



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Aprendizaje Automático aplicado al tratamiento de Epilepsia Refractaria

Tesis de Licenciatura en Ciencias de la Computación

Tomás Ariel Pastore

Director: Diego Fernández Slezak

Buenos Aires, 2020

RESUMEN

En casos de *epilepsia refractaria* a la medicación, la alternativa terapéutica más prometedora consiste en la *resección quirúrgica* de la zona epileptógena (EZ). El marcador "gold standard" para la identificación de la EZ es conocido como *seizure-onset zone* (SOZ), es decir, la zona en donde se originan las crisis. Las oscilaciones de alta frecuencia (HFOs) son eventos capturables mediante electroencefalografía (EEG), su utilidad como biomarcadores de epilepsia se encuentra bajo estudio. Investigaciones previas indican que sectores epileptógenos registran una mayor tasa de eventos por minuto (HFO rate). Sin embargo, existen limitaciones que deben ser resueltas para poder establecer a las HFOs como biomarcadores estándar en el contexto clínico. Una de las principales reside en que estos eventos también pueden ser observados en regiones sanas del cerebro, interviniendo en la cognición normal. Más aún, no se conoce un método para distinguir completamente las HFOs fisiológicos de los patológicos (pHFOs). Esta tesis tuvo como objetivo evaluar si, aplicando técnicas de *aprendizaje automático*, es posible distinguir HFOs características de la SOZ mejorando su capacidad como predictores de ésta.

Se analizó una base de datos de estéreo-electroencefalogramas (SEEGs) realizados a 91 pacientes con epilepsia refractaria. Se utilizaron propiedades morfológicas de las HFOs (duración, *contenido espectral* y *densidad de potencia*), el acoplamiento de amplitud de fase (PAC), y las coordenadas neuroanatómicas de su electrodo (MNI space), para predecir una probabilidad de que un nuevo evento pertenezca o no a la SOZ. Se implementaron clasificadores independientes para regiones neuroanatómicas específicas y distintos subtipos de HFOs. Luego, se aplicaron los modelos obtenidos para retener (filtrar) a nivel electrodo y Materiales únicamente los eventos predichos como pertenecientes a la SOZ, comparando la *performance* del HFO rate como predictor de la SOZ antes y después del filtrado. A partir de curvas ROC se evaluó el impacto de filtros de distinta granularidad. A su vez, se implementaron métricas combinando subtipos de HFOs. Utilizando la SOZ como marcador subrogado de la EZ, la hipótesis fue que un *filtro* óptimo conseguiría aumentar la diferencia entre HFO rates dentro y fuera de la SOZ.

Los FRonO (Fast Ripples on Oscillations) fueron únicamente relevantes como biomarcadores de la SOZ en el lóbulo temporal mesial. La *performance* observada por los modelos de aprendizaje automático para clasificar HFOs varió significativamente en función del subtipo de HFO y la región neuroanatómica. Por ejemplo: **RESULTADOS PARA DISTINTOS SUBTIPOS Y REGIONES.** En cuanto al impacto de la clasificación en el HFO rate como predictor de la SOZ, no se observó una mejora significativa, refutando la hipótesis. **RESULTADOS DISTINTAS REGIONES Y MODELOS.**

Se concluyó que en las zonas estudiadas (ACLARAR ZONAS) el enfoque de un *filtro* de pHFOs resulta ineficaz para mejorar la utilidad de los HFOs, ya que las features provistas por la base de datos no permiten distinguir lo suficiente entre la NSOZ y la SOZ.

Palabras clave: *epilepsia refractaria, resección quirúrgica, EZ, SOZ, HFOs, HFO rate, pHFOs, aprendizaje automático, contenido espectral, densidad de potencia, espacio MNI, PAC, ROC.*

Planteo

Objetivo

Hipótesis

Resultados

ABSTRACT

Para comentar con colores

D.F.S comment T.P comment

foo

Notas: Publicación de frontiers, traducir abs

TODO: Agregar la bibliografia

Índice general

1	Introducción	1
1.1.	Epilepsia	1
1.2.	Spikes, HFOs y HFO rate	4
1.3.	Literatura y planteo del problema	5
1.4.	Objetivos e Hipótesis	6
1.5.	Electroencefalografía	6
1.6.	Anatomía y Fisiología Cerebral	8
1.7.	Diagrama del trabajo	8
2	Materiales y Métodos	9
2.1.	Corpus	9
2.2.	Software	10
2.3.	Hardware	12
2.4.	Estadística	12
2.4.1.	Varianza y Desvío Estándar	12
2.4.2.	Kolmogorov-Smirnov	12
2.4.3.	Wilcoxon signed-rank	13
2.4.4.	Mann-Whitney U test	13
2.4.5.	P-Valor	13
2.4.6.	Error Cuadrático Medio	13
2.4.7.	R cuadrado	13
2.4.8.	Correlación de Pearson	13
2.4.9.	Correlación de Spearman	13
2.5.	Aprendizaje Automático	13
2.5.1.	Features y label	13
2.5.2.	Sesgo y Varianza de un modelo	13
2.5.3.	Validación cruzada	13
2.5.4.	Matriz de Confusión	13
2.5.5.	Sensibilidad y Especificidad	13
2.5.6.	Curva de Característica Operativa del Receptor (ROC)	13
2.5.7.	Curva de Precisión-Recall (PR)	13
2.5.8.	Regularización	13
2.5.9.	Hiperparámetros	13
2.5.10.	Bayes	13
2.5.11.	SVM	13
2.5.12.	XGBoost	13
3	Desarrollo	14
3.1.	Carga de datos	14
3.2.	Análisis preliminar de los datos	15
3.2.1.	Obtención de pacientes en estado de sueño	15
3.2.2.	Dimensión de los datos por localización neuroanatómica	15
3.3.	Cálculo del HFO rate	18

3.4.	Comparación del HFO rate en la SOZ VS NSOZ	19
3.4.1.	Todo el cerebro	19
3.4.2.	Localizado	21
3.5.	Predicción de la SOZ mediante event rates: Baselines	31
3.5.1.	Primeros pasos	31
3.5.2.	Todo el cerebro	32
3.5.3.	Localizado	34
3.6.	Clasificación de HFOs mediante aprendizaje automático	41
3.6.1.	Todo el cerebro	42
3.6.2.	Localizado	43
3.7.	Comparación del baseline y los modelos de rate filtrado	43
3.7.1.	Todo el cerebro	43
3.7.2.	Localizado	43
3.8.	Simulación del clasificador	44
3.9.	Driver	45
4	Conclusiones	47
	Bibliografía	48

1. INTRODUCCIÓN

En esta sección definimos los conceptos claves y la idea del trabajo. Comenzamos definiendo epilepsia y convulsión, damos una idea básica del funcionamiento eléctrico en el cerebro. Luego definimos *epilepsia refractaria*, planteamos el problema que conlleva el tratamiento quirúrgico, y mencionamos el papel de la seizure-onset zone (SOZ) como marcador del tejido epileptógeno. Introducimos los biomarcadores de epilepsia más conocidos: HFOs y Spikes. Resaltamos la importancia del HFO rate citando investigaciones recientes que asocian altas tasas de HFOs por minuto a sectores epileptógenos. Presentamos la idea central de la tesis y el rol del aprendizaje automático en la tarea, delimitando los objetivos y nuestra hipótesis. Finalmente comentamos algunos conceptos relevantes de electroencefalografía y anatomía cerebral para facilitar el entendimiento del trabajo.

1.1. Epilepsia

La epilepsia es una de las enfermedades neurológicas crónicas más frecuentes. Está fuertemente relacionada a problemas de sincronización en la actividad eléctrica cerebral y afecta a todas las edades. Se estima que el 1.2 % de la población mundial la padece. Se consensuó llamarla ‘enfermedad’ en el año 2014 (antes se hablaba de trastorno o desorden), tras el artículo “A practical clinical definition of epilepsy” [1].

Se habla de epilepsia a partir de la segunda convulsión espontánea. Una convulsión o crisis epiléptica es un evento súbito, caracterizado por actividad neuronal anormal y excesiva o bien sincrónica en el cerebro. Cabe mencionar que la frecuencia de las convulsiones puede variar desde menos de una al año hasta varias al día. En cuanto a su clasificación clínica, las convulsiones que afectan a una zona concreta, lóbulo particular, o un hemisferio, usualmente se llaman *focales* o *parciales*, mientras que las *generalizadas* son aquellas que afectan a ambos hemisferios. A su vez, se les denomina *simples* si no afectan la conciencia y *complejas* cuando si lo hacen.



Características de las crisis: Los *episodios* pueden ser desde muy breves, de ausencia o contracciones musculares, hasta convulsiones prolongadas y graves. Las características varían y dependen de en qué parte del cerebro comienza la alteración y cómo se propaga. Ocurren síntomas temporales, como pérdida del conocimiento o la conciencia, y alteraciones del movimiento, de los sentidos (incluyendo visión, audición y gusto), estado de ánimo u otras funciones cognitivas.

Problemas derivados: Entre las problemáticas que comprende, cabe mencionar que las personas con convulsiones tienden a padecer más problemas físicos (tales como fracturas y hematomas derivados de traumatismos relacionados con las convulsiones) y mayores tasas de trastornos psicosociales, entre ellos ansiedad y depresión. Tres cuartas partes de las personas que viven en países de ingresos bajos no reciben el tratamiento que necesitan.

Incluso aún, en muchos lugares del mundo, los pacientes y sus familias pueden ser víctimas de la estigmatización y la discriminación.

Potenciales de acción: Para poder interpretar lo que sucede a nivel celular es útil definir el concepto de potencial de acción. Nuestro sistema nervioso está formado por neuronas que se comunican entre sí a través de los espacios sinápticos. Se llama potenciales de acción (PA) a los impulsos nerviosos que transmiten las neuronas. Estos se desplazan gracias al flujo de iones que entran y salen a través de los canales proteicos que posee la membrana celular. Neurotransmisores se encargan de excitar o inhibir el potencial de acción hacia la post-sináptica. Se dice que en las crisis epilépticas se produce un “cambio de despolarización paroxística” en grupos de neuronas que se sincronizan entre ellas y envían un gran grupo de impulsos eléctricos de forma repetida (paroxísticas). Puede ser un exceso de excitación o déficit de inhibición.

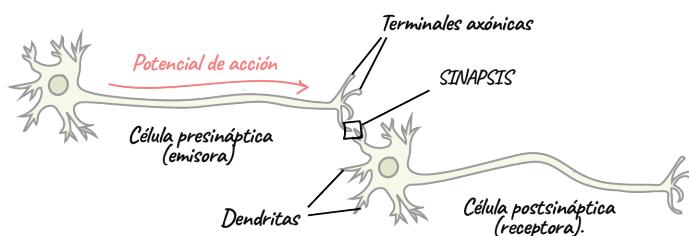


Fig. 1.1: Esquema general del proceso químico y eléctrico neuronal

Epilepsia refractaria: Se estima que entre 60 y 70 % de los casos de epilepsia pueden ser controlados mediante medicación por neurólogos especialistas, sin embargo, el 30 a 40 % restante no responde al tratamiento farmacológico ¹. Entre el 5 y el 10 % de esta población logra controlar las convulsiones cambiando o aumentando la medicación ². Caso contrario, cuando siguen ocurriendo tras probar varias medicaciones recetadas de forma correcta, ya se habla de *epilepsia refractaria*. El tratamiento alternativo de preferencia en este punto es la cirugía. En algunos tipos de epilepsia (generalmente las focales) se consigue la libertad total de las crisis o se mejora la calidad de vida considerablemente en hasta un 80 % de los casos. Por ejemplo la probabilidad de libertad de convulsiones para pacientes tratados por epilepsia refractaria focal del lóbulo temporal es aproximadamente 80 %. Sin embargo, para pacientes con epilepsia del lóbulo frontal, la libertad de crisis sólo se alcanza en la mitad de los pacientes que pasan por la cirugía, lo que destaca la necesidad de mejorar los procedimientos para poder planear una cirugía efectiva. [2]

Planeamiento quirúrgico: La cirugía consiste en la resección quirúrgica de la zona epileptógena (EZ). Sin embargo, ninguna herramienta de diagnóstico actual permite delimitar directamente la EZ, por lo que se la estima mediante un complejo proceso. El equipo médico realiza una variada gama de estudios entre los que se pueden incluir: tomografías computadas (CT), resonancias magnéticas (MRI), electroencefalogramas (EEG, iEEG, SEEG) y magnetoencefalogramas (MEG).

¹ Artículos en PubMed

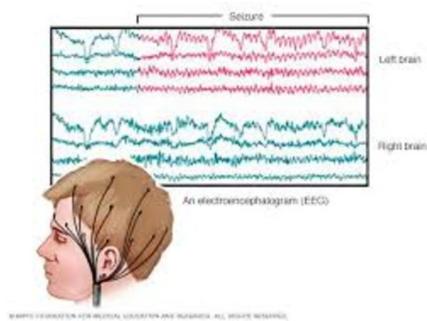
² Epilepsy.com



(a) Tomografía computada



(b) Resonancia magnética



(c) Electroencefalografía

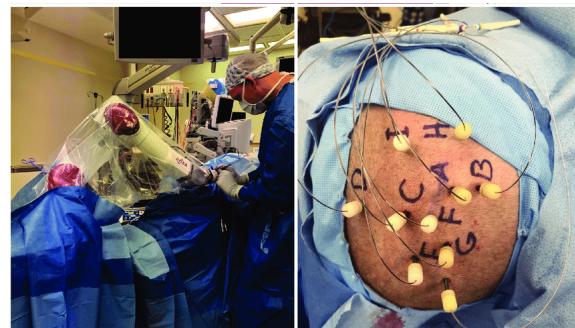


(d) Magnetoencefalografía

Fig. 1.2: Estudios diagnóstico más empleados para tratar la epilepsia.



(a) Scalp EEG (no invasivo)



(b) Estéreo-EEG (intracranial e invasivo)

Fig. 1.3: EEG invasivo y no invasivo

El marcador "gold standard" para la identificación de la EZ es conocido como *seizure-onset zone* (SOZ), es decir, la zona en donde se originan las crisis. Se utiliza para planificar qué sectores se van a resecar, y generalmente se localiza realizando una evaluación de EEG intracranial (iEEG). Sin embargo, sumado a los bajos porcentajes de éxito para algunos tipos de epilepsia, el enfoque clínico actual utilizando únicamente la SOZ tiene baja resolución temporal (estudio de poco tiempo) y puede resultar en una detección incompleta de los sitios epileptógenos. Una posible solución a este problema es incorporar biomarcadores del tejido epileptógeno que ocurran entre convulsiones (interictales).

1.2. Spikes, HFOs y HFO rate

Los biomarcadores interictales "gold standard" de epilepsia son las Spikes (descargas epileptiformes). Se utilizan actualmente para guiar resecciones en contexto intra-operatorio, sin embargo, tienen muchas desventajas como biomarcador. Son relativamente poco específicas y a veces incluso pueden tener baja sensibilidad. **Comentario**

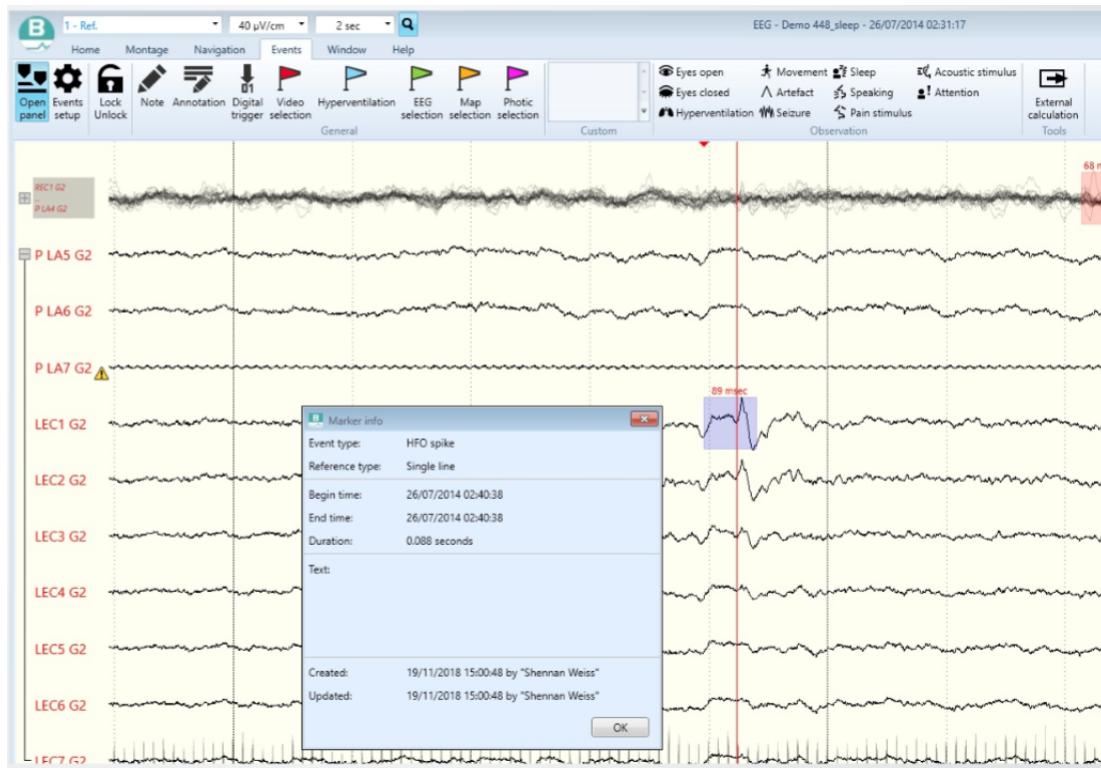


Fig. 1.4: Captura de pantalla de la forma de un Spike (azul). El programa de visualización de la señal se llama [Brain Quick](#)

Otro tipo de biomarcadores interictales son las oscilaciones de alta frecuencia (HFOs), ráfagas breves (15-100 ms) de energía con contenido espectral en el rango (80, 600 Hz). Los HFOs se subclasifican en Ripples (80-200 Hz) y Fast Ripples (200–600 Hz) de acuerdo a su frecuencia, y usualmente los primeros son considerados de mayor sensibilidad pero menor especificidad que los Fast Ripples[2]. Los HFOs se encuentran bajo estudio desde hace 2 décadas, y especialmente en los últimos años se resaltó su potencial como biomarcadores de epilepsia.

Tanto los Spikes como los HFOs son eventos capturables mediante electroencefalografía (EEG) y están fuertemente relacionados con el concepto de potencial de acción mencionado anteriormente. Una métrica que se considera útil para predecir qué sectores son epilépticos es la tasa de eventos por minuto, en el caso de los HFOs es llamada **HFO rate**. Por ejemplo: Podríamos establecer una linea de corte diciendo que un electrodo corresponde a un sector epileptógeno si su HFO rate supera un umbral determinado.

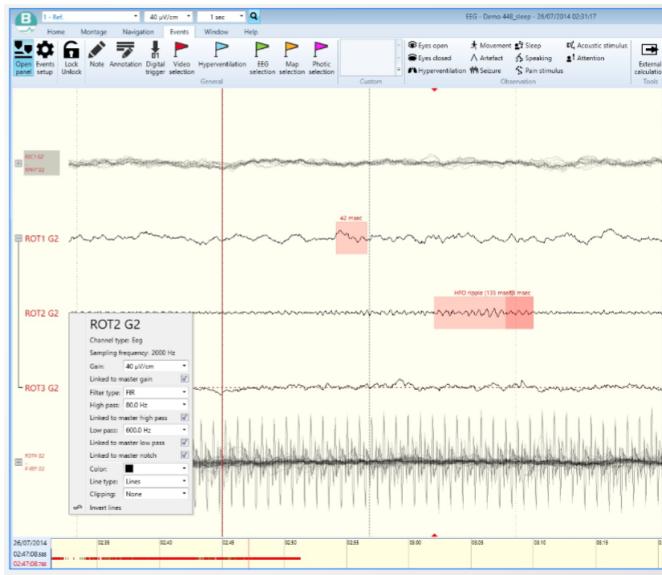


Fig. 1.5: Captura de pantalla de la forma de un Ripple (rojo)

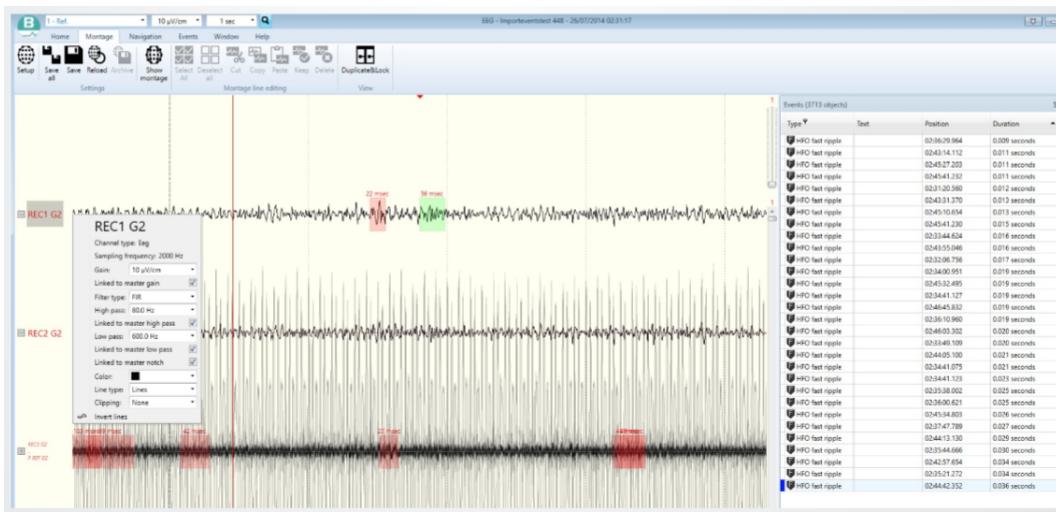


Fig. 1.6: Captura de pantalla de la forma de un Fast Ripple (verde)

1.3. Literatura y planteo del problema

TODO meter las referencias a los papers Estudios anteriores resaltan la importancia de los HFOs como biomarcadores de epilepsia. Particularmente, investigaciones previas señalaron que los sectores epileptógenos registran una mayor tasa de eventos por minuto (HFO rate). [2] Sin embargo, existen cuestiones que deben ser resueltas para poder establecer a las HFOs como biomarcadores estándar en el contexto clínico. Una de las principales reside en que estos eventos también pueden ser observados en regiones sanas del cerebro, interviniendo en la cognición normal. Más aún, no se conoce un método para distinguir completamente las HFOs fisiológicos de los patológicos (pHFOs). Este último tema tiene una influencia considerable en el campo de investigación de los HFOs. Tal distinción re-

sulta fundamental para revelar el verdadero potencial de los HFOs como predictores del resultado post operatorio. [2] En esta tesis planteamos el uso de técnicas de aprendizaje automático supervisado para intentar resolver esta tarea. Se analizó una numerosa base de datos de HFOs. Esta contenía *features* (propiedades) morfológicas obtenidas tras el procesamiento de la señal de electroencefalogramas intracraneales realizados a pacientes con epilepsia refractaria. Entre las distintas features se utilizó la duración, la frecuencia pico y promedio, la potencia pico y promedio, el acoplamiento de amplitud de fase (PAC) y una etiqueta que informaba si el electrodo había registrado una convulsión o no (etiqueta llamada SOZ). Utilizando la etiqueta SOZ como *label* del aprendizaje, es decir, utilizando a la SOZ como biomarcador subrogado de la EZ, entrenamos con las otras features mencionadas para predecir si los eventos pertenecían a la SOZ o no. Luego usamos el clasificador para retener únicamente a los eventos predichos como SOZ bajo la hipótesis de que los eventos patológicos sucederían con mayor frecuencia en la SOZ y que el *machine learning* podría diferenciar ambas clases. Se recalcó el HFO rate en diversas regiones neuroanatómicas para los distintos tipos de HFO y se comparó su rendimiento con el del HFO rate baseline (antes de filtrar los eventos con machine learning) para predecir los electrodos pertenecientes a la SOZ. Se evaluaron distintos algoritmos como XGBoost, Bayes, SVM y ADA Boost.

1.4. Objetivos e Hipótesis

Esta tesis tuvo como objetivo evaluar si, aplicando técnicas de *aprendizaje automático*, es posible distinguir HFOs características de la SOZ mejorando su capacidad como predictores de ésta. Utilizando la SOZ como marcador subrogado de la EZ, la hipótesis fue que un *filtro* óptimo conseguiría aumentar la diferencia entre HFO rates dentro y fuera de la SOZ.

1.5. Electroencefalografía

Como ya se mencionó, en este proyecto se trabajará a partir de registros de electroencefalogramas intracraneales. Particularmente obtenidos a partir de estéreo-electroencefalografías (SEEG), método que permite medir la actividad eléctrica del cerebro de forma invasiva, pero de una forma mucho más moderna evitando una craneotomía. Se realizan agujeros muy pequeños en el cráneo en los cuales se introducen electrodos. La señal del EEG es entonces, el voltaje asociado a la corriente resultante de la actividad neuronales en la corteza cerebral. Dicho fenómeno tiene origen en los procesos relacionados a los potenciales de acción, descriptos anteriormente en 1.1. El estudio registra actividad eléctrica espontánea del cerebro a lo largo de un período de tiempo.

Algunas de las técnicas más comunes para el análisis de la señal son potenciales evocados por eventos, análisis del espectro y estudios de conectividad entre regiones cerebrales. Muy resumidamente, el primero estudia los cambios en el potencial eléctrico asociado a un evento, tales como apretar un botón o recibir algún estímulo. El segundo es el que está aplicado en nuestros datos, estudia el rango de frecuencias en las que se transmite la información. Las distintas frecuencias están relacionadas a distintos estados de conciencia. En tercer lugar, los estudios de conectividad analizan cómo distintas partes del cerebro interactúan, cuales regiones del cerebro se comportan de manera sincronizada o no. A pesar de su limitada resolución espacial, los EEGs continúan siendo una herramienta muy



valiosa para la investigación y diagnóstico, ya que es uno de los pocos estudios que ofrecen resolución temporal en el rango de los milisegundos, lo cual no es posible con técnicas como MRI (Magnetic Resonance Imaging), PET (Positron Emission Tomography) o CT (Computed Tomography).

Mucha de la información en el cerebro se encuentra codificada en distintas frecuencias. Las oscilaciones suelen dividirse por bandas de frecuencias en Delta (δ), Theta (Θ), Alfa (α), Beta (β) y Gamma (γ). Cada una de estas bandas de frecuencia se manifiestan con mayor intensidad en situaciones particulares. A continuación se describen las características principales de cada una.

- **Delta:** Las ondas Delta (δ) son las que se vinculan al sueño profundo, siendo las que tradicionalmente se han vinculado al descanso y reparación de los tejidos nerviosos. Están en el rango de frecuencias que va de 0.1 a 3.9 Hz.
- **Theta:** Las ondas Theta (Θ) se observan al igual que las alfa cuando estamos relajados, pero en este caso son más frecuentes en momentos en que además de relajados estamos somnolientos, siendo el tipo de onda más predominante durante la fase dos del sueño no REM. Están en el rango de frecuencias que va de 4 a 7.5 Hz.
- **Alfa:** Las ondas Alfa (α) aparecen en momentos de relajación o ante tareas que no nos requieren concentración o esfuerzo. Están en el rango de frecuencias que va de 8 a 15 Hz.
- **Beta:** Las ondas Beta (β) suelen reflejar la realización de un esfuerzo mental intenso, apareciendo generalmente mientras estamos despiertos o bien durante el sueño REM. Están en el rango de frecuencias que va de 15 a 25 Hz.
- **Gamma:** Por arriba de los 25Hz se define la banda Gamma (γ). Esta banda está asociada con actividad cerebral producida en tareas de alta demanda cognitiva.

TODO translate info chat with shennan

In another Thesis I read : Delta (0.1 to 3.9 HZ), Theta (4 to 7.5 HZ), Alfa (8 to 15 hz), Beta (15 to 25 hz), Gamma (more than 25 hz)

classic oscillatory bands slow, delta, theta, alpha, beta, gamma last for extended durations hundreds of cycles. HFOs last only a couple of cycles. You proved this by looking at the duration of events

slow is 0.5 - 2 HZ, delta is 2-4 hz theta is 4-8 hz, alpha 8-12 hz this is by convention spindles only happen during stage 2,3 sleep they ar between 16-24 Hz

the HFOs are short bursts of range with spectral content between 200 to 600 hz occurs superimposed on the slow wave in some cases. The phase angle tells us what part of the slow wave the HFO occurred on It looks like from your analysis PAC is not all that helpful in the whole brain but may be helpful in certain brain regions slow angle tells you where on the slow wave (peak to trough) vs (trough to peak) the ripple occurred it is a hilbert transform of the slow wave

1.6. Anatomía y Fisiología Cerebral

TODO imagen lóbulos, MNI space y regiones neuroanatómicas

1.7. Diagrama del trabajo

El siguiente diagrama resume de manera concisa los conceptos claves del trabajo.
TODO cuadro sinóptico

2. MATERIALES Y MÉTODOS

TODO DID y completar secciones de teoria

2.1. Corpus

Se trabajó con una base de datos no relacional por documentos de MongoDB. La información corresponde a eventos capturados mediante estéreo-electroencefalografías realizadas a 115 pacientes con epilepsia refractaria, fue validada y provista por el equipo del neurociéntico [Shennan Weiss](#) quien realiza investigación en el campo de los HFOs y fue el experto de dominio consultado durante el trabajo.

Algunos exámenes pertenecían a procedimientos intraoperatorios y no habían sido completamente limpiados de artefactos musculares por lo que los descartó. El resto de los datos corresponden a exámenes realizados a pacientes en estado de sueño lo cual es común en epilepsia. La base contiene **3.550.000** eventos clasificados y caracterizados. Las dos colecciones utilizadas se llaman: Electrodes y HFOs, y contienen la información asociada a tres entidades de nuestro interés: Paciente, Electrodo y Evento.

El siguiente diagrama resume la información que nos aportan los documentos de ambas colecciones.

TODO DID

Entre la información asociada a los eventos (colección HFOs), se encuentra su duración, frecuencia pico y promedio, potencia pico y promedio, el ángulo de amplitud de acoplamiento de fase con distintas frecuencias, sus coordenadas neuroanatómicas y una etiqueta llamada SOZ, que indica si el electrodo registró el comienzo de una convulsión durante el examen. A continuación se detalla un poco el significado de cada propiedad utilizada o mencionada en el trabajo.

- **patient_id:** Es simplemente un código de examen anónimo en tipo string. Lo usamos como identificador para poder separar por paciente al crear set de entrenamiento, test y validación.
- **file_block:** Los exámenes de EEG fueron analizados en bloques de tiempo de forma paralela. Este atributo contiene el número relativo de bloque.
- **r_duration:** Duración de un bloqué en segundos.
- **electrode:** Nombre del electrodo.
- **x, y, z:** Coordenadas neuroanatómicas en relación al MNI Space.
- **loc_i:** Para i en 1, 2, 3, 4, 5, son tags asociados a regiones neuroanatómicas de acuerdo a un atlas cerebral conocido (MNI space). La i está asociada a granularidad, a mayor granularidad, más específica es la región. Un ejemplo de loc2 es "Temporal Lobe" un ejemplo de loc5 es "Hippocampus".
- **soz:** '1' si el electrodo registró el comienzo de una convulsión durante el examen (pertenece a la SOZ) '0' si no.

- **type:** String de id numérico referido a la categoría del evento '1': 'RonO', '2': 'RonS', '3': 'Spikes', '4': 'Fast RonO', '5': 'Fast RonS', '6': 'Sharp Spikes'
- **duration:** Duración de un evento en segundos.
- **freq_pk:** Pico de frecuencia.
- **freq_av:** Frecuencia promedio.
- **power_pk:** Pico de densidad de potencia.
- **power_av:** Densidad de potencia promedio.
- **phase_angle:** Ángulo de acoplamiento de amplitud de fase para distintas frecuencias del espectro (δ, γ, θ , slow, spindle).

Se normalizó los datos en memoria al cargarlos desde la base de datos solucionando las inconsistencias que había entre entradas asociadas al mismo par (paciente, electrodo). Nos basamos en la siguiente estructura.

- Un **Paciente** tiene un id y un conjunto de electrodos
- Un **Electrodo** tiene un nombre, información asociada a los bloques del examen (ya que algunos bloques de algunos electrodos fueron eliminados por poseer artefactos musculares), coordenadas, tags de localización, un flag de si pertenece a la SOZ, y un diccionario con los eventos de cada tipo.
- Un **Evento** es esencialmente un diccionario con los atributos del electrodo y los propios como duration, freq_pk, freq_av, power_pk, power_av, phase_angle

2.2. Software

Se desarrolló un entorno en lenguaje **Python** para analizar los datos. El código se encuentra modularizado y comentado. Por lo que resulta un recurso extra como especificación técnica de los procedimientos realizados. Se encuentra disponible para descargar en github.

La semántica de cada módulo es la siguiente:

- **main:** Punto de entrada del programa. Simplemente llama al Driver del programa.
- **driver:** Una interfaz básica que toma por parámetro el índice en la lista de experimentos y corre el experimento correspondiente con los parámetros que fueron utilizados.
- **config:** Constantes, variables globales, paths, parámetros de configuración.
- **tests:** Tests. Aunque también muchas cosas se testearon mediante aserciones y uso de excepciones
- **utils:** Funciones auxiliares útiles.
- **db_parsing:** Se encarga de la entrada salida entre la base de datos de MongoDB y Python, normaliza los datos llevándolos a un modelo de entidad relación consistente basado en las 3 entidades mencionadas en la sección anterior [2](#) y lo filtra respecto a variables de control como por ejemplo restringir a una región cerebral determinada.

- **patient**: La clase Paciente.
- **electrode**: La clase Electrodo.
- **event**: La clase Evento.
- **scratch**: Primeros pasos. Selección de pacientes en estado de sueño. Tabla de cantidad de pacientes, regiones epilépticas, electrodos, eventos.
- **stats**: Código para generar histogramas con KDE para comparar atributos de la SOZ VS la NSOZ. Calcula 3 estadísticos Kolmogorov-Smirnov, Wilcoxon signed-rank y Mann-Whitney U test. Se utilizó para comparar HFO rate para distintos tipos de HFO y diversas regiones neuroanatómicas. Análogamente se utilizó para comparar las distribuciones de las features disponibles para machine learning.
- **soz_predictor**: Se encarga del análisis del event rate como predictor de la SOZ para las diversas regiones neuroanatómicas y distintos tipos de HFOs, lo cual es nuestro baseline. Se emplearon curvas ROC para visualizar el rendimiento.
- **ml_hfo_classifier**: Se encarga de emplear modelos de aprendizaje automático a la clasificación de eventos en SOZ o no SOZ. Se utilizó principalmente la librería sklearn de Python y una interfaz del algoritmo XGBoost. Implementa la función que actúa como filtro para poder posteriormente recalcular el HFO rate sobre la distribución de eventos predicha como SOZ.
- **ml_hfo_simulator**: Implementa un modelo que emula la clasificación de HFOs en SOZ o NSOZ para recall y especificidades determinadas, empleando bootstrapping sobre distribuciones de probabilidades correspondientes a predicciones de otros clasificadores de la librería sklearn de Python.
- **graphics**: El código asociado a todos los gráficos del proyecto. Se utilizaron librerías como matplotlib, pandas y plotly.

El código permite principalmente analizar el rate de un tipo de evento y región cerebral dados y comparar sus curvas ROC asociadas para predecir la SOZ. Se pueden comparar múltiples tipos y regiones a la vez.

Se descartaron datos correspondientes a análisis intraoperatorios por poseer artefactos musculares sin procesar. Se trabajó por lo tanto con los 91 pacientes restantes que contaban con estudios en estado de sueño. Para mayor detalle ver el experimento 1 del driver.

2.3. Hardware

Los experimentos fueron corridos en una computadora con las siguientes características:

Especificación técnica

- SO: GNU/Linux Ubuntu
- CPU: Intel Core i7-7700HQ CPU @ 2.80GHz x 8
- RAM: Kingston HyperX FURY 32GB 2133MHz DDR4
- DISCO: SSD

Requerimientos de memoria para correr los experimentos

- **RAM:** El pico de memoria detectado corriendo los experimentos fue de TODO. Es un poco alto debido a que se priorizó optimizar el tiempo de ejecución minimizando la cantidad de llamados a la base de datos.
- **Disco:** El comprimido de mongo db pesa TODO

2.4. Estadística

La estadística es uno de los pilares del aprendizaje automático. En esta sección introducimos los conceptos y métodos estadísticos clave que fueron utilizados.

2.4.1. Varianza y Desvío Estándar

2.4.2. Kolmogorov-Smirnov

El test the Kolmogorov-Smirnov testeá si dos distribuciones arbitrarias son iguales. Puede ser utilizado para comparar dos distribuciones empíricas de datos, o para comparar una distribución empírica con otra de referencia (conocida). Está basado en la comparación de las funciones de distribución acumulada (CDFs). Una CDF de una variable aleatoria X es la probabilidad de que la variable sea menor o igual a algún valor. Precisamente, es una función F tal que

$$F(a) = P(X \leq a) \quad (2.1)$$

- 2.4.3. Wilcoxon signed-rank
- 2.4.4. Mann-Whitney U test
- 2.4.5. P-Valor
- 2.4.6. Error Cuadrático Medio
- 2.4.7. R cuadrado
- 2.4.8. Correlación de Pearson
- 2.4.9. Correlación de Spearman
- 2.5. Aprendizaje Automático
 - 2.5.1. Features y label
 - 2.5.2. Sesgo y Varianza de un modelo
 - 2.5.3. Validación cruzada
 - 2.5.4. Matriz de Confusión
 - 2.5.5. Sensibilidad y Especificidad
 - 2.5.6. Curva de Característica Operativa del Receptor (ROC)
 - 2.5.7. Curva de Precisión-Recall (PR)
 - 2.5.8. Regularización
 - 2.5.9. Hiperparámetros
 - 2.5.10. Bayes
 - 2.5.11. SVM
 - 2.5.12. XGBoost

3. DESARROLLO

Esta sección presenta los procedimientos y experimentos realizados detallando los detalles relevantes y presenta los resultados con una breve discusión.

3.1. Carga de datos

Al ser una base de datos no relacional y estar los datos desnormalizados lo primero que hicimos fue crear un módulo de Python encargado de la entrada salida entre MongoDB y Python. Es decir, responsable de traducir respuestas de la base de datos al modelo normalizado en memoria presentado anteriormente en 2.1: En el modelo un paciente contiene electrodos y estos a su vez contienen eventos de distinto tipo. Este módulo también cumplió el rol de chequear y corregir inconsistencias entre distintas entradas referidas al mismo par (paciente, electrodo). Para esto resultó de utilidad la herramienta gráfica MongoDB Compass, que permite visualizar resumidamente el esquema de los documentos de una colección, sus tipos y distribuciones. Nos ayudó por ejemplo a observar que algunos campos tenían entradas con distinto tipo de datos o el porcentaje de las clases en campos binarios.

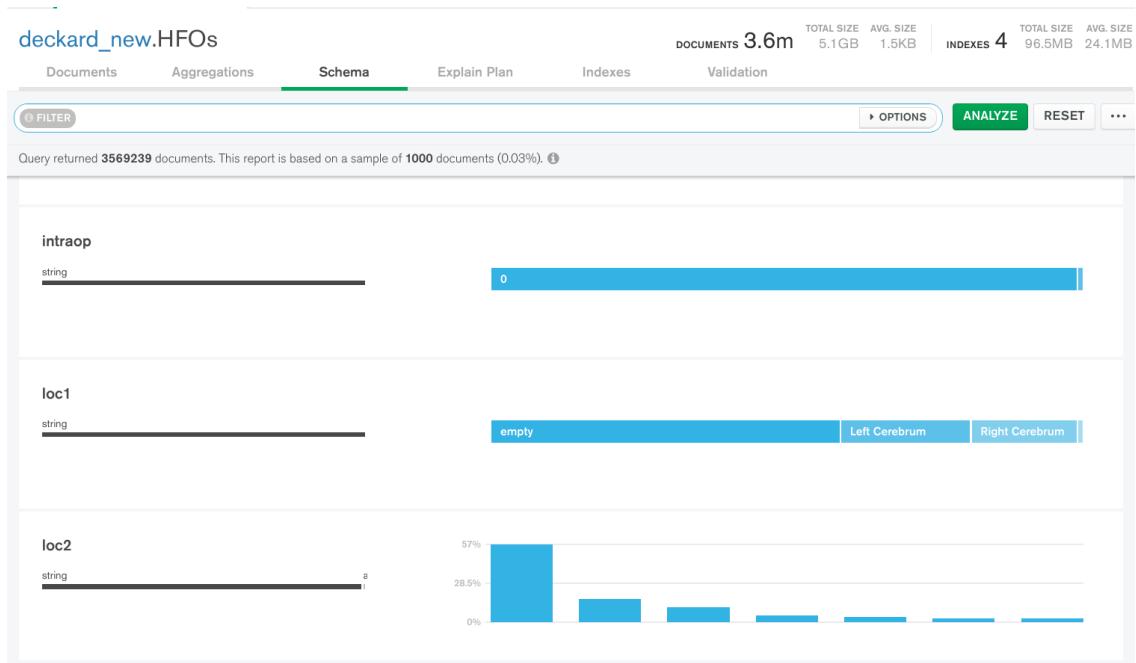


Fig. 3.1: Captura de pantalla de la interfaz de MongoDB Compass. En la sección Schema vemos por ejemplo que basados en una muestra de la colección HFOs, que el %1 de los documentos son de pacientes intraoperatorios.

En cuanto al tiempo de ejecución no resultó un punto particularmente relevante, se agregó índices en la base de datos para los campos patient_id, electrode, type, loc2, loc3, loc5 y intraop para optimizar las consultas más necesitadas.

3.2. Análisis preliminar de los datos

3.2.1. Obtención de pacientes en estado de sueño

Como se mencionó anteriormente en 2.1 algunos exámenes, aproximadamente el 1% pertenecen a procedimientos intraoperatorios y no habían sido completamente limpiados de artefactos musculares por lo que los descartó. Trabajamos entonces con el resto de los datos, que corresponden a exámenes realizados a pacientes en estado de sueño. Como el campo intraop sólo existe en la colección HFOs determinamos los pacientes en estado de sueño agrupando todos los pacientes que no tuviesen eventos con intraop == 1. El código se encuentra en la función print_non_intraop_patients del módulo scratch. La base cuenta con 91 exámenes de iEEG realizados en estado de sueño.

3.2.2. Dimensión de los datos por localización neuroanatómica

Trabajos anteriores remarcan la importancia de realizar un análisis discriminado en función de la región neuroanatómica [2]. Es por ello que nuestro siguiente paso fue detallar la cantidad de datos con los que contábamos por región. Para cada región que tuviese una muestra de al menos 12 exámenes y al menos 3 de ellos sean de pacientes con epilepsia (con electrodos SOZ), se computó el número de pacientes, el número de pacientes con epilepsia, el número de electrodos y porcentaje de ellos en la SOZ y el número de eventos por tipo.

Todo el cerebro:

Location	#Patients	#SOZPatients	#Elec	PSE	#Fast RonO	#Fast RonS	#RonO	#RonS	#Spikes
Whole Brain	91	88	9858	13.53	137442	29592	3044604	213791	49700

Tab. 3.1: Considerando todas las localizaciones posibles y permitiendo coordenadas neuroanatómicas en null.

En todo el cerebro la base provee información de 91 pacientes con examen de iEEG en estado de sueño para trabajar. 88 de ellos tienen electrodos SOZ, es decir, que registraron el comienzo de una convulsión durante el transcurso del estéreo-encefalograma. También tiene una buena cantidad total de electrodos (cerca de 10 mil). Es importante notar que el porcentaje de electrodos SOZ (PSE) se observó muy bajo (tan solo 13.5%) en todo el cerebro, aunque esto no necesariamente se cumple en regiones más localizadas. Notamos también que el tipo Ripple on Oscillation (RonO) sucedió con mucho más frecuencia que el resto de los tipos. Este tipo de HFO particularmente, se sabe que también sucede en procesos fisiológicos vinculados a la memoria en estado de sueño, por lo que resultaría muy útil que el aprendizaje automático pueda separar los RonO patológicos de los fisiológicos. Siguiendo a los RonO en frecuencia, aunque no tan de cerca, se observaron los Ripple on Spikes (RonS), luego los Fast RonO, Spikes y por último los Fast RonS.

Localizado:

Location	#Patients	#SOZPatients	#Elec	PSE	#Fast RonO	#Fast RonS	#RonO	#RonS	#Spikes
Frontal Lobe	48	18	1190	11.76	13456	1624	268611	17424	2584
Limbic Lobe	55	38	1435	19.51	22658	11938	405395	54954	14580
Occipital Lobe	28	9	210	18.1	3492	234	149927	2833	644
Parietal Lobe	41	11	536	15.11	8750	1087	199393	15393	1690
Temporal Lobe	53	33	1656	11.65	23152	1501	175678	15450	3132

Tab. 3.2: Tags utilizando el campo loc2 como criterio de restricción espacial.

En los lóbulos principales la base también cuenta con una buena cantidad de pacientes, en especial en el Temporal, Temporal mesial (Limbic Lobe) y Frontal, con 18 pacientes con epilepsia. A la hora de priorizar qué región investigar, nuestro criterio de fiabilidad respecto a los datos fue la cantidad de electrodos total y SOZ, por lo que optamos por comenzar haciendo foco en el temporal mesial, con 1435 electrodos y 19.5 % de ellos pertenecientes a la SOZ. Es importante destacar la diferencia en la cantidad de eventos para cualquiera de las categorías de HFO en esta región. En especial para RonS. En segundo lugar en cantidad de datos resultó ser el Temporal, y en tercero el Frontal.

Location	#Patients	#SOZPatients	#Elec	PSE	#Fast RonO	#Fast RonS	#RonO	#RonS	#Spikes
Cingulate Gyrus	41	9	523	6.31	4823	275	47917	2111	549
Clastrum	25	4	232	5.6	6303	76	29515	1023	270
Fusiform Gyrus	45	13	257	8.56	3618	223	32920	2356	301
Inferior Frontal Gyrus	36	6	262	8.78	2057	152	30759	2826	424
Inferior Parietal Lobule	31	3	137	5.84	668	145	41556	2009	301
Inferior Temporal Gyrus	44	6	202	12.38	6739	208	24397	1099	126
Insula	35	10	309	9.06	2593	25	55970	1596	321
Medial Frontal Gyrus	31	6	159	11.32	1401	521	28428	2352	422
Middle Frontal Gyrus	41	8	365	10.41	2910	548	66343	7184	998
Middle Occipital Gyrus	12	3	50	30	727	112	53773	1195	431
Middle Temporal Gyrus	50	18	510	9.41	6890	134	52340	2646	508
Paracentral Lobule	12	5	42	26.19	1778	34	19182	996	116
Parahippocampal Gyrus	48	31	621	31.24	15114	10723	308179	47848	12928
Postcentral Gyrus	28	4	168	17.86	6791	781	98682	9007	649
Precentral Gyrus	28	6	220	12.73	2361	71	77905	2298	368
Precuneus	23	7	137	26.28	1070	185	51871	4556	690
Sub-Gyral	56	19	320	15.31	1931	811	46395	5723	1178
Superior Frontal Gyrus	16	4	98	17.35	2711	269	38813	1595	231
Superior Temporal Gyrus	45	15	404	13.37	3594	130	48283	3672	965
Uncus	34	20	109	44.95	2297	908	23986	4594	1008

Tab. 3.3: Tags utilizando el campo loc3 como criterio de restricción espacial.

Yendo un nivel más en la granularidad de los tags, vimos que la cantidad de pacientes con electrodos SOZ disminuye bastante en algunas regiones. Priorizamos nuevamente con

el mismo criterio que en loc2 respecto a la cantidad de datos. Luego destacamos 6 zonas candidatas de prioridad para este nivel: Parahippocampal Gyrus, Middle Temporal Gyrus, Sub-Gyral, Superior Temporal Gyrus, Uncus y Fusiform Gyrus. A su vez se observó que para las regiones listadas se tenía una buena cantidad de observaciones para desarrollar aprendizaje automático con cualquiera de los tipos, quizas excluyendo a Fast RonS afuera del Parahippocampal Gyrus.

Location	#Patients	#SOZPatients	#Elec	PSE	#Fast RonO	#Fast RonS	#RonO	#RonS	#Spikes
Amygdala	38	21	164	32.93	2742	2055	69416	8495	2059
Brodmann area 11	23	3	66	7.58	248	18	10947	2440	277
Brodmann area 13	36	11	340	9.12	2666	13	55273	1440	308
Brodmann area 19	25	8	134	21.64	1779	95	70142	1650	418
Brodmann area 20	48	21	499	16.43	10475	1009	46487	6009	1038
Brodmann area 21	47	17	464	8.84	7630	168	26825	2578	527
Brodmann area 22	36	7	185	15.14	1589	92	21547	2261	536
Brodmann area 24	29	3	209	4.31	2098	121	15018	995	283
Brodmann area 28	25	14	43	46.51	1554	706	28668	3517	772
Brodmann area 3	16	3	73	21.92	4560	598	69176	5835	422
Brodmann area 31	37	9	300	11	2362	124	52765	2710	401
Brodmann area 32	37	3	182	4.95	1373	132	14204	680	159
Brodmann area 34	15	5	37	24.32	419	137	9190	922	291
Brodmann area 35	23	10	42	33.33	2412	997	19157	3591	581
Brodmann area 36	40	19	177	28.81	2579	1204	31872	6574	1337
Brodmann area 37	38	9	172	8.14	939	156	27320	1763	156
Brodmann area 38	27	9	123	9.76	1445	38	5971	577	99
Brodmann area 39	21	5	123	11.38	243	11	28908	534	83
Brodmann area 4	21	3	102	13.73	1212	48	58645	1318	216
Brodmann area 40	36	6	224	7.59	2751	322	68228	4685	491
Brodmann area 41	17	4	50	20	1105	6	17496	817	355
Brodmann area 6	27	10	337	13.95	3331	254	80800	4567	1054
Brodmann area 7	20	7	119	22.69	783	100	35844	3337	599
Brodmann area 8	19	4	102	10.78	1241	19	23307	597	62
Brodmann area 9	27	4	118	12.71	206	8	13042	338	53
Hippocampus	44	25	226	34.96	6407	6239	181159	28898	9008

Tab. 3.4: Tags utilizando el campo loc3 como criterio de restricción espacial.

En el último nivel de resolución espacial usando los tags, seleccionamos como 6 candidatas prioritarias a las siguientes regiones (con un criterio análogo a las dos tablas anteriores): Hippocampus, Amygdala, BA 20, BA 21, BA 28, BA 36.

3.3. Cálculo del HFO rate

El siguiente paso fue definir un método en la clase Electrodo para calcular el HFO rate para un tipo (o Event rate en caso de las Spikes). Como se mencionó anteriormente los exámenes están divididos en bloques temporales, por lo que el HFO rate de un electrodo se calculó como el promedio entre el rate de los bloques correspondientes. En este sentido el hecho de que la base fuera por documentos fue una ventaja ya que los bloques que eran defectuosos por poseer artefactos musculares ya habían sido removidos y no resultaron considerados en el promedio.

A continuación detallamos la implementación de su cómputo ya que es un detalle clave del trabajo.

```
def get_events_rate(self, event_types=EVENT_TYPES):
    block_rates = {block_id:[0, duration] for block_id, duration in self
                   .blocks.items()} #count per block
    for event_type in event_types:
        for block, count in self.evt_count[event_type].items():
            block_rates[block][0] += count

    # Note: rate[1] is block duration, may be None if no hfo was
    # registered for that block
    block_rates_arr = [(rate[0]/(rate[1]/60)) if rate[1] is not None
                       else 0.0 for rate in block_rates.values()]
    block_rates_arr.sort() #avoids num errors
    return sum(block_rates_arr)/len(self.blocks)
```

El parámetro es una lista de tipos de evento, permitiendo que se calcule la métrica considerando una combinación de varios subtipos como se verá más adelante en la sección [3.5.1](#)

3.4. Comparación del HFO rate en la SOZ VS NSOZ

TODO revisar visualización, comentar gráficos y conservar solo los mas importantes. Agregar tabla de localizado con resultados de los estadísticos.

El HFO rate es considerado como predictor de la SOZ en base a que estudios anteriores demostraron mayores de tasas de eventos por minutos dentro de la SOZ. En esta sección comparamos las distribuciones SOZ VS NSOZ del HFO rate para las regiones seleccionadas como prioritarias en la sección anterior (3.2.2). Es decir que realizamos una comparación independiente por cada región y tipo de HFO, por ejemplo RonS en el Hipocampo. A su vez, realizamos test estadísticos no paramétricos (Kolmogorov-Smirnov test, Wilcoxon signed-rank test y Mann-Whitney U test) para evaluar la diferencia entre las mismas. A continuación se detallan los resultados de los tests, se adjuntan los gráficos de distribuciones correspondientes a todo el cerebro e Hipocampo para las 4 tipos de HFO.

3.4.1. Todo el cerebro

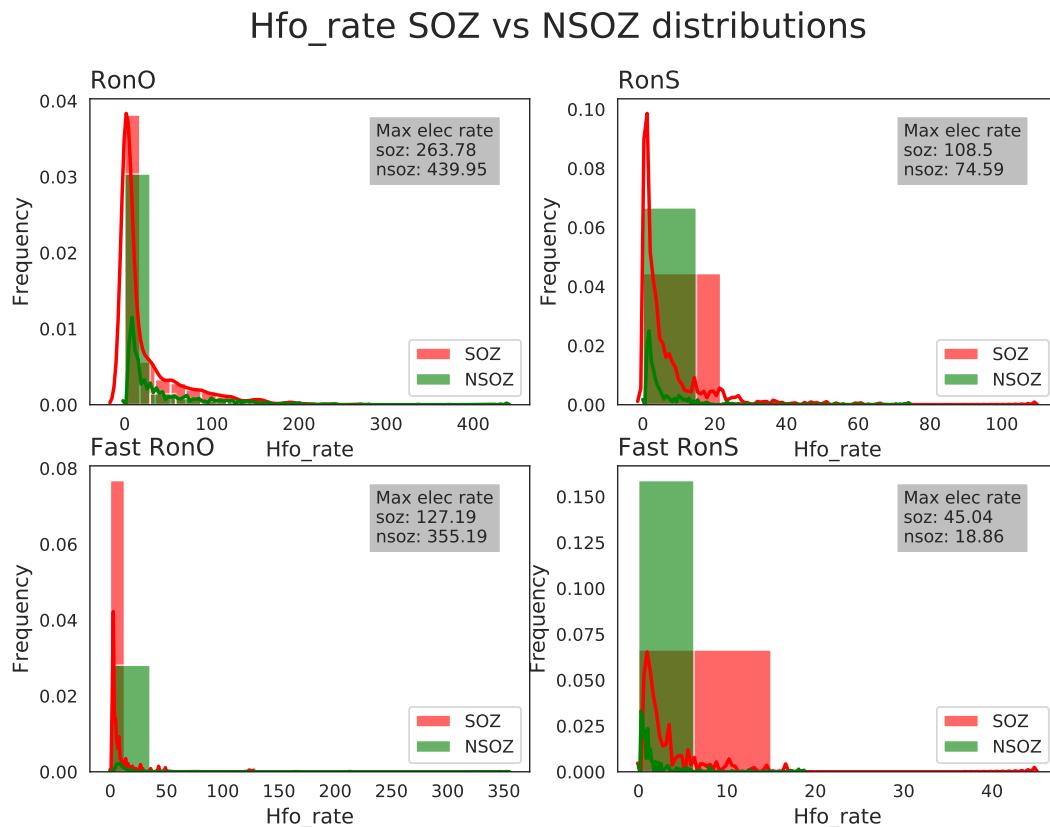


Fig. 3.2: HFO rate (events/min) en todo el cerebro

Location	HFO type	D	d_pval	W	w_pval	U	u_pval
Whole Brain	Fast RonO	0.2	1.00e+00	13.64	2.39e-42	4367356	1.02e-51
Whole Brain	Fast RonS	0.27	3.91e-76	16.79	2.91e-63	4062708	1.26e-179
Whole Brain	RonO	0.31	2.56e-98	23.24	1.90e-119	3439468.5	1.08e-121
Whole Brain	RonS	0.39	1.86e-150	26.08	5.48e-150	3164311.5	8.05e-209

Fig. 3.3: Kolmogorov-Smirnov (D), Wilcoxon (W) y Mann-Whitney (U) tests

3.4.2. Localizado

Repetimos el procedimiento para regiones de mayor granularidad, a continuación presentamos gráficos para las 3 regiones más relevantes de cada granularidad, según el criterio de cantidad de pacientes de la sección anterior.

Loc 2
Temporal Lobe

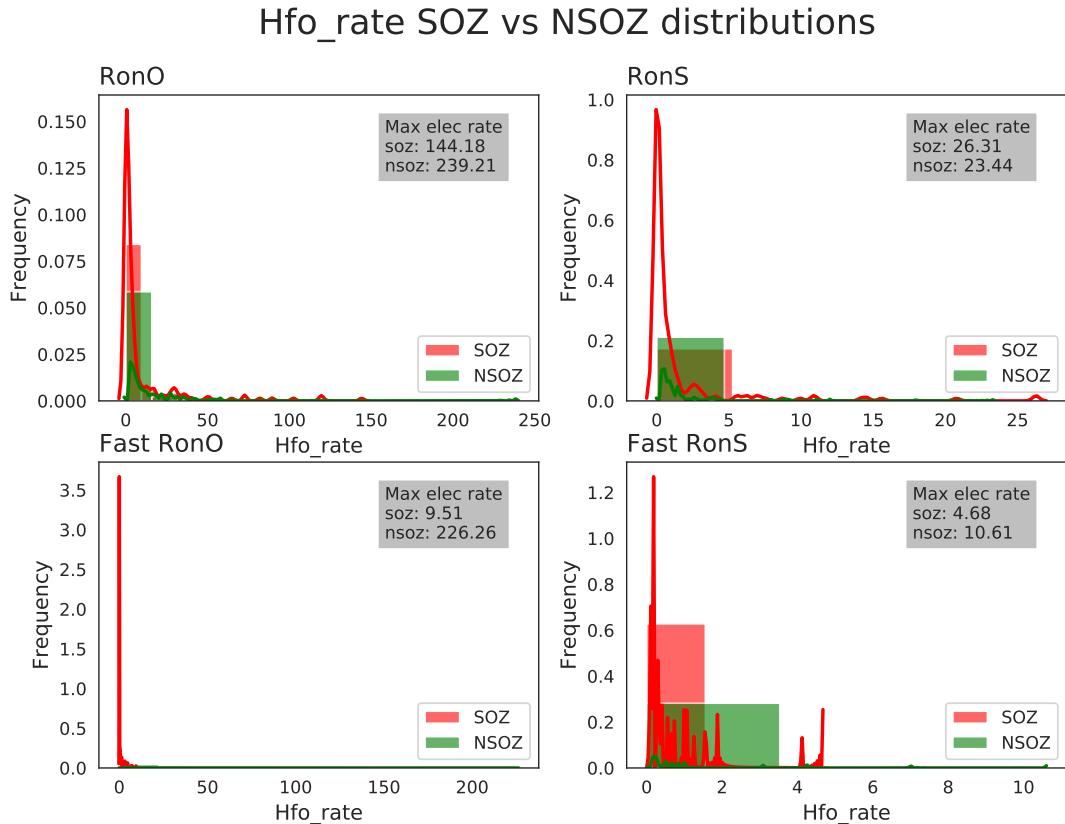


Fig. 3.4: HFO rate (events/min) en Temporal Lobe

Comentarios

Limbic Lobe

Hfo_rate SOZ vs NSOZ distributions

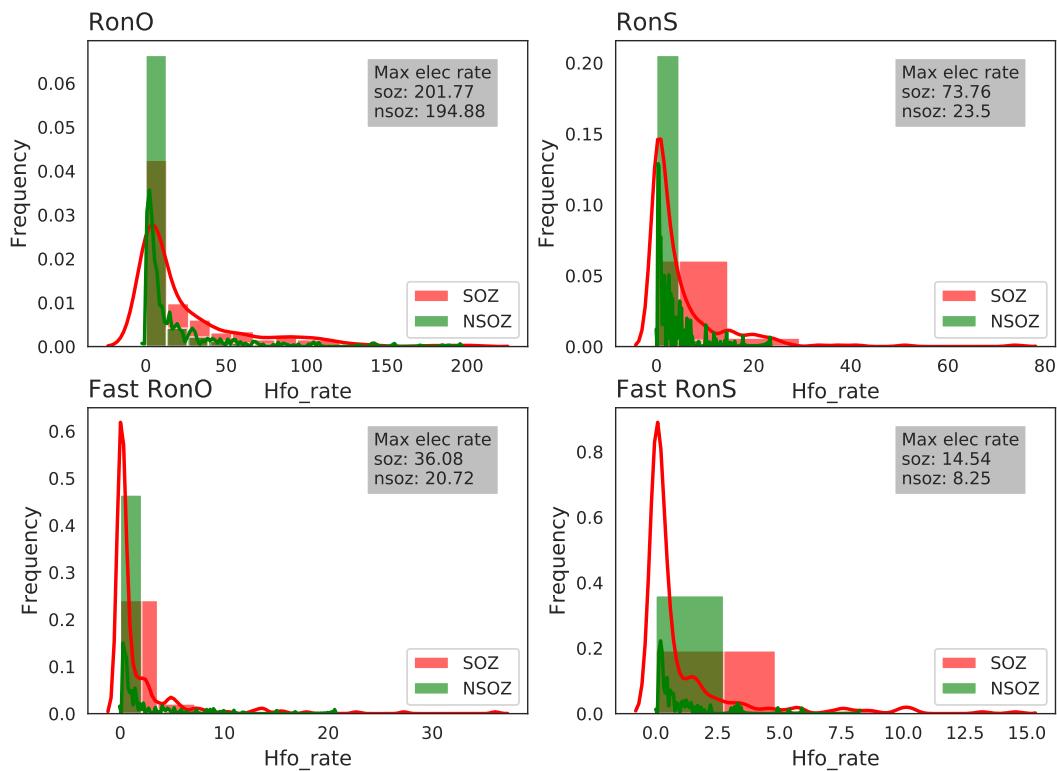


Fig. 3.5: HFO rate (events/min) en Limbic Lobe

Comentarios

Frontal Lobe

Hfo_rate SOZ vs NSOZ distributions

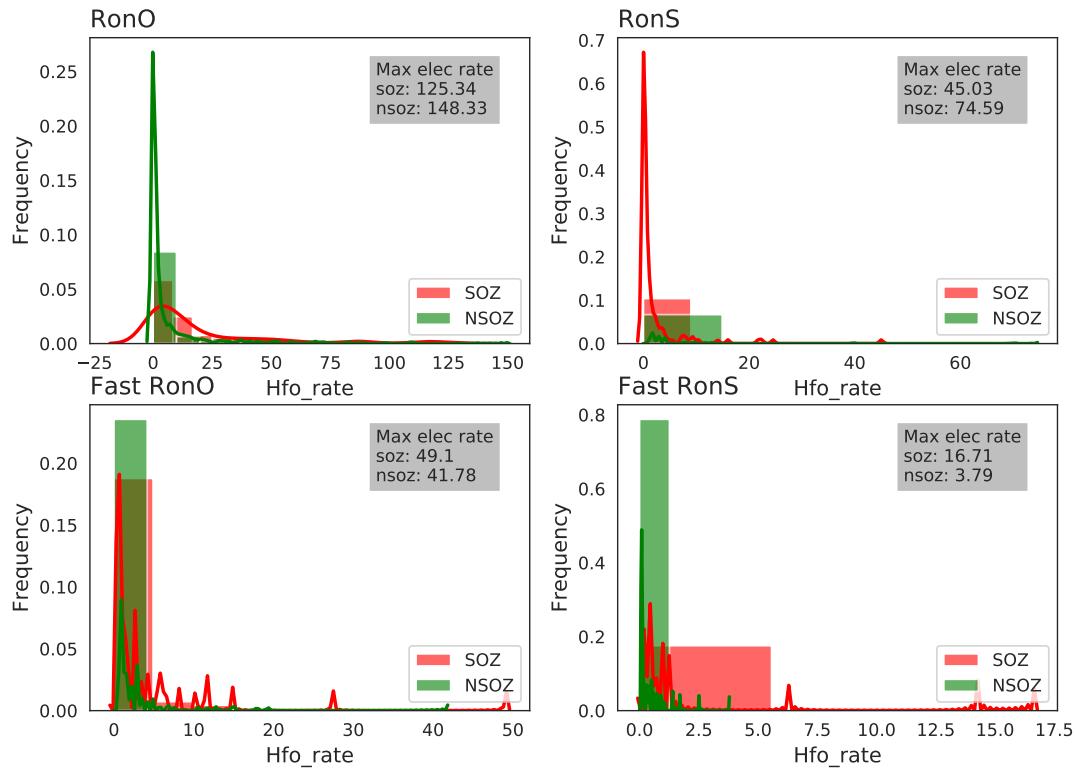


Fig. 3.6: HFO rate (events/min) en Frontal Lobe

Comentarios

Loc 3
Parahippocampal Gyrus

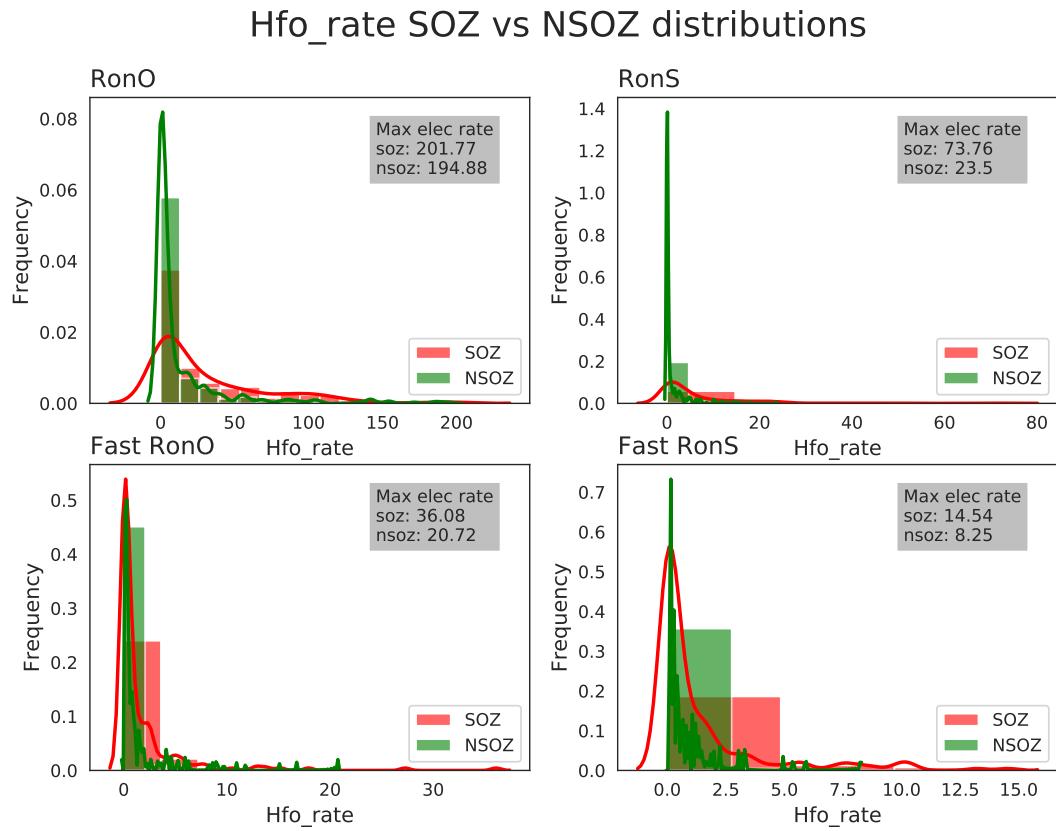
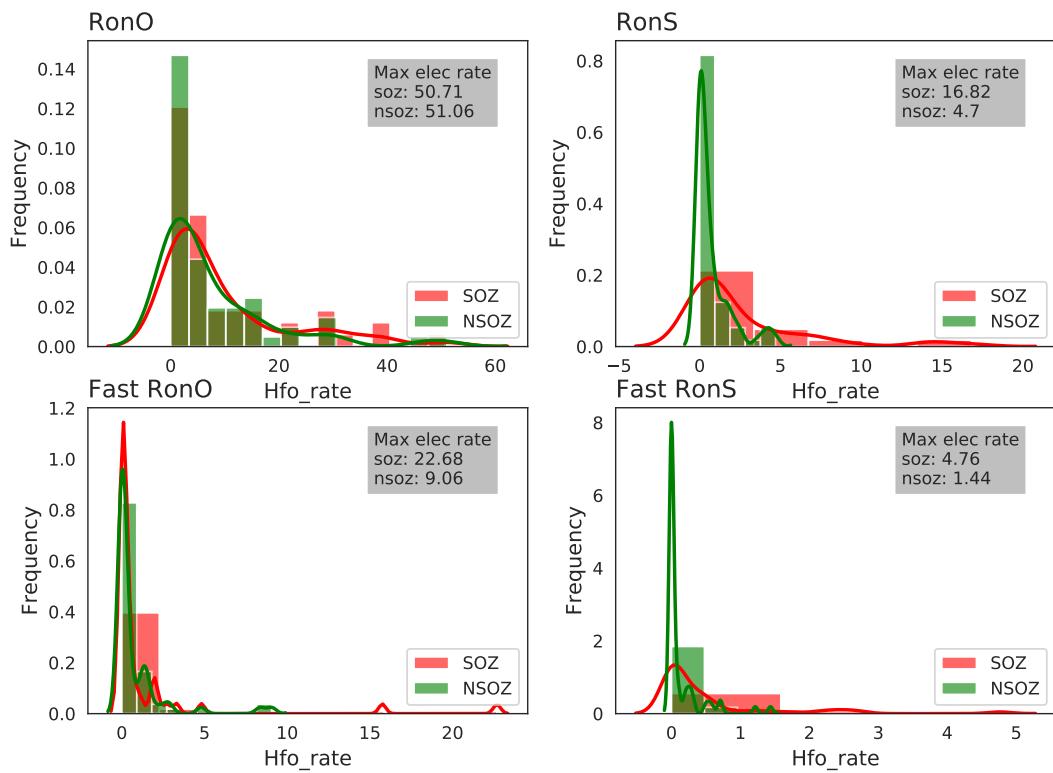


Fig. 3.7: HFO rate (events/min) en Parahippocampal gyrus

Comentarios

Uncus**Hfo_rate SOZ vs NSOZ distributions***Fig. 3.8: HFO rate (events/min) en Uncus***Comentarios**

Fusiform Gyrus

Hfo_rate SOZ vs NSOZ distributions

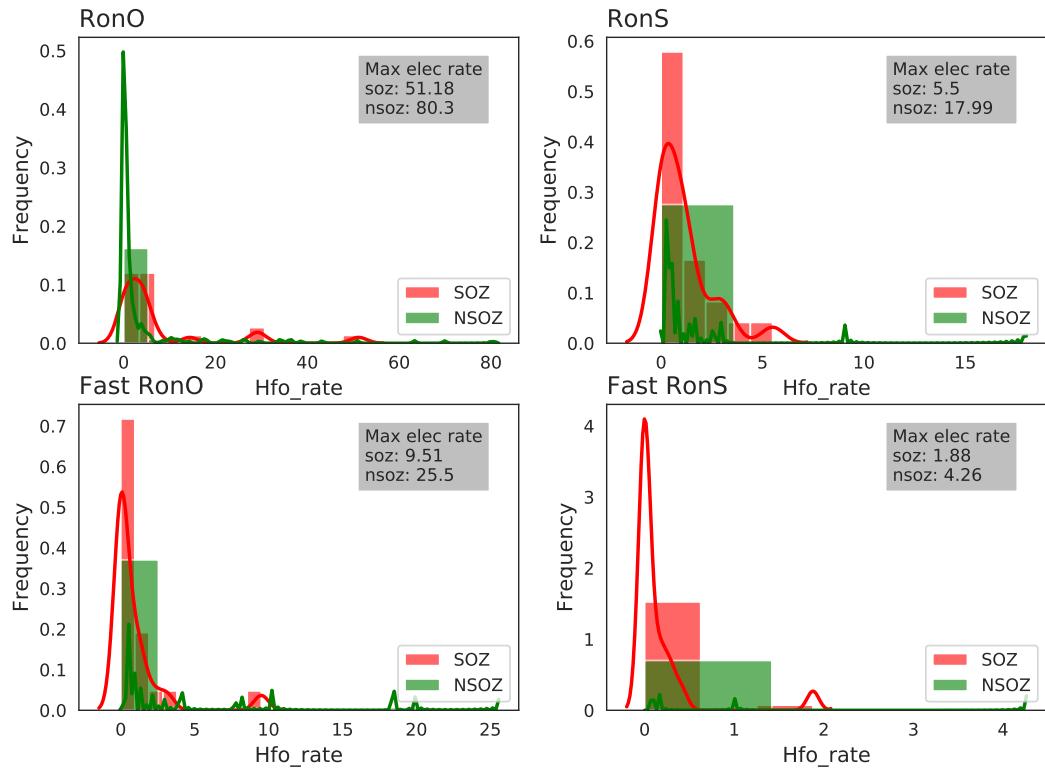


Fig. 3.9: HFO rate (events/min) en Fusiform Gyrus

Comentarios

Loc 5
Hippocampus

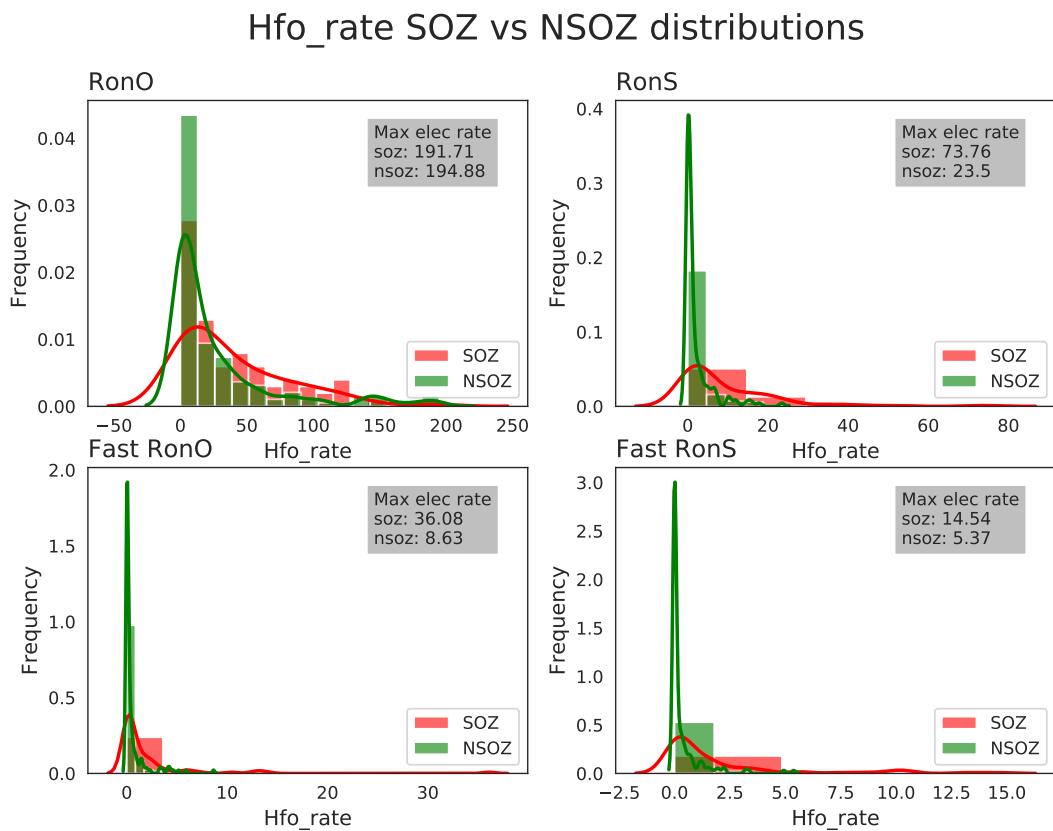


Fig. 3.10: HFO rate (events/min) en Hippocampus

Comentarios

Amygdala

Hfo_rate SOZ vs NSOZ distributions

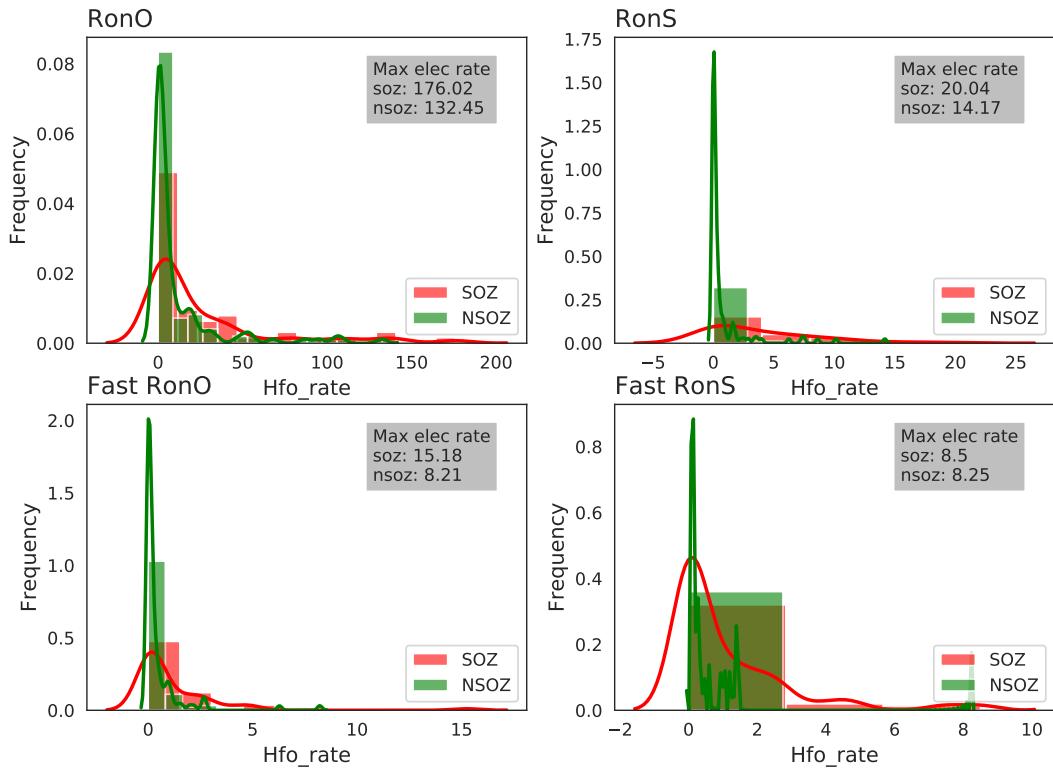


Fig. 3.11: HFO rate (events/min) en Amygdala

Comentarios

Brodmann area 28

Hfo_rate SOZ vs NSOZ distributions

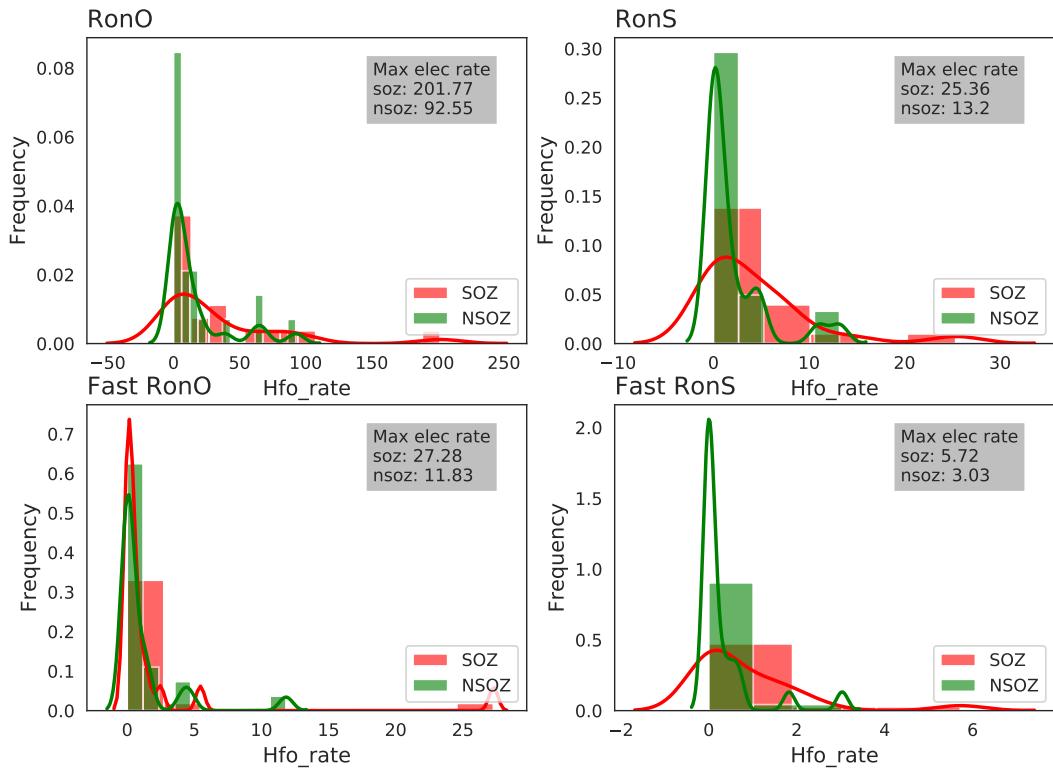


Fig. 3.12: HFO rate (events/min) en Brodmann area 28

Comentarios

Brodmann area 36

Hfo_rate SOZ vs NSOZ distributions

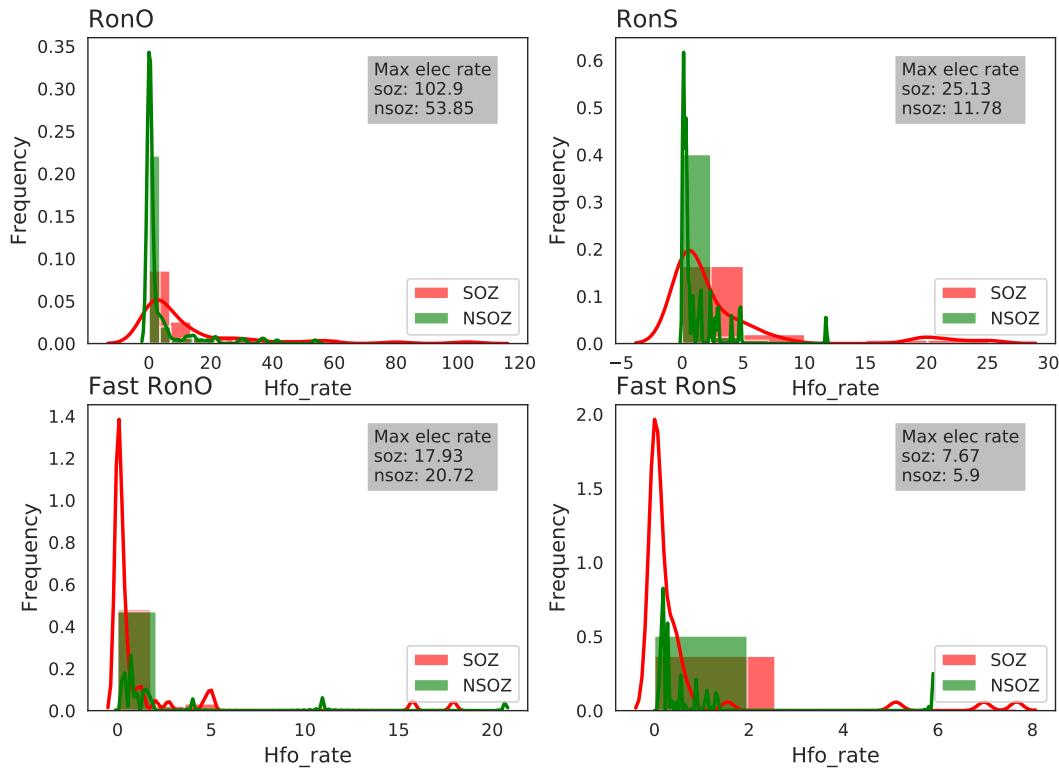


Fig. 3.13: HFO rate (events/min) en Brodmann area 36

Comentarios

3.5. Predicción de la SOZ mediante event rates: Baselines

En esta sección calculamos para cada tipo y región relevante, el baseline utilizando el HFO rate como predictor de la SOZ y graficamos las curvas ROC asociadas. Es decir:

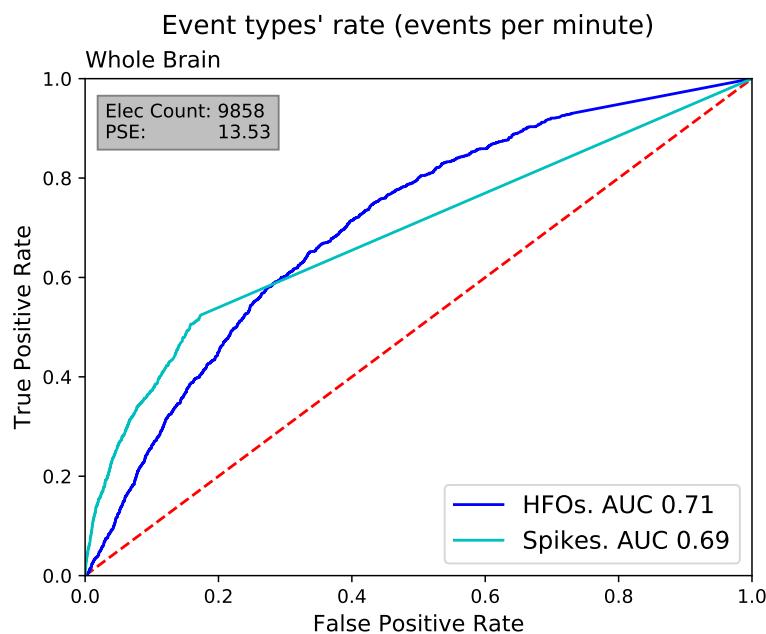
Dadas una región cerebral **R** y un tipo de evento **T**:

1. Se formó el conjunto de electrodos E asociados R (contemplando sólo pacientes "no intraoperatorios")
2. Se cargó en cada electrodo de E los eventos registrados de tipo T
3. Se calculó para cada electrodo el rate de T según lo detallado en la sección 3.3
4. Se construyeron curvas ROC donde la variable a predecir era si el electrodo era SOZ y el predictor era el event rate

A su vez, se graficó la relación entre la proporción de electrodos SOZ y el AUC de rate baseline para todas las Regiones con al menos 12 pacientes en total y mínimo 3 con electrodos en la SOZ.

3.5.1. Primeros pasos

Para simplificar un primer paso, unificamos las 4 categorías de HFOs y comparamos su rate contra el de las Spikes. El resultado fue el siguiente:

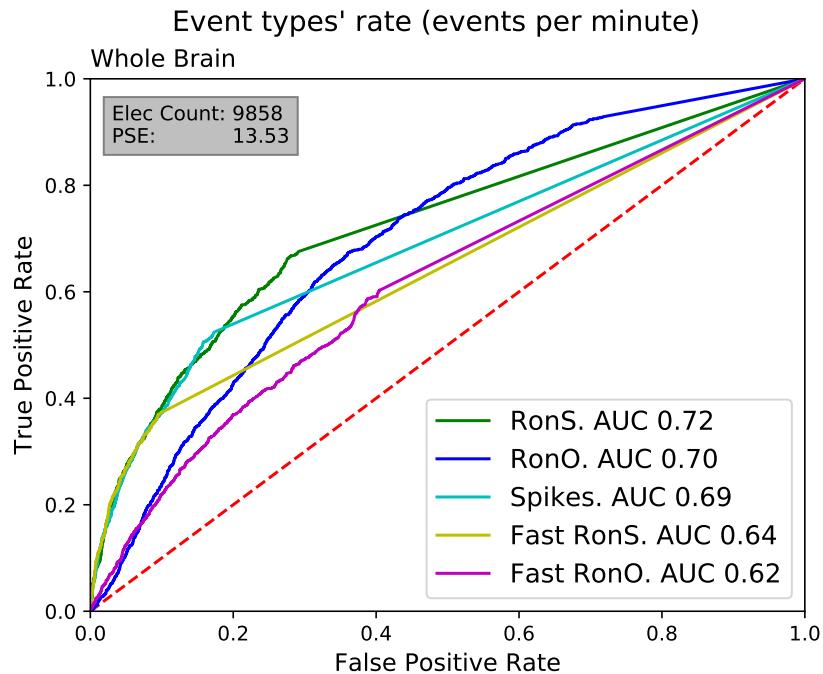


La imagen muestra que de esta manera, sin aprovechar la clasificación de tipos de los HFO, el HFO rate no resulta ser tan efectivo. Con un AUC de 0.71, supera por poco al 0.69 de Spikes, también se puede observar que la curva de Spikes es mejor para FPR pequeños, que son los más relevantes en nuestro problema ya que un falso positivo significa detectar como epileptógena una sección sana. Sin embargo el hecho de que el AUC de los HFOs sea mayor podría sugerir que en ciertas regiones es más un marcador más sensible.

3.5.2. Todo el cerebro

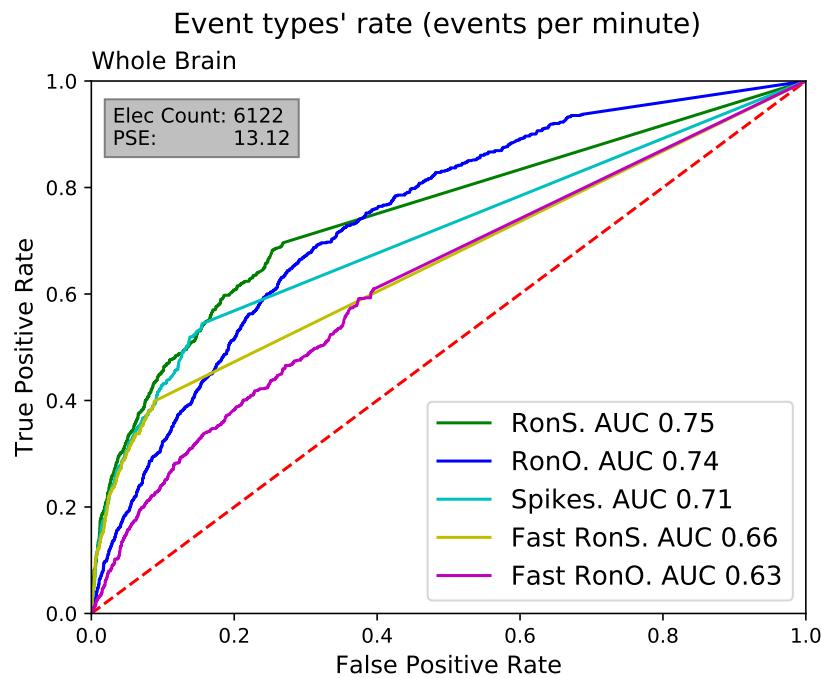
Como aproximadamente un 40 % de los electrodos tenían sus coordenadas neuroanatómicas y campos de localización en null a la hora de considerar todo el cerebro planteamos dos modelos, uno que cargara todos los electrodos, que imposibilita la utilización de dichos atributos posteriormente en el aprendizaje automático, y otro que sólo considerara electrodos tagueados, pero con la posibilidad de utilizar la localización como atributo en el entrenamiento del clasificador.

Permitiendo coordenadas en null



En esta figura podemos ver como utilizar una categoría independiente de HFO como los Ripple on Spikes pueden mejorar la performance de la métrica del rate aplicado a todos los HFO como una única clase (como vimos en la sección 3.5.1). RonS obtuvo valores de sensibilidad cercanos a 0.6 para un FPR de 0.2. A su vez, RonS resulta mejor que aplicar la técnica análoga a Spikes (usar spike rate) ya que tiene mejor sensibilidad global.

Permitiendo únicamente electrodos con coordenadas asignadas



En comparación con utilizar todos los electrodos, utilizar únicamente los que tuvieran coordenadas no nulas decrementa la cantidad de electrodos en todo el cerebro de 9658 a 6122, si bien nos quedan muchos datos para trabajar, es un porcentaje importante del corpus. A su vez las coordenadas podrían introducir sesgo ya que el número de pacientes no es tan elevado y dependería de las regiones epilépticas de cada paciente, por lo que consideramos no utilizar las coordenadas para los modelos localizados como el Hipocampo, sólo probamos ambos modelos para todo el cerebro en promedio.

3.5.3. Localizado

Primero calculamos el AUC ROC para cada región y subtipo de evento que tuviera más de 12 pacientes totales y 3 con epilepsia. Luego observamos si existía una relación con el porcentaje de electrodos soz, para ver si debíamos tener esto en cuenta a la hora de seleccionar regiones para implementar los clasificadores de HFOs utilizando aprendizaje automático. En tercer lugar graficamos las curvas ROC para la regiones que consideramos más relevantes.

LOC 2: El lóbulo Temporal registró los AUCs más bajos en promedio. Los AUCs de los **Ripples** se registraron cercanos a valores en el rango de 0.7 a 0.8, con un máximo de 0.81 en el Temporal mesial (Limbic Lobe). Asimismo, se los observó superiores a los **Fast Ripples** en todas las localizaciones. Como dato relevante para el campo de investigación de los HFOs, los Fast Ripples tuvieron relativamente buen rendimiento únicamente en el temporal mesial y en particular que los FronO. Esto se asocia a trabajos anteriores TODO referencia a los dos papers que menciono Shennan.

Granularity	Location	PSE	Fast RonO	Fast RonS	RonO	RonS
2	Temporal Lobe	11.65	0.53	0.58	0.65	0.69
2	Frontal Lobe	11.76	0.64	0.65	0.75	0.75
2	Parietal Lobe	15.11	0.59	0.63	0.75	0.72
2	Occipital Lobe	18.1	0.58	0.61	0.76	0.7
2	Limbic Lobe	19.51	0.72	0.74	0.78	0.81

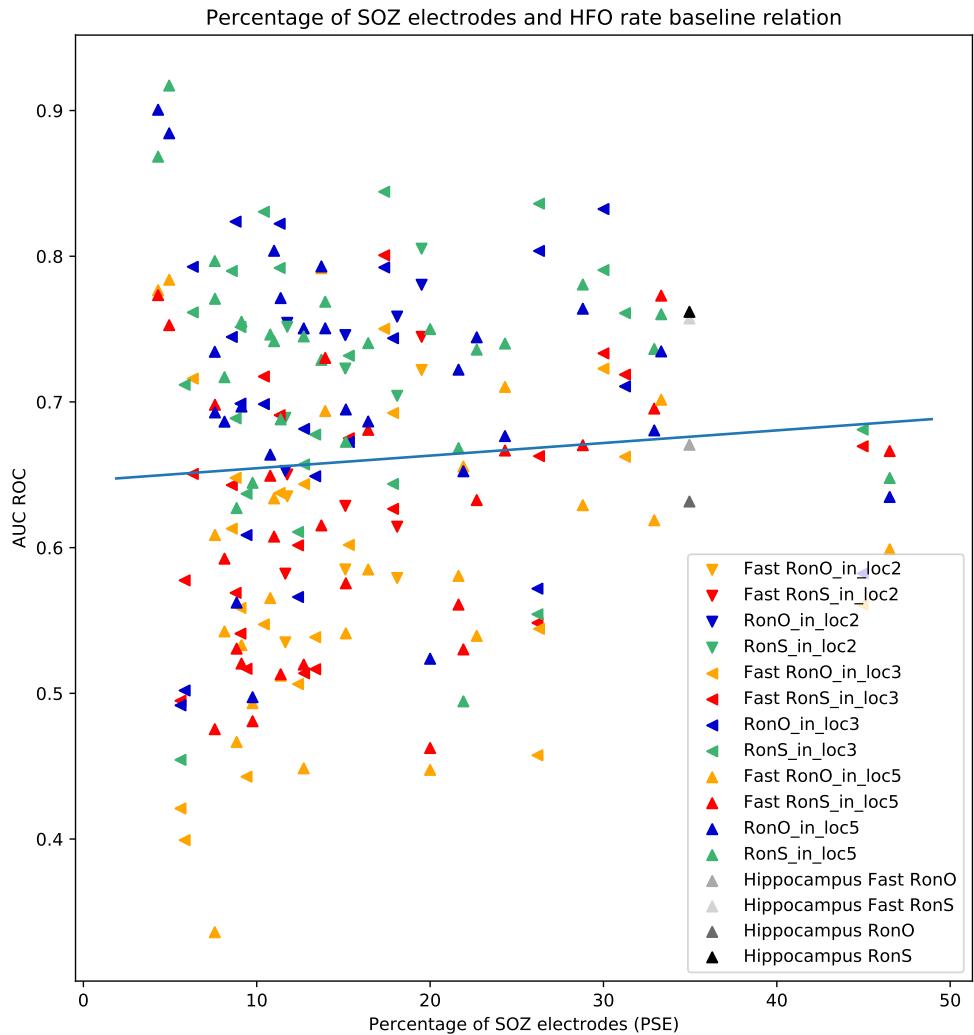
LOC 3: En general los RonS fueron el tipo con mejor rendimiento, es especial en regiones cercanas al hipocampo, con AUCs cercanos a 0.75. Los RonO superaron a los RonS en el Postcentral Gyrus, Inferior Frontal Gyrus, Medial Frontal Gyrus y Middle Occipital Gyrus, aunque mirando las tablas de dimensión de datos de la sección 3.2.2 vimos que eran regiones con pocos pacientes con electrodos SOZ (6 o menos). A su vez, observamos que los Fast RonS tuvieron rendimiento arriba del 0.7 en Parahippocampal Gyrus que es la región que en la sección 3.2.2 habíamos detectado cómo la única con una gran cantidad de FRonS (10 mil contra menos de mil en cualquier otra región). Otras regiones que registraron más de 0.7 de AUC para FRonS fueron Middle Frontal Gyrus, Middle Occipital Gyrus y Superior Frontal Gyrus, pero observando la tabla en 3.2.2 notamos nuevamente que no había tantos datos en dichas regiones como para aplicar machine learning. Por ejemplo Superior Frontal Gyrus registró el mejor AUC de 0.84 para RonS y 0.8 para FRonS, pero mirando la tabla de loc 3 en la sección 3.2.2, vimos que esta región sólo contaba con electrodos SOZ de 4 pacientes, más otros 16 estudios sin electrodos SOZ, y sólo 269 FRonS.

Granularity	Location	PSE	Fast RonO	Fast RonS	RonO	RonS
3	Clastrum	5.6	0.42	0.49	0.49	0.45
3	Paracentral Lobule	26.19	0.46	0.55	0.57	0.55
3	Inferior Temporal Gyrus	12.38	0.51	0.6	0.57	0.61
3	Middle Temporal Gyrus	9.41	0.44	0.52	0.61	0.64
3	Precentral Gyrus	12.73	0.64	0.51	0.68	0.66
3	Superior Temporal Gyrus	13.37	0.54	0.52	0.65	0.68
3	Uncus	44.95	0.56	0.67	0.58	0.68
3	Inferior Parietal Lobule	5.84	0.4	0.58	0.5	0.71
3	Sub-Gyral	15.31	0.6	0.68	0.67	0.73
3	Postcentral Gyrus	17.86	0.69	0.63	0.74	0.64
3	Insula	9.06	0.56	0.54	0.7	0.75
3	Parahippocampal Gyrus	31.24	0.66	0.72	0.71	0.76
3	Cingulate Gyrus	6.31	0.72	0.65	0.79	0.76
3	Fusiform Gyrus	8.56	0.61	0.64	0.74	0.79
3	Inferior Frontal Gyrus	8.78	0.65	0.57	0.82	0.69
3	Medial Frontal Gyrus	11.32	0.64	0.69	0.82	0.79
3	Middle Frontal Gyrus	10.41	0.55	0.72	0.7	0.83
3	Middle Occipital Gyrus	30	0.72	0.73	0.83	0.79
3	Precuneus	26.28	0.54	0.66	0.8	0.84
3	Superior Frontal Gyrus	17.35	0.75	0.8	0.79	0.84

LOC 5: Finalmente para las localizaciones de granularidad 5 (usando el campo loc5). Observamos nuevamente que los RonS fueron superiores en general por sobre los RonO, las regiones como BA 4, 24, 31, 32, y 39 registraron mayores AUC para RonO pero las tablas en [3.2.2](#) indicaron que tenían pocos pacientes con epilepsia, el más confiable de estos tres sería BA 31 con RonO, ya que tenía una buena cantidad de observaciones y al menos 9 pacientes con electrodos SOZ. Las áreas BA 35 y BA 6 fueron marcadas como considerables ya que tenían al menos 10 pacientes con electrodos SOZ y buena cantidad de datos y rendimiento en todos los tipos. Notar que son de las pocas regiones con valores significativos para los Fast Ripples. El Hipocampo fue seleccionado por tener muy buena cantidad de datos, buen rendimiento para RonS y Fast Ripples y porque se sabe que existen RonO fisiológicos en el Hipocampo vinculados a la fijación de memoria en sueño, por lo que resulta interesante evaluar el impacto de aplicar el aprendizaje automático. La Amígdala también resultó una zona de interés por su cantidad de datos, prevalencia de Fast Ripples. Las regiones BA 21 y 28 que habíamos mencionado anteriormente en [3.2.2](#) que tenían buena cantidad de datos no tuvieron buenos resultados, por lo que consideramos intercambiarlas por BA 6 y 35. BA 20 alcanzó los 0.74 con RonS y BA 36 los 0.78 y 0.76 con RonS y RonO respectivamente.

Granularity	Location	PSE	Fast RonO	Fast RonS	RonO	RonS
5	Brodmann area 21	8.84	0.47	0.53	0.56	0.63
5	Brodmann area 38	9.76	0.49	0.48	0.5	0.64
5	Brodmann area 3	21.92	0.66	0.53	0.65	0.49
5	Brodmann area 28	46.51	0.6	0.67	0.63	0.65
5	Brodmann area 22	15.14	0.54	0.58	0.69	0.67
5	Brodmann area 19	21.64	0.58	0.56	0.72	0.67
5	Brodmann area 37	8.14	0.54	0.59	0.69	0.72
5	Amygdala	32.93	0.62	0.7	0.68	0.74
5	Brodmann area 20	16.43	0.59	0.68	0.69	0.74
5	Brodmann area 34	24.32	0.71	0.67	0.68	0.74
5	Brodmann area 7	22.69	0.54	0.63	0.74	0.74
5	Brodmann area 41	20	0.45	0.46	0.52	0.75
5	Brodmann area 8	10.78	0.57	0.65	0.66	0.75
5	Brodmann area 9	12.71	0.45	0.52	0.75	0.74
5	Brodmann area 13	9.12	0.53	0.52	0.7	0.76
5	Hippocampus	34.96	0.67	0.76	0.63	0.76
5	Brodmann area 35	33.33	0.7	0.77	0.73	0.76
5	Brodmann area 39	11.38	0.51	0.51	0.77	0.69
5	Brodmann area 40	7.59	0.61	0.7	0.69	0.77
5	Brodmann area 6	13.95	0.69	0.73	0.75	0.77
5	Brodmann area 36	28.81	0.63	0.67	0.76	0.78
5	Brodmann area 4	13.73	0.79	0.62	0.79	0.73
5	Brodmann area 11	7.58	0.34	0.48	0.73	0.8
5	Brodmann area 31	11	0.63	0.61	0.8	0.74
5	Brodmann area 24	4.31	0.78	0.77	0.9	0.87
5	Brodmann area 32	4.95	0.78	0.75	0.88	0.92

A continuación presentamos un scatter plot que generamos para comprobar si existía una relación entre el porcentaje de electrodos SOZ y el AUC y visualizar la dispersión del AUC según el tipo.

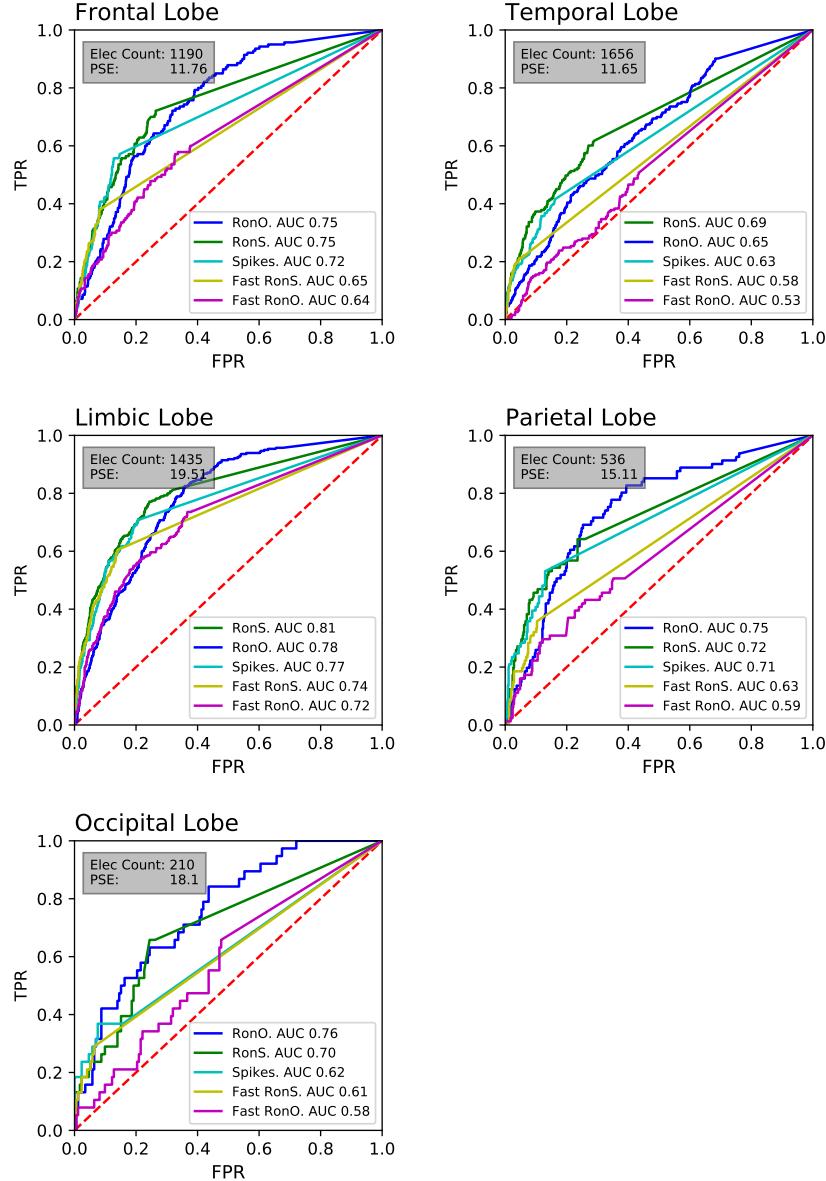


La figura mostró que no hay casi relación entre el porcentaje de electrodos con epilepsia y el AUC. A su vez, en general los RonS (triángulos verdes) tuvieron mejor rendimiento, seguidos por los RonO (triángulos azules), y luego los Fast Ripples (rojos y naranjas). Esto podría deberse a las referencias de la literatura que indican que los Fast Ripples suelen ser específicos pero poco sensibles, por lo que pueden no ser un gran marcador para todas las regiones del cerebro, representadas por cada triángulo.

ROCs

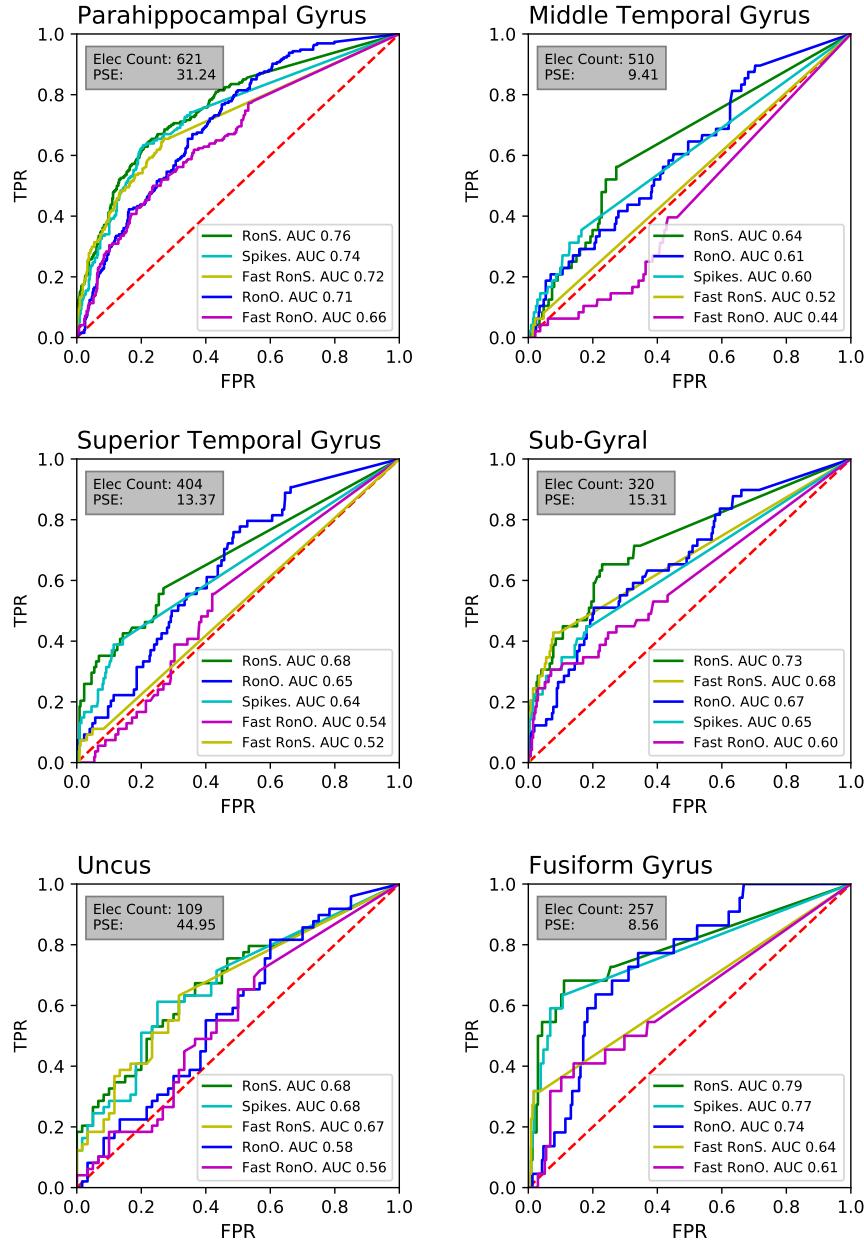
Elegimos 6 zonas para graficar de cada granularidad (excepto para loc 2 que tiene 5) en base al criterio de cantidad de datos mencionado en (3.2.2) y la sección anterior.

Event types' rate (events per minute)



Como valor agregado a las tablas, las curvas ROC nos permitieron ver qué tipos eran mejores para especificidades altas, que son las más relevantes a la hora de predecir la SOZ mediante el HFO rate. Observamos por ejemplo la importancia de los RonS y FRonS en Frontal y Limbic Lobe, mientras que para el Occipital Lobe resultaron con mejor rendimiento los RonO.

Event types' rate (events per minute)



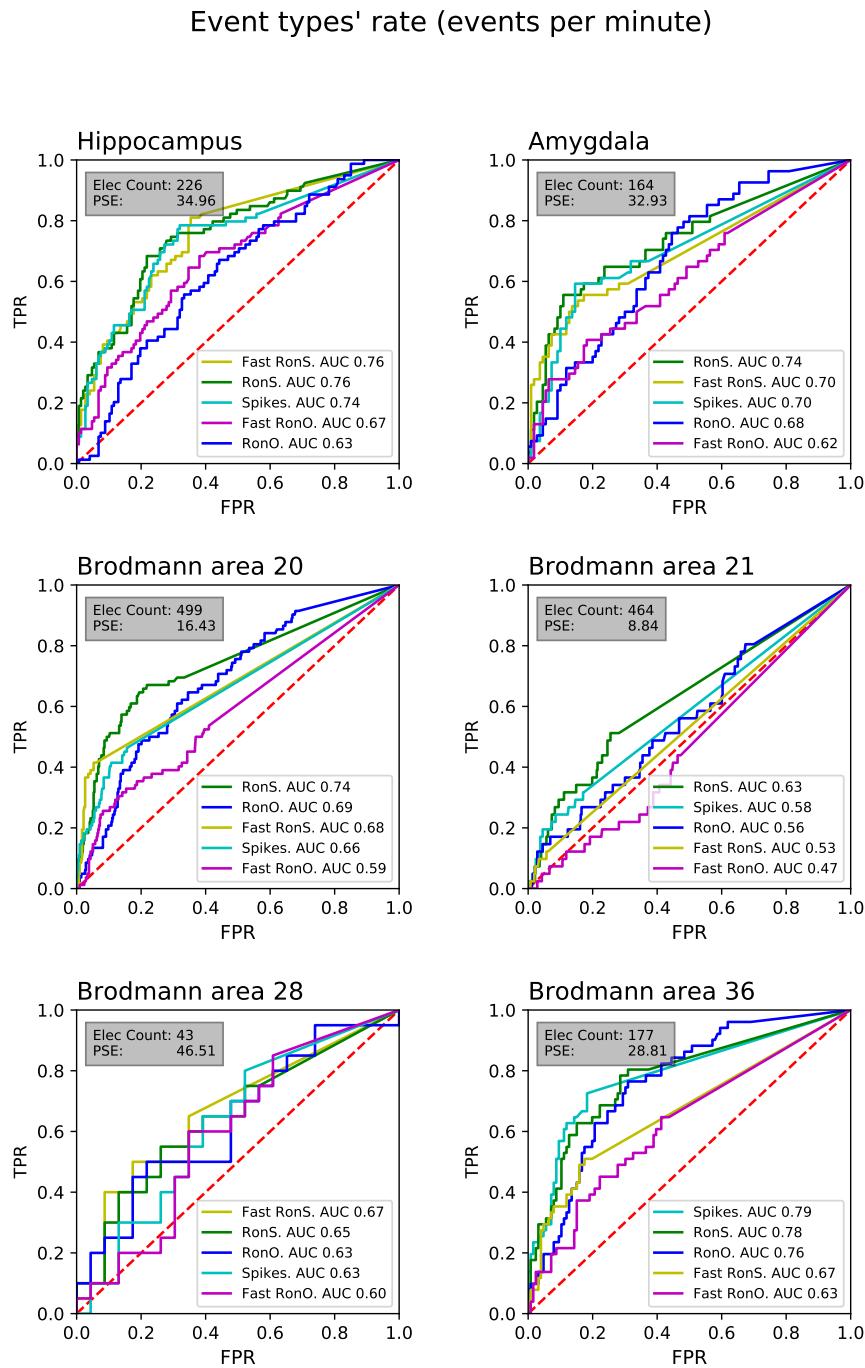


Fig. 3.14:

RonS alcanzó los mejores AUCs. BA 20 sólo tuvo buena performance con RonS. Mientras que para el Hipocampo también resultaron buenos marcadores los Fast Ripples on Spikes. Broadmann area 36 fue mejores valores obtuvo, con 0.78 para RonS y 0.76 para RonO.

3.6. Clasificación de HFOs mediante aprendizaje automático

Una vez calculados los baselines de la sección 3.5, se implementó clasificadores de HFOs mediante aprendizaje automático para regiones y tipos de HFOs que consideramos más significativos, bajo el criterio que mencionamos anteriormente: cantidad de datos en 3.2.2 y desempeño de la métrica como predictor de la SOZ 3.5.

En primer lugar se centró la atención en el armado de los conjuntos de pacientes que conformarían el grupo de validación, entrenamiento y test. Estos no debían tener eventos del mismo paciente entre grupos distintos de test, train o validación para evitar sesgo.

Se implementó una función para experimentar removiendo pacientes que estuvieran en dentro de un cuantil determinado en el ranking de cantidad de eventos. Por ejemplo remover a los pacientes que están en el primer cuartil en un ranking creciente en cantidad de eventos. Esta decisión fue debido a que notamos que algunos pacientes tenían muy pocos eventos comparados con otros, había un gran desvío estándar. Luego remover pacientes con muy pocas observaciones podría resultar beneficioso.

Validación: Dada una región R, separamos el 30 % de los pacientes con eventos en R separando primero pacientes con y sin electrodos en la SOZ, es decir, tomamos independientemente 30 % de los pacientes con cerebro sano en R y 30 % de los que habían registrado alguna convulsión en R. Esto fue para evitar la eventualidad de validar sólo con una clase, ya que generalmente la predominancia de una clase es significativa sobre la otra en una región dada. Los pacientes fueron elegidos de forma aleatoria si era la primera vez que se calculaban en dicha región y luego guardados en un json para persistir la selección. El resto de los pacientes quedarían disponibles para entrenar y testear el modelo.

Esto no genera sesgo en los pacientes que quedan para entrenar/testear?

Entrenamiento y Test: Una vez separados los pacientes de validación para R, usamos el resto de los pacientes para realizar cross validation. Se implementó dos funciones para crear conjuntos de nombres de pacientes, indicando con qué pacientes de debía testear en cada iteración y utilizando el complemento para entrenar. Se implementó k-fold cross validation y bootstrapping. La principal diferencia es que en K fold se predice una única vez cada evento y puede ser predicho por azar por un mal conjunto de pacientes, mientras que realizando bootstrapping podríamos tomar estadísticos como media, mediana o máximo entre varias predicciones. En cada iteración se testeó con 0.3 % de los pacientes disponibles (sin contar los ya excluidos para el conjunto de validación), lo que implica que se testea en cada iteración con un $0.3 * 0.7 = 0.21\%$ de los pacientes totales aproximadamente.

Selección de features: Definidos los conjuntos de Validación, Entrenamiento y Test para el modelo de R tradujimos los conjuntos de pacientes en conjuntos de features y labels (propiedades y etiquetas) para realizar las predicciones mediante algoritmos de aprendizaje automático. Para todos los tipos de HFOs consideramos la duración, frecuencia pico y promedio y densidad de potencia (power) pico y promedio. Los campos asociados a PAC (acoplamiento de amplitud de fase) varían según el tipo, para RonO y FRonO utilizamos slow, delta, theta y spindle, mientras que para RonS y FRonS sólo el spike angle. Los ángulos fueron utilizados mediante el seno y coseno para lidiar con los datos circulares. A su vez notamos que algunos grupos de entrenamiento tenían un desbalance significativo entre eventos SOZ y no SOZ por lo que se probó balancear el entrenamiento con algún algoritmo de resampling como SMOTETomek de la librería imblearn de Python, que realiza

oversampling de la clase minoritaria y subsampling de la mayoritaria. También se escaló los datos con la función StandardScaler o RobustScaler de sklearn.

Ver esto y
justificar

Se graficaron la distribución de las features en la SOZ y NSOZ y se realizaron los mismos tests estadísticos que en la sección 3.4. Esto reveló que los Fast RonO en todo el cerebro tenían una contaminación considerable de artefacto eléctrico para la frecuencia de 300 HZ, por lo que se consideró usar K-means para remover dichos eventos falsos, análogamente se observó este fenómeno aunque con menor contaminación para RonO en la frecuencia de 180 HZ. Las observaciones fueron tomadas en cuenta para seleccionar features según la zona y tipo.

Luego se experimentó con varios algoritmos de la librería sklearn y sus hiperparámetros para realizar las predicciones. Para evaluar el rendimiento del modelo se estudiaron las matrices de confusión y se utilizaron métricas como AUC, F1-score y AP de Precisión recall.

3.6.1. Todo el cerebro

Como explicamos en la sección 3.5 al considerar todo el cerebro planteamos dos modelos, permitiendo coordenadas null pero sin usarlas en el entrenamiento, o sólo utilizar electrodos con localización tagueada.

Utilizando coordenadas x, y, z en null

Distribución de los features y tests estadísticos

Duración

Frecuencia

Potencia

PAC

Entrenamiento permitiendo y utilizando coordenadas neuroanatómicas en null

Validación

Sin permitir coordenadas x, y, z en null

Distribución de los features y tests estadísticos

Duración

Frecuencia

Potencia

PAC

Entrenamiento sin permitir coordenadas neuroanatómicas en null***Validación*****3.6.2. Localizado**

Para los modelos localizados decidimos no utilizar las coordenadas neuroanatómicas para evitar sesgo en el caso de que la localización de los electrodos SOZ (ubicación de la epilepsia) entre los conjuntos de entrenamiento, test y validación difirieran.

Hipocampo***Distribución de los features y tests estadísticos******Duración******Frecuencia******Potencia******PAC******Entrenamiento******Validación*****3.7. Comparación del baseline y los modelos de rate filtrado****3.7.1. Todo el cerebro**

Coordenadas neuroanatómicas nulas permitidas

Coordenadas neuroanatómicas nulas no permitidas

3.7.2. Localizado

Esto claramente fue una limitación, ya que nuestro corpus de datos está basado en una parametrización de la señal y no contábamos con la señal cruda para extraer nuevas features.

3.8. Simulación del clasificador

Finalmente, se simuló la clasificación de HFOs a especificidades y sensibilidades pre-determinadas para entender la performance necesitada por el clasificador de HFOs para mejorar el baseline.

3.9. Driver

Para replicar los experimentos, a continuación se detallan los parámetros necesarios para la función `run_experiment` del módulo `driver.py`. Los parámetros de esta función son 3: número, número romano y letra, denotando un índice en la lista de experimentos presentada a continuación.

Lista de parámetros de la función `run_experiment()` del driver

1. **`run_experiment(number=1)`**: Imprime la lista y cantidad de pacientes en estado de sueño e intraoperatorios aludida en la sección [3.2.1](#). Guarda en los paths configurados en `config.py` las figuras de las tablas de cantidad de pacientes, electrodo y pacientes con epilepsia según la región neuroanatómica, referida en la sección [3.2.2](#)
2. I) **`run_experiment(number=2, roman_num='i')`**: Crea la figuras que comparan las distribuciones de los rate en la SOZ y NSOZ correspondientes a todo el cerebro a la vez, referidas en la sección [3.4.1](#).
II) **`run_experiment(number=2, roman_num='ii')`**: Crea la figuras que comparan las distribuciones de los rate en la SOZ y NSOZ correspondientes a regiones localizadas como el lóbulo Temporal o el Hipocampo, referidas en la sección [3.4.2](#).
3. I) a) **`run_experiment(number=3, roman_num='i', letter='a')`**: Crea las figuras correspondientes a las curvas ROC de la sección [3.5.2](#) Prediciendo la SOZ con los 4 tipos de HFOs discriminados y con el spike rate para determinar los baselines. Las coordenadas neuroanatómicas null se permiten pero no se utilizaran luego para el modelo de clasificación de HFOs que compare este baseline.
b) **`run_experiment(number=3, roman_num='i', letter='b')`**: Análogo al 3 i a sólo que permitiendo únicamente a los electrodos con coordenadas y localización neuroanatómica no nulos, para poder luego utilizarlos en los modelos de clasificación. Produce las figuras de la sección [3.5.2](#)
II) **`run_experiment(number=3, roman_num='ii')`**: Crea la tabla del AUC y porcentaje de electrodos SOZ por región y el gráfico de relación entre el PSE y el AUC ROC mencionado en la sección [3.5.3](#)
III) **`run_experiment(number=3, roman_num='iii')`**: Análogo al 3ib pero localizando la región de estudio en las relevantes según el experimento 3ii. Se presentan las figuras que genera en la sección [3.5.3](#)

Lista de parámetros de la función run_experiment() del driver (cont)

4. I) a) **run_experiment(number=4, roman_num='i', letter='a')**: Experimentos asociados al aprendizaje automático en todo el cerebro para los 4 tipos de HFOs, sin utilizar las coordenadas neuroanatómicas como features del modelo. Sección [3.6.1](#)
b) **run_experiment(number=4, roman_num='i', letter='b')**: Experimentos asociados al aprendizaje automático en todo el cerebro para los 4 tipos de HFOs, utilizando únicamente electrodos localizables y aprovechando las coordenadas neuroanatómicas como features para el aprendizaje del modelo. Sección [3.6.2](#)
II) **run_experiment(number=4, roman_num='ii')**: Análogo al 4 i b pero localizado por región. Presentado en la sección [3.6.2](#)
5. I) a) **run_experiment(number=5, roman_num='i', letter='a')**: Compara como predictores de la SOZ al baseline de la sección [3.5.2](#) con el modelo asociado al clasificador desarrollado en la sección [3.6.1](#). Genera las curvas ROC presentadas en la sección [3.7.1](#)
b) **run_experiment(number=5, roman_num='i', letter='a')**: Compara como predictores de la SOZ al baseline de la sección [3.5.2](#) con el modelo asociado al clasificador desarrollado en la sección [3.6.2](#). Genera las curvas ROC presentadas en la sección [3.7.1](#)
II) **run_experiment(number=5, roman_num='ii')**: Análogo al 5 i b pero localizado por región. Presentado en la sección [3.7.2](#)
6. **run_experiment(number=6)**: Genera las figuras presentadas en la sección [3.8](#), simula la clasificación de HFOs en el Hipocampo para valores predeterminados y grafica el impacto en el rate como predictor de la SOZ.

4. CONCLUSIONES

Se concluyó que en las zonas estudiadas (ACLARAR ZONAS) las features provistas por la base de datos no permiten distinguir lo suficiente entre la NSOZ y la SOZ, por lo cual el enfoque de un *filtro* de pHFOs resulta incapaz de mejorar la utilidad del HFO rate. No obstante, el simulador gradual demostró que para algunas regiones como el hipocampo, si encontráramos features más útiles para el ml, el baseline se superaría rápidamente. Ventajas y desventajas ? Trabajo futuro

Bibliografía

- [1] Alexis Arzimanoglou Alicia Bogacz J Helen Cross Christian E Elger Jerome Engel Jr Lars Forsgren Jacqueline A French Mike Glynn Dale C Hesdorffer B I Lee Gary W Mathern Solomon L Moshé Emilio Perucca Ingrid E Scheffer Torbjörn Tomson Masako Watanabe Samuel Wiebe Robert S Fisher 1, Carlos Acevedo. A practical clinical definition of epilepsy. *Epilepsia*, 2014.
- [2] Federico Raimondo Diego Slezak Mustafa Donmez Gregory Worrell Anatol Bragin Jerome Engel Richard Staba & Michael Sperling Shennan A Weiss*, Zachary Waldman. Localizing epileptogenic regions using high-frequency oscillations and machine learning. *Future Medicine*, 2019.