiguen en MASS-SS2-KC, seguida de cerca por MASS-SS2-E2SS. Por simplicidad, todos los análisis que siguen de MASS-SS2 se restringen al caso de validación cruzada completa con $N = 15$.

Al igual que varios otros métodos, REDv2 puede desplazar su punto de operación *precision-*

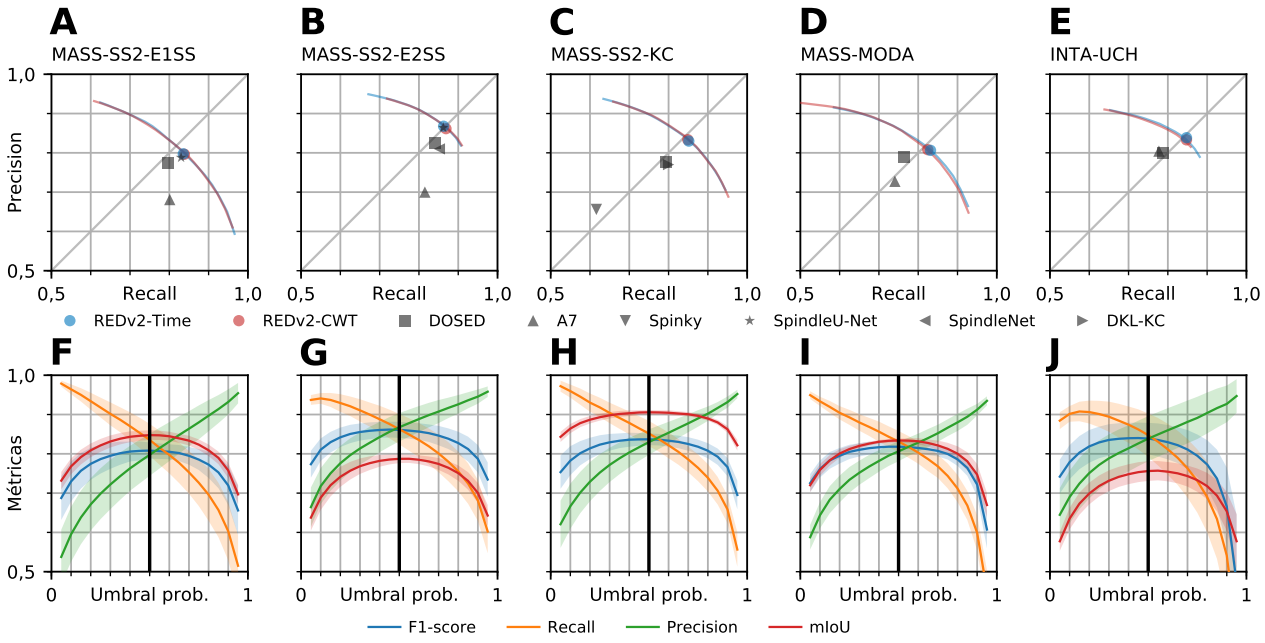

Figura 4.3: Efecto del umbral de probabilidad en el desempeño. (**A**-**E**) *Recall* y *precision* promedio, junto a la curva que se obtiene al variar el umbral de probabilidad ajustada a la salida de los modelos propuestos. Los símbolos usados para los distintos modelos están indicados al pie de los paneles A-E. (**F**-**J**) Promedio y desviación estándar de las métricas de desempeño de REDv2-Time (*F1-score*, *recall*, *precision* y mIoU) al variar su umbral de probabilidad ajustada. El umbral 0,5 es el punto de operación entregado por el entrenamiento.

recall ajustando un umbral. En este caso, se puede modificar el umbral de probabilidad ajustada, cuyo valor por defecto después del aprendizaje del modelo es 0,5. El efecto de modificar este umbral se muestra en la Figura 4.3. Como las diferencias entre MASS-SS2 con $N = 15$ y $N = 19$ son pequeñas, también se muestran SpindleU-net y DKL-KC. Por un lado, las Figuras 4.3A-E muestran que REDv2 admite un amplio rango de puntos de operación a través de este único hiperparámetro. Además, estos nuevos puntos de operación siguen Pareto-dominando aquellos determinados por los detectores de la literatura. Por otro lado, las Figuras 4.3F-J muestran que REDv2 admite modificar el umbral de probabilidad en un intervalo amplio alrededor de 0,5 con un efecto acotado en *F1-score* y mIoU. De hecho, existe una vecindad en torno a 0,5 en donde estas dos métricas se mantienen relativamente estables, particularmente para MASS-SS2-KC y MASS-MODA. Por lo tanto, se puede producir un cambio notorio en el compromiso entre *precision* y *recall* sin afectar tanto el desempeño.

Para investigar el efecto del umbral de IoU en el desempeño, la Figura 4.4 muestra el *F1-score* como función de este umbral, y el histograma de los valores de IoU de apareamiento (i.e., de las asociaciones válidas entre anotación y detección). En general, REDv2 es mejor en todos los umbrales de IoU según las Figuras 4.4A-E. En consecuencia, REDv2 consigue mayor AF1. Para todos los detectores evaluados, el *F1-score* varía recién a partir del umbral 0,2. Por lo tanto, las métricas de *F1-score*, *recall* y *precision* reportadas con umbral 0,2 son prácticamente invariantes a la exactitud con que se predice el inicio y fin de los eventos. Dicha exactitud se puede analizar usando los histogramas de IoU de las Figuras 4.4F-J y su colapso en la métrica mIoU. Tal como se mencionó antes, REDv2 tiene mejor mIoU. Los histogramas de IoU evidencian que REDv2 concentra un gran porcentaje de la distribución por sobre 0,8 mientras que mantiene una cola reducida hacia valores menores (i.e., baja dispersión). La excepción es MASS-SS2-E2SS e INTA-UCH, en donde

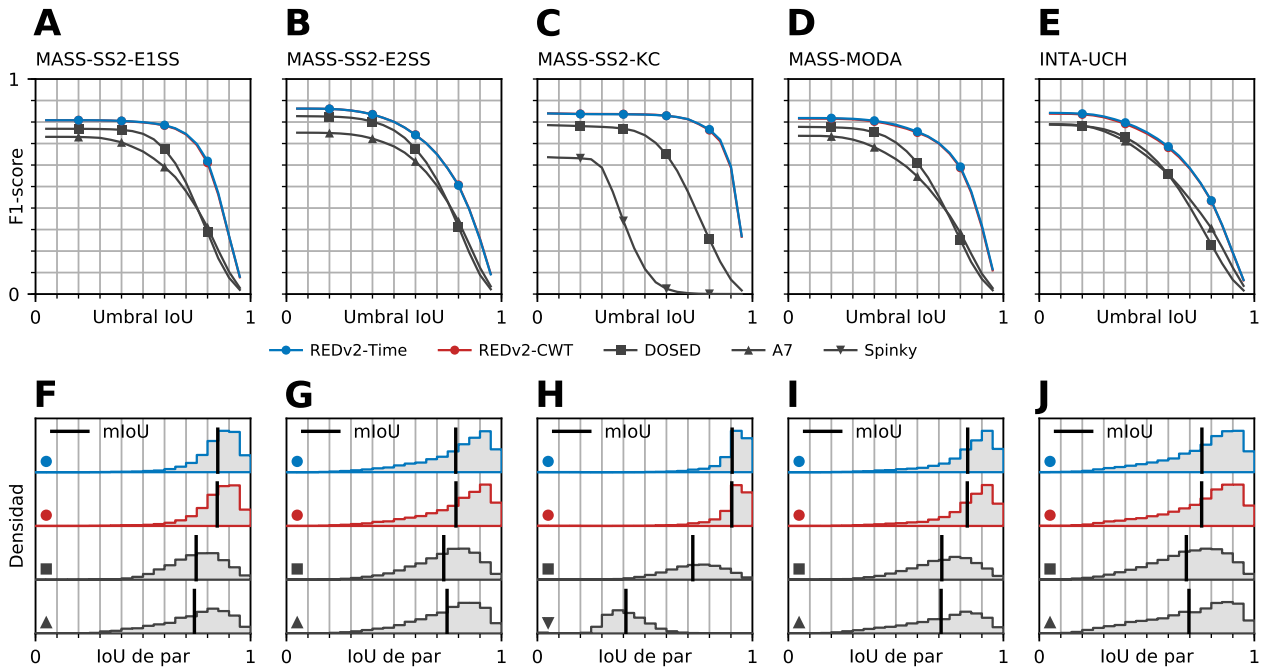

Figura 4.4: Efecto del umbral de IoU en el desempeño. (**A**-**E**) Curva promedio de *F1-score* en función del umbral de IoU. (**F**-**J**) Histograma promedio de los valores de IoU de apareamiento. Los símbolos usados para los distintos modelos están indicados al pie de los paneles A-E.

las dispersiones son mayores para todos los métodos. La mejora en IoU es particularmente importante en MASS-SS2-KC, en donde REDv2 es ≈18 % mejor que DOSED, el método que le sigue en desempeño. Por último, se observa que Spinky obtiene un desempeño notoriamente bajo en la localización de complejos K, una consecuencia directa de su falta de predicción de la duración.

4.2.2. Ajuste de parámetros por evento

La Figura 4.5 muestra la relación entre la duración experta y la detectada para las anotaciones que fueron apareadas con una detección (métricas complementarias se muestran en la Tabla E.1). Para mejorar la legibilidad y la comparación, los gráficos son acotados al rango 0–2 s, dejando fuera menos del 2 % de las anotaciones de MASS-MODA. No se observan diferencias significativas entre REDv2-Time y REDv2-CWT. Tanto para husos de sueño como para complejos K, REDv2 consigue un mejor ajuste, reflejado en una alta correlación, menor magnitud del error y menor sesgo. Por otro lado, DOSED tiene detecciones notoriamente sesgadas, resultando en predicciones que tienden a ser más largas de lo debido (≈ 0,16 s de sesgo en ambas bases). Los comportamientos anormales observados en A7 y Spinky se explican porque A7 tiene una resolución de 0,1 s mientras que Spinky asume una duración constante.

Para evaluar la calidad de las detecciones dentro de una base de datos completa, en la Figura 4.6 se muestra la distribución global experta de parámetros típicos junto a la distribución global inducida por cada detector. Específicamente, se evalúa la distribución de la duración, la amplitud PP y la frecuencia central de husos de sueño individuales, y la distribución de la duración y la amplitud PP de complejos K individuales. No se observan diferencias significativas entre REDv2-Time y REDv2-CWT. La duración es mejor ajustada por REDv2, mientras que DOSED posee un sesgo hacia duraciones mayores. Si bien el detector A7 posee una media y un decaimiento a partir de su

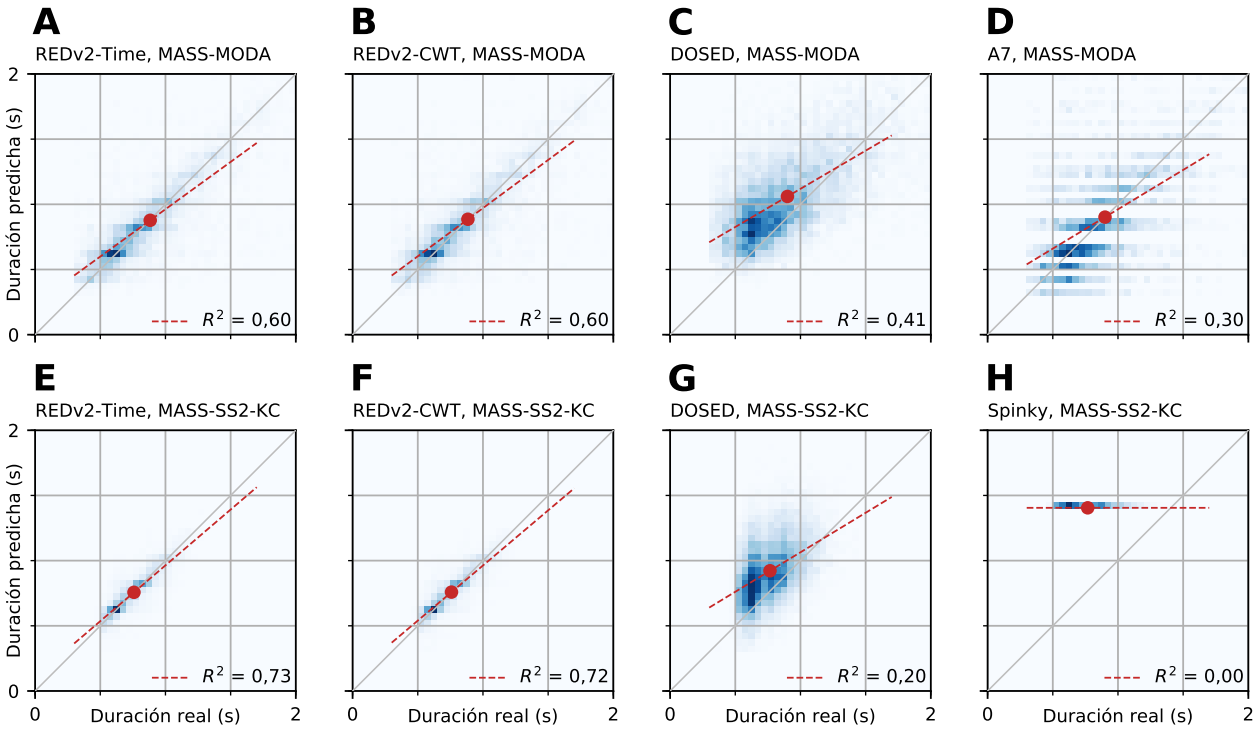

Figura 4.5: Desempeño del ajuste de duración por evento. Se muestra la duración real y la predicha través de un histograma 2D, junto al ajuste por regresión lineal y la media de las distribuciones. (**A**-**D**) Desempeño para husos de sueño (MASS-MODA). (**E**-**H**) Desempeño para complejos K (MASS-SS2-KC).

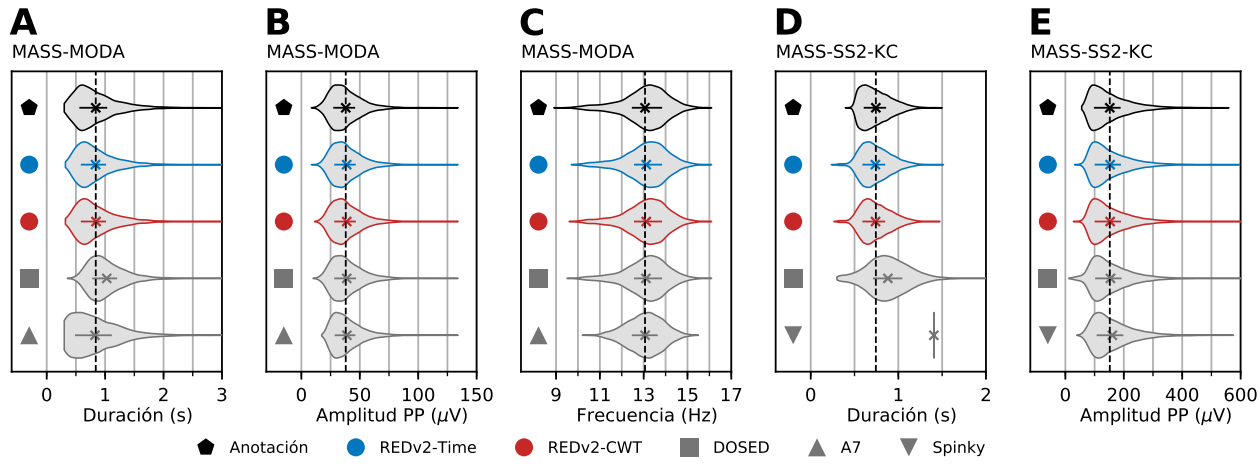

Figura 4.6: Desempeño del ajuste de la distribución de parámetros por evento. Se muestra la distribución real del parámetro (dada por las anotaciones) y la distribución obtenida a partir de las detecciones. En cada distribución, se marca el promedio con una cruz y el rango intercuartil con una línea horizontal. Se destaca con una línea vertical negra el promedio de las anotaciones. (**A**-**C**) Distribuciones para husos de sueño (MASS-MODA). (**D**-**E**) Distribuciones para complejos K (MASS-SS2-KC).

moda similares a los observados en la distribución experta de duración, posee muchas más detecciones cortas de husos de sueño. La distribución de amplitud PP de todos los detectores es similar a la experta salvo algunas diferencias en los extremos. En husos de sueño, REDv2 y DOSED casi no presentan diferencias con la distribución experta, mientras que A7 ignora un intervalo de amplitudes pequeñas. En complejos K, tanto REDv2 como DOSED presentan algunas detecciones con amplitudes menores a la esperada a partir de la distribución experta. Sin embargo, en REDv2 dicha cantidad es despreciable, mientras que en DOSED es significativa. Al contrario, Spinky presenta un sesgo hacia amplitudes mayores. Por último, la distribución de la frecuencia central en los husos de sueño detectados por REDv2 y DOSED posee diferencias despreciables con respecto a la distribución experta. En cambio, A7 induce una distribución más concentrada, con casi ninguna detección por debajo de 11 Hz, aunque con la misma media experta.

4.2.3. Ajuste de parámetros por sujeto

Siguiendo el análisis al nivel de sujetos planteado en [6], se calculan parámetros que representan un registro completo, a partir de todos los eventos pertenecientes a ese registro. Específicamente, se considera la densidad de eventos durante la etapa N2 (medida en eventos por minuto, epm), la duración promedio, la amplitud PP promedio, y, para el caso de husos de sueño, la frecuencia central promedio. Se espera que los parámetros obtenidos con las detecciones aproximen aquellos obtenidos por anotaciones expertas. Las Figuras 4.7 y 4.8 muestran la relación entre los parámetros expertos y los parámetros estimados para husos de sueño y complejos K, respectivamente (métricas complementarias se muestran en la Tabla E.2). Debido a las repeticiones de la validación cruzada, se tienen tres puntos por sujeto. Por lo tanto, se tienen 84 puntos en MASS-MODA y 45 puntos en MASS-SS2-KC.

En general, REDv2-Time y REDv2-CWT no muestran diferencias significativas. La excepción es el ajuste de duración promedio de los husos de sueño, en donde REDv2-Time tiene una correlación significativamente más alta que REDv2-CWT. En comparación con los otros detectores, REDv2 siempre es mejor o al menos tan bueno como la mejor alternativa. En husos de sueño, REDv2 es mejor estimando la duración promedio, la densidad, y la frecuencia promedio, aunque en este último parámetro es seguido de cerca por DOSED. Además, en correlación es al menos tan bueno como DOSED estimando la amplitud promedio, aunque REDv2 posee un menor sesgo (0,7–0,8 μV contra 1,5 μV) y una menor desviación estándar del error (1,0–1,1 μV contra 1,4 μV). En complejos K, REDv2 es mejor estimando la duración promedio y la amplitud PP, aunque en este último parámetro es seguido de cerca por DOSED. Además, es al menos tan bueno como DOSED estimando la densidad. Tanto en husos de sueño como en complejos K, la mayor ganancia de REDv2 con respecto a los otros detectores ocurre en la estimación de la duración promedio.

4.2.4. Desempeño por subconjuntos de parámetros

El desempeño de un detector podría ser sensible a algún parámetro de los eventos. Por ejemplo, el desempeño podría ser significativamente diferente entre los eventos más cortos y los más largos. Para medir estos efectos, se evalúa el cambio en el desempeño cuando el conjunto de verdaderos y falsos se restringe al interior de diversos intervalos (i.e., subconjuntos) de un parámetro dado. En husos de sueño, se consideran subconjuntos de duración, amplitud PP, frecuencia central y edad (fase de MASS-MODA). En complejos K, se consideran subconjuntos de duración y amplitud PP. La edad se divide en dos intervalos (adultos jóvenes y adultos mayores), mientras que los demás

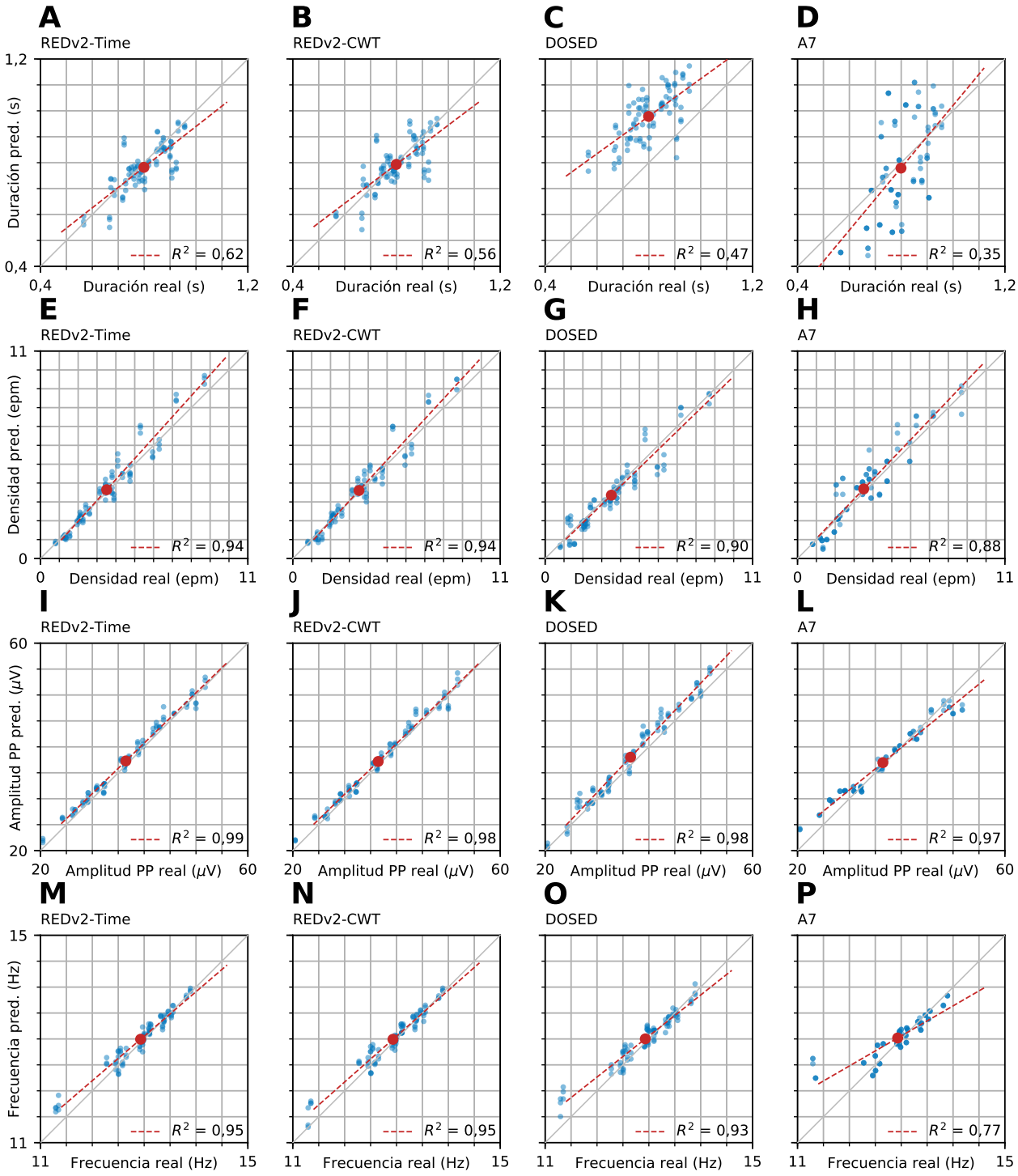

Figura 4.7: Desempeño del ajuste de parámetros por sujeto para husos de sueño. Para cada sujeto de MASS-MODA con suficientes datos, se muestra el parámetro real y el estimado a partir de las detecciones, junto al ajuste por regresión lineal y la media de las distribuciones.

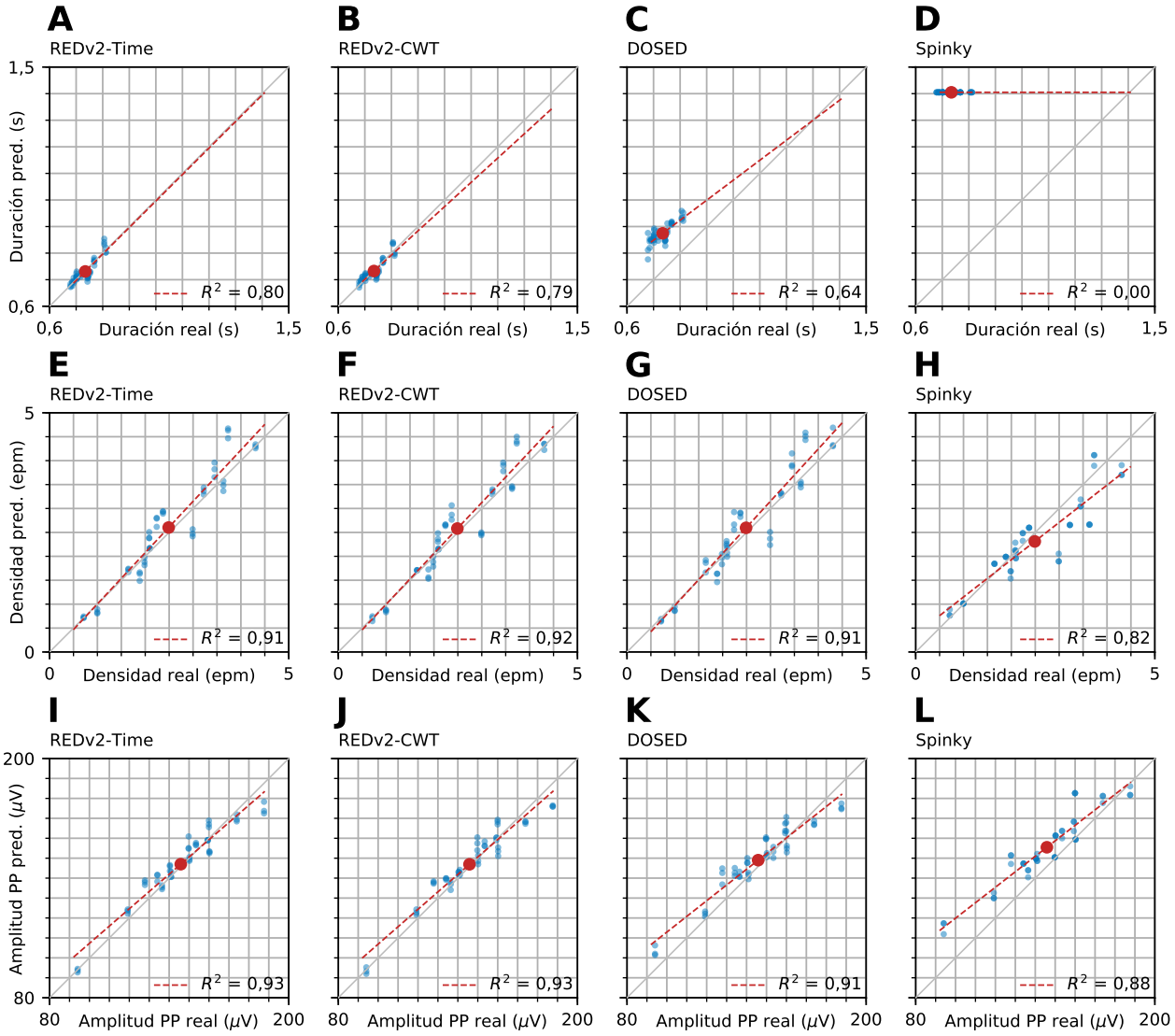

Figura 4.8: Desempeño del ajuste de parámetros por sujeto para complejos K. Para cada sujeto de MASS-SS2-KC, se muestra el parámetro real y el estimado a partir de las detecciones, junto al ajuste por regresión lineal y la media de las distribuciones.

parámetros se dividen en tres intervalos consecutivos que contienen aproximadamente el mismo número de anotaciones.

Para los apareamientos entre anotaciones y detecciones, la pertenencia a un subconjunto u otro del par completo se decide en base al parámetro de la anotación. A pesar de que en total cada intervalo contiene un número comparable de eventos y detecciones, esto no se asegura al nivel de cada sujeto. Por lo tanto, se calculan las métricas de desempeño por micro-promedio incluso en MASS-SS2-KC. Los resultados de este experimento se ilustran en la Figura 4.9 (mediciones precisas de *F1-score* se muestran en las Tablas E.3 y E.4). Se deja fuera del análisis por duración a Spinky.

En general, REDv2-Time y REDv2-CWT no muestran diferencias significativas. En comparación con los otros detectores, REDv2 es mejor en *F1-score* y mIoU en todos los subconjuntos excepto en los husos de sueño con duración menor a 0,6 s. En dicho subconjunto, DOSED tiene una pequeña ventaja en *F1-score* aunque no es significativa ($P > 0{,}05$). Dicha ventaja proviene de un *precision* notoriamente alto, consecuencia directa de la baja cantidad de detecciones cortas predichas por DOSED.

Más allá de las diferencias en desempeño entre REDv2 y los demás detectores, existen tendencias claras para los efectos de los parámetros que son comunes para todos los detectores. El *F1-score* crece con la duración y la amplitud PP en ambas clases de eventos. Además, el subconjunto de menor duración y el subconjunto de menor amplitud representan los peores desempeños, por un margen importante, de todos los subconjuntos evaluados. Al observar el efecto de la frecuencia en husos de sueño, el mejor *F1-score* se encuentra en el rango medio (12,8–13,5 Hz), con una baja pronunciada en las oscilaciones más lentas ($<$ 12,8 Hz) y una baja ligera en las oscilaciones más rápidas ($>$ 13,5 Hz). Por último, al observar el efecto de la edad de los sujetos en husos de sueño, el mejor *F1-score* se tiene para adultos jóvenes, con una diferencia de 5–5,5 % para REDv2 y 6,4–6,6 % para los demás detectores, con respecto a adultos mayores.

4.2.5. Desempeño ante transferencia directa

Para analizar la generalización de los detectores aprendidos, se evalúa el desempeño de la detección de husos de sueño cuando un detector, luego de ajustarse en una determinada base de datos, se usa directamente para inferir en otras bases de datos. Por simplicidad, el detector ajustado usando la partición i-ésima de la base de datos de entrenamiento solo se usa para inferir en la partición i-ésima de la base de datos de evaluación. De esta forma, también se asegura la independencia de los datos al transferir entre MASS-SS2-E1SS y MASS-SS2-E2SS. Los resultados de este experimento se ilustran en la Figura 4.10 (mediciones precisas de *F1-score* se muestran en la Tabla E.5).

Los detectores REDv2 y DOSED requieren normalizar las señales antes de procesarlas, y en un escenario de transferencia cada detector admite dos opciones naturales: usar la normalización que se usaría normalmente en la base de evaluación, o mantener una normalización cercana a la realizada en la base de entrenamiento. En REDv2, las señales pueden ser normalizadas usando la desviación estándar global de la base de evaluación o de la base de entrenamiento. En DOSED, como las señales se normalizan a media cero y varianza unitaria de forma independiente, al transferir se puede aplicar el modelo directamente o bien escalar antes la señal normalizada por un factor igual a la razón entre la desviación estándar global de la base de evaluación y la de la base de

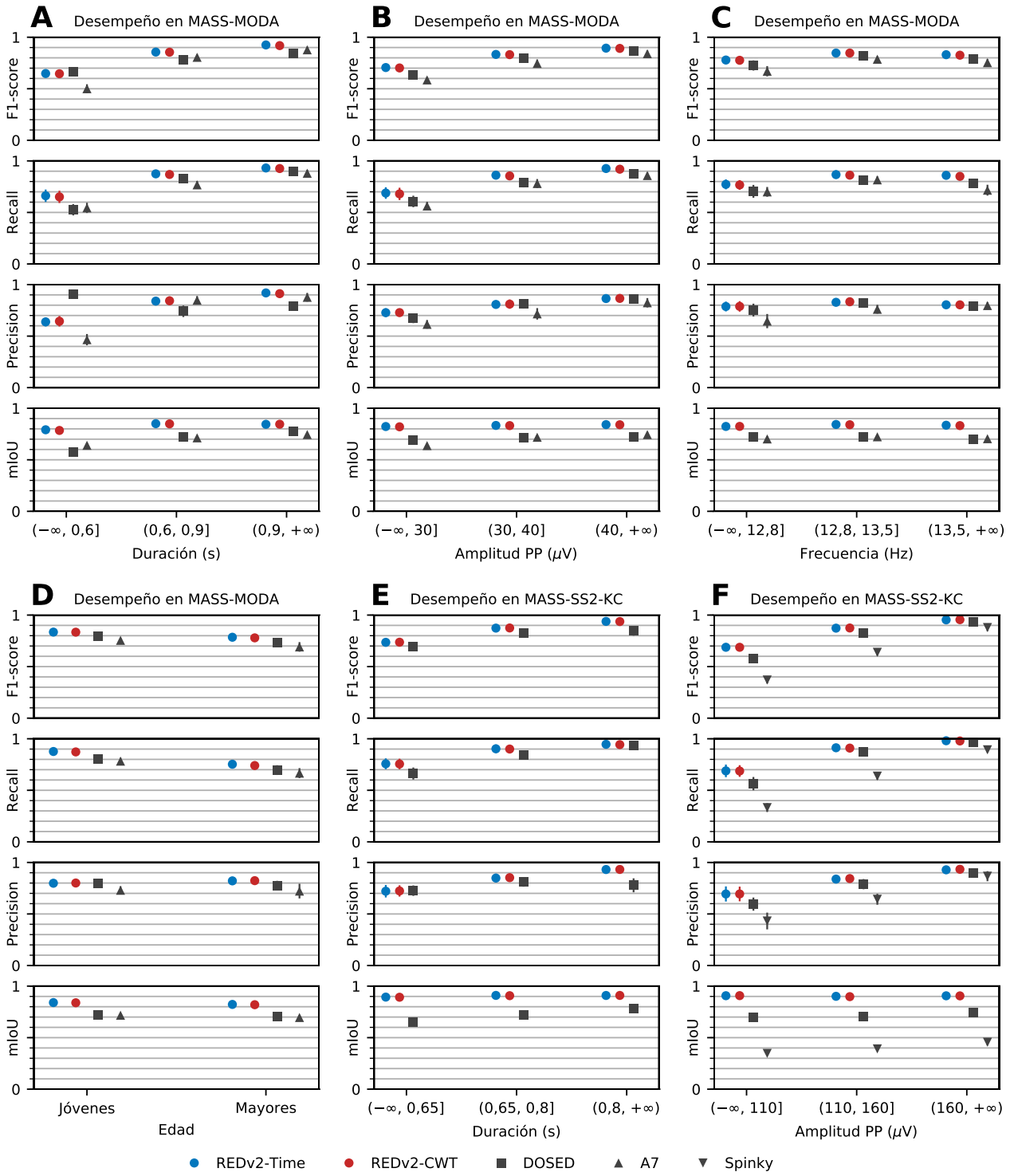

Figura 4.9: Desempeño de la detección cuando solo se consideran verdaderos y falsos que existen al interior de un intervalo de un parámetro dado (e.g., solo eventos y detecciones con duración entre 0,6 s y 0,9 s). Se indica la media con un marcador, y un rango de una desviación estándar con una línea vertical. Métricas calculadas por micro-promedio. (**A**-**D**) Desempeño en husos de sueño (MASS-MODA). (**E**-**F**) Desempeño en complejos K (MASS-SS2-KC).

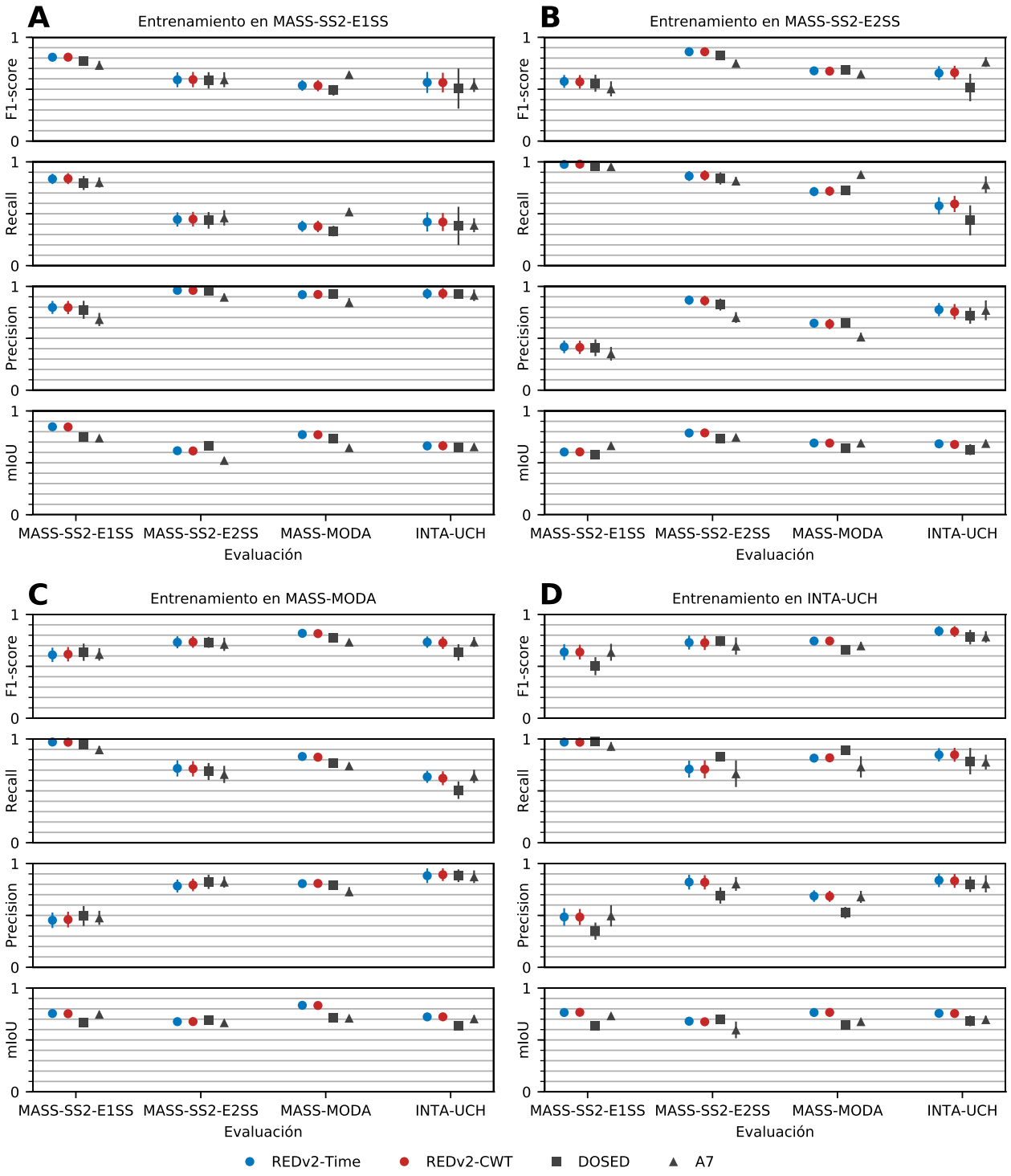

Figura 4.10: Desempeño de la detección de husos de sueño cuando el modelo se transfiere directamente de una base de datos a otra. Se indica la media con un marcador y un intervalo de una desviación estándar con una línea vertical.

entrenamiento. En ambos detectores, los mejores resultados se consiguen con la segunda opción. Por simplicidad, se omiten los resultados de la primera opción.

En general, REDv2-Time y REDv2-CWT no muestran diferencias significativas. Al transferir, REDv2 muestra un *F1-score* que es mejor o sin diferencias significativas con respecto a DOSED. En cambio, REDv2 tiene un *F1-score* significativamente menor al de A7 en dos transferencias: MASS-SS2-E1SS $\rightarrow$ MASS-MODA, y MASS-SS2-E2SS $\rightarrow$ INTA-UCH ($P < 0{,}05$). En las demás transferencias, REDv2 es mejor o sin diferencias significativas con respecto a A7.

A través de todos los detectores, las transferencias tienen efectos similares, lo que sugiere que el cambio del desempeño no está determinado principalmente por el modelo escogido. Además, MASS-SS2-E1SS y MASS-SS2-E2SS contienen los mismos registros, mientras que MASS-MODA posee registros provenientes del mismo cohorte (aunque de una demografía más amplia). Esto sugiere que la mayor contribución al cambio del desempeño proviene de un cambio en las anotaciones.

A través de todos los detectores, las peores transferencias son aquellas que involucran a MASS-SS2-E1SS, ya sea como base de entrenamiento o de evaluación, lo que sugiere que dicha base de datos es la que más dista de las demás. Cuando MASS-SS2-E1SS es la base de entrenamiento, en las demás bases se tiene un *recall* por debajo de 50 %, acompañado de un *precision* alto. Análogamente, cuando MASS-SS2-E1SS es la base de evaluación, se tiene un *precision* por debajo de 50 %, acompañado de un *recall* alto. Estos resultados sugieren que la distancia entre MASS-SS2-E1SS y las otras tres bases surge por un criterio de detección más estricto (i.e., admite menos eventos). Para REDv2 y DOSED, la peor transferencia es MASS-SS2-E1SS $\rightarrow$ MASS-MODA con un *F1-score* de 53,4–53,6 % y 48,9 %, respectivamente. Para A7, la peor transferencia es MASS-SS2-E2SS $\rightarrow$ MASS-SS2-E1SS con un *F1-score* de 50,3 %.

El mIoU también se ve afectado a través de las transferencias, consistente con un cambio en las anotaciones. De todas formas, para REDv2 el mIoU no baja de 60 %, mientras que para los demás detectores no baja de 50 %. Las caídas más intensas de mIoU se observan en las transferencias entre MASS-SS2-E1SS y MASS-SS2-E2SS, lo que es consistente con una gran diferencia entre ambos expertos al momento de determinar el inicio y el fin de las anotaciones.

4.3. Acuerdo entre los modelos propuestos

Los experimentos anteriores sugieren que existen muy pocas diferencias entre REDv2-Time y REDv2-CWT. Para analizar su similitud más directamente, se evalúa el acuerdo que existe entre sus predicciones en todas las bases de datos por medio del *F1-score* y el mIoU (Tabla E.6). Para obtener un acuerdo de referencia, se evalúa también el acuerdo que existe entre dos instancias del mismo modelo (acuerdo intra-modelo). Se simula una segunda instancia a través de la permutación de las predicciones de un mismo sujeto de prueba a través de las tres particiones en las que aparece en el esquema de validación cruzada. El acuerdo entre REDv2-Time y REDv2-CWT es significativamente menor al acuerdo intra-modelo ($P < 0{,}05$) tanto en *F1-score* como en mIoU, excepto en el *F1-score* obtenido en INTA-UCH ($P = 0{,}087$). En *F1-score*, la diferencia es de 1,2–2,9 %.

En base a esta pequeña pero significativa diferencia, se analiza el desempeño de diferentes ensambles entre los dos modelos, y se compara con el ensamble de dos instancias del mismo modelo

(Tabla E.7). Para ensamblar, se considera la unión de las detecciones (ensamble OR), la intersección de las detecciones (ensamble AND), y el promedio de las probabilidades ajustadas (ensamble AVG). La peor estrategia es el ensamble OR, de bajo *precision*, por lo que se omiten sus resultados. Por otro lado, el ensamble AND demuestra ser el más estricto, con alto *precision* pero bajo *recall*. En general, el ensamble AVG entrega las mejores medias. En *F1-score*, dicho ensamble aumenta la media en 0,2–0,4 %. Sin embargo, en ningún caso el ensamble entre REDv2-Time y REDv2-CWT es significativamente mejor que el ensamble entre dos instancias de una misma arquitectura.

4.4. Desempeño ante perturbaciones y datos artificiales

En esta sección, se analizan experimentos que buscan entender la respuesta y la capacidad de los modelos propuestos. En primer lugar, se analiza el desempeño de modelos entrenados al inferir en señales que han sido perturbadas de forma controlada. Después, se analiza la respuesta de modelos entrenados al inferir en señales aleatorias con estadísticas similares a un EEG. Por último, se analiza la capacidad de los modelos para imitar, por medio del entrenamiento, algoritmos de detección de husos de sueño usados tradicionalmente en la literatura.

4.4.1. Perturbaciones de la entrada

Para analizar lo aprendido por los modelos propuestos, se evalúa el desempeño de la detección de husos de sueño y complejos K cuando el modelo entrenado permanece intacto y la señal de entrada sufre una perturbación conocida. Se consideran tres tipos de perturbaciones: escalamiento, inversión de un eje, y filtrado rechaza-banda. Los resultados de este experimento se ilustran en la Figura 4.11 (mediciones precisas de *F1-score* se muestran en la Tabla E.8). En general, no se observan diferencias significativas entre REDv2-Time y REDv2-CWT.

El escalamiento tiene un efecto similar en husos de sueño y en complejos K, aunque con mayor intensidad en complejos K. El *recall* y el *precision* cambian en direcciones opuestas. Cuando se disminuye la escala de la señal, baja el *recall* y aumenta el *precision*. Lo contrario ocurre cuando se aumenta la escala de la señal. Es decir, cuando disminuye o aumenta la escala, disminuye o aumenta el número de detecciones, respectivamente. Esto sugiere que el detector considera la amplitud absoluta para decidir la existencia de un evento. Además, el efecto más intenso en complejos K sugiere que la amplitud absoluta es más importante en dicha clase. Por otro lado, el escalamiento prácticamente no afecta el mIoU. Esto sugiere que el detector da más importancia a la amplitud relativa en comparación con la amplitud absoluta para determinar el inicio y el fin de los eventos, una vez determina su existencia.

Las inversiones tienen efectos diferentes en husos de sueño y en complejos K. En husos de sueño, el desempeño prácticamente no cambia. La excepción es una pequeña pero significativa caída del mIoU ante una inversión temporal. La ausencia de cambio en el *F1-score* para ambas inversiones sugiere que el detector no considera una orientación preferencial en ninguno de los ejes. El efecto en el mIoU de una inversión temporal sugiere que el detector usa un criterio diferente para determinar los instantes de inicio y fin. Por otro lado, la detección de complejos K se ve notoriamente afectada por ambas inversiones, lo que sugiere que el detector considera una orientación preferencial.

Por último, el efecto de eliminar bandas de frecuencia también es diferente en husos de sueños

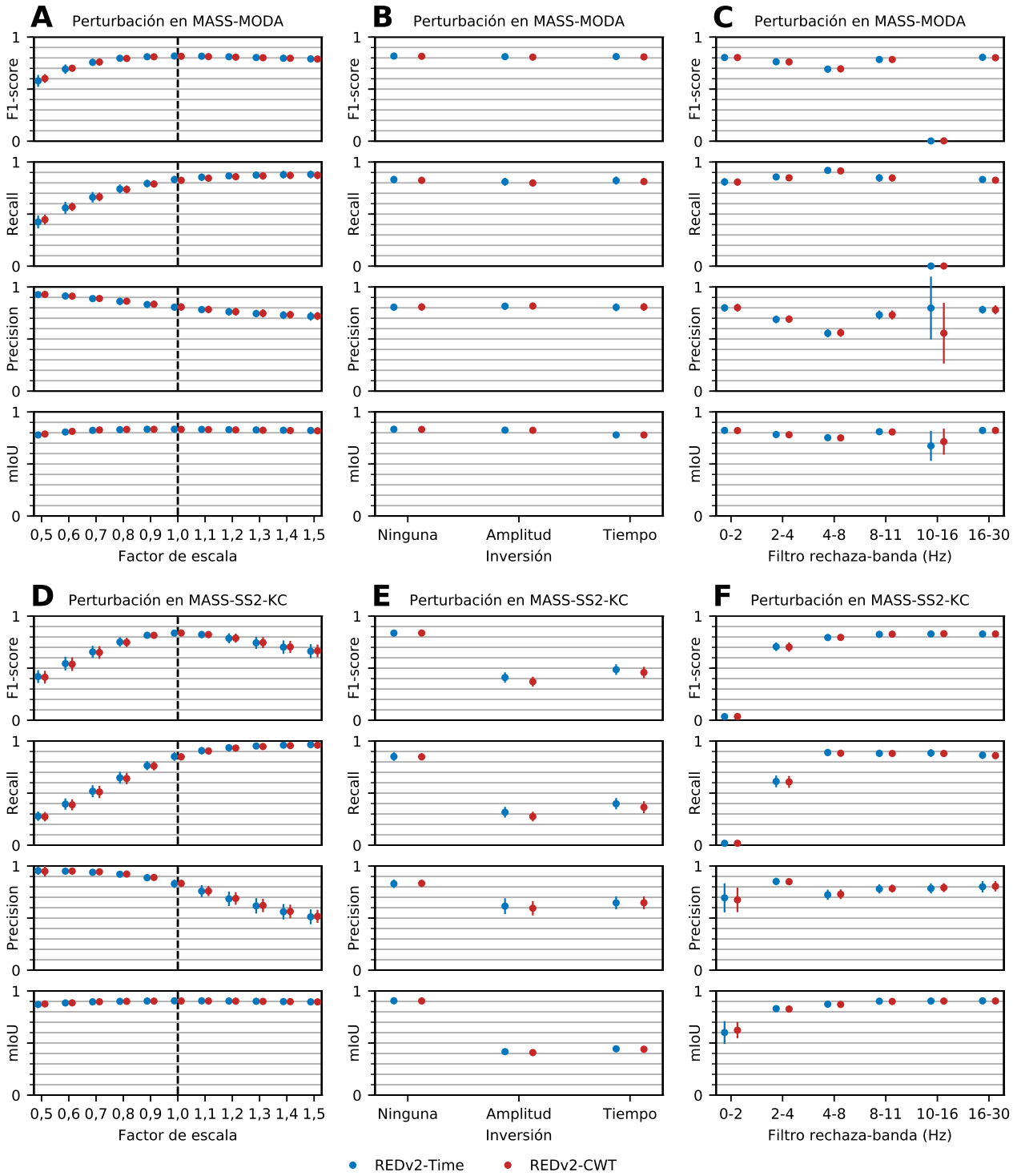

Figura 4.11: Desempeño de la detección de los modelos propuestos cuando la señal de entrada es perturbada. (**A**-**C**) Desempeño en husos de sueño (MASS-MODA). (**D**-**F**) Desempeño en complejos K (MASS-SS2-KC).

Tabla 4.3: Detecciones por hora de los modelos propuestos en PINK.

Datos de entrenamiento	Factor de escala de PINK	Detector	Detecciones por hora
MASS-MODA	1	REDv2-Time	≈ 0
	1	REDv2-CWT	0
	2	REDv2-Time	$1{,}4 \pm 0{,}6$
	2	REDv2-CWT	$1{,}1 \pm 0{,}7$
MASS-SS2-KC	1	REDv2-Time	$13{,}2 \pm 1{,}9$
	1	REDv2-CWT	$12{,}8 \pm 1{,}3$

y en complejos K. En husos de sueño, el *F1-score* es prácticamente nulo al eliminar la banda sigma (10–16 Hz). Además de este caso, el desempeño en *F1-score* y mIoU cae significativamente, en orden de mayor a menor intensidad, al eliminar la banda theta (4–8 Hz), delta rápida (2–4 Hz), y alfa (8–11 Hz), en donde la caída del *F1-score* ocurre principalmente a través de un empeoramiento del *precision*. Al contrario, el efecto de eliminar las bandas delta lenta (0–2 Hz) y beta (16–30 Hz) es muy pequeño. Esto sugiere que el detector requiere actividad sigma de forma excluyente y que, en caso de existir, busca además una baja actividad theta, delta rápida y alfa.

Por otro lado, el desempeño en complejos K se mantiene prácticamente invariante al eliminar las bandas alfa, sigma y beta. En cambio, el *F1-score* está muy cerca de cero ($\approx 3{,}5\,\%$) al eliminar la banda delta lenta. Tanto el *F1-score* como el mIoU sufren una caída significativa, en orden de mayor a menor intensidad, al eliminar la banda delta rápida y theta. La caída del *F1-score* en ambos casos ocurre por razones distintas. Para la banda delta rápida la causa principal es un empeoramiento del *recall*, mientras que para la banda theta la causa principal es un empeoramiento del *precision*. Esto sugiere que el detector se concentra en la banda 0–8 Hz y prácticamente ignora las frecuencias más rápidas. Además, los efectos distintos en el *recall* y en el *precision* sugieren que el detector requiere actividad en la banda delta (0–4 Hz) de forma excluyente y que, en caso de existir, busca una baja actividad theta.

4.4.2. Detecciones en ruido rosado

Para analizar la especificidad de los patrones detectados por modelos entrenados, se aplican dichos modelos a las señales artificiales de PINK (ver Sección 3.2.2) y se caracterizan las detecciones. Específicamente, se usan los 15 modelos entrenados que resultan de la validación cruzada en una determinada base de datos (MASS-MODA o MASS-SS2-KC) para inferir sobre todo PINK. Además, para el experimento se admite escalar las señales de PINK antes de la inferencia para disminuir o aumentar su amplitud. La tasa de detecciones por hora de los modelos propuestos se detalla en la Tabla 4.3. No se observan diferencias significativas entre REDv2-Time y REDv2-CWT.

Los modelos entrenados para detectar husos de sueño prácticamente no generan detecciones en PINK. De hecho, REDv2-CWT no genera ninguna detección mientras que solo una instancia de REDv2-Time genera una detección. Sin embargo, para forzar la generación de detecciones y permitir su análisis, se escalan las señales con un factor de 2. En cambio, los modelos entrenados para detectar complejos K generan detecciones en PINK sin un factor de escala. Se ilustran algunas detecciones en la Figura 4.12. La selección de casos es aleatoria excepto por la señal mostrada en la Figura 4.12A, que corresponde a la única detección de huso de sueño generada sin escalar la señal. En general, las detecciones mostradas comparten varias similitudes morfológicas con los eventos

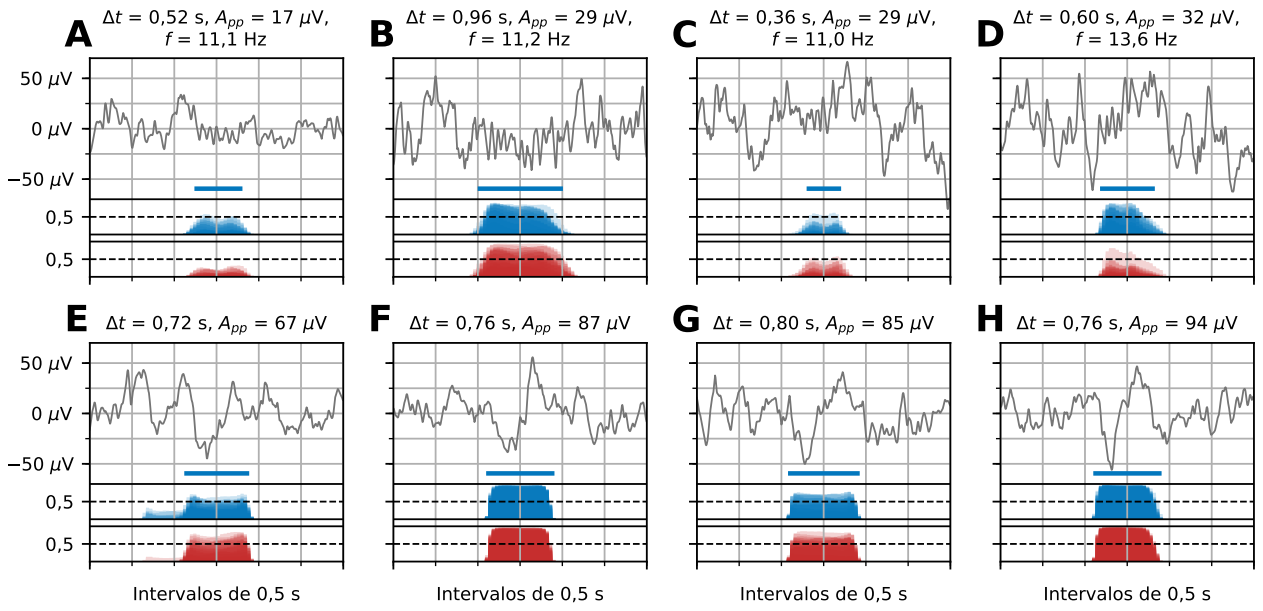

Figura 4.12: Visualización de detecciones en PINK. (**A**-**D**) Detecciones de husos de sueño al entrenar en MASS-MODA. (**E**-**H**) Detecciones de complejos K al entrenar en MASS-SS2-KC. En cada panel se indica una marca de referencia (línea azul), las probabilidades ajustadas de los 15 modelos REDv2-Time (áreas azules), y las probabilidades ajustadas de los 15 modelos REDv2-CWT (áreas rojas). En el título se indican los parámetros de la marca de referencia. Las áreas que indican probabilidad saturan su color cuando 5 o más modelos se superponen.

reales. A continuación, se analizan las estadísticas del conjunto completo de detecciones.

Las distribuciones de parámetros de todas las detecciones se muestran en las Figuras 4.13 y 4.14 para husos de sueño y complejos K, respectivamente. En cada panel, las distribuciones obtenidas en PINK son comparadas con las distribuciones obtenidas en la base de entrenamiento, y el rango de algunos gráficos ha sido recortado por legibilidad. Las distribuciones se muestran con un diagrama de cajón cuyas patillas cubren desde el percentil 1 al 99, en donde además se indica el promedio con un círculo verde. En general, no se observan diferencias significativas entre REDv2-Time y REDv2-CWT. Tanto para husos de sueño como complejos K, todos los parámetros calculados en PINK se encuentran al interior del rango de referencia, aunque más concentrados hacia el límite inferior. Es decir, las detecciones en PINK son cortas, de baja amplitud, y con menor probabilidad. En husos de sueño, la frecuencia tiende a ser más lenta pero con un rango intercuartil al interior del intervalo 11–13 Hz. En complejos K, la diferencia en el rango de duraciones es menos marcada.

Para caracterizar la morfología de los complejos K detectados en PINK, se calcula la señal promedio de todos los eventos luego de alinearse en su valor mínimo. Este perfil promedio de complejo K se muestra en la Figura 4.15. Como referencia, también se muestra el perfil promedio obtenido por las anotaciones en MASS-SS2-KC. No se observan diferencias significativas entre REDv2-Time y REDv2-CWT. Consistente con las distribuciones observadas anteriormente, el perfil promedio en PINK es más corto y de menor amplitud que la referencia experta. A pesar de ello, se observan similitudes morfológicas, tales como que el mínimo del evento es un pico pronunciado, y dicho pico negativo tiene mayor amplitud que el pico positivo siguiente (76 % más en PINK y 88 % más en la referencia experta). Además, el pico negativo aparece más rápido de lo que demora el pico positivo en decaer. Sin embargo, estas relaciones son menos pronunciadas en PINK en comparación a la referencia experta, particularmente en las diferentes velocidades con que ocurren

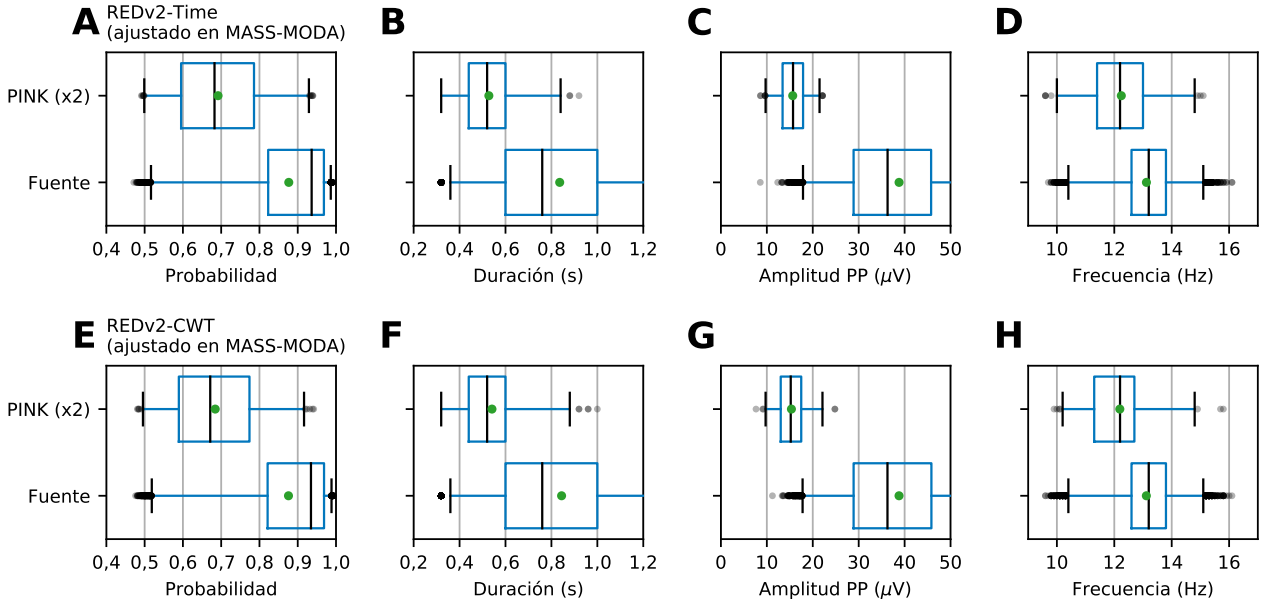

Figura 4.13: Distribución de parámetros por evento de las detecciones de husos de sueño de los modelos propuestos. Ambos modelos son entrenados en MASS-MODA y aplicados en PINK. Las señales de PINK fueron amplificadas por 2. En cada panel se compara la distribución en PINK con la distribución de las detecciones en la base de entrenamiento. La probabilidad mostrada en (**A**) y (**E**) es la probabilidad del evento.

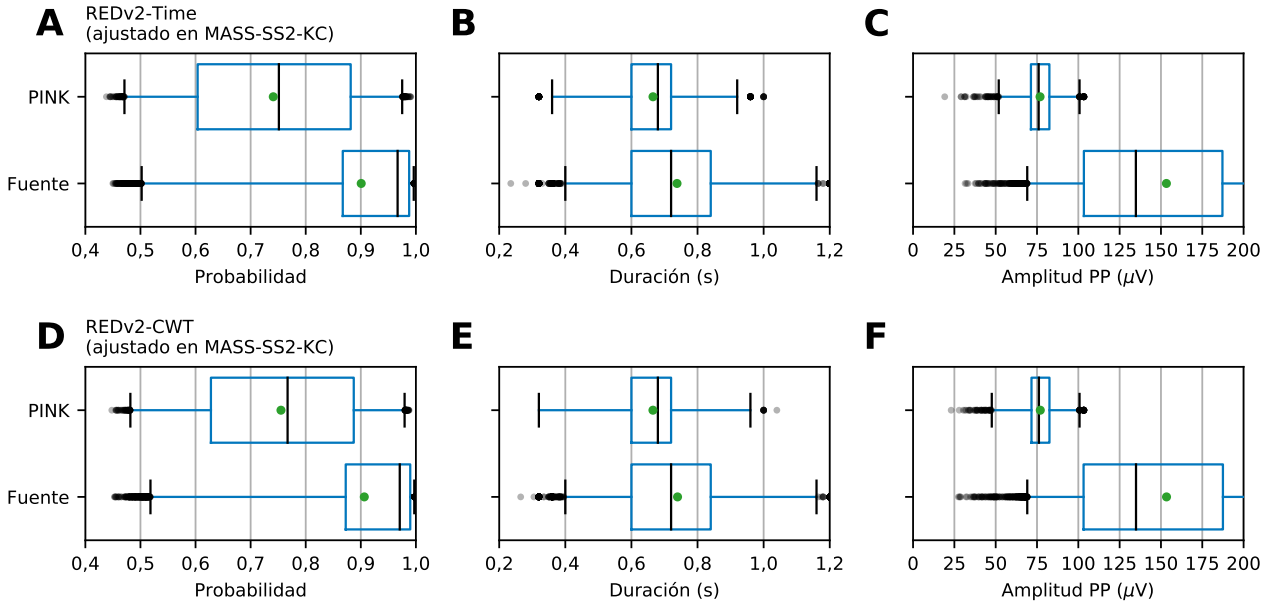

Figura 4.14: Distribución de parámetros por evento de las detecciones de complejos K de los modelos propuestos. Ambos modelos son entrenados en MASS-SS2-KC y aplicados en PINK. En cada panel se compara la distribución en PINK con la distribución de las detecciones en la base de entrenamiento. La probabilidad mostrada en (**A**) y (**D**) es la probabilidad del evento.

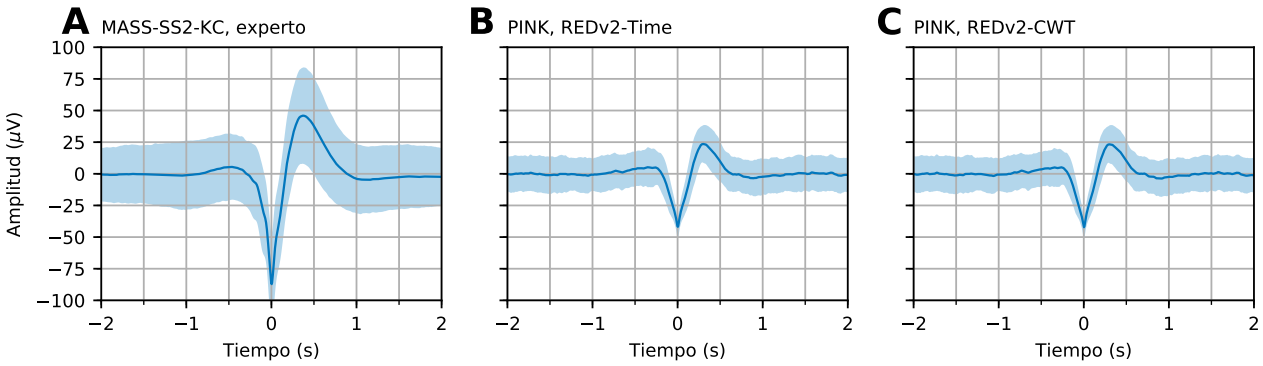

Figura 4.15: Señal promedio, con una desviación estándar, de todos los complejos K alineados en el pico negativo. (**A**) Promedio generado con las anotaciones en MASS-SS2-KC. (**B**) Promedio generado con las detecciones de REDv2-Time en PINK. (**C**) Promedio generado con las detecciones de REDv2-CWT en PINK.

las dinámicas del pico negativo y del pico positivo. Por ejemplo, en PINK el pico positivo demora menos en volver a cero (demora 0,38 s en PINK y 0,51 s en la referencia experta), y dicha demora es más similar al tiempo de subida del pico positivo (0,30 s desde el pico negativo en PINK y 0,37 s en la referencia experta).

4.4.3. Desempeño en etiquetas artificiales

Para analizar la capacidad de los modelos propuestos, se evalúa el desempeño en las bases de datos con etiquetas artificiales de husos de sueño basadas en las señales de CAP: CAP-S1, CAP-S2 y CAP-A7 (ver Sección 3.2.2). A diferencia de las bases de datos expertas, estas bases artificiales permiten analizar los resultados bajo condiciones controladas, en donde se considera un volumen de datos notoriamente mayor (80 registros de noche completa) con anotaciones que son una función determinística de las señales. Además, las funciones de anotación seleccionadas reflejan esquemas de detección de husos de sueño usados tradicionalmente en la literatura. Por lo tanto, estos experimentos reflejan la capacidad de los modelos propuestos de aprender a codificar dichos criterios tradicionales solo a partir de los datos.

El desempeño general de los modelos propuestos se reporta en la Tabla 4.4, mientras que el desempeño por sujeto se ilustra en la Figura 4.16. Si bien REDv2-CWT muestra un *F1-score* promedio ligeramente mayor en las tres bases, ninguna de las diferencias es significativa (aunque en CAP-S1 se tiene $P = 0{,}062$). De mayor a menor desempeño, las bases se ordenan como CAP-S1, CAP-A7 y CAP-S2, tanto en promedio como en dispersión.

En CAP-S1, las anotaciones dependen exclusivamente de la amplitud sigma absoluta, por lo que las otras bandas de frecuencia y las características individuales de cada sujeto son irrelevantes. En este caso, tanto el *F1-score* como el mIoU bordean el 95 %, un desempeño cercano al ideal. La nube de sujetos es muy compacta y en aquellos que se podrían considerar anómalos se sigue observando un excelente desempeño. Esto sugiere que REDv2 puede capturar fácilmente el criterio S1 de ser necesario.

Por otro lado, en CAP-S2, las anotaciones dependen exclusivamente de la amplitud sigma relativa a la amplitud media de la noche completa, por lo que si bien las otras bandas siguen siendo irrelevantes, las características individuales de cada sujeto (en este caso su actividad sigma media)

Tabla 4.4: Desempeño de la detección en las bases de datos artificiales basadas en CAP.

Datos	Detector	F1-score (%)	Recall (%)	Precision (%)	mIoU (%)
CAP-S1	REDv2-Time	$95{,}5 \pm 0{,}4$	$96{,}1 \pm 0{,}5$	$94{,}9 \pm 0{,}5$	$94{,}7 \pm 0{,}2$
	REDv2-CWT	$96{,}1 \pm 0{,}3$	$96{,}4 \pm 0{,}4$	$95{,}8 \pm 0{,}5$	$95{,}2 \pm 0{,}2$
CAP-S2	REDv2-Time	$82{,}2 \pm 1{,}0$	$85{,}2 \pm 4{,}5$	$82{,}3 \pm 3{,}3$	$88{,}0 \pm 0{,}6$
	REDv2-CWT	$82{,}4 \pm 0{,}9$	$84{,}4 \pm 4{,}5$	$83{,}3 \pm 3{,}9$	$88{,}5 \pm 0{,}2$
CAP-A7	REDv2-Time	$88{,}1 \pm 0{,}3$	$89{,}5 \pm 1{,}8$	$87{,}2 \pm 1{,}4$	$88{,}3 \pm 0{,}4$
	REDv2-CWT	$88{,}2 \pm 0{,}4$	$89{,}5 \pm 2{,}2$	$87{,}5 \pm 1{,}7$	$88{,}2 \pm 0{,}5$

son extremadamente importantes. En este caso, el desempeño de REDv2 empeora fuertemente y la dispersión se dispara. Debido a dispersión, la nube de sujetos puebla casi homogéneamente el rectángulo $[0{,}6,\ 1{,}0] \times [0{,}6,\ 1{,}0]$ en el plano de *recall* y *precision*. Además, aparece un sujeto anómalo muy distante a la nube, con *precision* menor a 40 %. Esto sugiere que es muy difícil para REDv2 capturar el criterio S2 de ser necesario, incluso con un gran volumen de datos. Dado que las anotaciones están condicionadas a la noche completa del sujeto, esta limitación en la capacidad es consistente con el contexto restringido a 20 s de REDv2.

Por último, en CAP-A7, las anotaciones provienen del detector A7, por lo que son relevantes tanto las otras bandas de frecuencia como las características individuales de cada sujeto. En este caso, el desempeño de REDv2 es intermedio, tanto en promedio como en dispersión. La nube de sujetos está agrupada principalmente en el rectángulo $[0{,}8,\ 1{,}0] \times [0{,}8,\ 1{,}0]$ en el plano de *recall* y *precision*. Además, se puede identificar una pequeña agrupación de sujetos anómalos de bajo *precision* ($< 70\,\%$). Esto sugiere que REDv2 puede capturar gran parte de los criterios usados por A7 de ser necesario, aunque con limitaciones significativas. El detector A7 considera un contexto de 30 s para personalizar (normalizar) algunos de sus estadísticos, un contexto comparable al utilizado por REDv2. Sin embargo, la existencia de sujetos con desempeño anómalo sugiere que REDv2 no es capaz de capturar de forma robusta las personalizaciones consideradas por A7, incluso con un gran volumen de datos.

Para analizar los sujetos que muestran un desempeño anómalo, se calcula su espectro promedio durante la etapa N2. Dichos espectros se ilustran en la Figura 4.17, en donde también se muestran los espectros de los demás sujetos de CAP en el fondo a modo de referencia. Los sujetos anómalos en CAP-S1 (*precision* $< 90\,\%$) tienen actividad sigma muy tenue en comparación a los demás. En consecuencia, tienen baja densidad de anotaciones (1,0–1,3 epm), provocando que sus métricas de desempeño sean particularmente sensibles al número de falsos positivos (e.g., el sujeto ID rbd19 tiene 136 anotaciones, 18 falsos positivos, y *precision* de 87,9 %). Por otro lado, el sujeto anómalo en CAP-S2 (*precision* $< 40\,\%$) tiene alta actividad en las bandas adyacentes al pico sigma en comparación a los demás. En consecuencia, la amplitud media de la señal filtrada en 11–16 Hz es anormalmente alta, provocando un umbral de amplitud más estricto que REDv2 no puede anticipar e incurre en muchos falsos positivos.

Los sujetos anómalos en CAP-A7 (*precision* $< 70\,\%$) tienen una alta densidad de anotaciones (4,7–5,8 epm), consistente con la intensa actividad sigma en sus espectros. En comparación a los demás, sus espectros poseen baja actividad theta (4–8 Hz) y beta (16–30 Hz). Esto sugiere que sus husos de sueño poseen una potencia sigma relativa mayor de lo típico (son más notorios), lo que sumado a su alta densidad podría elevar el umbral personalizado de A7 para dicha caracte-

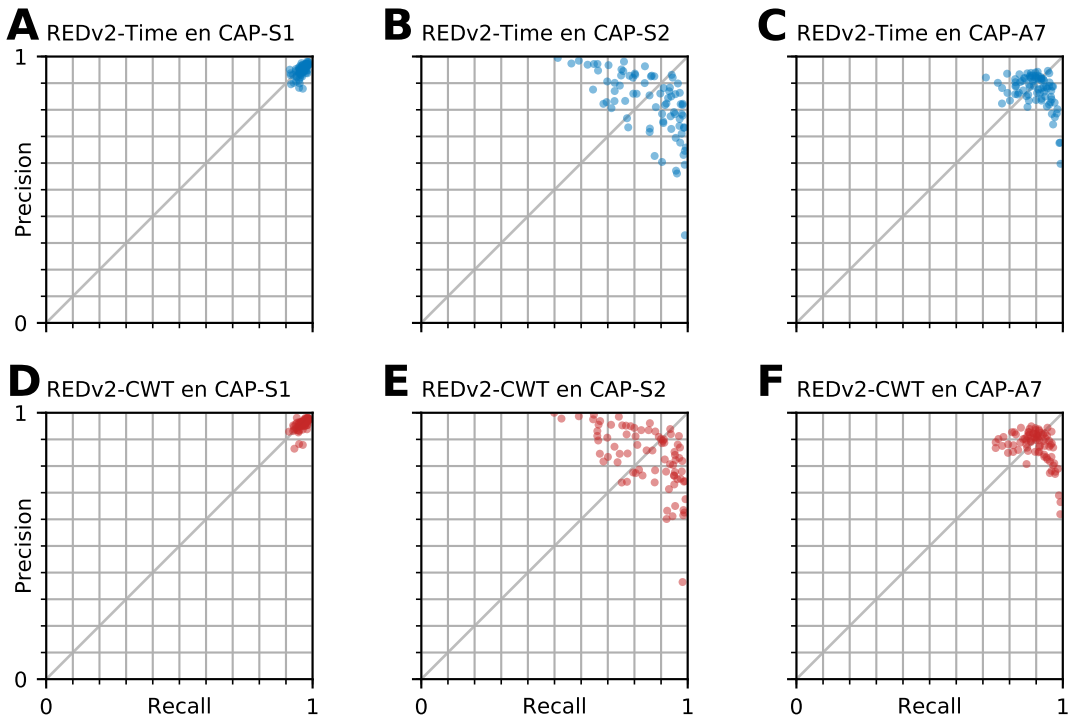

Figura 4.16: Desempeño de la detección por sujeto en las bases artificiales basadas en CAP. (**A**-**C**) Desempeño de REDv2-Time. (**D**-**F**) Desempeño de REDv2-CWT.

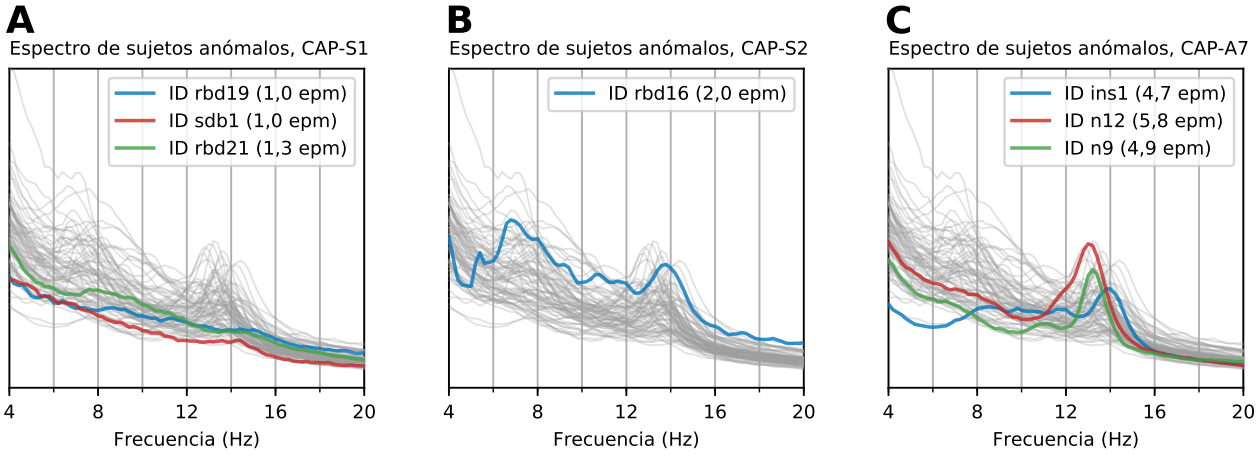

Figura 4.17: Espectro de los sujetos de CAP durante la etapa N2. En cada panel se destacan aquellos sujetos en donde el desempeño de los modelos propuestos es anómalo en (**A**) CAP-S1, (**B**) CAP-S2, y (**C**) CAP-A7.

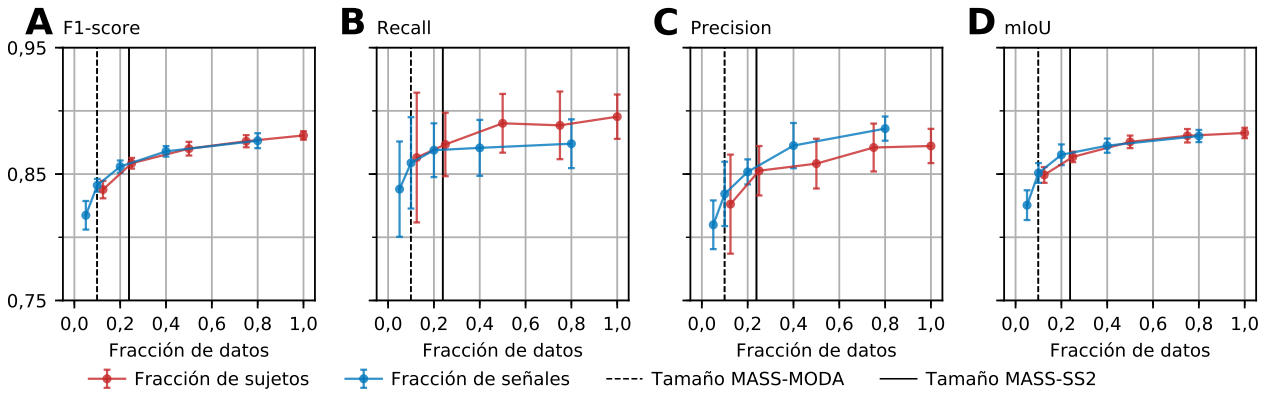

Figura 4.18: Desempeño de la detección de REDv2-Time en CAP-A7 cuando los conjuntos de entrenamiento y validación son una fracción del total disponible. Se indican con líneas verticales negras las fracciones que equivalen, en número de horas, a los datos anotados disponibles en MASS-MODA y MASS-SS2.

rística espectral. Como consecuencia, REDv2 podría identificar como husos algunos eventos cuya potencia sigma relativa es aceptable de acuerdo a estadísticas globales pero no de acuerdo a las características propias del sujeto. Debido a esta conjetura, se mide la razón de potencia entre la banda 11–16 Hz y la banda 4,5–30 Hz y se compara su distribución entre los verdaderos positivos y falsos positivos de cada sujeto. En efecto, los sujetos ID ins1 e ID n12, los de peor *precision*, poseen las diferencias más significativas entre dichas distribuciones, con el sujeto ID n9 ocupando el séptimo lugar (de 80 sujetos en total). Además, los tres sujetos tienden a mostrar razones de potencia grandes. Por lo tanto, la anomalía en estos sujetos podría estar causada por una limitada capacidad de REDv2 para adaptar su sensibilidad al perfil espectral de los sujetos, sobre todo en aquellos con espectros alejados de la población en general.

Para analizar el requerimiento de datos, se mide la curva de aprendizaje de REDv2-Time en CAP-A7 y se muestra en la Figura 4.18. Específicamente, se mide la evolución de las métricas cuando el conjunto de prueba permanece intacto y solo se usa una fracción del conjunto de entrenamiento y validación. La fracción puede generarse por medio de un subconjunto de sujetos, o bien por medio de un subconjunto de señal de todos los sujetos del conjunto. Además, los datos en una fracción contienen siempre a los datos de las fracciones menores. Para el caso de la fracción de señales, las métricas de entrenamiento y validación usan micro-promedio. Sin embargo, en todos los casos se usa macro-promedio para las métricas de prueba.

La curva de aprendizaje muestra que el desempeño aumenta rápidamente al agregar más datos cuando se usan fracciones cercanas al tamaño de MASS-MODA. Sin embargo, una vez se sobrepasa un tamaño comparable al de MASS-SS2, la tasa de ganancia es mucho menor. De todas formas, la curva de *F1-score* sugiere que el desempeño podría seguir mejorando ante un volumen de datos mayor al total disponible. Adicionalmente, la mejora en desempeño que ocurre más allá del tamaño de MASS-SS2 parece deberse principalmente a mejoras en *precision* (i.e., descartar mejor los falsos positivos), ya que el *recall* se satura antes, particularmente para las fracciones de señales. Lo anterior es consistente con la facilidad de obtener un alto *recall* en comparación a un alto *precision* al detectar husos de sueño, ya que basta detectar aumentos de amplitud en la señal filtrada en 11–16 Hz para capturar todos los husos de sueño verdaderos (y muchos falsos).

En general, la dispersión decrece al aumentar el volumen de datos. Sin embargo, al usar frac-

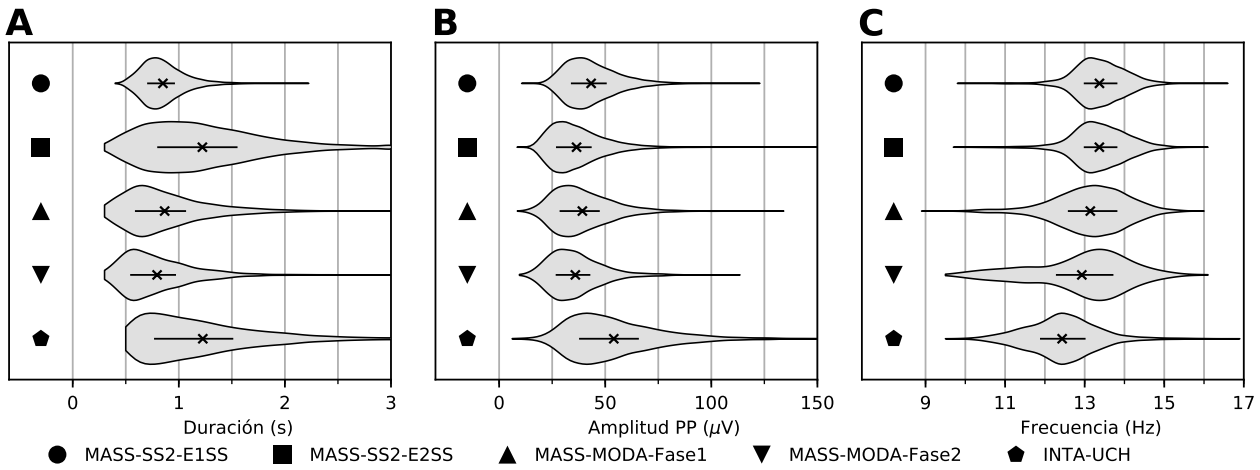

Figura 4.19: Distribución de parámetros por evento de las bases de datos de husos de sueño para diferentes expertos y demografías.

ciones de señales se consigue disminuir la dispersión más rápido e incluso a menores magnitudes, como se evidencia en las curvas de *recall* y *precision*. Esto es consistente con la mayor diversidad inyectada. Además, usando fracciones de señales se consigue un mejor desempeño en el régimen de pocos datos (tamaños menores al de MASS-SS2), se pueden alcanzar fracciones más pequeñas que usando fracciones de sujetos, y se favorece la mejora del *precision*. Esto sugiere que, en el régimen de pocos datos, es más eficiente construir una base de datos usando la estrategia de fracción de señales (como en MASS-MODA).

4.5. Transferencia a sujetos nuevos

En esta sección, se analizan experimentos que buscan compensar la caída del desempeño al transferir un modelo entrenado a otra distribución de datos. Se consideran dos escenarios de transferencia. El primero es la transferencia de una base de datos a otra (transferencia externa), cuyo desempeño sin ajuste ya se exploró en la Sección 4.2.5. El segundo es la transferencia que ocurre dentro de una misma base de datos cuando el modelo se aplica a diferentes sujetos (transferencia interna).

4.5.1. Transferencia externa: Distinto criterio de anotación

Los resultados de la Sección 4.2.5 sugieren que las bases de datos tienen distintos criterios para anotar husos de sueño, provocando una caída importante en el desempeño al momento de transferir un modelo entrenado. Para verificar dicha diferencia, la Figura 4.19 muestra la distribución de parámetros por evento que existe en cada base de datos experta, en donde se ha separado MASS-MODA en sus dos fases demográficas.

En efecto, si bien MASS-MODA-Fase1, MASS-SS2-E1SS y MASS-SS2-E2SS provienen del mismo cohorte y poseen una demografía similar (adultos jóvenes), sus anotaciones provienen de diferentes expertos y muestran diferencias notorias. En duración, MASS-SS2-E2SS muestra una distribución más dispersa con un promedio más alto, mientras que MASS-SS2-E1SS y MASS-MODA muestran una distribución más compacta y con una media similar. La distribución es parti-

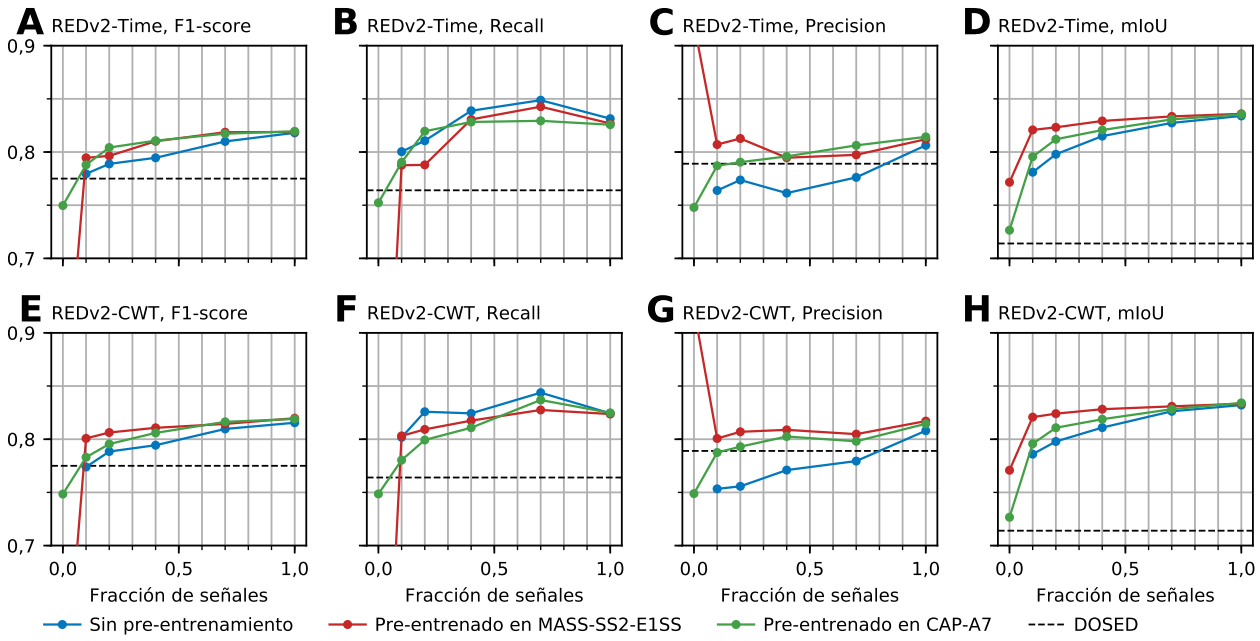

Figura 4.20: Desempeño de la detección en MASS-MODA cuando los conjuntos de entrenamiento y validación son una fracción del total disponible, y se usan diferentes inicializaciones del modelo. (**A**-**D**) Desempeño de REDv2-Time. (**E**-**H**) Desempeño de REDv2-CWT. Se indica el desempeño de DOSED [8] como referencia.

cularmente compacta en MASS-SS2-E1SS, en donde además se anotan menos eventos cortos. En amplitud, MASS-SS2-E2SS y MASS-MODA muestran distribuciones similares, aunque MASS-SS2-E2SS muestra una media ligeramente menor, mientras que MASS-SS2-E1SS muestra una media significativamente mayor, en donde además se anotan menos eventos de baja amplitud. En frecuencia, MASS-SS2-E1SS y MASS-SS2-E2SS muestran distribuciones similares, mientras que MASS-MODA contiene un mayor porcentaje en eventos más lentos. Las observaciones anteriores sugieren que existen diferencias en la amplitud requerida y en el criterio para determinar los instantes de inicio y fin (probablemente correlacionado con la sensibilidad a la amplitud).

Debido a estas diferencias, se evalúa el uso de ajuste fino como estrategia para compensar la caída en el desempeño ante una transferencia. Por simplicidad, el análisis se limita al desempeño en MASS-MODA, en donde se considera el caso regular sin pre-entrenamiento, la inicialización por pre-entrenamiento en MASS-SS2-E1SS (la base de datos experta más distante), y la inicialización por pre-entrenamiento en CAP-A7 (datos artificiales). En todos los casos, se analiza el desempeño ante distintos tamaños de datos disponibles para el ajuste fino en MASS-MODA, por medio de fracciones de señales (mismo procedimiento descrito en la Sección 4.4.3). Una fracción de 1 indica el uso de todo MASS-MODA, mientras que una fracción de 0 indica ausencia de ajuste fino, i.e., transferencia directa (inválido sin pre-entrenamiento). Para las fracciones evaluadas distintas de 0 y 1, no se ajusta el umbral de probabilidad a la salida de REDv2, i.e., se fija siempre en 0,5, ya que se encontró que ajustar el umbral con muy pocos datos induce demasiada varianza.

Los resultados de este experimento se muestran en la Figura 4.20 (mediciones precisas de *F1-score* se muestran en la Tabla E.9). Como referencia, en cada métrica se indica el desempeño de DOSED (principal detector alternativo) sin pre-entrenamiento y con acceso a todos los datos. El eje vertical se ha recortado por legibilidad, dejando fuera la transferencia directa MASS-SS2-E1SS $\rightarrow$ MASS-MODA en varios paneles.

REDv2-Time y REDv2-CWT tienen un desempeño similar sin pre-entrenamiento para todas las fracciones. Además, para ambos modelos se consigue un desempeño significativamente mejor ($P < 0{,}05$) usando algún pre-entrenamiento ante fracciones pequeñas ($< 0{,}5$). En cuanto a *F1-score*, en REDv2-Time es mejor pre-entrenar en CAP-A7, mientras que en REDv2-CWT es mejor pre-entrenar en MASS-SS2-E1SS. Esta mejora en el desempeño proviene principalmente de un mejor *precision*. En cuanto a mIoU, ambos modelos se benefician más de un pre-entrenamiento en MASS-SS2-E1SS, en donde con tan solo una fracción de 0,1 se acercan mucho al máximo. De todas formas, gran parte de los beneficios de pre-entrenar en MASS-SS2-E1SS también se observan al pre-entrenar en CAP-A7 en todas las métricas, a pesar de ser una base compuesta de anotaciones artificiales. Con el 100 % de los datos, el pre-entrenamiento no mejora significativamente el desempeño con respecto a no usarlo.

Si bien no existen diferencias significativas entre ambos modelos, REDv2-CWT consigue subir la media y disminuir la desviación estándar más rápido que REDv2-Time cuando se trata de transferir desde MASS-SS2-E1SS. Además, REDv2-CWT es el único en conseguir un *F1-score* superior a 80 % con una fracción de 0,1, lo cual ocurre en la transferencia desde MASS-SS2-E1SS y que resulta en un *F1-score* de $80{,}1 \pm 1{,}3$ %. Dicha ventaja desaparece al progresar a fracciones más grandes.

Usando solo un 10 % de los datos, el punto de operación del modelo transferido desde MASS-SS2-E1SS, la base experta más distante, mejora notablemente. El *F1-score* sube desde 53,4–53,6 % a 79,5–80,1 %, con un compromiso entre *recall* y *precision* más balanceado. Esto es consistente con que la caída de desempeño se debe principalmente a diferencias de anotación en lugar de diferencias en las características extraídas. Así, un 10 % de los datos bastan para corregir el criterio de anotación a la salida del modelo, a la vez que reutiliza gran parte de su extractor de características. De todas formas, la subida lenta pero sostenida del desempeño posterior a la fracción de 0,1 sugiere la necesidad de correcciones adicionales. Por ejemplo, podría requerirse corregir la extracción de características para hacer frente a una mayor diversidad en las señales ocasionada por la demografía más extensa de MASS-MODA con respecto a MASS-SS2-E1SS.

El *F1-score* de referencia de DOSED es alcanzado por REDv2 sin usar pre-entrenamiento y con una fracción de apenas 0,1. Sin embargo, en dicho caso REDv2 muestra un *precision* por debajo de DOSED y solo lo supera usando más del 70 % de los datos. Esta desventaja se resuelve usando pre-entrenamiento, resultando en un mejor desempeño que DOSED en todas las métricas desde una fracción de 0,1. Esto es cierto incluso para la transferencia desde CAP-A7, a pesar de la ausencia de etiquetas expertas.

El desempeño de transferencia directa desde CAP-A7 (i.e., sin ajuste fino) en ambos modelos es de aproximadamente 75 % en *F1-score*, *recall* y *precision*. En este escenario, se está utilizando directamente un modelo que aprendió a imitar al detector A7, por lo que un buen desempeño de REDv2 es consistente con el buen desempeño de A7. Sin embargo, el desempeño que muestra REDv2 es, de forma inesperada, significativamente mayor que el desempeño de A7 en MASS-MODA (*F1-score* $73{,}3 \pm 1{,}9$ %). Esto sugiere que, a pesar de que fue entrenado para imitar, REDv2 aprendió mejores características, lo que podría deberse a un sesgo inductivo de la arquitectura neuronal.

4.5.2. Transferencia interna: Mismo criterio de anotación

En todos los detectores evaluados se evidencia una importante dispersión en el desempeño entre sujetos de una misma base de datos, i.e., que fueron anotados con el mismo criterio (ver Figura 4.2). Si bien REDv2 muestra una dispersión menor, sigue siendo importante. Esta dispersión se ilustra para la detección de husos de sueño en las Figuras 4.21A-H, en donde también se observan sujetos con desempeño anómalo. Esta dispersión sugiere que el modelo no es efectivo en capturar las características individuales de cada sujeto. Debido a que el punto de operación del modelo admite un desplazamiento a través de un cambio en el umbral de salida, se analiza el potencial de la personalización del umbral de salida como estrategia para disminuir la dispersión.

En primer lugar, se evalúa el mejor caso posible: a cada sujeto se le aplica el umbral que maximiza su AF1. A esta referencia sobre-ajustada se le llama *umbral oráculo*, y su efecto se ilustra en las Figuras 4.21I-P. En todos los casos se observa una disminución de la dispersión acompañada de un aumento en el *F1-score* promedio. Sin embargo, la efectividad del *umbral oráculo* es diferente en cada caso. La modificación es particularmente efectiva en el caso de MASS-SS2-E1SS, resultando en un aumento del 2 % en *F1-score* y la desaparición de sujetos anómalos.

Para analizar los sujetos que muestran el peor desempeño, se calcula su espectro promedio durante la etapa N2. Dichos espectros se ilustran en la Figura 4.22, en donde también se muestran los espectros de los demás sujetos de cada base de datos en el fondo, a modo de referencia. En MASS-SS2-E2SS, el sujeto ID 01-02-006 tiene el peor *recall*, un pico sigma muy pequeño, y relativamente baja densidad; mientras que el sujeto ID 01-02-0018 tiene el peor *precision*, un pico sigma muy pronunciado y extenso, y relativamente alta densidad. Es decir, en un sujeto de baja actividad sigma, hay muchos falsos negativos, mientras que en un sujeto de alta actividad sigma, hay muchos falsos positivos. Esto sugiere que REDv2 tiene dificultades en esta base de datos para aprender un comportamiento robusto ante los extremos de amplitud. Por otro lado, en MASS-MODA los tres sujetos de peor desempeño corresponden a adultos mayores (Fase 2) que muestran un bajo *recall*, y cuyos espectros muestran una muy baja actividad sigma. Esto podría corresponder a un efecto análogo al observado en MASS-SS2-E2SS. En INTA-UCH, si bien se observa que el sujeto de peor *recall* tiene baja actividad sigma mientras que el sujeto de peor *precision* tiene alta actividad sigma, sugiriendo también un efecto de la amplitud, dichas diferencias espectrales no son tan notorias como en las otras dos bases, por lo que la anomalía podría estar causada por factores externos al espectro.

En MASS-SS2-E1SS, el peor desempeño lo tiene el sujeto ID 01-02-0014 con *precision* menor a 63 %. Su espectro contiene un pico sigma mayor que los demás sujetos, lo que contrasta con su densidad intermedia, y posee alta actividad alfa. A diferencia de las otras bases de datos, la anomalía no se explica satisfactoriamente por amplitud sigma, ya que los siguientes dos sujetos de peor *precision* tienen una amplitud intermedia y baja en dicha banda. Al igual que en CAP-A7, se mide la distribución de la razón de potencia sigma en dichos sujetos, tanto en sus verdaderos como en sus falsos positivos, y se observa que en los sujetos de peor *precision* dicha razón tiende a ser más grande en ambos tipos de eventos, en comparación a los demás sujetos. De forma análoga a lo descrito en CAP-A7, esto sugiere que la anomalía proviene de una incapacidad de REDv2 para capturar las características espectrales de cada sujeto, como puede ser la relación de amplitud entre la banda sigma y las demás bandas, que no le permite volverse más estricto al momento de detectar en el sujeto ID 01-02-0014.

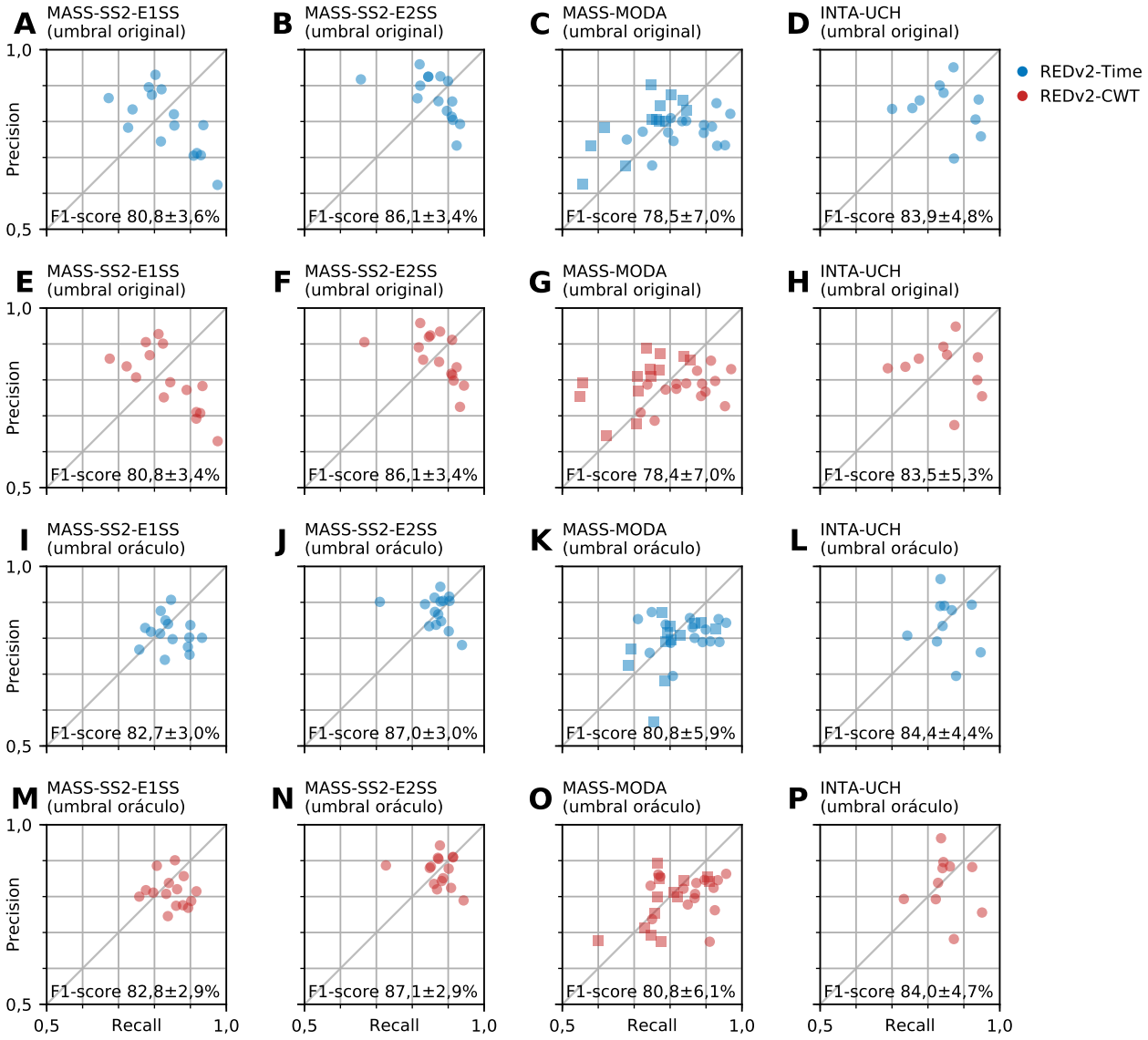

Figura 4.21: Desempeño de la detección por sujeto en las bases expertas de husos de sueño para REDv2-Time (azul) y REDv2-CWT (rojo). En MASS-MODA solo se muestran sujetos con suficientes datos, en donde el marcador cuadrado indica un sujeto de MASS-MODA-Fase2 (adultos mayores). (**A**-**H**) Desempeño original para REDv2-Time (paneles A-D) y REDv2-CWT (paneles E-H). (**I**-**P**) Desempeño cuando el umbral de salida se personaliza con un oráculo para REDv2-Time (paneles I-L) y REDv2-CWT (paneles M-P).

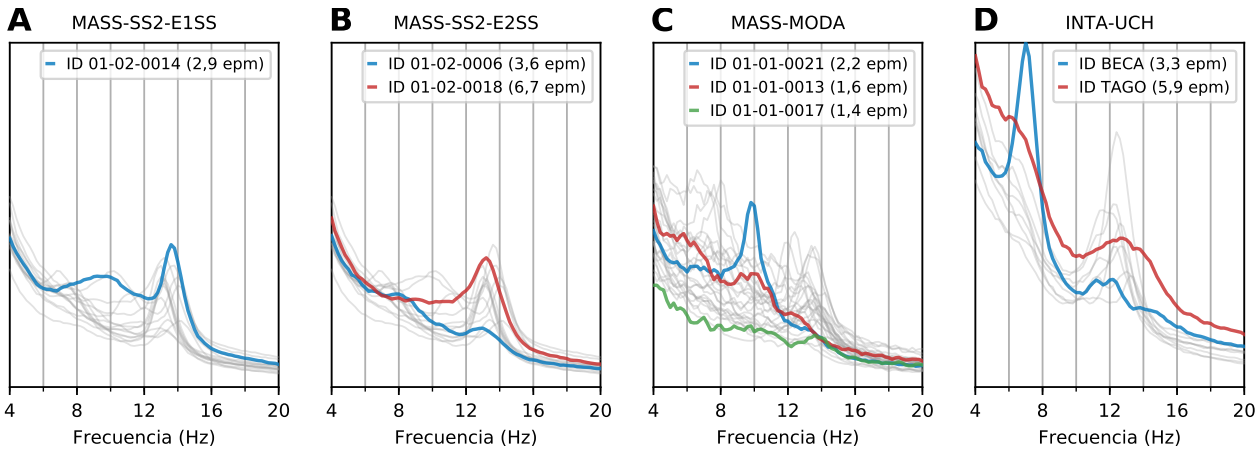

Figura 4.22: Espectros de los sujetos de cada base de datos durante la etapa N2. En cada panel se destacan aquellos sujetos en donde el desempeño de los modelos propuestos es anómalo.

A continuación, se analiza la capacidad de aproximar el *umbral oráculo* usando solo una porción de señal anotada. Para esto, en cada sujeto se aíslan los primeros $N_{\max}$ minutos de señal anotada, y se usa el resto para evaluar el desempeño, tanto de los ajustes como de los casos de referencia (umbral original y *umbral oráculo*) para mantener siempre el mismo conjunto de prueba. Luego, para personalizar el umbral se usan distintas fracciones de $N_{\max}$. En MASS-MODA, $N_{\max} = 10$, mientras que en las demás bases $N_{\max} = 30$. Los resultados de este experimento se muestran en la Figura 4.23, en donde se usa la dispersión entre sujetos (mediciones precisas de *F1-score* se muestran en la Tabla E.10). Por simplicidad, no se muestran los cambios de mIoU, pero se mantiene relativamente constante en todos los casos.

En general, las mejores aproximaciones del umbral se consiguen en MASS-SS2-E1SS, seguido de MASS-SS2-E2SS. En MASS-SS2-E1SS, la aproximación aún podría mejorar después de usar los 30 minutos disponibles, pero se alcanza una disminución importante de la dispersión. En cambio, en MASS-SS2-E2SS se requiere de solo 10 minutos para estar muy cerca del desempeño oráculo. Por otro lado, en MASS-MODA la aproximación no ofrece mejoras, probablemente por la poca cantidad de datos. Por último, en INTA-UCH la aproximación es limitada, permitiendo obtener mejoras principalmente en la dispersión, pero no en la media.

4.6. Tendencias en datos sin etiquetas

En esta sección, se analizan experimentos que buscan validar las detecciones de husos de sueño de REDv2 por medio de la reproducción de comportamientos esperados desde la literatura. En primer lugar, se caracterizan los parámetros de los husos detectados y se analizan los cambios provocados por la edad y el sexo de los sujetos. En segundo lugar, se analizan las detecciones para profundizar en la interpretación de REDv2.

Para ello, se usa como detector el ensamble AVG (promedio de probabilidades ajustadas) de los cinco modelos de REDv2-Time que corresponden al primer ciclo de validación cruzada en MASS-MODA. Este ensamble se usa para detectar husos de sueño en NSRR (base de datos grande y sin anotaciones) durante la etapa N2. Gracias al ensamble, se usa información de todo MASS-MODA y se obtienen estadísticas más robustas para el análisis.

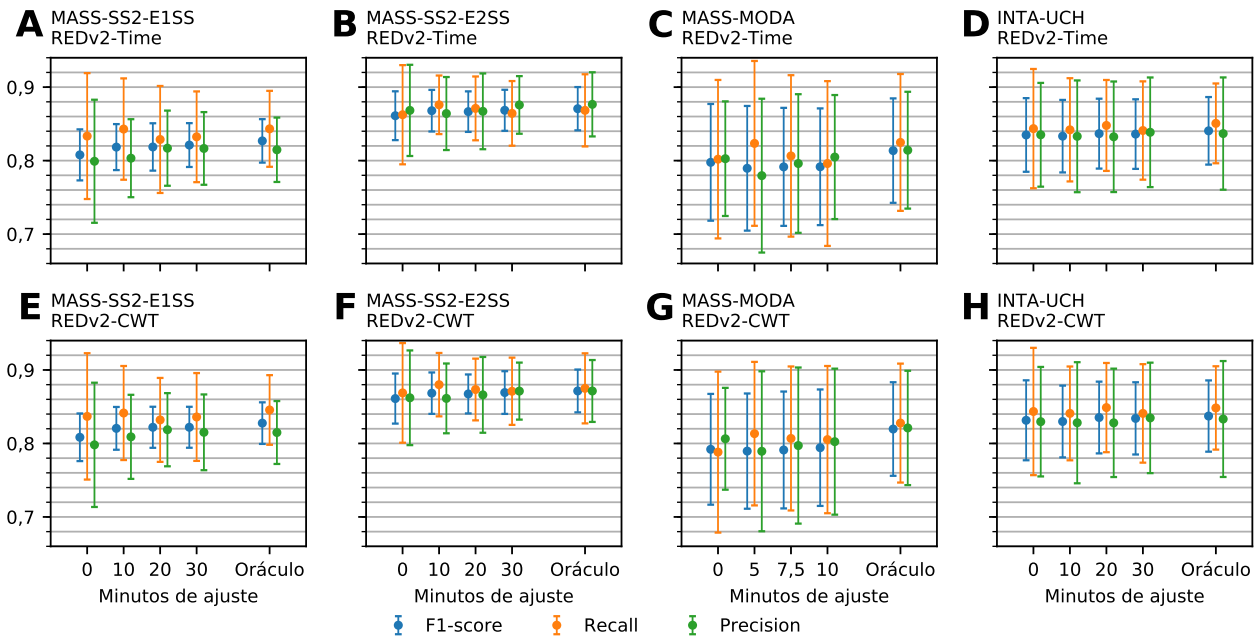

Figura 4.23: Desempeño de la detección por sujeto cuando se personaliza el umbral de salida usando una porción de señal N2 de cada sujeto.

Las señales se pre-procesan igual que en MASS-MODA, y las detecciones se post-procesan de forma similar salvo por dos excepciones: no se reparan las detecciones más largas que 3 s, y se ignoran detecciones con amplitud PP superior a 134,12 μV (amplitud máxima observada en MASS-MODA). Estos cambios tienen un impacto pequeño ($<$ 0,1 % de las detecciones) y permiten remover artefactos del análisis. Como resultado, se tienen 4.388.910 detecciones de husos de sueño en NSRR para analizar.

4.6.1. Caracterización de las detecciones

La distribución global de los parámetros de las detecciones se muestra en la Figura 4.24, en donde la densidad se mide por sujeto y los demás parámetros por evento. El promedio, el intervalo del 50 % central, y el intervalo del 90 % central de los parámetros se detalla en la Tabla 4.5. En general, las distribuciones de duración, amplitud PP y frecuencia son similares a las observadas en las bases de datos expertas. La distribución de densidad, en cambio, muestra un rango similar pero con mayor porcentaje cerca de cero.

Para profundizar en el análisis de la densidad, se verifica su correlación con la potencia promedio en la banda sigma, tanto absoluta como relativa a las otras bandas, en la Figura 4.25. Para diferentes intervalos de densidad, se muestra la distribución de la potencia con un diagrama de cajón cuyas patillas cubren desde el percentil 1 al 99 (valores fuera de las patillas se indican con círculos negros). En cada potencia, se compara la relación obtenida en NSRR con la determinada por las anotaciones de MASS-MODA. Se observa que ambas relaciones son similares tanto en tendencia como en dispersión. Además, la correlación es más intensa con la potencia sigma relativa que con la absoluta.

Para analizar las tendencias de los parámetros por sujeto con robustez, solo se consideran sujetos con al menos 10 detecciones. Como resultado, se ignoran 110 sujetos (0,98 % del total de 11.224),

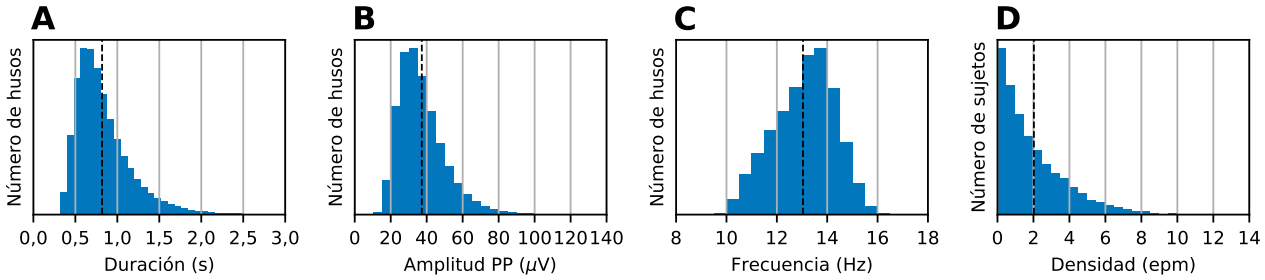

Figura 4.24: Distribución de parámetros de las detecciones de husos de sueño en NSRR (todos los sujetos). (**A**-**C**) Parámetros por evento. (**D**) Densidad por sujeto.

Tabla 4.5: Estadísticas de parámetros de las detecciones de husos de sueño en NSRR (todos los sujetos). La duración, amplitud PP y frecuencia son parámetros por evento, mientras que la densidad es un parámetro por sujeto. Los intervalos de 50 % y 90 % corresponden al rango determinado por los percentiles 25–75 y 5–95, respectivamente.

Parámetro	Promedio	Intervalo 50 %	Intervalo 90 %
Duración (s)	0,82	0,56–0,96	0,44–1,48
Amplitud PP (μV)	37,3	27,7–43,8	20,6–62,9
Frecuencia (Hz)	13,0	12,2–13,9	10,9–14,9
Densidad (epm)	2,02	0,60–2,94	0,14–5,84

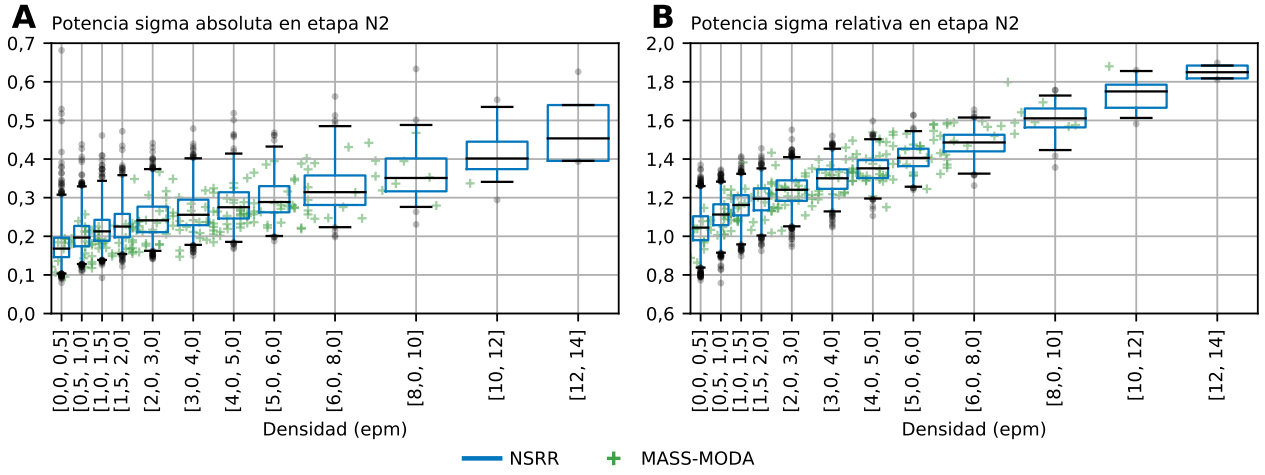

Figura 4.25: Relación entre la densidad predicha de husos de sueño y la potencia sigma del espectro en N2 para todos los sujetos de NSRR. Como referencia, se indican con cruces verdes las posiciones de todos los sujetos de MASS-MODA usando sus densidades expertas. (**A**) Promedio de la potencia en 11–16 Hz. (**B**) Promedio de la potencia en 11–16 Hz dividido por el promedio de la potencia en 4,5–30 Hz.

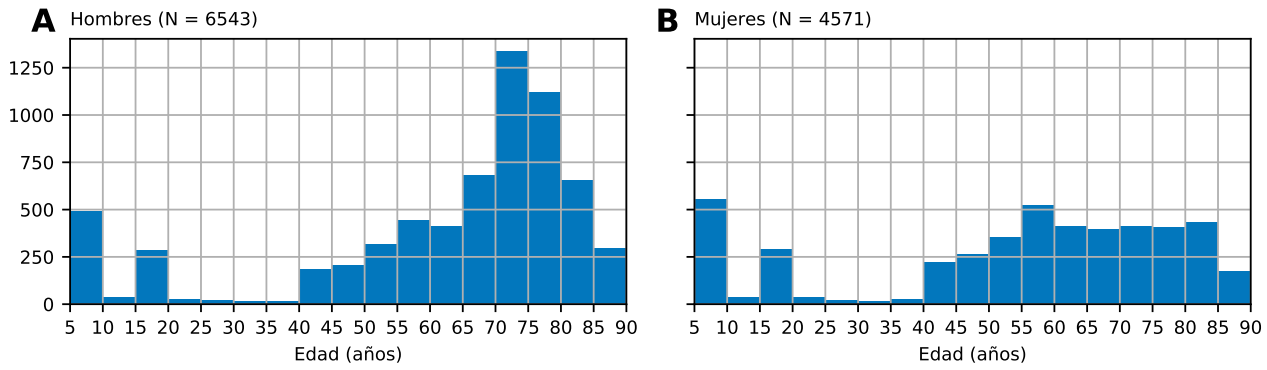

Figura 4.26: Demografía de los sujetos de NSRR con al menos 10 detecciones de husos de sueño. (**A**) Histograma de edad para hombres. (**B**) Histograma de edad para mujeres.

de los cuales uno no tiene detecciones (ID mros-visit1-aa5843 con potencia sigma relativa de 0,82). La demografía de los sujetos que se consideran se detalla en la Figura 4.26. Por simplicidad, los tres sujetos con edad entre 4 y 5 años se consideran de 5 años. En esta demografía, se tienen datos suficientes para ambos sexos (41 % de mujeres) y para todos los rangos de edad entre 5 y 90 años (el menos poblado es 30–35 años con 15 hombres y 14 mujeres). Sin embargo, los datos muestran baja representatividad del rango entre 20 y 40 años, y están sesgados hacia los adultos mayores, particularmente hacia los hombres mayores de 65 años.

Para observar el efecto de la edad con alta resolución, en la Figura 4.27 se mide la distribución de los parámetros en intervalos de 5 años sin diferenciar por sexo. Por otro lado, para también observar el efecto del sexo, en la Figura 4.28 se mide la distribución de los parámetros según sexo y agrupados en tres rangos etarios marcados por la menarquia (12 años) y la menopausia (50 años). En ambos casos, las distribuciones se muestran con diagramas de cajón cuyas patillas cubren del percentil 1 al 99 (valores fuera de las patillas se indican con círculos negros), en donde también se señala el promedio con un círculo verde.

En general, se observan comportamientos relativamente continuos a través de los años en la Figura 4.27. La duración aumenta hasta alcanzar un pico en 10–15 años, decae progresivamente hasta los 65–70 años, y permanece estable a partir de ahí. La amplitud es máxima en 5–10 años, decae fuertemente hasta los 20–25 años, se mantiene relativamente estable hasta los 60–65 años, y decae en los 65–70 años para nuevamente permanecer estable. La frecuencia es mínima en 5–10 años, crece progresivamente hasta los 20–25 años, permanece relativamente estable hasta los 40–45 años, y decae muy lentamente a partir de ahí. Por último, la densidad aumenta hasta alcanzar un pico en 15–20 años, decae fuertemente hasta los 30–35 años, permanece relativamente estable hasta los 55–60 años, y decae fuertemente a partir de ahí.

Los efectos de la edad observados en la Figura 4.27 también son visibles, a una menor resolución, en la Figura 4.28. Sin embargo, al agrupar por sexo, se observa que las mujeres presentan parámetros con un promedio significativamente mayor que los hombres ($P < 0{,}001$). La única excepción es la amplitud PP en 5–12 años, sin diferencia significativa ($P > 0{,}05$). La diferencia de los promedios es particularmente intensa en la amplitud PP y la densidad en 50–90 años. De hecho, al contrario de los hombres, la amplitud PP en mujeres no disminuye al pasar de 12–50 años a 50–90 años, sino que permanece estable.

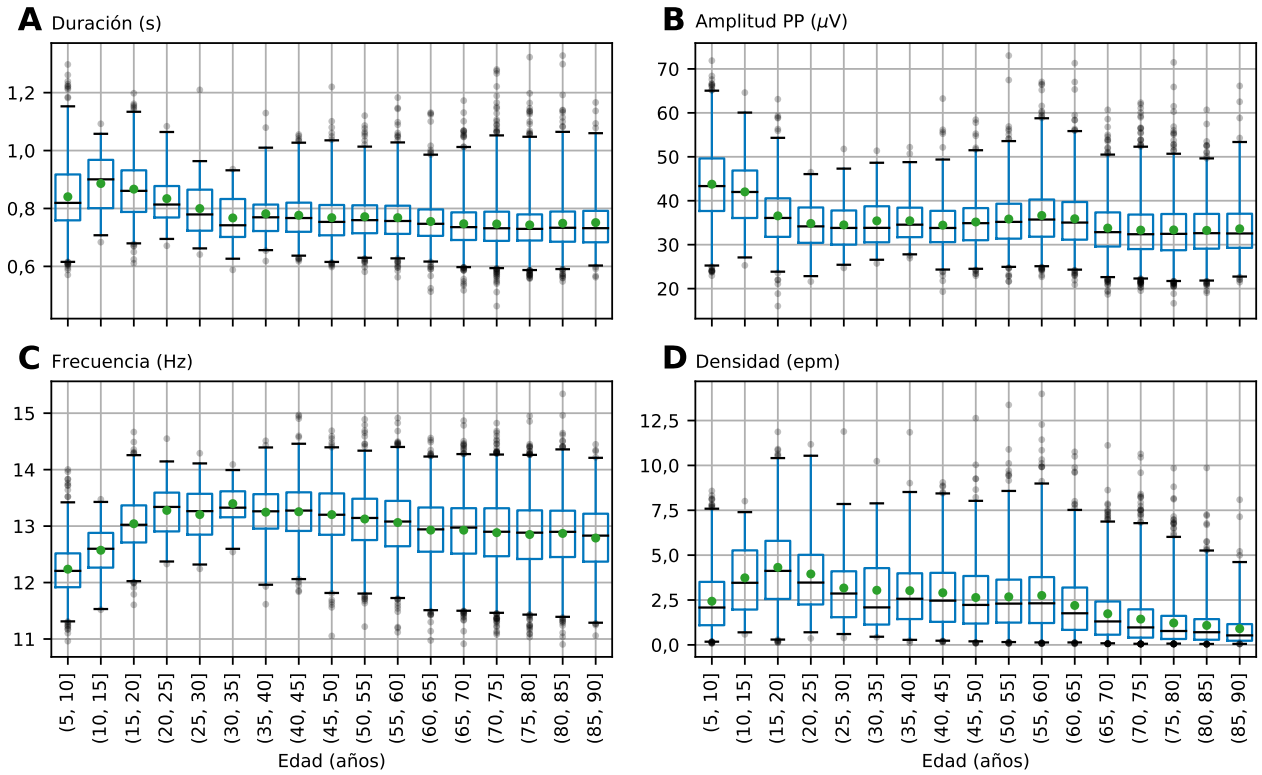

Figura 4.27: Efecto de la edad en la distribución de parámetros por sujeto de las detecciones de husos de sueño en NSRR (sujetos con al menos 10 detecciones).

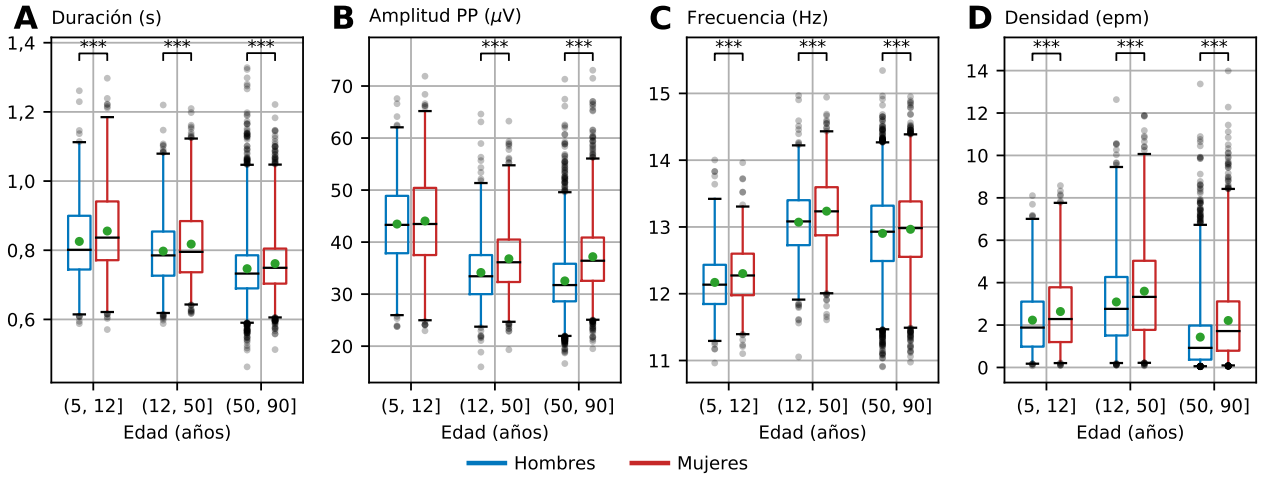

Figura 4.28: Efecto del sexo en la distribución de parámetros por sujeto de las detecciones de husos de sueño en NSRR (sujetos con al menos 10 detecciones). En todos los casos, excepto en la amplitud PP en 5–12 años, existen diferencias estadísticamente significativas entre ambos sexos ($P < 0{,}001$, indicado con ***).

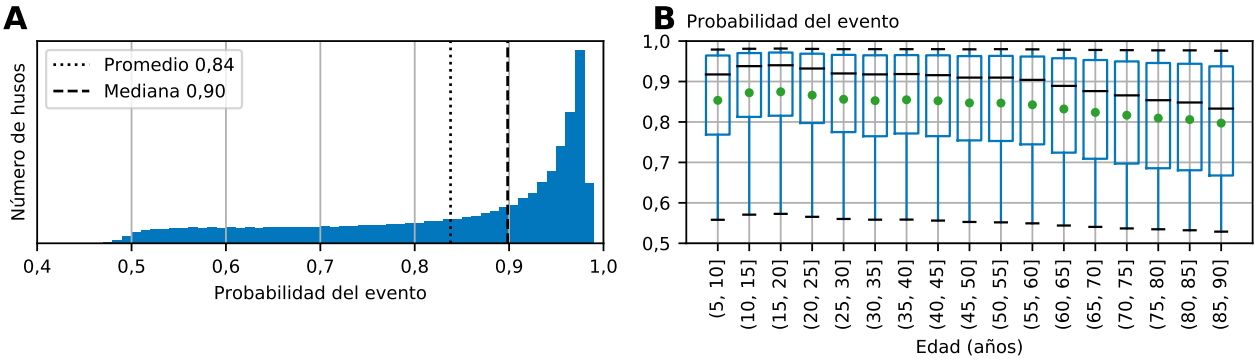

Figura 4.29: Probabilidad asignada a las detecciones de husos de sueño en NSRR (todos los sujetos). (**A**) Distribución de la probabilidad de los eventos. (**B**) Efecto de la edad en la distribución.

4.6.2. Interpretación de la probabilidad predicha

La Figura 4.29 muestra la distribución global, y su dependencia con la edad, de la probabilidad de los eventos detectados por el ensamble de REDv2-Time. Las distribuciones se muestran con diagramas de cajón cuyas patillas cubren del percentil 1 al 99 y cuyo promedio se señala con un círculo verde. Se observa que la probabilidad de los eventos se dispersa hacia valores menores a medida que los sujetos envejecen, sobre todo a partir de los 60 años, lo que sugiere que los adultos mayores muestran husos de sueño menos ideales bajo el estándar aprendido por el detector. Tomando como referencia la distribución de probabilidad en MASS-MODA (se puede consultar en la Figura 4.13A), la distribución en NSRR sugiere una buena generalización del detector. De hecho, las diferencias en el promedio, la mediana y el IQR son pequeñas en la demografía menor a 60 años. En NSRR, los sujetos sobre 60 años están sobre-representados, lo que sugiere que la disminución en 3–4 % del promedio y la mediana de la probabilidad con respecto a MASS-MODA es causada principalmente por una diferencia en la composición etaria.

Para analizar cualitativamente el estándar aprendido por el detector, en la Figura 4.30 se visualizan, con un contexto de 20 s, los cinco primeros y últimos casos que se obtienen al ordenar todas las detecciones de mayor a menor probabilidad. Dado el gran volumen de señales en NSRR, elegir las cinco detecciones con la mayor probabilidad es una forma de aproximar prototipos de REDv2 por medio de la fuerza bruta en lugar de por optimización iterativa de la entrada. Los cinco prototipos encontrados tienen una forma de huso nítida que sobresale claramente del fondo, inicio y fin abrupto, gran amplitud, duración que ronda 1 s, y baja actividad de otras bandas salvo frecuencias muy lentas. En su contexto inmediato, la señal luce limpia de frecuencias cercanas a la sigma. En un contexto más amplio, la mayoría ocurre en presencia de otros husos de sueño, generalmente también de alta probabilidad. En cambio, las detecciones de menor probabilidad no se distinguen claramente de su vecindad, tienen baja amplitud, y corta duración. A excepción de la segunda, no ocurren en presencia de otros husos de sueño.

Para analizar cuantitativamente la probabilidad predicha, se mide la distribución de diversas características para diferentes valores de probabilidad. Para uniformar la distribución observada en la Figura 4.29A, se usan intervalos equiespaciados en el logit de la probabilidad en lugar de la probabilidad. La relación entre el logit y la distribución de ocho características del huso se muestra en la Figura 4.31, mientras que la relación entre el logit y la distribución de dos características del contexto de 20 s del huso se muestra en la Figura 4.32. En ambos casos, las distribuciones

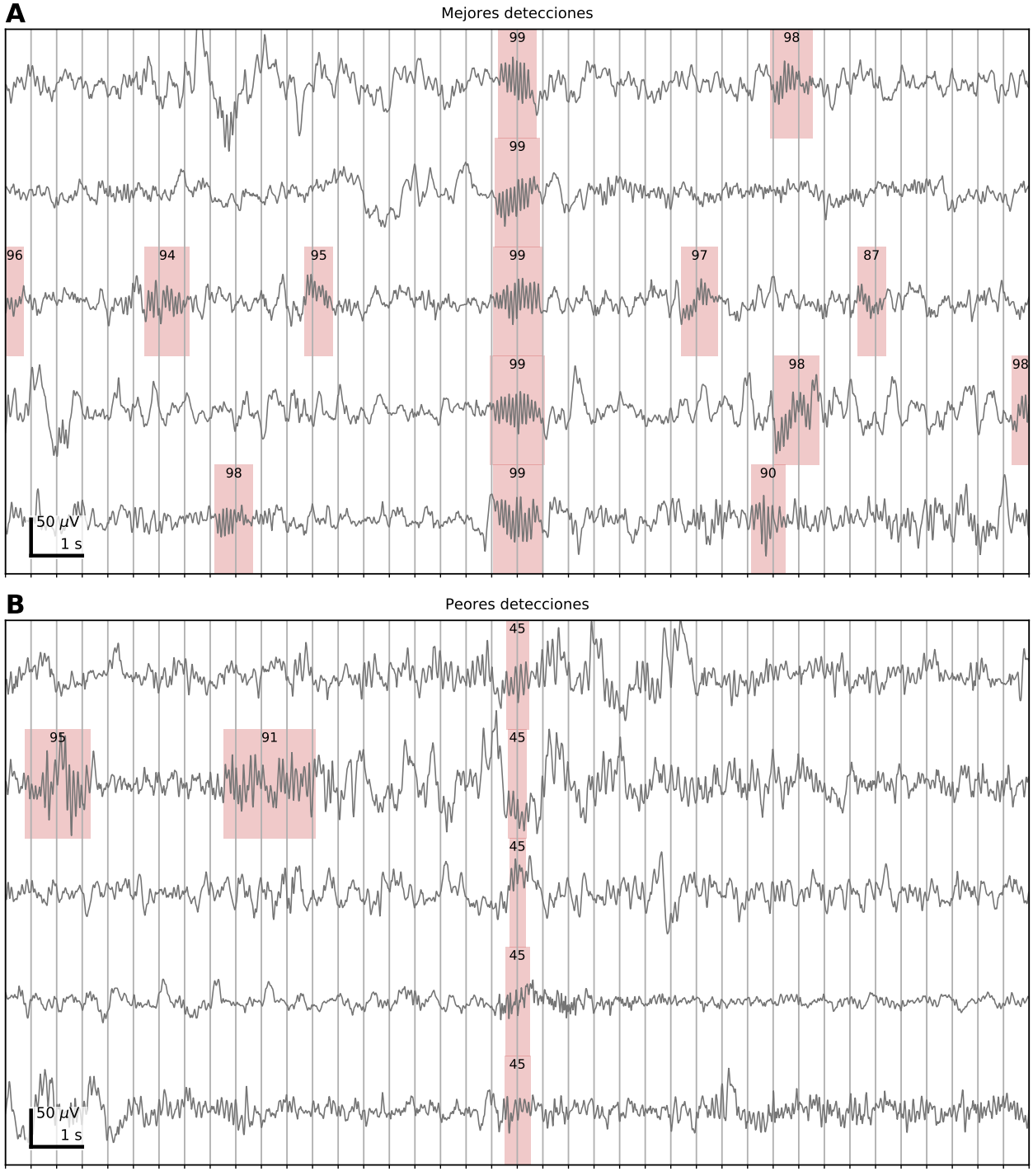

Figura 4.30: Visualización de las cinco detecciones de husos de sueño con mayor (arriba) y menor (abajo) probabilidad en NSRR. En cada panel se muestran cinco señales de 20 s, en cuyo centro se encuentra la detección de interés. Cada detección se indica en rojo con su probabilidad de evento anotada como porcentaje.

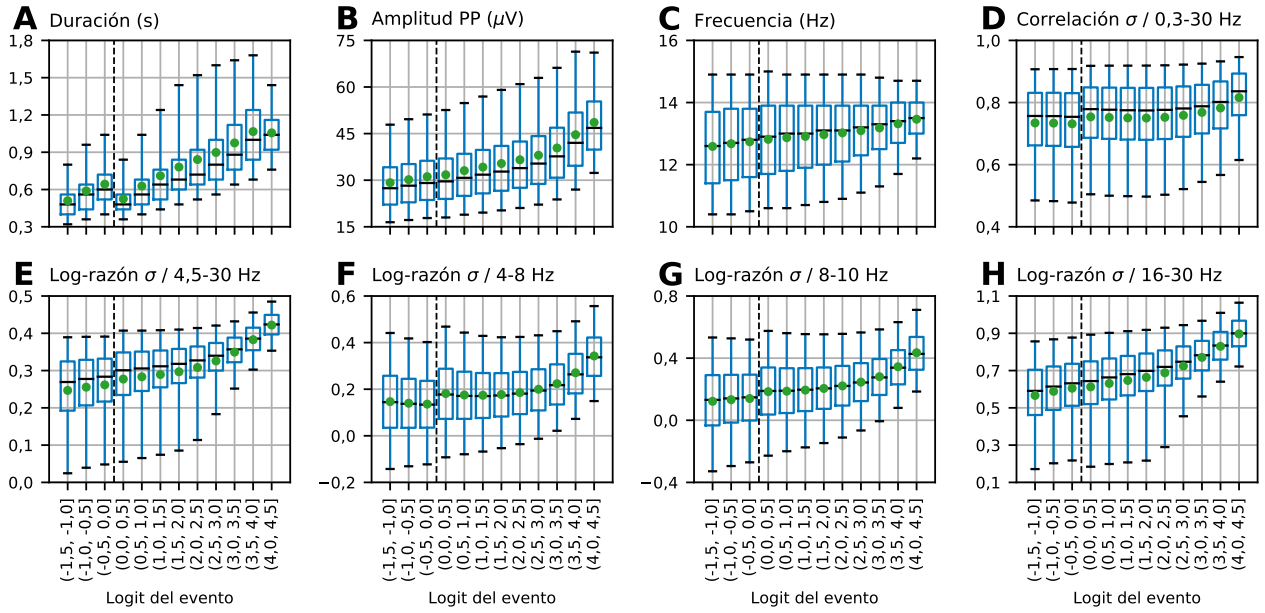

Figura 4.31: Correlación entre el logit del evento y características medidas dentro del intervalo detectado. Se indica con una línea vertical negra la probabilidad 0,5.

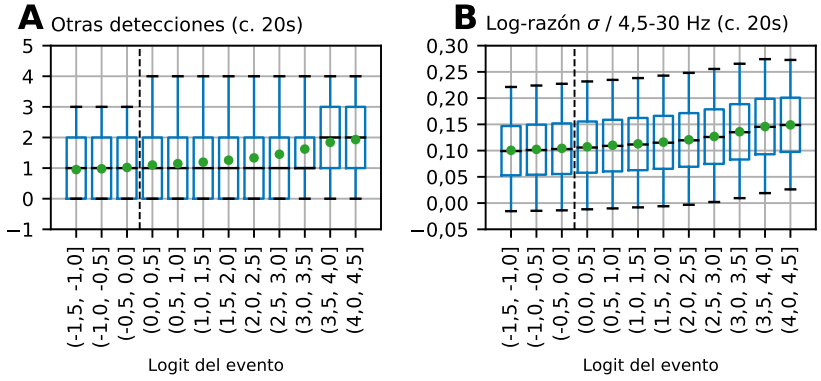

Figura 4.32: Correlación entre el logit del evento y características medidas en su contexto de 20 s. Se indica con una línea vertical negra la probabilidad 0,5.

se muestran con diagramas de cajón cuyas patillas cubren del percentil 5 al 95 y cuyo promedio se señala con un círculo verde.[1] Para extrapolar las tendencias de los gráficos hacia intervalos de señal con probabilidades muy pequeñas para la detección, además se consideran *detecciones falsas*, i.e., eventos que no se detectan con el umbral 0,5 pero sí con el umbral 0,25. Como resultado, el conjunto de eventos con probabilidades menores a 0,5 (logits menores a 0) se puebla con 1.367.930 detecciones falsas.

Debido a que el umbral de 0,25 también deja pasar más borde en cada evento, ocurren discontinuidades en varias de las relaciones medidas cuando se pasa de las detecciones reales a las detecciones falsas. Sin embargo, al ignorar dichas discontinuidades las tendencias son consistentes. En general, todas las características medidas se correlacionan, en mayor o menor medida, con el logit. A medida que la probabilidad asignada aumenta, las detecciones tienden a mostrar una

[1] Las patillas cubren los percentiles 5–95 en lugar de los percentiles 1–99 como en los anteriores diagramas de cajón para no tener rangos tan amplios que dificulten la visualización de la tendencia general.

mayor duración, una mayor amplitud, una frecuencia más concentrada en 12–15 Hz, una mayor potencia sigma relativa, y una mayor actividad sigma en su contexto de 20 s (observada tanto en el número de detecciones reales vecinas como en la potencia sigma relativa del contexto). La correlación entre la banda sigma y la señal original parece hacer una diferencia solo en los niveles más exigentes de probabilidad. Las Figuras 4.31F-H sugieren que la potencia sigma relativa de las detecciones aumenta con la probabilidad a través de una menor actividad beta, alfa, y theta (en orden de importancia). El decrecimiento de la actividad beta se observa en todo el rango de probabilidades, mientras que el decrecimiento de la actividad alfa y theta parece hacer una diferencia solo en los niveles más altos de probabilidad.

Capítulo 5

Discusión

5.1. Diferencias entre REDv2-Time y REDv2-CWT

En general, no se encontraron diferencias significativas entre ambos modelos. Tampoco se encontraron beneficios de ensamblar ambos modelos en comparación con dos instancias del mismo modelo. Podría ser que la separación previa realizada por la CWT no introduce ninguna capacidad extra a la red neuronal, y que la red neuronal puede aprender fácilmente a extraer bandas de frecuencia. Esta evidencia sugiere que es falsa la hipótesis de que es mejor usar la señal transformada a un espacio de tiempo-frecuencia en lugar de usar directamente la señal en el tiempo. Alternativamente, podría ser aún verdadera en alguna dimensión no explorada en este trabajo. De todas formas, se observa que no se cumple en ninguna de las siguientes dimensiones: métricas de detección, aproximación de parámetros, respuesta frente a perturbaciones y señales artificiales, transferencia a otros sujetos, y transferencia a otras bases de datos. Ante la similitud, se sugiere usar REDv2-Time, ya que se evita el cálculo extra de la CWT.

5.2. Efectividad de la perturbación de la entrada durante el entrenamiento

La adición de ondas y anti-ondas, el método propuesto de aumento de datos, tuvo un efecto útil en husos de sueño e insignificante en complejos K. En husos de sueño, el método actúa principalmente subiendo el *recall*, sobre todo en sujetos anómalos de bajo *recall*. Además, es en el *recall* en husos de sueño donde más disminuye la dispersión. En cambio, el método no fue efectivo para compensar las anomalías de bajo *precision*. Por otro lado, en complejos K, el método prácticamente no tuvo efecto. De todas formas, no se observan sujetos anómalos en la forma en que se observan en husos de sueño, implicando que la dispersión inicial (previa a la aplicación del método) ya es buena. Como el método parece afectar más a los sujetos anómalos, la ausencia de un efecto significativo en complejos K es esperable.

Aunque el cambio observado en husos de sueño no es estadísticamente significativo, se observó la mejora en *recall* consistentemente a lo largo de los experimentos. Sin embargo, existe mucha dispersión (desviación estándar) en las métricas calculadas, lo que hace difícil cuantificar las mejoras

pequeñas en la media. Probablemente existe espacio para mejorar los resultados porque no se optimizaron los hiperparámetros del método de aumento de datos por simplicidad. De todas formas, se espera que se requiera de una metodología diferente para compensar los sujetos anómalos en *precision*, porque ni las ondas ni las anti-ondas introducen potenciales falsos positivos para aumentar la robustez de REDv2 frente a ellos. Además, se conjetura que la anomalía en dichos sujetos es causada por una falta de percepción de sus características individuales, en lugar de una falta de discriminación de falsos positivos en general (se discute en mayor profundidad en la Sección 5.5).

5.3. Desempeño de la detección

5.3.1. Comparación del desempeño

Los resultados muestran que REDv2 alcanza el estado del arte en la detección de husos de sueño y complejos K, tanto en métricas de detección como en la aproximación de parámetros. Aunque parece haber un empate en *F1-score* con SpindleU-Net en MASS-SS2, la diferencia en la duración mínima del post-procesamiento (0,3 s en REDv2 y 0,5 s en SpindleU-net) no permite concluir de forma precisa. Consistente con la literatura, se observó que el desempeño es generalmente peor en husos cortos (e.g., $<$0,6 s, ver Figura 4.9A), lo que pone a REDv2 en desventaja al compararse con el desempeño reportado por SpindleU-Net. De todas formas, el mIoU alcanzado por REDv2 es mucho mejor.

REDv2 ofrece varias ventajas al compararse con la literatura. Al contrario de los métodos tradicionales, no fue necesario usar ninguna característica o regla experta en el detector. Solo se usó dicho conocimiento en el post-procesamiento, en donde usó principalmente el conocimiento de la duración y la combinación de eventos, y en el aumento de datos, en donde se usaron razones de potencia. Esto permite generalizar el detector fácilmente, ya que basta modificar o sencillamente quitar el post-procesamiento o el aumento de datos. Notar que si fuese beneficioso descartar detecciones usando características que constituyen una condición necesaria (e.g., amplitud mínima), es directo hacerlo a través del post-procesamiento. Por otro lado, las detecciones son cualitativamente superiores ya que ajustan mejor los parámetros por sujeto, permitiendo mediciones más precisas de dichos parámetros en investigaciones, tales como la asociación de un cambio en la densidad o duración con alguna condición o tratamiento. Adicionalmente, el desempeño sigue siendo superior al desagregarlo por subconjuntos de parámetros. Aún cuando hay espacio de mejora, ya que el desempeño cae en subconjuntos complicados (e.g., eventos cortos o de baja amplitud), REDv2 muestra un buen desempeño. En husos de sueño, el mejor desempeño está en los eventos con frecuencia 12–14 Hz, que coincide con el rango de frecuencia más característico de estos eventos. Tanto en los husos más lentos como en los más rápidos el desempeño baja, probablemente por el mayor riesgo de actividad confusa proveniente de las bandas de frecuencia adyacentes (alfa y beta).

Los resultados evidencian que aún cuando los husos de sueño y los complejos K son patrones sobresalientes y localizados en una banda de frecuencia, son difíciles de detectar con métodos tradicionales. El detector A7 sufre de los falsos positivos, consistente con lo reportado para otros detectores tradicionales. Spinky, por otro lado, tiene un desempeño notoriamente bajo, aún ignorando el mIoU, lo que sugiere que a pesar de que el complejo K sea un patrón de gran amplitud y con una morfología característica, es difícil discriminarlo con reglas sencillas de otros intervalos de señal con picos de gran amplitud. Tanto DOSED como DKL-KC mejoran por mucho el desempeño de Spinky, demostrando los beneficios del aprendizaje profundo, y REDv2 lo mejora aún más, so-

bre todo en mIoU. Si bien ni DKL-KC ni SpindleNet se reprodujeron en este trabajo, el desempeño obtenido para DOSED es cercano al reportado por estos dos métodos, lo que sugiere que estarían bien representados.

A pesar de su desempeño, REDv2 tiene una desventaja importante frente a los métodos tradicionales: la interpretabilidad. Por ser REDv2 basado en aprendizaje profundo, un experto no puede atribuir los resultados de la predicción a características específicas de la señal, al contrario del caso de métodos basados en características. Este problema también está presente en todos los detectores basados en aprendizaje profundo de la literatura. Una forma de compensar esto es validar extensamente la respuesta del detector para así evaluar diversos comportamientos esperados (e.g., que sea sensible a cierta característica). De esta forma, el experto puede confiar en la robustez del detector para ser usado en una gran diversidad de señales, y puede entender sus límites generales de operación. En este trabajo, esta validación se realiza por primera vez para un detector basado en aprendizaje profundo, y los resultados sugieren que es un detector confiable. Aspectos específicos de esta validación, como la respuesta ante escenarios artificiales, la transferencia del aprendizaje, la replicación de tendencias demográficas conocidas, y las características que afectan la probabilidad predicha, se discuten más adelante.

5.3.2. Métricas

Se observó que las métricas de detección para todos los detectores prácticamente no cambian antes del umbral de IoU de 0,2. En vista de ello, se justifica mantener ese umbral, propuesto inicialmente en [6], para medir el desempeño de determinar la *existencia* de eventos. El mIoU, la métrica complementaria impulsada en este trabajo, es útil para evaluar y comparar de forma ortogonal la dimensión de *localización* (i.e., ajustar el instante de inicio y fin). Por lo tanto, se recomienda en trabajos futuros reportar el mIoU en conjunto con las demás métricas que usan umbral de IoU 0,2.

El uso del mIoU como una métrica aparte se desvía de lo hecho en otros trabajos, en donde se opta por usar un umbral más estricto de IoU para considerar la localización en el desempeño. La desventaja directa es que la elección del umbral es arbitraria, y con un umbral muy alto se enfatiza demasiado la localización. En su lugar, en este trabajo se recomienda mezclar ambas dimensiones, existencia y localización, usando el *Average F1-score* (AF1), que resume el desempeño integrando todos los umbrales de IoU. En este sentido, el AF1 integra información tanto del *F1-score* de un umbral bajo (existencia) como del mIoU (localización). Gracias a esto, el AF1 se puede usar para diseñar detectores o ajustar hiperparámetros (como un umbral de salida) sin decisiones arbitrarias en la métrica, tal como se hizo en este trabajo.

5.3.3. La importancia del procesamiento del contexto

El detector más relevante de la literatura para este trabajo es SpindleU-Net, por su procesamiento intenso del contexto y su predicción de segmentos de 20 s a través de segmentación densa. Sin embargo, SpindleU-Net reporta un mIoU muy bajo en comparación a REDv2. Durante el desarrollo de REDv2, se evaluó el uso de una red convolucional 1D tipo U-Net para la etapa de contextualización, similar pero independiente a lo hecho en SpindleU-Net. Se observó un menor mIoU en comparación a las alternativas secuenciales, consistente con los resultados reportados por SpindleU-Net. Esto sugiere que la inclusión de un procesamiento secuencial mejoró el desempeño, probablemente a través de una mayor capacidad de analizar la dinámica temporal con alta resolución, que permite,

por ejemplo, predecir con más exactitud el instante de inicio y fin de los eventos.

En los experimentos de desarrollo se observó que un contexto grande es requerido para un buen desempeño, del orden de 20 s. Tanto DOSED como SpindleU-Net seleccionan dicho contexto, probablemente porque coincide con el tamaño de una página de los registros. De todas formas, la evidencia experimental de este trabajo confirma que dicho contexto es adecuado y que más allá de eso el desempeño se satura. Al mismo tiempo, estos resultados confirman que los enfoques de clasificación aislada de ventanas pequeñas ($\leq$ 1 s) son limitados. Además, las ventanas deslizantes requieren la elección de un ancho de ventana y son redundantes en sus cálculos, lo que no ocurre en un enfoque de segmentación densa de segmentos grandes, como REDv2 y SpindleU-Net, haciendo el método más sencillo, general y rápido.

5.3.4. Efecto del desbalance de clases

La mayoría de las muestras en el EEG, incluso al restringirse solo a la etapa N2, pertenecen a la clase negativa (actividad de fondo), provocando un gran desbalance. Sin embargo, los resultados sugieren que el desbalance no representa un problema al entrenar REDv2. Aunque fue necesario generar batches representativos (como se hizo para DOSED en [8]), no fue necesario usar una función de pérdida diferente a la entropía cruzada, ni fue necesario aplicar ponderadores de clase en ella. De hecho, no se encontraron ganancias en el desempeño al evaluar varias alternativas de función de costo y ponderadores. Esto es consistente con los resultados de SpindleU-Net en [10], en donde proponen una función de costo diseñada para balancear las clases pero no resulta en mejoras en comparación con la entropía cruzada.

Que no sea necesario compensar el desbalance es contraintuitivo, y difiere de lo que se ha observado en detectores previos basados en clasificar ventanas por medio de aprendizaje de máquinas tradicional e incluso aprendizaje profundo. Es posible que predecir un segmento de 20 s en lugar de ventanas pequeñas contribuya a este efecto. El hecho de que en REDv2 y en SpindleU-Net baste usar entropía cruzada, mientras que en DOSED sea necesario adaptar la función de costo, aún cuando los tres métodos predicen segmentos de 20 s, sugieren que la ventaja proviene del enfoque de segmentación densa. La causa podría ser que la mayoría de las muestras negativas no compiten con las muestras positivas por ser rápidas de discriminar usando criterios simples (en husos de sueño, la ausencia de la actividad sigma). Esto evitaría que el entrenamiento colapse a la clase negativa, permitiendo optimizar la discriminación entre la clase positiva y los intervalos negativos desafiantes (en husos de sueño, actividad en banda sigma no asociada a husos de sueño). La explicación es consistente con que incluso haya sido dañino asignarle un mayor peso a la clase positiva en los experimentos. La reducción de la frecuencia *efectiva* de la clase negativa a los intervalos desafiantes es análogo al proceso de dos pasos seguido por otros detectores, en donde primero detectan ventanas candidatas usando condiciones necesarias para luego aplicar clasificadores. Sin embargo, los resultados sugieren que realizar dos pasos de forma explícita es innecesario en redes neuronales.

5.4. Validación extensa de REDv2

5.4.1. Respuesta ante escenarios artificiales

Los experimentos muestran que REDv2 se comporta adecuadamente en los escenarios artificiales considerados, que incluyen someter el detector a señales perturbadas, señales compuestas por

ruido (PINK), y señales etiquetadas artificialmente (CAP-S1, CAP-S2 y CAP-A7). Los cambios en el desempeño frente a las diversas perturbaciones muestran que lo aprendido por REDv2 es consistente con un buen detector. Tal como en detectores tradicionales, la amplitud absoluta de los eventos es importante. Curiosamente, importa principalmente para determinar la existencia de los eventos, no para su duración, lo que es consistente con la literatura que busca puntos de inflexión en la señal para determinar los puntos de inicio y fin, o una disminución relativa de la amplitud con respecto a la amplitud máxima alcanzada.

Por otro lado, las inversiones de amplitud o temporales no afectan la detección de husos de sueño, lo que es consistente con su naturaleza oscilatoria y relativamente simétrica. Esto contrasta con los complejos K, en donde la secuencia ordenada de un pico negativo seguido de uno positivo es parte de su definición. Consistente con esto, el desempeño cae ante cualquiera de las dos inversiones, puesto que se altera la secuencia. Sin embargo, el *recall* no es nulo, lo que sugiere que se siguen detectando algunos complejos K verdaderos, aunque con un bajo IoU. Existen dos posibles explicaciones. La primera es que REDv2 no aprende a ser completamente específico a la fase, pero esta explicación se puede descartar al observar la especificidad de las formas detectadas como complejos K al predecir en PINK. Por lo tanto, queda la segunda explicación, que consiste en que dichas detecciones ocurren porque una de las dos componentes del complejo K verdadero invertido se completa usando señal circundante para formar algo compatible con un complejo K. Esto daría lugar a una detección falsa pero con una intersección parcial con el intervalo correcto, lo que es consistente con el IoU observado.

La respuesta a la eliminación de bandas de frecuencia también es consistente con las características reconocidas en la literatura. En husos de sueño, el detector aprende a ser específico a la banda sigma, y aprende a tomar en cuenta la actividad de las bandas vecinas para descartar falsos positivos, sobre todo de la banda theta. Además, el desempeño es invariante a la banda delta lenta, consistente con el hecho de que los husos de sueño pueden montarse en ondas lentas. El desempeño prácticamente no cambia al quitar la banda beta, contrario a lo esperado por la literatura, en donde la contaminación beta puede indicar la presencia de artefactos que generan eventos falsos. Como MASS-MODA está compuesta de segmentos de EEG libres de artefactos, es posible que la eliminación de la banda beta no haya tenido efecto en estos datos. Sin embargo, las correlaciones medidas en NSRR muestran que REDv2 aprende a asignar una mayor probabilidad a eventos con menor actividad beta. En complejos K, el detector aprende a utilizar principalmente las frecuencias lentas, que es justamente en donde se localiza la actividad de estos patrones. El detector aprende a ignorar las frecuencias por sobre 8 Hz, consistente con la literatura que admite la superposición con actividad de frecuencia mayor, como husos de sueño. Si bien el complejo K tiene un período principal que se encuentra al interior de la banda delta lenta, su morfología abrupta introduce componentes más rápidas que explican la contribución al desempeño de frecuencias sobre 2 Hz.

Los experimentos en PINK sugieren que REDv2 tiene una respuesta específica al evento de interés, según se observa en los parámetros de dichas detecciones, en la visualización de casos, y en la plantilla promedio de complejo K. Es decir, la gran variedad de formas de onda aleatorias en PINK no logra inducir detecciones con morfologías inesperadas. Este comportamiento deseable no es trivial de asegurar en métodos de aprendizaje profundo, en donde es difícil conocer su respuesta ante entradas fuera de la distribución de los datos de entrenamiento, justificando la importancia de su evaluación. El decaimiento continuo de la amplitud con respecto a la frecuencia, sin aumentos transitorios de la actividad sigma, hacen esperable la necesidad de aumentar la amplitud de la señal

para generar detecciones de husos de sueño y que dichas detecciones ocurran en el rango inferior de la banda sigma. Además, la dificultad de mantener una oscilación en una señal ruidosa hacen esperable que estas detecciones sean cortas.

Los experimentos en las bases de datos con señales de CAP, i.e., CAP-S1, CAP-S2 y CAP-A7, tienen por objetivo evaluar la capacidad de REDv2 de imitar detectores de husos de sueño de la literatura. Parece razonable esperar que una red neuronal tenga en su espacio de búsqueda las funciones implementadas por detectores basados en características ampliamente reconocidas, y que además sea capaz de dar con dichas funciones cuando se entrena con datos anotados con esos criterios. Así, los expertos pueden confiar en que, si dichos criterios son útiles, la red neuronal será capaz de aprenderlos.

Los resultados sugieren que REDv2 es capaz de aproximar los criterios expertos en base a ejemplos, pero existe espacio de mejora. En general, la tarea de imitación es más difícil (i.e., menor desempeño promedio y mayor dispersión) mientras mayor sea la contribución de la normalización de las características usando las estadísticas de cada sujeto (CAP-S1 $<$ CAP-A7 $<$ CAP-S2). Existen sujetos con desempeño anómalo, que se puede atribuir a sus espectros poco comunes que provocan que la interacción entre sus formas de señal y las reglas de anotación, sobre todo las normalizaciones, no esté bien representada en el conjunto de entrenamiento. Por lo tanto, se infiere que REDv2 no es eficaz para aprender dichas normalizaciones, y en general no es eficaz para aprender a *personalizar* las características. En base a ello, se infiere que la mayor dispersión observada en MASS-SS2-E1SS en comparación a MASS-SS2-E2SS y MASS-SS2-KC sería causada por una mayor contribución de las estadísticas individuales en la generación de anotaciones de MASS-SS2-E1SS. Si bien los resultados en bases de datos expertas evidencian la necesidad de una personalización del detector a cada sujeto, particularmente en husos de sueño (más sobre esto en la Sección 5.5), no es claro si las normalizaciones usadas en CAP-S2 y CAP-A7 son el camino para ello. Por ejemplo, aún cuando A7 normaliza sus características, REDv2 alcanza un mayor promedio y una menor dispersión en el desempeño en las bases de datos expertas.

De los tres detectores tradicionales usados en CAP, el detector A7 es el más confiable por ser el de mayor desempeño, por lo que CAP-A7 constituye la base de datos artificial más importante para evaluar la capacidad de personalización de un detector basado en aprendizaje profundo. Además, CAP-A7 permite tener un gran control en el experimento, ya que las reglas de anotación son conocidas y consistentes, y se cuenta con un gran volumen de datos. Aún cuando la normalización podría no ser el camino, CAP-A7 puede ser una prueba para evaluar si el detector reconoce que es necesario personalizar y si descubre el mecanismo de personalización, al estilo de un *sanity check*. En cambio, los detectores S1 y S2 tienen un desempeño bajo, con *F1-score* menor a 65 % en MASS-MODA, aún cuando dicho desempeño está sobre-ajustado y considera eventos de al menos 0,5 s. Por lo tanto, las detecciones de S1 y S2 no son suficientemente representativas de husos de sueño reales, y no se recomienda su uso.

5.4.2. Transferencia del aprendizaje

En la transferencia, tanto REDv2 como DOSED se benefician de mantener una normalización cercana a la realizada en la base de entrenamiento. En otras palabras, es conveniente mantener fijo en cuánto se escala 1 μV, lo que es consistente con la observación de que la amplitud absoluta es una característica importante. En la mayoría de las transferencias sin ajuste fino, REDv2 tiene un

F1-score mejor o igual que los otros detectores, sugiriendo que el sobre-ajuste no es un problema para REDv2 a pesar de alcanzar el estado del arte en las bases evaluadas. Sin embargo, existen caídas considerables, y que se observan en todos los detectores evaluados. Los resultados de la transferencia directa y las distribuciones de parámetros de las distintas anotaciones, sugieren que la caída del desempeño no responde a una mala generalización a señales externas, sino que a cambios en las reglas de anotación al cambiar de base de datos. En otras palabras, el detector aprende un criterio distinto al usado para evaluar su desempeño en la transferencia. A partir de esta premisa, debería ser posible corregir este problema a través de un ajuste fino de la red neuronal con pocos datos, ya que se estarían reutilizando casi todas las características aprendidas y solo se estarían modificando las reglas de clasificación de las últimas capas. Justamente, es esto lo que se observa en los experimentos de ajuste fino hacia MASS-MODA, confirmando la suposición.

Los desplazamientos en la distribución de las anotaciones en este tipo de tareas son casi imposibles de compensar automáticamente. Por lo tanto, su existencia obliga a realizar un ajuste fino del detector cada vez que se desee cambiar de criterio de anotación (ver Sección 2.4.4). De todas formas, sería conveniente si dicho ajuste fino fuese más fácil de hacer, con menos datos o sin reentrenamiento. Por ejemplo, el detector basado en aprendizaje profundo podría integrar umbrales interpretables asociados a la sensibilidad a diversos aspectos de las señales, al modo de los detectores tradicionales como A7. De esta forma, el ajuste fino se podría realizar modificando dichos umbrales en lugar de los pesos de las capas neuronales. Además, permitiría a un experto explorar diversas configuraciones por su cuenta, entregándole mayor control y mejorando su experiencia de uso.

Además de la transferencia a otras bases de datos (transferencia externa) también se encontraron deficiencias al transferir el modelo a sujetos no vistos en la misma base de datos (transferencia interna). Esto se debe a que, al menos en husos de sueño, es necesario personalizar los criterios de detección según las características de cada sujeto, pero REDv2 no es capaz de lograrlo (más sobre esto en la Sección 5.5). A modo de compensación, se evaluó personalizar el umbral de probabilidad a cada sujeto usando una fracción de señal anotada. Esto solo es eficaz en MASS-SS2-E1SS. Probablemente, el detector que se aprende en dicha base de datos predice una probabilidad que está fuertemente correlacionada con la característica que el experto personaliza (e.g., alguna razón de potencia). Por lo tanto, la eficacia de personalizar el umbral de probabilidad en dicha base de datos es más bien una coincidencia. De todas formas, el experimento ilustra la ventaja de aislar características que, al personalizar sus umbrales, permitan personalizar el detector.

La base de datos INTA-UCH presenta una dispersión muy grande en sus resultados, causada por la combinación de pocos sujetos ($N = 10$) y una gran dispersión entre sujetos. Esta dispersión, en lugar de responder a una limitación de los detectores, podría ser causada por desplazamientos en la distribución de las anotaciones dentro de la misma base de datos. Varios expertos generaron el conjunto de las anotaciones (sin consenso, sino que cada uno anotó diferentes porciones). Además, algunos sujetos han sido sometidos a revisiones manuales parciales para resolver conflictos de anotación.

5.4.3. Requerimiento de datos

Las curvas de aprendizaje de CAP-A7 (Figura 4.18) y MASS-MODA (Figura 4.20) sugieren que el desempeño actual puede seguir mejorando con más datos, ya que las curvas de *F1-score* no

alcanzan a mostrar un estancamiento. Aún así, REDv2 es capaz de alcanzar un buen desempeño incluso en MASS-MODA sin pre-entrenamiento (*F1-score* 81,5–81,8 %), a pesar de que dicha base de datos es relativamente pequeña para aplicaciones de aprendizaje profundo. Probablemente, el hecho de que cada segmento de EEG de 20 s tiene varias etiquetas en lugar de solo una (por no tratarse de un problema de clasificar segmentos) contribuye a un mejor entrenamiento. Esto es consistente con el menor requerimiento de datos necesarios para entrenar, con buenos resultados, redes neuronales para segmentar imágenes médicas [67]. De todas formas, en la curva de aprendizaje de CAP-A7 el desempeño aumenta significativamente con más datos alrededor del tamaño de MASS-MODA, lo que sugiere que el desempeño obtenido en MASS-MODA podría ser significativamente mejor si el detector se diseña para aprender más eficientemente en dicho régimen.

El requerimiento de datos puede bajar aún más gracias al pre-entrenamiento en otra base de datos más grande (ver Figura 4.20). Con tan solo el 10 % de MASS-MODA, equivalente a 2 h de señal anotada para la combinación del conjunto de entrenamiento y validación, un modelo pre-entrenado alcanza un *F1-score* de 79,5–80,1 %. La evolución de las distintas métricas con y sin pre-entrenamiento muestra que el mayor beneficio del pre-entrenamiento proviene de un mejor posicionamiento inicial del *precision*. Es decir, el requerimiento de datos baja porque el modelo ya tiene información incorporada de cómo discriminar falsos positivos. Sorprendentemente, el pre-entrenamiento en CAP-A7 también entrega estos beneficios, sugiriendo que el detector A7 aporta criterios útiles. Durante el desarrollo de REDv2, se evaluó entregar al modelo directamente las características usadas por A7, sin resultar en ganancias, lo que sugiere que es más efectivo el pre-entrenamiento. Se conjetura que dicho pre-entrenamiento podría ser mejor usando etiquetas suaves de un ensamble de diferentes configuraciones de A7.

La efectividad del pre-entrenamiento en CAP-A7 sugiere una estrategia prometedora para entrenar un detector de eventos basado en una red neuronal profunda, en donde se cuenta con pocas anotaciones y no se cuenta con otra base de datos anotada para pre-entrenar:

1. Recolectar señales sin anotaciones del evento y con características similares (e.g., misma etapa de sueño, demografía similar). En este trabajo, fue la base de datos CAP, con más de 250 h de etapa N2.
2. Determinar un detector del evento basado en características y reglas expertas reconocidas y que entregue un buen desempeño, ya sea de la literatura o de desarrollo propio. En este trabajo, fue el detector A7 de husos de sueño.
3. Ajustar los parámetros de dicho detector en una pequeña porción de los datos etiquetados. En este trabajo, se usaron los parámetros publicados de A7, ajustados en datos privados.
4. Construir una base de datos con anotaciones artificiales al generar detecciones en la base de datos sin anotaciones. En este trabajo, esto da lugar a CAP-A7.
5. Pre-entrenar la red neuronal en la base artificial.
6. Realizar un ajuste fino de la red neuronal en los datos con anotaciones expertas.

En base a los resultados de la transferencia CAP-A7$\rightarrow$ MASS-MODA, se espera que la estrategia anterior alcance un mejor desempeño que entrenar directamente en los datos con anotaciones expertas. Es probable que esta ventaja provenga de dos efectos. En primer lugar, las detecciones del detector simple serían un medio para transferir conocimiento experto a la red neuronal en su inicialización. En segundo lugar, aún cuando dicho conocimiento tenga fallas, gran parte de las representaciones aprendidas por la red neuronal serían una buena inicialización para aprender des-

pués mejores características y criterios de clasificación, regularizando así el entrenamiento.

5.4.4. Reproducción de tendencias demográficas en husos de sueño

La inferencia en NSRR intenta replicar el estudio realizado en [49] para validar cualitativamente (i.e., usando características de la agregación de varias detecciones, en lugar de cada detección por sí sola) el detector desarrollado. Existen tres diferencias metodológicas importantes respecto a dicho estudio. La primera es el detector utilizado, ya que en [49] usan un detector simple basado en wavelet que considera solo la potencia en la banda sigma, con un umbral proporcional a la media del registro. Es decir, se elige un método similar al detector S2 usado para generar CAP-S2. Aún cuando el uso de enfoques similares a S1 y S2 está extendido en las investigaciones de husos de sueño, se mostró anteriormente que tienen un mal desempeño, por lo que sus detecciones no son suficientemente representativas de husos de sueño reales. La segunda diferencia es el post-procesamiento de duración de las detecciones. En [49], probablemente debido a las limitaciones del detector usado, se fija una duración mínima de 0,5 s y una separación mínima de 1 s, mientras que en este trabajo se usa 0,3 s para ambos parámetros, siguiendo recomendaciones recientes [6, 23]. La tercera diferencia se relaciona con el manejo de los artefactos. Mientras que en [49] se usan varios procedimientos para corregir las señales, en este trabajo se optó solo por remover del análisis las páginas con características espectrales anómalas y las detecciones con amplitud PP anómala, según los rangos observados en MASS-MODA. Por un lado, esta decisión se tomó para ahorrar tiempo y reducir variables, pero también está motivada por la mayor robustez observada en REDv2 en comparación a los detectores que dependen del cálculo de características o del filtrado pasa-banda en la banda sigma. Por ejemplo, dichos filtros son altamente excitados por picos tipo impulso que pueden aparecer por artefactos. Debido a esta simplificación, es muy probable que existan datos afectados por artefactos en el análisis, pero se espera que sean la minoría y que las tendencias reales sean bien representadas por la mayoría de los datos.

Debido a las tres diferencias anteriores, las estadísticas obtenidas en [49] se deben tratar con precaución al momento de comparar, aún cuando es el estudio más cercano. Para compensar esta brecha, también se considera en la comparación las estadísticas basadas en consensos de anotaciones expertas, tanto las obtenidas usando señales privadas del Wisconsin Sleep Cohort (WSC) en [6] como las obtenidas usando señales de MASS (MASS-MODA) en [23]. Justamente por ser MASS-MODA el conjunto de entrenamiento de REDv2 para inferir en NSRR, se espera que las tendencias obtenidas en NSRR se acerquen a lo reportado en [23]. Por brevedad, los datos analizados en [49] se indican por NSRR-Purcell, mientras que los datos analizados en este trabajo se indican por NSRR-REDv2.

En general, la distribución global de los parámetros por evento (duración, amplitud PP y frecuencia) de NSRR-REDv2 es similar a lo observado en NSRR-Purcell y WSC. La distribución de duración es más cercana a lo reportado en WSC, la referencia experta, tal como se esperaba por las razones expuestas anteriormente. Por otro lado, en NSRR-REDv2 hay menos eventos con amplitudes PP menores a 15 μV en comparación a NSRR-Purcell y WSC, lo que podría indicar un sesgo de REDv2 contra eventos con amplitudes muy pequeñas. Esto es consistente con el menor desempeño en las amplitudes pequeñas que se observó en el análisis por subconjuntos. En frecuencia, los tres casos son similares al interior de la banda sigma. Sin embargo, tanto la distribución obtenida en NSRR-REDv2 como la reportada en WSC tiene un conteo significativo hasta 16 Hz mientras que en NSRR-Purcell el conteo es significativo solo hasta 15 Hz. En las frecuencias menores, ni WSC

ni NSRR-Purcell tienen conteo significativo en 10–11 Hz, como sí ocurre en NSRR-REDv2. Esto podría deberse a que en NSRR hay una porción importante de niños, cuyos husos tienden a ser más lentos, pero en WSC los sujetos tienen 57 ± 8 años y en NSRR-Purcell solo se detectaron eventos en 11–15 Hz.

La distribución global de la densidad de husos en cada sujeto decae exponencialmente hacia las densidades altas, y en algunos casos la densidad supera los 10 epm, tal como en WSC y MASS-MODA. En cambio, en NSRR-Purcell la distribución de densidad tiene un conteo relativamente uniforme en 0,5–3 epm y corte abrupto en 5 epm, alejándose de la distribución experta. Si bien existe una diferencia demográfica importante, el hecho de que REDv2 se acerque más a las distribuciones de WSC y MASS-MODA sugiere que esa no es la principal causa de la diferencia, sino que el detector utilizado, por lo que las densidades medidas en NSRR-Purcell no son confiables. A pesar de la ventaja de los resultados de REDv2, la distribución obtenida tiene un gran porcentaje cerca de cero, lo que no ocurre en las referencias expertas. Debido a que la correlación entre la potencia sigma relativa y la densidad en NSRR ajusta bien la correlación obtenida por las anotaciones de MASS-MODA (ver Figura 4.25), el mayor porcentaje cerca de 0 epm parece corresponder a una diferencia demográfica. En efecto, la demografía de NSRR está sesgada hacia los adultos mayores, y varios de ellos presentan muy baja actividad sigma en su espectro, sugiriendo que también tendrían, correctamente, muy baja densidad.

Las tendencias demográficas por edad y sexo son bien replicadas en NSRR-REDv2. Tanto la duración como la amplitud disminuyen con la edad y son mayores en mujeres, tal como en NSRR-Purcell, WSC y MASS-MODA. Sin embargo, a lo largo de la vida, el decaimiento obtenido en NSRR-REDv2 para la duración y la amplitud solo ocurre con intensidad hasta la adultez, luego de la cual ocurre con una pendiente mucho menor. En cambio, en NSRR-Purcell el decaimiento es siempre pronunciado, y es particularmente intenso para la amplitud al entrar en la tercera edad. Esta diferencia ocurriría por el detector usado en NSRR-Purcell, ya que al usar un umbral de actividad sigma proporcional a la amplitud media, se detectarían muchos eventos de muy baja amplitud en los adultos mayores con una baja actividad sigma en su espectro. No es claro si dichas detecciones son representativas de husos de sueño reales, o si en cambio sobran. Por otro lado, la evolución obtenida para la frecuencia es similar a la reportada en NSRR-Purcell, en donde aumenta lineal y pronunciadamente con la edad hasta la adultez para luego decaer con una pendiente pequeña. El pico de frecuencia en ambos casos ocurre cerca de los 20 años con un valor medio que ronda los 13,2 Hz, pero la frecuencia media tanto en el extremo inferior como superior de la edad es más baja en NSRR-REDv2 que en NSRR-Purcell (aproximadamente 0,3 Hz de diferencia en ambos extremos), consistente con la menor presencia de husos lentos en las detecciones de NSRR-Purcell. En WSC y MASS-MODA no se reporta la evolución de la frecuencia. Por último, la densidad disminuye con la edad a partir de los 18 años, y es mayor en mujeres, tal como en NSRR-Purcell, WSC y MASS-MODA. En el rango 12–50 años, NSRR-REDv2 muestra densidades con un IQR que cubre aproximadamente 2–5 epm, similar al reportado en MASS-MODA. Además, en NSRR-REDv2 la densidad decae con mayor pendiente a partir de los 60 años, efecto que también se observa en WSC pero no en NSRR-Purcell. En el rango 5–18 años la densidad en NSRR-REDv2 crece con la edad y alcanza un pico alrededor de los 20 años, efecto que también se observa en NSRR-Purcell pero de forma menos pronunciada y con un pico más bajo, consistente con su distribución de densidad alejada de la referencia experta.

5.4.5. Interpretación de lo aprendido por REDv2

Si bien no se tiene un mecanismo de detección interpretable en REDv2, a diferencia de los detectores tradicionales, sí se puede ofrecer una interpretación parcial de lo aprendido por REDv2 a través de un análisis de sus respuestas. Los experimentos en donde modelos entrenados predicen sobre señales perturbadas (discutidos en la Sección 5.4.1) permiten generar explicaciones del mecanismo de predicción a través de los efectos que las perturbaciones tienen en las métricas de desempeño. Esto es análogo al método de identificar entradas relevantes por medio de su oclusión. Tanto en husos de sueño como en complejos K se encontró que REDv2 es afectado de forma consistente con el conocimiento del problema en la literatura. Además, se identificaron algunos comportamientos llamativos. Por ejemplo, en husos de sueño, se identificó que REDv2 es sensible a la amplitud absoluta para detectar pero no para definir la duración del evento detectado, lo cual va en contra de algunos detectores tradicionales de la literatura. Por ejemplo, el detector A7 decide la duración del evento detectado en base a exceder el umbral de sus dos características de unidades absolutas (amplitud sigma, y covarianza entre banda sigma y señal original).

El análisis de perturbaciones en las bases de datos anotadas por expertos se complementa con el análisis de la probabilidad predicha en el gran volumen de detecciones en NSRR. En primer lugar, las cinco detecciones con la mayor probabilidad asignada (ver Figura 4.30), seleccionadas de entre más de cuatro millones de detecciones, aproximan lo que sería una entrada óptima para REDv2, i.e., *prototipos* de la clase. Esto es análogo a la generación de prototipos por medio de la optimización de la entrada para maximizar la probabilidad de una clase de interés [85], pero en lugar de una optimización iterativa por gradiente, se tiene aquí una optimización por fuerza bruta. Los prototipos encontrados muestran formas de onda, tanto en su interior como en su vecindad, que coinciden con la descripción canónica de lo que constituye un huso de sueño, apoyando el aprendizaje exitoso del patrón por parte de REDv2. En segundo lugar, la relación entre la probabilidad predicha y diversas características reconocidas en la literatura permite inspeccionar con más detalle las características que REDv2 aprendió a considerar. Al comparar la distribución de características de las mejores detecciones (probabilidad cercana a 1) contra la de las peores (probabilidad cercana a 0,5), se repiten las observaciones cualitativas realizadas a partir de la visualización de los cinco prototipos. En el rango completo de probabilidad, aunque todas las características evaluadas muestran algún grado de correlación, las características más influyentes son la duración, la amplitud, y la potencia sigma relativa. Estas características son populares en los detectores tradicionales. Curiosamente, también hay una correlación entre la potencia sigma relativa del contexto de 20 s y la probabilidad predicha para el evento, lo que sugiere una interacción con los eventos cercanos que podría contribuir a la mejora del desempeño que ocurre cuando se usa un contexto de este tamaño.

5.5. Robustez frente a la variabilidad entre sujetos

A pesar de que REDv2 muestra una menor dispersión del desempeño entre sujetos en comparación a otros métodos (ver Figura 4.2), dicha variabilidad aún es significativa. Además, en la detección de husos de sueño, particularmente en MASS-SS2-E1SS, se observan sujetos anómalos, ya sea de muy bajo *recall* o muy bajo *precision* (ver Figura 4.21). Que existan sujetos anómalos al detectar husos de sueño pero no al detectar complejos K es consistente con que la banda sigma se encuentre dentro de la porción del espectro con mayor variabilidad entre sujetos, mientras que la banda delta no. Además, la baja variabilidad observada al detectar husos de sueño en MASS-SS2-E2SS (con anotaciones dependientes principalmente de la amplitud sigma) en lugar de

MASS-SS2-E1SS, sugiere que el problema principal no proviene de la variabilidad de la amplitud sigma, sino de la variabilidad de su relación con las otras bandas de frecuencia.

Esta variabilidad sería más difícil de abarcar en el conjunto de entrenamiento para el problema de husos de sueño en comparación al de complejos K, dando lugar a un desempeño anómalo en los casos más cercanos al borde de la distribución. Este efecto se observa incluso en la tarea artificial de CAP-A7 (ver Figura 4.16), a pesar del gran volumen de datos y de la consistencia en las reglas de anotación, en donde el detector lo hace bien para la gran mayoría de lo sujetos excepto unos pocos sujetos anómalos. En el caso de CAP-A7, debido a la transparencia del mecanismo de anotación, se puede inferir que las anomalías se producen en sujetos en donde existen muchos eventos candidatos falsos con características cercanas al umbral de detección. Para REDv2, sería difícil aprender el punto de transición a partir de los datos de entrenamiento, sobre todo de combinaciones poco comunes.

Por la baja variabilidad esperada entre sujetos en complejos K, la dispersión entre sujetos observada en MASS-SS2-KC constituye una buena referencia de lo que sería ideal observar al detectar husos de sueño en MASS-SS2-E1SS. De hecho, esta es la dispersión que se alcanza aproximadamente al ajustar el umbral de salida en cada sujeto de MASS-SS2-E1SS con un oráculo (ver Figura 4.21). Además, en la Tabla 4.2 se observa que la desviación estándar en MASS-SS2-KC es comparable a la de MASS-MODA, una base de datos con segmentos de 180 sujetos, apoyando la idea de que en complejos K no es necesario considerar tantos sujetos para alcanzar un desempeño promedio estable como sí lo es en husos de sueño.

En consecuencia, se recomienda enfocar los esfuerzos en aumentar la robustez del detector a la variabilidad entre sujetos en la detección de husos de sueño. En particular, se sugiere atacar las anomalías de bajo *precision*, que son las más desafiantes tanto en MASS-SS2-E1SS como en CAP-A7. Curiosamente, las anomalías de estas dos bases de datos presentan similitudes (ver Figuras 4.17 y 4.22). En ambos casos, los verdaderos positivos tienden a tener una potencia sigma relativa alta en comparación a los demás sujetos, y los falsos positivos tienden a tener una potencia sigma relativa más baja que los verdaderos pertenecientes al sujeto en cuestión pero de valor comparable a los verdaderos pertenecientes a otros sujetos. Esto sugiere que en dichos sujetos, los husos de sueño son más notorios y en consecuencia la detección debe ser más exigente con las características requeridas. Por lo tanto, estas anomalías serían causadas por una incapacidad de REDv2 para adaptar su sensibilidad a las características individuales de la señal del sujeto, como puede ser su perfil espectral poco común.

La necesidad de desarrollar un detector de husos de sueño que se adapte a cada sujeto ha sido varias veces reconocida en la literatura. En métodos tradicionales, la personalización se ha intentado, por ejemplo, a través de un umbral proporcional a la media de alguna característica del sujeto. Sin embargo, ningún método logra el estado del arte en el desempeño promedio a la vez que disminuye la dispersión entre sujetos a un nivel aceptable (e.g., la observada en complejos K). Alcanzar esto es un problema abierto. La incapacidad de personalización de REDv2 es esperable de su contexto restringido al orden de 20 s. Se conjetura que para lograr la personalización es necesario usar información de un contexto mucho más grande, o inducir mayor estructura en la red neuronal que permita capturar con menos datos aquellas características que usan los expertos para personalizar sus criterios. En los experimentos de diseño, se intentó usar un contexto mayor a 20 s de varias formas, sin obtener mejoras. Esto se podría deber a que los mecanismos evaluados no son adecuados

para promover la personalización, o bien, a que MASS-SS2 (la base de datos usada para el diseño) no provee de suficientes sujetos en el conjunto de entrenamiento para aprender una personalización, implicando la necesidad de desarrollar un mejor detector usando una base de datos con más sujetos (e.g., MASS-MODA).

Si bien REDv2, al detectar husos de sueño, no asegura que en todos los sujetos se tiene un desempeño cercano al promedio por la existencia de posibles anomalías, estas anomalías son la minoría, y en general sus resultados son buenos y mejores que el de otros detectores. Aparte de las anomalías, los resultados muestran que REDv2 es sensible a los efectos de la edad en husos de sueño, ya que en adultos mayores el detector tiene un menor desempeño que en adultos jóvenes (ver Figura 4.9). Esto es consiste con que en adultos mayores los husos de sueño tienden a ser de peor calidad (e.g., menor duración y amplitud). De hecho, los expertos también sufren de esta caída de desempeño según [23]. Curiosamente, en el plano de *recall* y *precision* para REDv2, los adultos mayores tienden a ocupar la región de menor *recall*, mientras que los jóvenes la de menor *precision* (ver Figura 4.21), lo que sugiere que, ante la incapacidad de personalizar, REDv2 aprende un comportamiento intermedio. Una consecuencia importante de esto es que REDv2 podría sub-estimar la densidad en adultos mayores y sobre-estimar la densidad en jóvenes. Sin embargo, este efecto parece ser pequeño, ya que la correlación entre la densidad real y la predicha para REDv2 en MASS-MODA es grande y mejor que los otros detectores, con $R^2 = 0{,}94$ (ver Figura 4.7).

Capítulo 6

Conclusión

En este trabajo, se llevaron a cabo con éxito todos los objetivos específicos planteados. Como resultado, se propuso REDv2, un método basado en aprendizaje profundo para detectar eventos transitorios en el EEG del sueño, que alcanza el estado del arte en la detección de husos de sueño y complejos K. Gracias a sus capas recurrentes que procesan secuencialmente un gran segmento de EEG, el método tiene una mayor capacidad para modelar dependencias temporales que los detectores previos basados en aprendizaje profundo, mejorando no solo su capacidad para discriminar sino que también para predecir con exactitud el instante de inicio y fin de los patrones. El método propuesto no define particiones arbitrarias de la entrada ni requiere de anchos de ventanas deslizantes o candidatas, haciéndolo simple y más probable de poder ser aplicado para la detección de otros eventos transitorios. De hecho, el mismo diseño se utilizó para la detección de husos de sueño y complejos K, con diferencias solo en el post-procesamiento y en el aumento de datos.

El método propuesto se validó extensamente a través de varios experimentos, que incluyen el uso de datos artificiales y datos no etiquetados. Como resultado, REDv2 demostró ser un detector confiable, específico al patrón de interés, que generaliza bien a señales externas, y con un comportamiento alineado con criterios expertos típicamente encontrados en la literatura. Además, se identificó y analizó en profundidad la principal limitación de REDv2, también presente en los demás detectores, consistente en la incapacidad para personalizar sus criterios de detección según las estadísticas de las señales de cada sujeto. Esto es particularmente importante en la detección de husos de sueño, y se mantiene como un problema abierto. En este sentido, REDv2 aún no es adecuado para garantizar un desempeño estable al nivel de un solo sujeto. A pesar de esta limitación, REDv2 es capaz de entregar un excelente desempeño promedio, lo que permite su uso para investigar diferencias entre grupos. Consistente con ello, al inferir en un gran volumen de datos sin anotaciones, fue capaz de replicar tendencias demográficas reportadas en la literatura para los parámetros principales de los husos de sueño, sobre todo aquellas reportadas usando anotaciones de alta calidad gracias a un consenso de expertos. En cierta forma, REDv2 es capaz de tomar el rol de un consenso de expertos una vez se entrena en sus ejemplos. Adicionalmente, no es necesaria una gran cantidad de datos para alcanzar un buen desempeño, sobre todo si primero se realiza un pre-entrenamiento.

De las hipótesis planteadas, solo la hipótesis H2 no pudo ser validada, relacionada al uso de la representación tiempo-frecuencia. Al menos considerando el universo de todos los experimentos

realizados en este trabajo, usar la CWT en lugar de la señal EEG directamente en el tiempo no entrega beneficios. Por lo tanto, se recomienda usar la variante REDv2-Time por simplicidad. Aún cuando no hay diferencias en el desempeño, es probable que usar la CWT como entrada permita obtener mapas de relevancia más interpretables si así se desea.

Resultados preliminares de esta tesis fueron publicados y presentados en la *2020 International Joint Conference on Neural Networks* (IJCNN) [12].

6.1. Recomendaciones de investigación futura

Mejorar la adaptabilidad a cada sujeto. Como ya se mencionó anteriormente, la principal limitación actual es la dispersión del desempeño entre sujetos al detectar husos de sueño, en donde incluso a veces se tienen sujetos con desempeño anómalo con respecto al grupo. Se espera que corregir este problema sea el mejor camino para aumentar significativamente el desempeño.

Ajustar el aumento de datos o explorar mejores alternativas. Los hiperparámetros de la adición de ondas aleatorias y anti-ondas aleatorias no se optimizaron, por lo que existe espacio de mejora.

Simplificar el ajuste fino. Encontrar alternativas más interpretables para ajustar el detector a nuevos criterios de anotación. Es particularmente interesante ofrecer a los expertos un conjunto de umbrales con los cuales pueden explorar diversas configuraciones.

Pre-entrenar con una tarea auxiliar en datos no etiquetados. El pre-entrenamiento en CAP-A7 ilustra una forma de aprovechar datos no etiquetados, al pre-entrenar para predecir detecciones de un detector tradicional. No es claro si existen mejores tareas auxiliares, sobre todo que permitan construir representaciones generales que sirvan como punto de partida para un ajuste fino de distintos eventos transitorios. Por ejemplo, se podrían aprovechar las etiquetas de etapa de sueño que tienen prácticamente todas las bases de datos públicas, como las que componen NSRR, ya que patrones como los husos de sueño y los complejos K aparecen más en ciertas etapas.

Validar el detector en otras demografías. En este trabajo se validó el detector principalmente en sujetos sanos mayores a 5 años (10 años si solo se consideran las bases expertas). Queda pendiente validar el detector en sujetos menores a 5 años (e.g., bebés o infantes) o en sujetos con patologías (e.g., esquizofrénicos), ya que en ambas situaciones podrían cambiar significativamente las propiedades de los patrones.

Validar la generalización del diseño del detector en otros eventos transitorios. Sería interesante evaluar qué otros eventos se podrían detectar sin realizar modificaciones importantes, ya sea en el EEG del sueño o durante la vigilia. Dadas las suposiciones del detector, se espera que dicha generalización solo se pueda hacer en eventos relativamente cortos, de un orden similar a los husos de sueño y complejos K, ya que debe ser posible observar que inician y terminan dentro del contexto de entrada. Se podrían explorar eventos más largos si el contexto de entrada es más grande. En dicho caso, para disminuir el costo computacional, probablemente haya que submuestrear la señal, o bien modificar las capas convolucionales para disminuir más la resolución antes

del procesamiento secuencial. Por ejemplo, DOSED fue evaluado en la detección de momentos de excitación durante el sueño (*arousals* en inglés), cuya duración media ronda los 10 s, usando como entrada segmentos de 120 s y obteniendo como resultado un *F1-score* cercano a 70 % [8]. Alternativamente, la generalización podría evaluarse hacia otras etapas del sueño, ya que en este trabajo se limitó el estudio a la etapa N2, pero tanto los husos de sueño como los complejos K se pueden encontrar en la etapa N3. Dado que la etapa N3 tiene una actividad de fondo distinta a la etapa N2, no es claro si el detector tiene un buen desempeño al ser aplicado directamente, o si por el contrario requiere ser entrenado con ejemplos de dicha etapa.

Bibliografía

[1] R. B. Berry, C. L. Albertario, S. M. Harding, R. M. Uoyd, D. T. Plante, S. F. Quan, M. M. Troester, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.5*. American Academy of Sleep Medicine, 2018.

[2] A. Wauquier, L. Aloe, and A. Declerck, "K-complexes: Are they signs of arousal or sleep protective?" *Journal of Sleep Research*, vol. 4, no. 3, pp. 138–143, 1995.

[3] D. Coppieters, P. Maquet, and C. Phillips, "Sleep spindles as an electrographic element: Description and automatic detection methods," *Neural Plasticity*, vol. 2016, p. 6783812, 2016.

[4] B. Clawson, J. Durkin, and S. Aton, "Form and function of sleep spindles across the lifespan," *Neural Plasticity*, vol. 2016, p. 6936381, 2016.

[5] J. El Helou, V. Navarro, C. Depienne, E. Fedirko, E. LeGuern, M. Baulac, I. An-Gourfinkel, and C. Adam, "K-complex-induced seizures in autosomal dominant nocturnal frontal lobe epilepsy," *Clinical Neurophysiology*, vol. 119, no. 10, pp. 2201–2204, 2008.

[6] S. C. Warby, S. L. Wendt, P. Welinder, E. G. S. Munk, O. Carrillo, H. B. D. Sorensen, P. Jennum, P. E. Peppard, P. Perona, and E. Mignot, "Sleep-spindle detection: Crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature Methods*, vol. 11, no. 4, pp. 385–392, 2014.

[7] B. Lechat, K. Hansen, P. Catcheside, and B. Zajamsek, "Beyond K-complex binary scoring during sleep: Probabilistic classification using deep learning," *Sleep*, vol. 43, no. 10, p. zsaa077, 2020.

[8] S. Chambon, V. Thorey, P. Arnal, E. Mignot, and A. Gramfort, "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal," *Journal of Neuroscience Methods*, vol. 321, pp. 64–78, 2019.

[9] P. M. Kulkarni, Z. Xiao, E. J. Robinson, A. S. Jami, J. Zhang, H. Zhou, S. E. Henin, A. A. Liu, R. S. Osorio, J. Wang, and Z. Chen, "A deep learning approach for real-time detection of sleep spindles," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036004, 2019.

[10] J. You, D. Jiang, Y. Ma, and Y. Wang, "SpindleU-Net: An adaptive U-Net framework for sleep spindle detection in single-channel EEG," *IEEE Transactions on Neural Systems and*

Rehabilitation Engineering, vol. 29, pp. 1614–1623, 2021.

[11] P. Addison, *The illustrated wavelet transform handbook: Introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[12] N. I. Tapia and P. A. Estévez, "RED: Deep recurrent neural networks for sleep EEG event detection," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[13] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, and S. M. Williams, Eds., *Neuroscience*. Sinauer Associates, 2001, ch. Electrical Synapses, https://www.ncbi.nlm.nih.gov/books/NBK11164/.

[14] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: The neurophysics of EEG*. Oxford University Press, USA, 2006.

[15] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[16] D. L. Schomer and F. L. Da Silva, *Niedermeyer's Electroencephalography: Basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2012.

[17] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalography and Clinical Neurophysiology*, vol. 66, no. 4, pp. 376–382, 1987.

[18] J. M. Stern, *Atlas of EEG patterns*. Lippincott Williams & Wilkins, 2005.

[19] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, technique and scoring system for sleep stages of human sleep," *Brain Information Service, Los Angeles*, 1968.

[20] M. A. Carskadon and W. C. Dement, "Chapter 2 - Normal human sleep: An overview," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Elsevier, 2017, pp. 15–24.e3.

[21] C. Guilleminault, A. Tilkian, and W. Dement, "The sleep apnea syndromes," *Annual Review of Medicine*, vol. 27, no. 1, pp. 465–484, 1976.

[22] F. Glasauer, "Restless legs syndrome," *Spinal Cord*, vol. 39, no. 3, pp. 125–133, 2001.

[23] K. Lacourse, B. Yetton, S. Mednick, and S. C. Warby, "Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data," *Scientific Data*, vol. 7, no. 1, p. 190, 2020.

[24] L. M. J. Fernandez and A. Lüthi, "Sleep spindles: Mechanisms and functions," *Physiological Reviews*, vol. 100, no. 2, pp. 805–868, 2020.

[25] S. Shinomiya, K. Nagata, K. Takahashi, and T. Masumura, "Development of sleep spindles in young children and adolescents," *Clinical Electroencephalography*, vol. 30, no. 2, pp. 39–43,

1999.

[26] L. De Gennaro, C. Marzano, F. Fratello, F. Moroni, M. C. Pellicciari, F. Ferlazzo, S. Costa, A. Couyoumdjian, G. Curcio, E. Sforza, A. Malafosse, L. A. Finelli, P. Pasqualetti, M. Ferrara, M. Bertini, and P. M. Rossini, "The electroencephalographic fingerprint of sleep is genetically determined: A twin study," *Annals of Neurology*, vol. 64, no. 4, pp. 455–460, 2008.

[27] C. Berthomier, V. Muto, C. Schmidt, G. Vandewalle, M. Jaspar, J. Devillers, G. Gaggioni, S. L. Chellappa, C. Meyer, C. Phillips, E. Salmon, P. Berthomier, J. Prado, O. Benoit, R. Bouet, M. Brandewinder, J. Mattout, and P. Maquet, "Exploring scoring methods for research studies: Accuracy and variability of visual and automated sleep scoring," *Journal of Sleep Research*, vol. 29, no. 5, p. e12994, 2020.

[28] G. Bremer, J. R. Smith, and I. Karacan, "Automatic detection of the K-complex in sleep electroencephalograms," *IEEE Transactions on Biomedical Engineering*, vol. 17, no. 4, pp. 314–323, 1970.

[29] S. Devuyst, T. Dutoit, P. Stenuit, and M. Kerkhofs, "Automatic K-complexes detection in sleep EEG recordings using likelihood thresholds," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 4658–4661.

[30] T. Lajnef, S. Chaibi, J.-B. Eichenlaub, P. Ruby, P.-E. Aguera, M. Samet, A. Kachouri, and K. Jerbi, "Sleep spindle and K-complex detection using tunable Q-factor wavelet transform and morphological component analysis," *Frontiers in Human Neuroscience*, vol. 9, p. 414, 2015.

[31] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[33] K. Kotowski, K. Stapor, and J. Ochab, "Deep learning methods in electroencephalography," in *Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications*, G. A. Tsihrintzis and L. C. Jain, Eds., 2020, pp. 191–212.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[35] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[36] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[37] L. Bottou, "Stochastic learning," in *Summer School on Machine Learning. ML 2003: Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds., 2004,

pp. 146–168.

[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[40] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[41] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1310–1318.

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[45] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2342–2350.

[46] S. Devuyst, T. Dutoit, P. Stenuit, and M. Kerkhofs, "Automatic sleep spindles detection—overview and development of a standard proposal assessment method," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 1713–1716.

[47] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal Archive of Sleep Studies: An open-access resource for instrument benchmarking and exploratory research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 2014.

[48] A. Tsanas and G. D. Clifford, "Stage-independent, single lead EEG sleep spindle detection using the continuous wavelet transform and local weighted smoothing," *Frontiers in Human Neuroscience*, vol. 9, p. 181, 2015.

[49] S. M. Purcell, D. S. Manoach, C. Demanuele, B. E. Cade, S. Mariani, R. Cox, G. Panagiotaropoulou, R. Saxena, J. Q. Pan, J. W. Smoller, S. Redline, and R. Stickgold, "Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource," *Nature Communications*, vol. 8, no. 1, p. 15930, 2017.

[50] C. Held, L. Causa, P. Estevez, C. Perez, M. Garrido, C. Algarin, and P. Peirano, "Dual approach for automated sleep spindles detection within EEG background activity in infant polysomnograms," in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2004, pp. 566–569.

[51] L. Causa, C. M. Held, J. Causa, P. A. Estévez, C. A. Perez, R. Chamorro, M. Garrido, C. Algarín, and P. Peirano, "Automated sleep-spindle detection in healthy children polysomnograms," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 9, pp. 2135–2146, 2010.

[52] B. Babadi, S. M. McKinney, V. Tarokh, and J. M. Ellenbogen, "Diba: A data-driven bayesian algorithm for sleep spindle detection," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 483–493, 2011.

[53] A. Erdamar, F. Duman, and S. Yetkin, "A wavelet and teager energy operator based method for automatic detection of K-complex in sleep EEG," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1284–1290, 2012.

[54] S. Yazdani, S. Fallet, and J.-M. Vesin, "A novel short-term event extraction algorithm for biomedical signals," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 754–762, 2017.

[55] X. Zhao, C. Chen, W. Zhou, Y. Wang, J. Fan, Z. Wang, S. Akbarzadeh, and W. Chen, "An energy screening and morphology characterization-based hybrid expert scheme for automatic identification of micro-sleep event K-complex," *Computer Methods and Programs in Biomedicine*, vol. 201, p. 105955, 2021.

[56] S. Ulloa, P. A. Estevez, P. Huijse, C. M. Held, C. A. Perez, R. Chamorro, M. Garrido, C. Algarín, and P. Peirano, "Sleep-spindle identification on EEG signals from polysomnographie recordings using correntropy," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 3736–3739.

[57] D. Jiang, Y. Ma, and Y. Wang, "A robust two-stage sleep spindle detection approach using single-channel EEG," *Journal of Neural Engineering*, vol. 18, no. 2, p. 026026, 2021.

[58] S. A. Imtiaz, S. Saremi-Yarahmadi, and E. Rodriguez-Villegas, "Automatic detection of sleep spindles using teager energy and spectral edge frequency," in *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2013, pp. 262–265.

[59] K. Lacourse, J. Delfrate, J. Beaudry, P. Peppard, and S. Warby, "A sleep spindle detection algorithm that emulates human expert spindle scoring," *Journal of Neuroscience Methods*, vol. 316, pp. 3–11, 2019.

[60] P. Estévez, R. Zilleruelo-Ramos, R. Hernández, L. Causa, and C. Held, "Sleep spindle detection by using merge neural gas," in *The 6th International Workshop on Self-Organizing Maps (WSOM 2007)*, 2007.

[61] H. Q. Vu, G. Li, N. S. Sukhorukova, G. Beliakov, S. Liu, C. Philippe, H. Amiel, and A. Ugon, "K-complex detection using a hybrid-synergic machine learning method," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp.

1478–1490, 2012.

[62] D. Lachner-Piza, N. Epitashvili, A. Schulze-Bonhage, T. Stieglitz, J. Jacobs, and M. Dümpelmann, "A single channel sleep-spindle detector based on multivariate classification of EEG epochs: MUSSDET," *Journal of Neuroscience Methods*, vol. 297, pp. 31–43, 2018.

[63] A. Parekh, I. W. Selesnick, D. M. Rapoport, and I. Ayappa, "Detection of K-complexes and sleep spindles (DETOKS) using sparse optimization," *Journal of Neuroscience Methods*, vol. 251, pp. 37–46, 2015.

[64] A. Parekh, I. Selesnick, R. Osorio, A. Varga, D. Rapoport, and I. Ayappa, "Multichannel sleep spindle detection using sparse low-rank optimization," *Journal of Neuroscience Methods*, vol. 288, pp. 1–16, 2017.

[65] T. Lajnef, C. O'Reilly, E. Combrisson, S. Chaibi, J.-B. Eichenlaub, P. M. Ruby, P.-E. Aguera, M. Samet, A. Kachouri, S. Frenette, J. Carrier, and K. Jerbi, "Meet Spinky: an open-source spindle and K-complex detection toolbox validated on the open-access Montreal Archive of Sleep Studies (MASS)," *Frontiers in Neuroinformatics*, vol. 11, p. 15, 2017.

[66] J. LaRocco, P. Franaszczuk, S. Kerick, and K. Robbins, "Spindler: A framework for parametric analysis and detection of spindles in EEG with application to sleep spindles," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066015, 2018.

[67] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[69] P. Chen, D. Chen, L. Zhang, Y. Tang, and X. Li, "Automated sleep spindle detection with mixed EEG features," *Biomedical Signal Processing and Control*, vol. 70, p. 103026, 2021.

[70] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[71] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, 2020.

[72] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[73] B. Welch, "The generalization of 'Student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[74] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, "Atlas,

rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep," *Sleep Medicine*, vol. 2, no. 6, pp. 537–553, 2001.

[75] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: Towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.

[76] C. L. Marcus, R. H. Moore, C. L. Rosen, B. Giordani, S. L. Garetz, H. G. Taylor, R. B. Mitchell, R. Amin, E. S. Katz, R. Arens, S. Paruthi, H. Muzumdar, D. Gozal, N. H. Thomas, J. Ware, D. Beebe, K. Snyder, L. Elden, R. C. Sprecher, P. Willging, D. Jones, J. P. Bent, T. Hoban, R. D. Chervin, S. S. Ellenberg, and S. Redline, "A randomized trial of adenotonsillectomy for childhood sleep apnea," *New England Journal of Medicine*, vol. 368, no. 25, pp. 2366–2376, 2013.

[77] C. L. Rosen, E. K. Larkin, H. L. Kirchner, J. L. Emancipator, S. F. Bivins, S. A. Surovec, R. J. Martin, and S. Redline, "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: Association with race and prematurity," *The Journal of Pediatrics*, vol. 142, no. 4, pp. 383–389, 2003.

[78] S. Redline, P. V. Tishler, T. D. Tosteson, J. Williamson, K. Kump, I. Browner, V. Ferrette, and P. Krejci, "The familial aggregation of obstructive sleep apnea," *American Journal of Respiratory and Critical Care Medicine*, vol. 151, no. 3, pp. 682–687, 1995.

[79] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[80] T. Blackwell, K. Yaffe, S. Ancoli-Israel, S. Redline, K. E. Ensrud, M. L. Stefanick, A. Laffan, and K. L. Stone, "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: The osteoporotic fractures in men sleep study," *Journal of the American Geriatrics Society*, vol. 59, no. 12, pp. 2217–2225, 2011.

[81] A. P. Spira, T. Blackwell, K. L. Stone, S. Redline, J. A. Cauley, S. Ancoli-Israel, and K. Yaffe, "Sleep-disordered breathing and cognition in older women," *Journal of the American Geriatrics Society*, vol. 56, no. 1, pp. 45–50, 2008.

[82] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[83] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[84] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[85] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *International Conference on Learning Representations*, 2014.

[86] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Artificial Neural Networks — ICANN'97*, 1997, pp. 583–588.

Anexos

Anexo A

Generación de PINK

El ruido rosado, por definición, tiene un espectro que decae como una ley de potencia, i.e., bf^{-a}. Por otro lado, la actividad cerebral sigue dicha tendencia solo de forma aproximada. Por ejemplo, el espectro promedio de las señales de MASS-SS2-Train durante la etapa N2 no es bien ajustado por una ley de potencia durante todo el rango de frecuencias. En cambio, el espectro observado en la Figura 3.1A sugiere una ley de potencia entre 1 Hz y el inicio de la banda sigma que luego cambia, con un exponente de mayor magnitud, a partir del final de la banda sigma. Por esta razón, se decide flexibilizar la definición del ruido rosado para imitar de mejor forma el comportamiento observado en MASS-SS2-Train, y así conseguir señales artificiales de mejor calidad.

En primer lugar, se determina el espectro deseado. Para ello, se extraen los segmentos en etapa N2 de cada señal en bruto (i.e., sin ningún preprocesamiento) de MASS-SS2-Train. En cada segmento se calcula la FFT en ventanas de 5 s sin traslape con una función de ventana de Hann. Se promedia la magnitud de todas las FFT obtenidas y se preserva el rango de frecuencias 0–100 Hz, generando una curva de referencia dada por n pares ordenados $\{(f_i, s_i^{\mathrm{MASS}})\}_{i=1}^{n}$. Después, se ajusta una ley de potencia $s^{\mathrm{Fit}}(f) = bf^{-a}$ al espectro restringido al intervalo $f \in [17,\ 100]$ Hz. Para preservar el comportamiento real previo a la banda sigma y para remover el pico de la banda sigma, el espectro deseado s^{Target} comienza igualando al espectro real s^{MASS} y transita continuamente al ajuste s^{Fit} en $f = 10$ Hz. Es decir,

$$s^{\mathrm{Target}}(f) = (1 - \lambda(f))s^{\mathrm{MASS}}(f) + \lambda(f)s^{\mathrm{Fit}}(f), \quad f \in \{f_i\}_{i=1}^{n}, \tag{A.1}$$

en donde $s^{\mathrm{MASS}}(f_i) = s_i^{\mathrm{MASS}}$ y

$$\lambda(f) = \begin{cases} 0 & \text{si } f \leq 9 \text{ Hz}, \\ \dfrac{f-9}{11-9} & \text{si } f \in (9, 11) \text{ Hz}, \\ 1 & \text{si } f \geq 11 \text{ Hz}. \end{cases} \tag{A.2}$$

De esta forma, la curva objetivo para el espectro (s^{Target}) es especificada discretamente por los pares ordenados $\{(f_i, s_i^{\mathrm{Target}})\}_{i=1}^{n}$. Para evaluar s^{Target} en cualquier frecuencia del rango 0–100 Hz, se usa interpolación lineal.

Con el espectro deseado ya determinado, se procede a la generación de las señales. Sea N el número de señales a simular y sea T el número de muestras requeridas para cada señal. Sea

$i \in \{0, \ldots, N-1\}$ el identificador de una señal a simular. Se genera un ruido Gaussiano estándar $\mathbf{x}^{\text{Gauss}} \in \mathbb{R}^T$ de T muestras independientes y semilla aleatoria fijada en i. Luego, se calcula $\mathbf{y} = \text{FFT}(\mathbf{x}^{\text{Gauss}})$, en donde cada y_j es la componente asociada a una frecuencia f_j. Como el espectro $\mathbf{y}$ tiene una magnitud con valor esperado constante en todas sus componentes (i.e., es *plano*), el espectro objetivo se alcanza haciendo

$$y_j \leftarrow y_j s^{\text{Target}}(f_j), \quad \forall j. \tag{A.3}$$

Después, haciendo $\mathbf{x} = \text{InverseFFT}(\mathbf{y})$ se obtiene una señal con el espectro deseado pero con otra escala. La señal obtenida se filtra en 0.1–35 Hz usando el mismo filtro descrito en el preprocesamiento (Sección 3.3.2), y se escala haciendo $\mathbf{x} \leftarrow c\mathbf{x}$, en donde $c \in \mathbb{R}$ es tal que la potencia promedio en la banda delta y theta (0–8 Hz) iguala al valor encontrado en el espectro s^{MASS}. Es decir,

$$c = \frac{\text{Mean}(s^{\text{MASS}}(\text{0–8 Hz}))}{\text{Mean}(s[\mathbf{x}](\text{0–8 Hz}))}. \tag{A.4}$$

Anexo B

Partición de datos en MASS-SS2

En la base de datos MASS-SS2, de 19 sujetos (un registro por sujeto), se seleccionan únicamente los 15 sujetos con anotaciones de ambos expertos. A partir de estos 15 sujetos, se seleccionan y apartan 4 sujetos de prueba. Dicho proceso deja 11 sujetos para ser utilizados en los conjuntos de entrenamiento y validación según el esquema de validación cruzada a utilizar.

Se seleccionan los 4 sujetos de prueba para que el conjunto de prueba sea representativo del resto de los sujetos. Para conseguir esto, se inspecciona visualmente la proyección de la FFT de las señales EEG. En primer lugar, se recolecta cada época en etapa N2 del canal C3-LE desde todos los sujetos. Para cada época x_i, se calcula su FFT. Se determinan cinco bandas de frecuencia B_j con frecuencias f entre f_j^L y f_j^U en base a las bandas utilizadas en la medicina del sueño [1], obteniendo

$$(f_j^L, f_j^U) \in \{(1,4),(4,8),(8,12),(12,15),(15,30)\} \text{ Hz}. \tag{B.1}$$

A continuación, se calcula la potencia promedio dentro de cada banda como

$$\bar{b}_{ij} = \log\left(\frac{1}{|B_j|}\sum_{f\in B_j}|\text{FFT}[x_i](f)|\right). \tag{B.2}$$

Estas características son estandarizadas y proyectadas en un plano 2D usando Kernel PCA [86] con kernel RBF. El coeficiente γ del kernel se selecciona a partir de $\gamma \in \{0{,}01, 0{,}1, 1, 10\}$ para

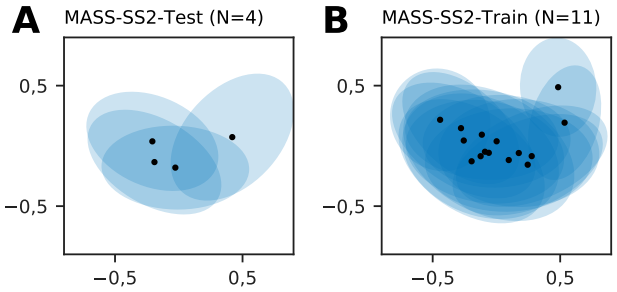

Figura B.1: Proyección de Kernel PCA de los sujetos de MASS-SS2 basado en características espectrales. (**a**) Distribuciones Gaussianas ajustadas de los sujetos de prueba 2, 6, 12, y 13. (**b**) Distribuciones Gaussianas ajustadas del resto de los sujetos.

obtener una buena visualización, resultando en $\gamma = 0{,}1$. En este espacio proyectado, se ajusta una distribución Gaussiana 2D a cada conjunto de épocas que pertenece al mismo sujeto. Después, se generan conjuntos de prueba aleatorios y se juzga visualmente si la distribución combinada del conjunto de prueba es representativa de la distribución combinada del resto de los sujetos. Finalmente, se obtiene un conjunto de prueba representativo con los sujetos 2, 6, 12 y 13 como se observa en la Figura B.1, en donde las distribuciones Gaussianas se muestran con su media y sus elipses de 95 % de confianza. Fijando esta separación se forman los subconjuntos MASS-SS2-Test y MASS-SS2-Train para evaluación y diseño, respectivamente.

Anexo C

Amplitud máxima por bandas

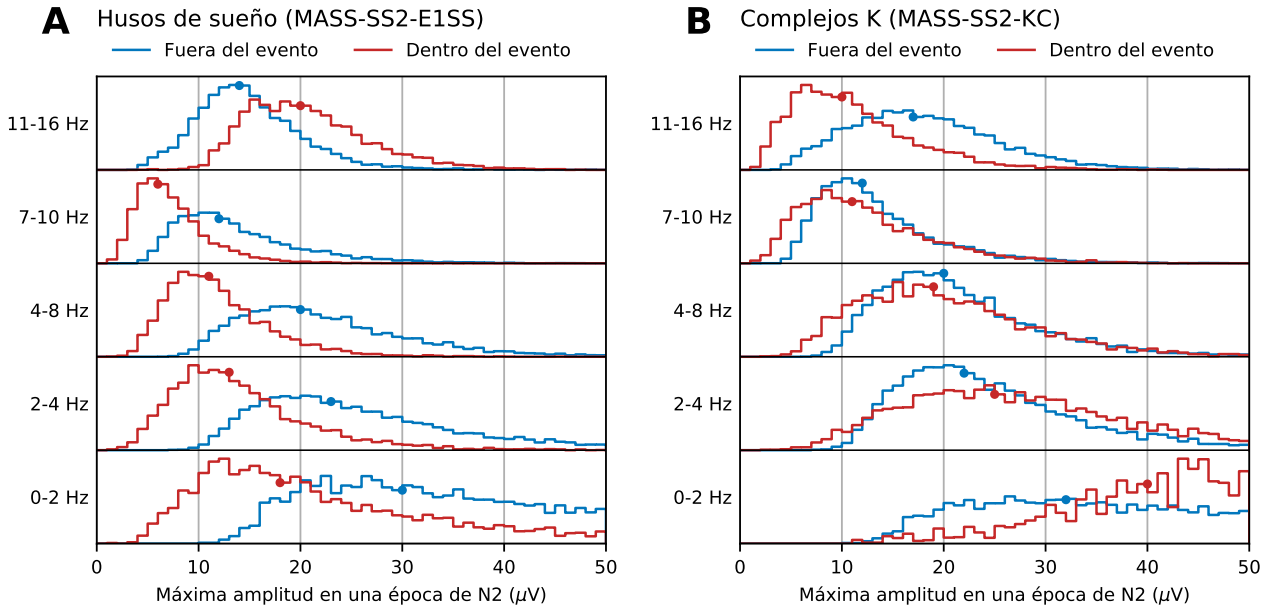

Figura C.1: Histogramas de amplitud máxima en señales de MASS-SS2-Train, mencionados en la Sección 3.5.2, tanto para husos de sueño (**A**) como para complejos K (**B**). En cada histograma se destaca la mediana con un marcador circular.

Tabla C.1: Mediana de las distribuciones de amplitud máxima de la Figura C.1, útiles para fijar la amplitud máxima de *ondas aleatorias*. «SS» indica huso de sueño y «KC» indica complejo K.

Banda	Dentro de SS (μV)	Fuera de SS (μV)	Dentro de KC (μV)	Fuera de KC (μV)
0–2 Hz	18	30	40	32
2–4 Hz	13	23	25	22
4–8 Hz	11	20	19	20
7–10 Hz	6	12	11	12
11–16 Hz	20	14	10	17

Anexo D

Ejemplos de casos de detección

En este anexo se muestran segmentos de señal que ilustran cuatro tipos de casos de detección en el conjunto de prueba: falsos negativos, falsos positivos, verdaderos positivos con IoU bajo, y verdaderos positivos con IoU alto. En este contexto, los casos de IoU bajo se definen como aquellos de IoU no nulos dentro del primer cuartil (i.e., dentro del peor 25 %), mientras que los casos de IoU alto se definen como aquellos de IoU no nulos dentro del cuarto cuartil (i.e., dentro del mejor 25 %). Por simplicidad, se muestran 4 ejemplos de cada categoría, seleccionados aleatoriamente, con un contexto de 3 s de señal, y para una sola instancia de REDv2-Time. Casos similares se obtienen para REDv2-CWT.

En la Figura D.1 se ilustran casos para la detección de husos de sueño, a partir de la base de datos MASS-MODA. Además, en la Figura D.2 se ilustran casos para la detección de complejos K, a partir de la base de datos MASS-SS2-KC.

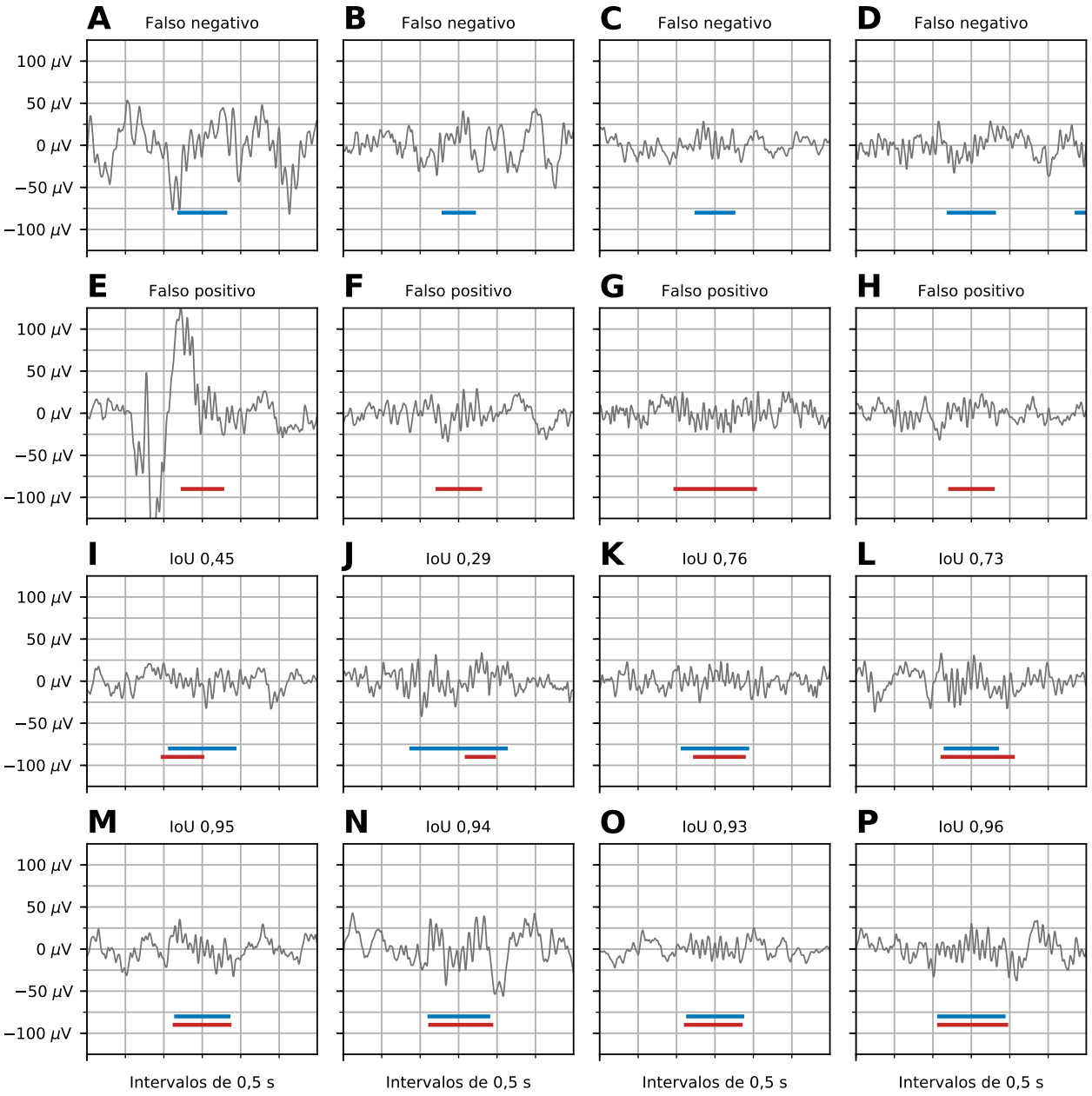

Figura D.1: Ejemplos de casos para la detección de husos de sueño. Detecciones de REDv2-Time en la base de datos MASS-MODA. En cada panel se indica la anotación experta por una línea horizontal azul, y la detección por una línea horizontal roja. (**A**-**D**) Ejemplos de falsos negativos. (**E**-**F**) Ejemplos de falsos positivos. (**I**-**L**) Ejemplos de traslape con IoU bajo (peor 25 %). (**M**-**P**) Ejemplos de traslape con IoU alto (mejor 25 %).

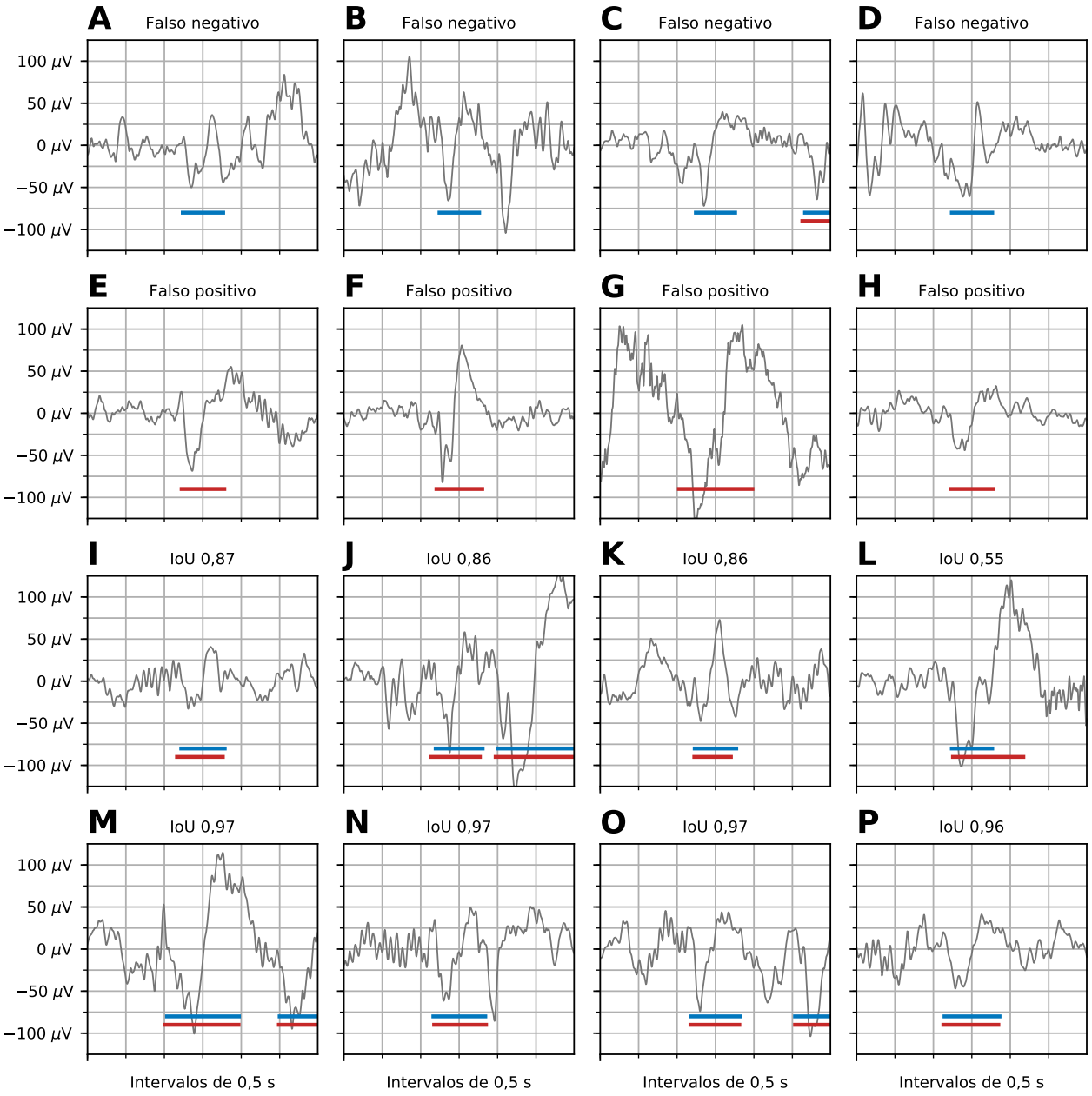

Figura D.2: Ejemplos de casos para la detección de complejos K. Detecciones de REDv2-Time en la base de datos MASS-SS2-KC. En cada panel se indica la anotación experta por una línea horizontal azul, y la detección por una línea horizontal roja. (**A**-**D**) Ejemplos de falsos negativos. (**E**-**F**) Ejemplos de falsos positivos. (**I**-**L**) Ejemplos de traslape con IoU bajo (peor 25 %). (**M**-**P**) Ejemplos de traslape con IoU alto (mejor 25 %).

Anexo E

Resultados complementarios

En este anexo se muestran tablas que permiten consultar mediciones exactas de algunos de los resultados principales mostrados y analizados en el cuerpo.

Tabla E.1: Desempeño del ajuste de duración por evento. Se muestra el desempeño para husos de sueño (MASS-MODA) y complejos K (MASS-SS2-KC). Error definido como valor predicho menos real.

Datos	Detector	R^2	$\mathbb{E}(\text{error})$	$\sqrt{\text{Var}(\text{error})}$
MASS-MODA	**REDv2-Time**	0,60	$-0{,}0057$ s	0,24 s
	REDv2-CWT	0,60	0,0010 s	0,24 s
	DOSED [8]	0,41	0,1602 s	0,30 s
	A7 [59]	0,30	0,0012 s	0,38 s
MASS-SS2-KC	**REDv2-Time**	0,73	$-0{,}0016$ s	0,09 s
	REDv2-CWT	0,72	$-0{,}0001$ s	0,09 s
	DOSED [8]	0,20	0,1560 s	0,21 s
	Spinky [65]	0,00	0,6373 s	0,17 s

Tabla E.2: Desempeño del ajuste de parámetros por sujeto. Se muestra el desempeño para husos de sueño (MASS-MODA) y complejos K (MASS-SS2-KC). Error definido como valor predicho menos real.

Datos	Parámetro	Detector	R^2	$\mathbb{E}(\text{error})$	$\sqrt{\text{Var}(\text{error})}$
MASS-MODA	Duración (s)	**REDv2-Time**	0,62	$-0{,}017$	0,06
		REDv2-CWT	0,56	$-0{,}006$	0,07
		DOSED [8]	0,47	0,179	0,08
		A7 [59]	0,35	$-0{,}020$	0,15
	Densidad (epm)	**REDv2-Time**	0,94	0,136	0,55
		REDv2-CWT	0,94	0,095	0,55
		DOSED [8]	0,90	$-0{,}154$	0,60
		A7 [59]	0,88	0,177	0,74
	Amplitud PP (μV)	**REDv2-Time**	0,99	0,815	1,00
		REDv2-CWT	0,98	0,713	1,08
		DOSED [8]	0,98	1,542	1,44
		A7 [59]	0,97	0,496	1,90
	Frecuencia (Hz)	**REDv2-Time**	0,95	0,063	0,15
		REDv2-CWT	0,95	0,062	0,14
		DOSED [8]	0,93	0,071	0,18
		A7 [59]	0,77	0,087	0,32
MASS-SS2-KC	Duración (s)	**REDv2-Time**	0,80	$-0{,}003$	0,02
		REDv2-CWT	0,79	$-0{,}002$	0,02
		DOSED [8]	0,64	0,140	0,02
		Spinky [65]	0,00	0,671	0,04
	Densidad (epm)	**REDv2-Time**	0,91	0,112	0,35
		REDv2-CWT	0,92	0,091	0,32
		DOSED [8]	0,91	0,106	0,36
		Spinky [65]	0,82	$-0{,}179$	0,43
	Amplitud PP (μV)	**REDv2-Time**	0,93	1,157	5,99
		REDv2-CWT	0,93	1,103	5,84
		DOSED [8]	0,91	3,179	7,04
		Spinky [65]	0,88	9,642	7,89

Tabla E.3: Desempeño de la detección de husos de sueño (MASS-MODA) cuando se restringen los eventos a un subconjunto de un parámetro. Métricas calculadas por micro-promedio.

		F1-score (%)			
Parámetro	Subconjunto	**REDv2-Time**	**REDv2-CWT**	DOSED [8]	A7 [59]
Duración (s)	$(-\infty$, 0,6]	$64{,}8 \pm 2{,}6$	$64{,}6 \pm 3{,}2$	$66{,}3 \pm 3{,}4$	$50{,}0 \pm 3{,}4$
	(0,6, 0,9]	$85{,}6 \pm 1{,}7$	$85{,}5 \pm 1{,}7$	$78{,}2 \pm 2{,}7$	$80{,}5 \pm 1{,}7$
	(0,9, $+\infty$)	$92{,}5 \pm 1{,}0$	$91{,}9 \pm 1{,}2$	$83{,}9 \pm 1{,}6$	$87{,}8 \pm 2{,}0$
Amplitud PP (μV)	$(-\infty$, 30]	$70{,}6 \pm 2{,}6$	$70{,}2 \pm 2{,}7$	$63{,}6 \pm 2{,}9$	$58{,}6 \pm 1{,}7$
	(30, 40]	$83{,}3 \pm 1{,}5$	$83{,}1 \pm 1{,}8$	$79{,}8 \pm 2{,}2$	$74{,}6 \pm 2{,}9$
	(40, $+\infty$)	$89{,}4 \pm 2{,}0$	$89{,}2 \pm 1{,}9$	$86{,}5 \pm 2{,}5$	$84{,}0 \pm 2{,}3$
Frecuencia (Hz)	$(-\infty$, 12,8]	$77{,}9 \pm 2{,}6$	$77{,}6 \pm 2{,}5$	$72{,}6 \pm 4{,}0$	$67{,}0 \pm 4{,}1$
	(12,8, 13,5]	$84{,}6 \pm 1{,}6$	$84{,}7 \pm 1{,}6$	$81{,}7 \pm 1{,}1$	$78{,}7 \pm 1{,}9$
	(13,5, $+\infty$)	$83{,}0 \pm 1{,}2$	$82{,}6 \pm 1{,}5$	$78{,}6 \pm 1{,}5$	$75{,}3 \pm 2{,}6$
Edad	Jóvenes	$83{,}5 \pm 1{,}5$	$83{,}4 \pm 1{,}4$	$79{,}8 \pm 1{,}6$	$75{,}5 \pm 1{,}5$
	Viejos	$78{,}5 \pm 2{,}4$	$77{,}9 \pm 2{,}3$	$73{,}2 \pm 2{,}8$	$69{,}1 \pm 3{,}8$

Tabla E.4: Desempeño de la detección de complejos K (MASS-SS2-KC) cuando se restringen los eventos a un subconjunto de un parámetro. Métricas calculadas por micro-promedio.

		F1-score (%)			
Parámetro	Subconjunto	**REDv2-Time**	**REDv2-CWT**	DOSED [8]	Spinky [65]
Duración (s)	$(-\infty$, 0,65]	$73{,}6 \pm 1{,}8$	$73{,}7 \pm 2{,}0$	$69{,}1 \pm 3{,}1$	No aplica
	(0,65, 0,8]	$87{,}4 \pm 1{,}2$	$87{,}5 \pm 1{,}2$	$82{,}4 \pm 1{,}8$	No aplica
	(0,8, $+\infty$)	$93{,}9 \pm 1{,}2$	$93{,}7 \pm 1{,}3$	$84{,}7 \pm 3{,}8$	No aplica
Amplitud PP (μV)	$(-\infty$, 110]	$68{,}8 \pm 3{,}0$	$68{,}8 \pm 2{,}8$	$57{,}6 \pm 3{,}0$	$37{,}0 \pm 2{,}5$
	(110, 160]	$87{,}3 \pm 1{,}2$	$87{,}5 \pm 1{,}0$	$82{,}8 \pm 2{,}0$	$63{,}9 \pm 2{,}3$
	(160, $+\infty$)	$95{,}4 \pm 1{,}3$	$95{,}5 \pm 1{,}2$	$93{,}1 \pm 2{,}5$	$87{,}9 \pm 2{,}6$

Tabla E.5: Desempeño de la detección de husos de sueño cuando el modelo se transfiere directamente de una base de datos a otra.

Datos de	Datos de	F1-score (%)			
entrenamiento	evaluación	**REDv2-Time**	**REDv2-CWT**	DOSED [8]	A7 [59]
MASS-SS2-E1SS	**MASS-SS2-E1SS**	$80{,}8 \pm 2{,}1$	$80{,}8 \pm 2{,}0$	$76{,}8 \pm 2{,}9$	$73{,}0 \pm 3{,}4$
	MASS-SS2-E2SS	$59{,}1 \pm 6{,}3$	$59{,}3 \pm 6{,}5$	$58{,}5 \pm 7{,}1$	$59{,}1 \pm 6{,}3$
	MASS-MODA	$53{,}6 \pm 4{,}3$	$53{,}4 \pm 4{,}5$	$48{,}9 \pm 4{,}3$	$64{,}1 \pm 1{,}8$
	INTA-UCH	$56{,}5 \pm 9{,}2$	$56{,}4 \pm 8{,}6$	$50{,}6 \pm 18{,}5$	$54{,}0 \pm 5{,}8$
MASS-SS2-E2SS	MASS-SS2-E1SS	$57{,}5 \pm 5{,}3$	$57{,}1 \pm 5{,}7$	$55{,}7 \pm 7{,}3$	$50{,}3 \pm 6{,}4$
	MASS-SS2-E2SS	$86{,}1 \pm 2{,}0$	$86{,}1 \pm 2{,}1$	$82{,}5 \pm 2{,}5$	$74{,}9 \pm 2{,}8$
	MASS-MODA	$67{,}7 \pm 2{,}5$	$67{,}4 \pm 2{,}5$	$68{,}3 \pm 2{,}1$	$64{,}6 \pm 2{,}5$
	INTA-UCH	$65{,}5 \pm 5{,}9$	$66{,}0 \pm 5{,}8$	$51{,}6 \pm 12{,}4$	$76{,}2 \pm 3{,}8$
MASS-MODA	MASS-SS2-E1SS	$61{,}1 \pm 6{,}0$	$61{,}7 \pm 6{,}1$	$63{,}6 \pm 7{,}4$	$61{,}5 \pm 5{,}2$
	MASS-SS2-E2SS	$73{,}2 \pm 4{,}9$	$73{,}4 \pm 4{,}7$	$73{,}1 \pm 4{,}4$	$71{,}2 \pm 5{,}5$
	MASS-MODA	$81{,}8 \pm 1{,}4$	$81{,}5 \pm 1{,}3$	$77{,}5 \pm 1{,}7$	$73{,}3 \pm 1{,}9$
	INTA-UCH	$73{,}3 \pm 4{,}6$	$72{,}7 \pm 5{,}1$	$63{,}4 \pm 7{,}0$	$73{,}5 \pm 3{,}9$
INTA-UCH	MASS-SS2-E1SS	$63{,}7 \pm 6{,}8$	$63{,}7 \pm 6{,}2$	$50{,}1 \pm 7{,}8$	$63{,}5 \pm 7{,}3$
	MASS-SS2-E2SS	$72{,}9 \pm 5{,}9$	$72{,}7 \pm 6{,}1$	$74{,}3 \pm 4{,}0$	$69{,}5 \pm 7{,}5$
	MASS-MODA	$74{,}4 \pm 2{,}4$	$74{,}4 \pm 2{,}6$	$66{,}0 \pm 3{,}4$	$69{,}8 \pm 2{,}8$
	INTA-UCH	$83{,}9 \pm 4{,}0$	$83{,}5 \pm 4{,}3$	$78{,}1 \pm 6{,}2$	$78{,}4 \pm 4{,}6$

Tabla E.6: Acuerdo entre modelos propuestos.

Datos	Comparación	F1-score (%)	mIoU (%)
MASS-SS2-E1SS	REDv2-Time vs REDv2-Time	$95{,}9 \pm 2{,}0$	$96{,}4 \pm 1{,}5$
	REDv2-CWT vs REDv2-CWT	$95{,}2 \pm 2{,}4$	$96{,}1 \pm 1{,}9$
	REDv2-Time vs REDv2-CWT	$93{,}5 \pm 1{,}1$	$93{,}9 \pm 0{,}5$
MASS-SS2-E2SS	REDv2-Time vs REDv2-Time	$96{,}7 \pm 1{,}7$	$95{,}3 \pm 2{,}2$
	REDv2-CWT vs REDv2-CWT	$96{,}5 \pm 1{,}7$	$95{,}1 \pm 2{,}3$
	REDv2-Time vs REDv2-CWT	$95{,}3 \pm 0{,}4$	$92{,}5 \pm 0{,}6$
MASS-SS2-KC	REDv2-Time vs REDv2-Time	$96{,}0 \pm 2{,}0$	$97{,}9 \pm 0{,}9$
	REDv2-CWT vs REDv2-CWT	$95{,}6 \pm 2{,}1$	$97{,}7 \pm 0{,}9$
	REDv2-Time vs REDv2-CWT	$93{,}1 \pm 1{,}1$	$96{,}1 \pm 0{,}5$
MASS-MODA	REDv2-Time vs REDv2-Time	$96{,}1 \pm 0{,}7$	$95{,}7 \pm 0{,}6$
	REDv2-CWT vs REDv2-CWT	$95{,}7 \pm 0{,}8$	$95{,}2 \pm 0{,}7$
	REDv2-Time vs REDv2-CWT	$93{,}3 \pm 0{,}7$	$92{,}1 \pm 0{,}4$
INTA-UCH	REDv2-Time vs REDv2-Time	$97{,}3 \pm 2{,}4$	$95{,}6 \pm 3{,}0$
	REDv2-CWT vs REDv2-CWT	$96{,}9 \pm 2{,}6$	$95{,}0 \pm 3{,}3$
	REDv2-Time vs REDv2-CWT	$95{,}6 \pm 0{,}9$	$91{,}6 \pm 0{,}9$

Tabla E.7: Desempeño de la detección de diferentes ensambles de dos modelos REDv2.

Datos	Detector	F1-score (%)	mIoU (%)
MASS-SS2-E1SS	REDv2-Time	$80{,}8 \pm 2{,}1$	$84{,}8 \pm 1{,}2$
	REDv2-CWT	$80{,}8 \pm 2{,}0$	$84{,}5 \pm 1{,}2$
	AND(REDv2-Time, REDv2-Time)	$80{,}8 \pm 1{,}9$	$84{,}9 \pm 1{,}2$
	AND(REDv2-CWT, REDv2-CWT)	$80{,}9 \pm 1{,}8$	$84{,}6 \pm 1{,}1$
	AND(REDv2-Time, REDv2-CWT)	$80{,}8 \pm 1{,}9$	$84{,}8 \pm 1{,}1$
	AVG(REDv2-Time, REDv2-Time)	$81{,}0 \pm 2{,}1$	$84{,}8 \pm 1{,}2$
	AVG(REDv2-CWT, REDv2-CWT)	$81{,}1 \pm 2{,}0$	$84{,}6 \pm 1{,}2$
	AVG(REDv2-Time, REDv2-CWT)	$81{,}0 \pm 2{,}1$	$84{,}8 \pm 1{,}2$
MASS-SS2-E2SS	REDv2-Time	$86{,}1 \pm 2{,}0$	$78{,}7 \pm 1{,}1$
	REDv2-CWT	$86{,}1 \pm 2{,}1$	$78{,}8 \pm 1{,}0$
	AND(REDv2-Time, REDv2-Time)	$86{,}1 \pm 2{,}0$	$78{,}9 \pm 1{,}1$
	AND(REDv2-CWT, REDv2-CWT)	$86{,}0 \pm 2{,}0$	$79{,}0 \pm 1{,}0$
	AND(REDv2-Time, REDv2-CWT)	$86{,}0 \pm 2{,}1$	$78{,}9 \pm 1{,}0$
	AVG(REDv2-Time, REDv2-Time)	$86{,}3 \pm 2{,}0$	$78{,}9 \pm 1{,}0$
	AVG(REDv2-CWT, REDv2-CWT)	$86{,}3 \pm 2{,}0$	$78{,}9 \pm 1{,}0$
	AVG(REDv2-Time, REDv2-CWT)	$86{,}4 \pm 2{,}1$	$78{,}9 \pm 1{,}1$
MASS-SS2-KC	REDv2-Time	$83{,}7 \pm 1{,}5$	$90{,}6 \pm 0{,}6$
	REDv2-CWT	$83{,}8 \pm 1{,}4$	$90{,}4 \pm 0{,}5$
	AND(REDv2-Time, REDv2-Time)	$83{,}6 \pm 1{,}5$	$90{,}6 \pm 0{,}5$
	AND(REDv2-CWT, REDv2-CWT)	$83{,}6 \pm 1{,}5$	$90{,}4 \pm 0{,}5$
	AND(REDv2-Time, REDv2-CWT)	$83{,}6 \pm 1{,}5$	$90{,}6 \pm 0{,}5$
	AVG(REDv2-Time, REDv2-Time)	$83{,}9 \pm 1{,}5$	$90{,}6 \pm 0{,}6$
	AVG(REDv2-CWT, REDv2-CWT)	$84{,}1 \pm 1{,}4$	$90{,}5 \pm 0{,}6$
	AVG(REDv2-Time, REDv2-CWT)	$84{,}1 \pm 1{,}3$	$90{,}7 \pm 0{,}6$
MASS-MODA	REDv2-Time	$81{,}8 \pm 1{,}4$	$83{,}4 \pm 0{,}5$
	REDv2-CWT	$81{,}5 \pm 1{,}3$	$83{,}2 \pm 0{,}5$
	AND(REDv2-Time, REDv2-Time)	$81{,}9 \pm 1{,}4$	$83{,}4 \pm 0{,}6$
	AND(REDv2-CWT, REDv2-CWT)	$81{,}5 \pm 1{,}1$	$83{,}3 \pm 0{,}5$
	AND(REDv2-Time, REDv2-CWT)	$81{,}7 \pm 1{,}2$	$83{,}4 \pm 0{,}6$
	AVG(REDv2-Time, REDv2-Time)	$82{,}0 \pm 1{,}4$	$83{,}5 \pm 0{,}5$
	AVG(REDv2-CWT, REDv2-CWT)	$81{,}7 \pm 1{,}2$	$83{,}5 \pm 0{,}5$
	AVG(REDv2-Time, REDv2-CWT)	$82{,}0 \pm 1{,}4$	$83{,}6 \pm 0{,}5$
INTA-UCH	REDv2-Time	$83{,}9 \pm 4{,}0$	$75{,}6 \pm 2{,}5$
	REDv2-CWT	$83{,}5 \pm 4{,}3$	$75{,}5 \pm 2{,}4$
	AND(REDv2-Time, REDv2-Time)	$83{,}9 \pm 3{,}9$	$75{,}8 \pm 2{,}5$
	AND(REDv2-CWT, REDv2-CWT)	$83{,}5 \pm 4{,}3$	$75{,}7 \pm 2{,}3$
	AND(REDv2-Time, REDv2-CWT)	$83{,}5 \pm 4{,}2$	$75{,}8 \pm 2{,}4$
	AVG(REDv2-Time, REDv2-Time)	$83{,}9 \pm 4{,}0$	$75{,}7 \pm 2{,}4$
	AVG(REDv2-CWT, REDv2-CWT)	$83{,}8 \pm 4{,}1$	$75{,}5 \pm 2{,}4$
	AVG(REDv2-Time, REDv2-CWT)	$83{,}9 \pm 4{,}2$	$75{,}7 \pm 2{,}5$

Tabla E.8: Desempeño de la detección de husos de sueño (MASS-MODA) y complejos K (MASS-SS2-KC) de los modelos propuestos cuando la señal de entrada es perturbada.

		F1-score (%)			
		MASS-MODA		MASS-SS2-KC	
Perturbación	Valor	REDv2-Time	REDv2-CWT	REDv2-Time	REDv2-CWT
Factor de escala	0,5	57,9 ± 4,9	60,2 ± 3,4	41,9 ± 5,3	41,4 ± 5,2
	0,6	69,2 ± 3,7	70,1 ± 2,3	54,4 ± 5,7	53,8 ± 5,7
	0,7	75,7 ± 2,5	76,0 ± 1,8	65,6 ± 5,0	65,1 ± 5,2
	0,8	79,6 ± 1,8	79,5 ± 1,4	75,2 ± 3,7	74,8 ± 3,6
	0,9	81,1 ± 1,4	81,1 ± 1,1	81,6 ± 1,9	81,6 ± 1,9
	1,0 (original)	81,8 ± 1,4	81,5 ± 1,3	83,7 ± 1,5	83,8 ± 1,4
	1,1	81,6 ± 1,4	81,3 ± 1,4	82,3 ± 2,5	82,3 ± 2,3
	1,2	81,1 ± 1,4	80,7 ± 1,5	78,7 ± 4,0	78,8 ± 3,5
	1,3	80,4 ± 1,4	80,2 ± 1,7	74,4 ± 5,0	74,6 ± 4,4
	1,4	79,7 ± 1,5	79,6 ± 1,4	70,2 ± 5,6	70,4 ± 5,0
	1,5	79,1 ± 1,7	78,9 ± 1,5	66,3 ± 5,9	66,6 ± 5,2
Inversión	Amplitud	81,2 ± 1,5	80,8 ± 1,5	41,0 ± 4,1	37,1 ± 3,9
	Tiempo	81,3 ± 1,6	81,0 ± 1,5	48,6 ± 4,2	45,9 ± 4,6
Filtro rechaza-banda	0–2 Hz	80,4 ± 1,4	80,4 ± 1,5	3,5 ± 1,6	3,6 ± 1,3
	2–4 Hz	76,3 ± 1,6	76,1 ± 1,6	70,6 ± 3,3	70,2 ± 3,6
	4–8 Hz	69,2 ± 2,4	69,5 ± 2,4	79,5 ± 2,2	79,6 ± 2,0
	8–11 Hz	78,5 ± 2,1	78,4 ± 2,0	82,5 ± 1,8	82,6 ± 1,6
	10–16 Hz	0,2 ± 0,3	0,3 ± 0,2	82,9 ± 1,8	83,1 ± 1,6
	16–30 Hz	80,6 ± 1,4	80,2 ± 1,6	82,8 ± 2,2	82,9 ± 2,0

Tabla E.9: Desempeño de la detección en MASS-MODA cuando los conjuntos de entrenamiento y validación son una fracción del total disponible, y se usan diferentes inicializaciones del modelo.

Fracción de	Datos de	F1-score (%)	
señales (%)	pre-entrenamiento	REDv2-Time	REDv2-CWT
0	MASS-SS2-E1SS	53,6 ± 4,3	53,4 ± 4,5
	CAP-A7	75,0 ± 1,5	74,8 ± 1,7
10	Ninguno	77,9 ± 1,5	77,4 ± 1,9
	MASS-SS2-E1SS	79,5 ± 2,4	80,1 ± 1,3
	CAP-A7	78,8 ± 1,5	78,3 ± 1,7
20	Ninguno	78,9 ± 1,6	78,8 ± 1,3
	MASS-SS2-E1SS	79,6 ± 2,0	80,6 ± 1,3
	CAP-A7	80,4 ± 1,1	79,5 ± 1,4
40	Ninguno	79,5 ± 2,6	79,4 ± 2,1
	MASS-SS2-E1SS	81,0 ± 1,9	81,1 ± 1,5
	CAP-A7	81,1 ± 1,3	80,6 ± 1,6
70	Ninguno	81,0 ± 1,5	81,0 ± 1,5
	MASS-SS2-E1SS	81,9 ± 1,3	81,4 ± 1,5
	CAP-A7	81,7 ± 1,1	81,6 ± 1,2
100	Ninguno	81,8 ± 1,4	81,5 ± 1,3
	MASS-SS2-E1SS	81,9 ± 1,3	82,0 ± 1,2
	CAP-A7	81,9 ± 1,2	81,9 ± 1,1

Tabla E.10: Desempeño de la detección por sujeto cuando se personaliza el umbral de salida usando una porción de señal N2 de cada sujeto.

		F1-score (%)	
Datos	Minutos de ajuste	REDv2-Time	REDv2-CWT
MASS-SS2-E1SS	0 (original)	$80{,}8 \pm 3{,}5$	$80{,}8 \pm 3{,}2$
	10	$81{,}8 \pm 3{,}1$	$82{,}1 \pm 2{,}9$
	20	$81{,}9 \pm 3{,}2$	$82{,}2 \pm 2{,}8$
	30	$82{,}1 \pm 3{,}0$	$82{,}2 \pm 2{,}8$
	Oráculo	$82{,}7 \pm 3{,}0$	$82{,}8 \pm 2{,}8$
MASS-SS2-E2SS	0 (original)	$86{,}1 \pm 3{,}3$	$86{,}1 \pm 3{,}4$
	10	$86{,}8 \pm 2{,}8$	$86{,}8 \pm 2{,}8$
	20	$86{,}7 \pm 2{,}8$	$86{,}7 \pm 2{,}7$
	30	$86{,}8 \pm 2{,}8$	$86{,}9 \pm 2{,}9$
	Oráculo	$87{,}1 \pm 2{,}9$	$87{,}1 \pm 2{,}9$
MASS-MODA	0 (original)	$79{,}8 \pm 7{,}9$	$79{,}2 \pm 7{,}5$
	5	$79{,}0 \pm 8{,}5$	$79{,}0 \pm 7{,}8$
	7.5	$79{,}1 \pm 8{,}0$	$79{,}1 \pm 8{,}0$
	10	$79{,}2 \pm 7{,}9$	$79{,}4 \pm 7{,}9$
	Oráculo	$81{,}4 \pm 7{,}1$	$82{,}0 \pm 6{,}4$
INTA-UCH	0 (original)	$83{,}5 \pm 5{,}0$	$83{,}2 \pm 5{,}4$
	10	$83{,}3 \pm 4{,}9$	$83{,}0 \pm 4{,}9$
	20	$83{,}7 \pm 4{,}8$	$83{,}5 \pm 4{,}9$
	30	$83{,}6 \pm 4{,}7$	$83{,}4 \pm 4{,}9$
	Oráculo	$84{,}1 \pm 4{,}6$	$83{,}7 \pm 4{,}8$