DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES

TEMA: PREDICCIÓN DE CRISIS EPILÉPTICAS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE SEÑALES ELECTROENCEFALOGRÁFICAS Y APRENDIZAJE DE MÁQUINA

AUTOR: QUINGA QUINGA, FRANCISCO DAVID

DIRECTOR: ING. CARRERA ERAZO, ENRIQUE V. PhD

SANGOLQUÍ

2017

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"PREDICCIÓN DE CRISIS EPILÉPTICAS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE SEÑALES ELECTROENCEFALOGRÁFICAS Y APRENDIZAJE DE MÁQUINA"** realizado por el señor **FRANCISCO DAVID QUINGA QUINGA**, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas – ESPE, por lo tanto me permito acreditarlo y autorizar al señor **FRANCISCO DAVID QUINGA QUINGA** para que lo sustente públicamente.

Sangolquí, 21 de Agosto del 2017

Ing. Enrique Vinicio Carrera Erazo, PhD

DIRECTOR

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORÍA DE RESPONSABILIDAD

Yo, **FRANCISCO DAVID QUINGA QUINGA**, con cédula de identidad Nº 1717265308 declaro que este trabajo de titulación "**PREDICCIÓN DE CRISIS EPILÉPTICAS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE SEÑALES ELECTROENCEFALOGRÁFICAS Y APRENDIZAJE DE MÁQUINA"** ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que es trabajo de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 21 de Agosto del 2017

Francisco David Quinga Quinga

C.C.: 1717265308

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **FRANCISCO DAVID QUINGA QUINGA**, autorizo a la Universidad de las Fuerzas Armadas – ESPE publicar en la biblioteca virtual de la institución el presente trabajo de titulación **"PREDICCIÓN DE CRISIS EPILÉPTICAS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE SEÑALES ELECTROENCEFALOGRÁFICAS Y APRENDIZAJE DE MÁQUINA"** cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 21 de Agosto del 2017

Francisco David Quinga Quinga

C.C.: 1717265308

DEDICATORIA

A Dios, porque sin él nada de esto habría sido posible, sus bendiciones siempre han sido parte de mis días y nunca me ha dejado rendir. Siempre me ha cuidado y no ha permitido que algo malo me suceda, porque sabe que siempre hay alguien esperándome en casa. Me ha demostrado que no estoy solo, que está conmigo en cada paso que doy. A él, por cada día, por cada experiencia, por cada lección, por mi familia, por mis facultades y por la vida. Me considero muy afortunado por todas las bendiciones que ha derramado sobre mí.

A la persona más importante en mi vida, por la que siempre he dado mi mejor esfuerzo, mi inspiración para salir adelante, mi apoyo en tiempos difíciles, la mujer que ha luchado día tras día y segundo a segundo por darme un mejor porvenir, aquella que nunca se rindió ante las adversidades de la vida, ni se dejó amedrentar por las circunstancias. Aquella que con fuerza, valor y enorme voluntad supo sacarme adelante sola, forjándome como un hombre de bien, con valores, ética y moral, enseñándome la importancia de luchar, de la perseverancia, de la valentía, del sacrificio, de ser diferente, de no conformarme, de ser mejor cada día, de valorar la vida y la familia, y sobre todo, de llevar siempre a Dios conmigo. Para María de los Ángeles Quinga, mi Madre, va dedicado este logro tan importante en mi vida.

"Solamente los que se arriesgan a llegar demasiado lejos son los que descubren hasta dónde pueden llegar" -T. S. Eliot-

"Ningún mar en calma hizo experto a un marinero" -Proverbio inglés-

Francisco David Quinga Quinga

AGRADECIMIENTO

A mi familia. A Jorge Q., por haber hecho las veces de padre durante mi infancia, y junto a Elsa D. siempre fueron un apoyo para nosotros a lo largo de este camino, tanto en los buenos como en los malos momentos. A Rafael Q., una de las personas que más admiro y respeto, por su trabajo duro y perseverancia por salir adelante, estoy enormemente agradecido por todo lo bueno que ha hecho por nosotros, y junto a Nancy M. por el cariño y aprecio que me han demostrado desde que era niño. A Pilar L., por haber sido parte de este camino largo y sacrificado, y haber contribuido a mi formación durante mi niñez, cuidándome durante muchos años, siendo para mí como una segunda madre. A mis primos, Christian, Lenin, Maxamed, Jhardel, Carolina y Robert, que al crecer juntos, para mí son como mis hermanos, gracias por cada momento compartido. Finalmente, a todas las personas que han aportado un grano de arena en mi vida.

A Belén, por haber sido mí confidente, mi apoyo y haber compartido conmigo todo este tiempo de etapa universitaria, viviendo momentos hermosos, bellas experiencias y recuerdos que vivirán para siempre conmigo.

A mis amigos, por cada momento compartido, cada anécdota vivida, cada sonrisa provocada, cada ocurrencia expresada, en fin, por cada momento gozado, que ahora son gratos recuerdos y hermosas experiencias.

De manera especial al Dr. Vinicio Carrera, por haberme guiado a lo largo de todo este proceso, con su experiencia, sabiduría, ayuda y consejos. Es una persona a la que admiro mucho, y me siento orgulloso de haber sido su alumno durante todo este trayecto. Sentimientos de gratitud, respeto y admiración hacia su persona.

A la vida, porque cada vivencia, cada experiencia y cada caída han hecho de mí una mejor persona, dicen que los errores son necesarios, porque traen consigo sabiduría y son la puerta a nuevas oportunidades. Este proyecto de investigación para mí fue un reto más en mi vida, demostrando las bases que desde niño he ido creando, con inteligencia, perseverancia, autosuficiencia y trabajo duro. Esto es apenas el inicio de un largo camino por recorrer, que con determinación sé que lo lograré.

ÍNDICE DE CONTENIDOS

CERTIFICACIÓN ii

AUTORÍA DE RESPONSABILIDAD iii

AUTORIZACIÓN iv

DEDICATORIA v

AGRADECIMIENTO vi

ÍNDICE DE CONTENIDOS vii

ÍNDICE DE TABLAS xi

ÍNDICE DE FIGURAS xii

RESUMEN xiii

ABSTRACT xiv

CAPÍTULO I 1

1. PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN 1

1.1 Antecedentes 1

1.2 Justificación e Importancia 2

1.3 Alcance del Proyecto 3

1.4 Estado del arte 4

1.5 Objetivos 5

1.5.1 Objetivo General 5

1.5.2 Objetivos Específicos 6

CAPÍTULO II 7

2. MARCO TEÓRICO 7

2.1 La Epilepsia 7

2.1.1 Causas 8

2.2 Electroencefalograma (EEG) 9

2.2.1 EEG del cuero cabelludo 9
2.2.2 EEG intracraneal 10
2.2.3 Bandas EEG 11
2.2.4 Artefactos 12
2.2.5 EEG en la epilepsia 12
2.3 Procesamiento de señales EEG 13
2.3.1 Método tiempo-escala y tiempo-frecuencia 14
2.3.2 Transformada *wavelet* discreta 15
2.3.3 Propiedades transformada *wavelet* discreta 17
2.4 Extracción de características 19
2.4.1 Características lineales 19
2.4.2 Característica no-lineal 23
2.5 Reducción de dimensionalidad 25
2.6 Aprendizaje de máquina 26
2.6.1 Regresión logística 27
2.6.2 Análisis discriminante 27
2.6.3 *k*-Vecinos más cercanos 29
2.6.4 Árboles de decisión 30
2.6.5 Máquinas de vectores de soporte 31
2.6.6 Conjunto de clasificadores redundantes 32
2.7 Validación cruzada *k-fold* 33
2.8 Medidas de desempeño 34
2.8.1 Matriz de confusión 34
2.8.2 Curva ROC 36
CAPÍTULO III 38
3. DISEÑO E IMPLEMENTACIÓN 38

3.1 *Software* y *hardware* 38

3.2 Esquema general del algoritmo 39

3.3 Base de datos EEG 40

3.4 Pre-procesamiento 42

3.4.1 Fragmentación de segmentos EEG en subsegmentos EEG 43

3.4.2 Selección de subsegmentos EEG por porcentaje de pérdidas 44

3.4.3 Balanceo del número de segmentos y subsegmentos EEG 45

3.5 Procesamiento 46

3.5.1 Enventanado de los subsegmentos EEG 46

3.5.2 Descomposición *wavelet* 48

3.5.3 Extracción de características 51

3.6 Etapa de clasificación y evaluación 54

3.6.1 Validación de resultados 55

3.6.2 Reducción de dimensionalidad 55

3.6.3 Configuración y entrenamiento de modelos 56

3.6.4 Evaluación de modelos 58

CAPÍTULO IV 59

4. RESULTADOS Y ANÁLISIS 59

4.1 Introducción 59

4.1 Casos de estudio 60

4.2 Selección de características (análisis lineal y no-lineal) 61

4.2.1 Análisis lineal 63

4.2.2 Análisis no-lineal 66

4.2.3 Resumen y análisis de resultados 69

4.3 Análisis en registros EEG de una hora de duración 71

4.4 Algoritmo empleado en una base de datos alternativa 73

CAPÍTULO V 76

5. CONCLUSIONES 76

5.1 Conclusiones 76

5.2 Recomendaciones 77

5.3 Trabajos Futuros 77

REFERENCIA 79

ÍNDICE DE TABLAS

Tabla 1 Definición matriz de confusión 34
Tabla 2 Elección de parámetros DWT óptimos 49
Tabla 3 Parámetros de configuración probados en cada clasificador 57
Tabla 4 Exactitud obtenida Análisis Lineal - Caso A - Pacientes 1 y 2 64
Tabla 5 Exactitud obtenida Análisis Lineal - Caso B - Pacientes 1 y 2 65
Tabla 6 Exactitud obtenida Análisis Lineal - Caso C 65
Tabla 7 Exactitud obtenida Análisis Lineal - Caso D 65
Tabla 8 Matriz de confusión Análisis Lineal - Caso D 66
Tabla 9 Exactitud obtenida Análisis No-lineal - Caso A - Pacientes 1 y 2 67
Tabla 10 Exactitud obtenida Análisis No-lineal - Caso B - Pacientes 1 y 2 68
Tabla 11 Exactitud obtenida Análisis No-lineal - Caso C 68
Tabla 12 Exactitud obtenida Análisis No-lineal - Caso D 68
Tabla 13 Matriz de confusión Análisis No-lineal - caso D 69
Tabla 14 Resumen de resultados Análisis Lineal y No-Lineal 70
Tabla 15 Número de segmentos EEG de 10 minutos totales analizados 72
Tabla 16 Resultados totales para clips EEG de una hora 72
Tabla 17 Número de segmentos EEG analizados en base de datos alternativa 74
Tabla 18 Resultados para clips EEG de una hora en base de datos alternativa 75

ÍNDICE DE FIGURAS

Figura 1. Posicionamiento sEEG y muestra de EEG de 10 segundos. 9
Figura 2. Malla de electrodos miniatura colocada en la superficie del cerebro. 11
Figura 3. Descomposición de sub-bandas en implementación DWT 16
Figura 4. Funciones *wavelet*: (a) Haar y (b) Daubechies 18
Figura 5. *k*-Vecinos más cercanos de un punto *x* para *k*=1,2,3 29
Figura 6. Partes de un árbol de decisión 30
Figura 7. Máquina de vectores de soporte en dos dimensiones 31
Figura 8. Ejemplo de límite de decisión utilizando *Ensemble* 33
Figura 9. Ejemplo de curva ROC 36
Figura 10. Esquema general del algoritmo de predicción de crisis epilépticas 39
Figura 11. Muestra clip preictal de 1 hora 41
Figura 12. Algoritmo de pre-procesamiento de los segmentos EEG 42
Figura 13. Representación subsegmento EEG 43
Figura 14. Algoritmo de procesamiento de los subsegmentos seleccionados 46
Figura 15. Muestra de ventanas de 1 segundo con solapamiento del 50% 47
Figura 16. (a) Descomposición *DWT* (b) Comparación con bandas EEG 50
Figura 17. Algoritmo de clasificación y evaluación de modelos 54
Figura 18. Representación caso de estudio A 60
Figura 19. Representación caso de estudio B 61
Figura 20. Representación caso de estudio C 61
Figura 21. Representación caso de estudio D 61
Figura 22. Curva ROC Análisis Lineal - caso D 66
Figura 23. Curva ROC Análisis No-lineal - caso D 69

RESUMEN

La epilepsia es una enfermedad que se caracteriza por un mal funcionamiento repentino y recurrente del cerebro, denominado "convulsión". Las convulsiones epilépticas reflejan los signos clínicos de una actividad excesiva e hipersincrónica de las neuronas en el cerebro, y pueden ir acompañadas de un deterioro o pérdida de la conciencia, síntomas psíquicos, sensoriales, o fenómenos motores. Uno de los aspectos más incapacitantes de la epilepsia es la naturaleza aparentemente impredecible de las convulsiones. Un método capaz de predecir la aparición de crisis epilépticas a partir del electroencefalograma (EEG) de pacientes con epilepsia abriría nuevas posibilidades terapéuticas. Con este enfoque, el presente proyecto contempla el desarrollo de un algoritmo de predicción de crisis epilépticas, usando técnicas de procesamiento de señales electroencefalográficas y aprendizaje de máquina. El algoritmo identifica el estado preictal (pre-crisis), y lo diferencia del estado interictal (sin crisis), basado en los cambios dinámicos espacio-temporales en el registro EEG, que empiezan varios minutos antes de la convulsión. Se emplea una base de datos que consta de registros EEG intracraneales (iEEG) de dos pacientes con epilepsia. Estos registros son analizados en el dominio tiempo-escala empleando la transformada *wavelet*, y extrayendo una serie de características. Para la clasificación se empleó aprendizaje de máquina supervisado, por medio de los siguientes algoritmos: regresión logística, análisis discriminante, k-vecinos más cercanos (kNN), árboles de decisión, máquina de vectores de soporte (SVM) y conjunto de clasificadores redundantes (*ensemble*). El algoritmo obtiene una exactitud total de 99.1%, sensibilidad de 99%, especificidad de 99.1% y capacidad predictiva (AUR) de 0.98.

Palabras claves

- **EPILEPSIA**
- **ELECTROENCEFALOGRAMA**
- **PROCESAMIENTO DE SEÑALES**
- **APRENDIZAJE DE MÁQUINA**

ABSTRACT

Epilepsy is a disease characterized by a sudden and recurring malfunction of the brain, called "seizure". Epileptic seizures reflect clinical signs of excessive and hypersynchronous activity of neurons in the brain, and may be accompanied by deterioration or loss of consciousness, psychic, sensory symptoms, or motor phenomena. One of the most disabling aspects of epilepsy is the seemingly unpredictable nature of seizures. A method capable of predicting the occurrence of epileptic seizures from the electroencephalogram (EEG) of patients with epilepsy would open new therapeutic possibilities. With this approach, the present project contemplates the development of an algorithm for predicting epileptic seizures, using electroencephalographic signal processing techniques and machine learning. The algorithm identifies the preictal (pre-seizure) state, and differentiates it from the interictal (seizure free) state, based on the spatiotemporal dynamic changes in the EEG, which begin several minutes before the seizure. A database consisting of intracranial EEG (iEEG) registers of two patients with epilepsy is used. These registers are analyzed in the time-scale domain using the wavelet transform, and extracting a group of features. For the classification, we used supervised machine learning, using the following algorithms: logistic regression, discriminant analysis, k-nearest neighbors (kNN), decision trees, support vector machine (SVM) and ensemble. The algorithm obtains overall accuracy 99.1%, sensitivity 99%, specificity 99.1% and predictive capacity (AUR) 0.98.

Keywords:

- **EPILEPSY**
- **ELECTROENCEPHALOGRAM**
- **SIGNAL PROCESSING**
- **MACHINE LEARNING**

CAPÍTULO I

1. PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1 Antecedentes

La epilepsia es un grupo de trastornos cerebrales caracterizados por la recurrencia de crisis con descargas eléctricas en la corteza cerebral, que dan lugar a alteraciones irregulares de las funciones cerebrales (Gevins & Remond, 1987). Esta enfermedad no es solo una condición; más bien es una familia diversa de trastornos que comprenden muchos tipos de convulsiones. El 10% de la población está en riesgo de sufrir una convulsión durante su vida (Epilepsy Australia, 2015). De tal manera que alrededor del mundo este trastorno está siendo motivo de estudio en ámbitos de investigación biomédica, con el propósito de entender mejor sus mecanismos, causas y posibles opciones terapéuticas.

En el diagnóstico de la epilepsia se realiza un examen neurológico que generalmente incluye una prueba para medir la actividad eléctrica del cerebro por medio del electroencefalograma (EEG) (Epilepsy Australia, 2015). En una crisis epiléptica, los cambios dinámicos espacio-temporales en el registro EEG empiezan varios minutos antes y terminan varios minutos después de la convulsión (Gevins & Remond, 1987). Existe evidencia emergente de que la dinámica temporal de la actividad cerebral en una crisis puede clasificarse en cuatro estados: Interictal (entre convulsiones), Preictal (antes de la convulsión), Ictal (convulsión) y Postictal (después de la convulsión). El reto principal en el pronóstico de las convulsiones es diferenciar entre los estados preictal e interictal (Gadhoumi, Lina, Mormann & Gotman, 2016).

La mayoría de los estudios presentan solución al problema de detección de convulsiones en el contexto de un sistema de apoyo en la toma de decisiones para un experto en neurología. Debido a la existencia de muchos tipos de ataques epilépticos, esto es a veces una tarea difícil, teniendo en cuenta la naturaleza, longitud y singularidades de cada tipo de convulsión. Una tarea aún más exigente, que es todavía

considerada una pregunta científica abierta, es la predicción de una convulsión, que mejoraría profundamente la calidad de vida de las personas que sufren convulsiones epilépticas severas. Además, la comprensión de los mecanismos subyacentes que conducen a las convulsiones y el origen de las mismas están todavía bajo estudio (Winterhalder et al., 2003).

1.2 Justificación e Importancia

Actualmente, más de 50 millones de personas en todo el mundo están afectadas por la epilepsia, asociada a un trastorno cerebral con consecuencias neurobiológicas, cognitivas, psicológicas y sociales. Es conocida por transiciones repentinas e inesperadas desde estados normales a patológicos llamados convulsiones epilépticas, y puede afectar igualmente a todos los grupos de edad (Dua, De Boer, Prilipko & Saxena, 2006).

La epilepsia es causada por la predisposición debido a factores genéticos, tumores, uso indebido de drogas, etc. Aunque hay medicamentos disponibles para prevenir las convulsiones, sólo aproximadamente el 70% de los pacientes responden a estos fármacos (Epilepsy Australia, 2015). Otros tratamientos, como la cirugía, neuroestimulación, y el cambio en la dieta no siempre tienen éxito (Kossoff, 2011). Además de los problemas de salud debido a la epilepsia, los pacientes sufren ansiedad persistente de experimentar futuras convulsiones. También tienen muchas restricciones en su vida cotidiana, como la incapacidad de conducir un automóvil (Mormann, Andrzejak, Elger & Lehnertz, 2007).

La predicción fiable de futuras convulsiones beneficiaría en gran medida a muchos pacientes. Las primeras aplicaciones probablemente estarían situadas en entornos clínicos, donde el equipo de medición y personal capacitado esté disponible. Conociendo la probabilidad de producirse una convulsión dentro de un tiempo determinado, los médicos y enfermeras podrían adaptar mejor los métodos de tratamiento al paciente.

Es así que la posible predicción de crisis epilépticas sería una herramienta sumamente útil para el monitoreo y posterior tratamiento de pacientes con epilepsia,

ya sea en ambientes hospitalarios o caseros, por medio de dispositivos portátiles de alarma. Permitiendo al médico o paciente tomar las medidas necesarias para evitar riesgos antes de que se presente una crisis.

1.3 Alcance del Proyecto

El presente trabajo de investigación tiene como objetivo, basado en registros de EEG intracraneal (iEEG) y empleando la herramienta de software *MatLab*®: identificar el estado preictal, que precede a una crisis epiléptica y diferenciarlo del estado interictal, presente entre convulsiones, extrayendo las características de cada estado, por medio de técnicas de procesamiento de señales y realizando su correcta clasificación, empleando algoritmos de aprendizaje de máquina.

Para el desarrollo del algoritmo de predicción de crisis epilépticas se utilizó una base de datos de dominio público, que consta de varias horas de señales iEEG de dos pacientes, divididas en registros preictales e interictales.

Los registros iEEG se someten a una etapa de pre-procesamiento, con el fin de eliminar del estudio aquellos que tengan excesivo número de pérdidas o ceros (*drops*). Una vez pre-procesados los registros, se procede a su extracción de características, empleando para ello el dominio tiempo-escala, por medio de la transformada *wavelet*, que es una herramienta matemática que puede utilizarse para el análisis de señales no estacionarias tales como el EEG (Crowe, Gibson, Woolfson & Somekh, 1992). Cada registro contiene un número N de ventanas de l-segundos, y cada ventana es descompuesta en n sub-bandas *wavelet*. De cada sub-banda se extrae las características que mejor desempeño produjeron de entre varias.

La clasificación se realiza empleando los siguientes algoritmos de aprendizaje de máquina: Regresión logística, Análisis discriminante, k-Vecinos más cercanos (kNN), Árboles de decisión, Máquina de Vectores de Soporte (SVM) y Conjunto de clasificadores redundantes (*Ensemble*), y verificando cuál de estos presenta mejor desempeño.

Finalmente para evaluar el rendimiento del algoritmo se utiliza las siguientes medidas de desempeño: exactitud, sensibilidad, especificidad y área bajo la curva ROC (AUR). Se plantea que el algoritmo obtenga una exactitud, sensibilidad y especificidad mayor al 80%.

1.4 Estado del arte

Con este enfoque, se han propuesto muchos algoritmos de análisis de registros EEG. El análisis lineal ha sido ampliamente utilizado, basado principalmente en características de sincronización como un primer enfoque. Aunque estos métodos pueden revelar en algunos casos la existencia de ataques epilépticos, tienen sus limitaciones si alguien toma en cuenta la naturaleza real de los datos del EEG humano (Kantz & Schreiber, 2004). De esta manera, se han realizado también estudios de la dinámica no lineal del EEG humano (Faure & Korn, 2001).

Después de algunos primeros trabajos sobre la previsibilidad de las convulsiones que se remontan a la década de 1970, intentos de extraer precursores de convulsiones de registros EEG del cuero cabelludo (sEEG) fueron llevados a cabo por diferentes grupos empleando enfoques lineales (Viglione & Walsh, 1975).

El estudio de la predicción de las crisis epilépticas se puede dividir en tres clases: dominio del tiempo, frecuencia y métodos no lineales. En el dominio del tiempo, los patrones periódicos y rítmicos en el EEG se buscan para analizar la autocorrelación del EEG para proporcionar una medida rítmica, que es similar a lo que se produce durante la actividad de convulsión. En el dominio de la frecuencia, la predicción de convulsiones se basa en el análisis del espectro dividiendo las diferentes frecuencias del EEG. Estos dos métodos son lineales. Pero las investigaciones afirman que las señales EEG poseen dinámica no lineal, no estacionaria, e incluso caótica (Chaovalitwongse et al., 2011).

Nasehi & Pourghassem (2011) proponen un algoritmo de predicción de convulsiones basado en clasificadores *Kernel Fisher Discriminant* (KFD). En este algoritmo, las características espectrales son extraídas usando la transformada *wavelet* de señales EEG convulsivas y no-convulsivas. Se emplea el método de validación

cruzada *leave-one-out*. El rendimiento del algoritmo es evaluado en base a cuatro medidas: precisión, tasa de falsas detecciones (FDR), tasa de correctas detecciones (GDR) y retraso. Los resultados ilustran que el algoritmo puede reconocer 81 ataques epilépticos de un total de 87, con un retraso medio de 3,7 segundos.

Zeljkovic et al. (2013) presentan un enfoque que implica la colocación de electrodos en las regiones críticas de la cabeza del paciente con epilepsia que se comunicaría con un dispositivo registrador EEG y la unidad que realiza la detección automatizada del estado preictal de la señal EEG obtenida. La posible alerta temprana del paciente proporcionaría una alarma oportuna sobre un posible ataque, para que este pueda tomar precauciones de seguridad. La información sobre la fase preictal se obtiene aplicando varias transformaciones matemáticas y modelos a la señal EEG original.

Elgohary, Eldawlatly & Khalil (2016) proponen un nuevo método específico por paciente para la predicción de convulsiones, aplicado a señales de electroencefalografía del cuero cabelludo (sEEG). El método propuesto se basa en el recuento de cruces por cero de los coeficientes detallados de las señales EEG como la característica principal. Esto es seguido por un clasificador binario que discrimina entre estados preictal e interictal. El método propuesto es práctico para aplicaciones en tiempo real dada su eficiencia computacional. Aplicado a los datos de ocho pacientes, el método propuesto alcanza alta precisión y sensibilidad, con una precisión media del 94% y una sensibilidad media del 96%. Estos resultados se obtuvieron usando sólo diez minutos de datos de entrenamiento, en contraposición a usar horas de grabaciones típicamente usadas en enfoques tradicionales.

1.5 Objetivos

1.5.1 Objetivo General

Diseñar, implementar y evaluar un algoritmo de predicción de crisis epilépticas, utilizando técnicas de procesamiento de señales y aprendizaje de máquina en registros electroencefalográficos, para la correcta detección del estado preictal.

1.5.2 Objetivos Específicos

- Estudiar las técnicas actuales e investigaciones realizadas en el campo de la predicción de crisis epilépticas.
- Diseñar los componentes del algoritmo e implementarlos en el software *MatLab,* utilizando procesamiento de señales electroencefalográficas y aprendizaje de máquina.
- Evaluar el rendimiento del algoritmo, usando medidas estadísticas de desempeño como: sensibilidad, exactitud, especificidad y AUR.

CAPÍTULO II

2. MARCO TEÓRICO

2.1 La Epilepsia

La epilepsia es un término general para varias condiciones en las que dos o más convulsiones son observadas. Una convulsión ocurre cuando se da una hiperexcitabilidad de las células de la corteza cerebral (neuronas), debido a la perturbación eléctrica repentina en el cerebro, y es considerada el segundo trastorno neurológico más poderoso después del accidente cerebrovascular. Hasta el siglo XIX la epilepsia no fue reconocida como un trastorno cerebral (Annegers et al., 1980).

La diferencia entre convulsiones y la epilepsia está en que se puede definir a la epilepsia como una serie de convulsiones repetitivas. Las convulsiones pueden causar daños al cerebro, pero la mayoría de ellas no tienen un efecto sobre este, y pueden ocurrir a cualquier edad, con una duración que va desde unos pocos segundos hasta unos minutos (Hauser & Hesdorffer, 1990).

Las convulsiones pueden ser divididas en diferentes categorías. Esta división se basa en el área del cerebro que se ve afectada como consecuencia de la convulsión. Las convulsiones se clasifican en términos generales como crisis focales y convulsiones generalizadas. Las convulsiones focales o convulsiones parciales se producen en una determinada parte del cerebro, no en el cerebro completo. En un ataque focal simple, una persona permanece consciente pero experimenta una sensación inusual de alegría, ira o tristeza. Sin embargo, en un ataque focal complejo, la persona pierde su conciencia.

La crisis producida en ambos lados del cerebro se denomina convulsión generalizada. Crisis tónico-clónica, crisis de ausencia, crisis mioclónica, crisis atónica son algunos tipos de convulsiones generalizadas. Hay un límite bastante delgado entre convulsiones focales y generalizadas. A veces una convulsión focal inicial puede extenderse por todo el cerebro para formar una convulsión generalizada. Algunas

personas pueden tener ambos tipos de convulsiones, pero sin un patrón claro entre ellas. Las estadísticas muestran que el 60% de los pacientes epilépticos tienen convulsiones en las cuales los músculos del cuerpo se contraen y expanden rápidamente sin intervención del cerebro humano. Tales crisis se llaman ataques convulsivos (Brown et al., 1998).

2.1.1 Causas

El número de casos de epilepsia reportados cada año es de aproximadamente 180.000, de los cuales el 30% son niños. Diferentes causas de la epilepsia pueden ser observadas en diferentes edades. La principal causa de las crisis epilépticas son trastornos genéticos, accidentes cerebrovasculares y tumores cerebrales, aunque las causas de las convulsiones son a veces de naturaleza no determinista. Alrededor del 30% de personas con epilepsia la adquieren debido al cambio en la estructura del cerebro. El trastorno del espectro autista (TEA) en niños conduce a la epilepsia, sin embargo sus causas todavía están por determinar. La falta de oxígeno en los recién nacidos y las infecciones en el caso de los niños y bebés también se convierte en una razón para la epilepsia (National Institute of Neurological Disorders, & Stroke, 2004).

La epilepsia se clasifica de la siguiente manera: epilepsia sintomática, epilepsia idiopática y epilepsia criptogénica. Algunas de las causas específicas como las drogas, el alcohol y las anormalidades de nacimiento pueden también causar epilepsia sintomática. La epilepsia idiopática se produce debido a causas no conocidas. Un daño cerebral puede resultar en una epilepsia criptogénica. La epilepsia también puede ocurrir como resultado de niveles anormales de sodio, azúcar o sangre (Shorvon, 2011).

Las convulsiones y la epilepsia pueden ser evitadas evadiendo lo siguiente: uso de alcohol, abuso de drogas, la falta de sueño adecuado, ansiedad/estrés. La epilepsia afecta negativamente a la salud del paciente. La ansiedad, depresión, pérdida de memoria y lesiones mentales son síntomas de esta enfermedad (Adewuya & Ola, 2005).

2.2 Electroencefalograma (EEG)

El cerebro humano tiene dos partes: el hemisferio izquierdo y el hemisferio derecho. El hemisferio izquierdo tiene el control del lado derecho del cuerpo, mientras que el hemisferio derecho controla el lado izquierdo del cuerpo. Una parte del cerebro está presente para controlar la visión y alguna parte del cerebro controla la memoria. Las neuronas presentes en el cerebro producen ciertas señales eléctricas y químicas para llevar a cabo ciertas funciones. Estas señales son captadas por el electroencefalograma (EEG) (Annett, 1985).

Los electrodos utilizados en el proceso de EEG gobiernan el tipo de grabación a ser examinada. Hay dos tipos de lecturas de EEG: EEG de cuero cabelludo y EEG intracraneal.

2.2.1 EEG del cuero cabelludo

El EEG del cuero cabelludo (*scalp* EEG, sEEG) es una medida no invasiva de los potenciales eléctricos generados por la actividad de decenas de millones de neuronas dentro del cerebro. El sEEG suele medirse a través de electrodos que están dispuestos simétricamente en el cuero cabelludo como se muestra en la figura 1.

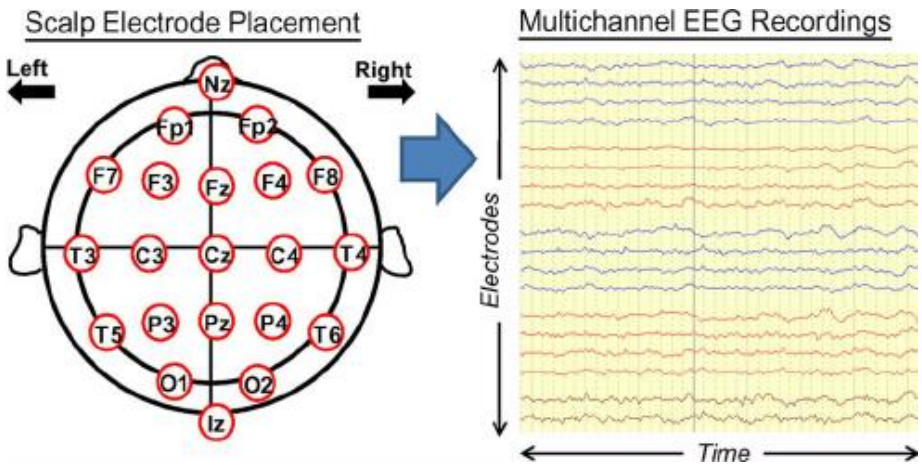


Figura 1. Posicionamiento sEEG y muestra de EEG de 10 segundos.
Fuente: (Chaovalitwongse, Pottenger, Wang, Fan & Iasemidis, 2011)

Un canal sEEG, se forma tomando la diferencia entre los potenciales medidos en dos electrodos. Por ejemplo, el canal FP1-F7 se forma tomando la diferencia entre los potenciales medidos en los electrodos FP1 y F7. Cada canal sEEG resume la actividad localizada dentro de una región del cerebro, así, el canal FP1 - F7 refleja la actividad neuronal que se origina dentro del lóbulo frontal del hemisferio izquierdo. El inicio de una convulsión focal implica un cambio en la actividad en los pocos canales del sEEG que se encuentran por encima o cerca del sitio del cerebro que da lugar a una convulsión. Por otro lado, el inicio de una convulsión generalizada implica la actividad en todos los canales del sEEG (Guger et al., 2001).

La física de la generación del sEEG restringe el origen y las características de la actividad neuronal visible dentro del EEG. En particular, las neuronas que contribuyen más al sEEG son las más cercanas a la superficie del cuero cabelludo. En contraste, la actividad de las neuronas enterradas dentro de las estructuras profundas del cerebro no es observable. Además, el líquido cefalorraquídeo y el cráneo que rodea el cerebro actúan como atenuadores que disminuyen en gran medida la amplitud de las oscilaciones neuronales de mayor frecuencia (Guger et al., 2001).

Una consecuencia importante de estas limitaciones físicas es que ciertos tipos de convulsiones que implican una región pequeña y profunda dentro del cerebro no pueden ser observadas usando el sEEG.

2.2.2 EEG intracraneal

Similar al EEG del cuero cabelludo (sEEG), el EEG intracraneal (iEEG) o Electrocorticograma (ECoG) proporciona un resumen espacial y temporal de la actividad eléctrica de una población de neuronas. Dado que los electrodos del iEEG se colocan en la corteza cerebral o profunda dentro de las estructuras cerebrales, la señal EEG resume la actividad de una población más pequeña de neuronas y es inmune a la corrupción por muchos de los factores fisiológicos y ambientales (artefactos) que plagan el sEEG. Debido a que son altamente invasivos, los electrodos iEEG sólo pueden implantarse en un número limitado de sitios cerebrales en un momento dado.

Por lo que el iEEG ofrece mayor resolución espacial, pero la cobertura espacial es menor que la del sEEG (Pacia & Ebersole, 1997).

La mayor resolución espacial de iEEG permite notar la hipersincronía neuronal asociada con una convulsión decenas de segundos antes de que el mismo fenómeno se note en el sEEG (Pacia & Ebersole, 1997).

Al mismo tiempo, la mayor resolución espacial de iEEG permite el registro de una gama más amplia de actividad anormal no convulsiva que no es visible dentro del sEEG (Jirsch et al, 2006). Esta actividad puede confundir el proceso de detección de ataques epilépticos.

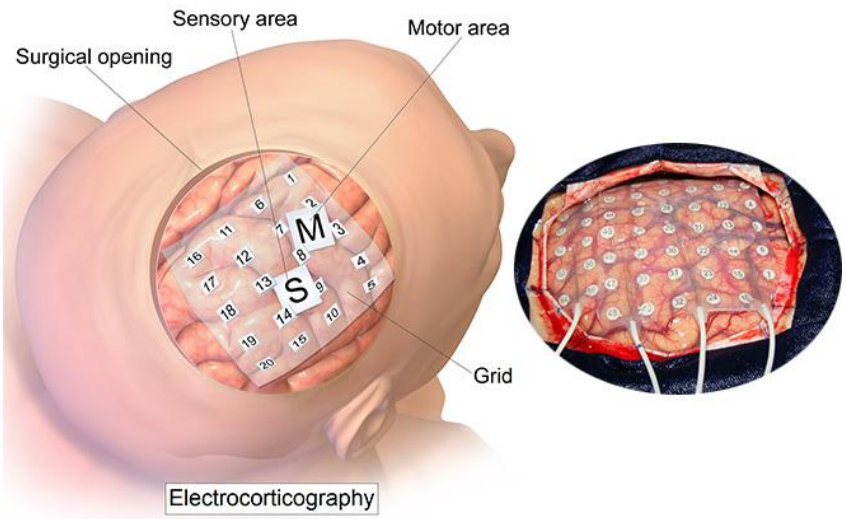


Figura 2. Malla de electrodos miniatura colocada en la superficie del cerebro.
Fuente: (MedicWiz, 2016)

La figura 2 ilustra la ubicación de los electrodos en la corteza cerebral para la grabación de registros iEEG.

2.2.3 Bandas EEG

Por lo general, se observan diferentes bandas de frecuencias donde cada banda está ligada a una actividad especial del cuerpo humano. Estos grupos y sus actividades se exponen a continuación:

La actividad más lenta, aunque más alta en amplitud, se llama Delta. Su frecuencia es 3 Hz o menos. Es prominente en el lóbulo frontal en adultos y bebés. La actividad lenta se denomina Teta. Su rango de frecuencia va de 3,5 a 7 Hz. Aparece en los niños normales, hasta la edad de 13 años, y en el sueño. La actividad Alfa aparece al cierre de los ojos o en estado relajación y desaparece al abrir los ojos o en estado de alerta. Su frecuencia es normalmente de 7,5 a 13 Hz. La actividad rápida con un rango de frecuencia de 14 a 30 Hz se llama actividad Beta. Se encuentra generalmente ausente en el área con lesión cerebral. La actividad Gama gobierna las funciones motoras. Su frecuencia varía de 30 a 100 Hz (Başar, Başar-Eroglu, Karakaş & Schürmann, 2001).

2.2.4 Artefactos

En un registro EEG normal pueden aparecer distintas señales denominadas artefactos, que son señales eléctricas producto de múltiples causas, tales como: movimientos musculares, oculares, respiración, pulso, resistencia de la piel, sudor, electrodos, problemas técnicos, línea de alimentación, etc. Esto no significa que haya patología.

Los artefactos, especialmente los relacionados con el movimiento ocular, parecen ser el principal inconveniente con el que los investigadores y los médicos tienen que lidiar todos los días (Park, Luo, Parhi & Netoff, 2011).

Aunque los artefactos en el EEG intracraneal no son tan importantes como en el EEG del cuero cabelludo, los datos que el iEEG proporciona también se someten a algunos artefactos, que afectaría el procesamiento de las señales durante el entrenamiento y la prueba (Park et al., 2011).

2.2.5 EEG en la epilepsia

El EEG en la epilepsia da una clara representación de una descarga brusca, excesiva, desordenada y recurrente, de las neuronas corticales. El EEG es la herramienta más sensible a la hora de diagnosticar a un paciente con epilepsia, pero debe ser complementado con el diagnóstico clínico. Muchos pacientes epilépticos

tienen un EEG interictal perfectamente normal, incluso durante una convulsión parcial compleja. Por otro lado, en algunas personas normales se presentan paroxismos epileptiformes en el EEG, y algunos de ellos tienen antecedentes familiares de epilepsia, por lo que pueden desarrollar convulsiones más tarde (Gadhoumi et al, 2016).

El EEG es de gran ayuda en el diagnóstico de la epilepsia y su clasificación. Sin embargo, es necesario conocer y delimitar el alcance y limitaciones que presenta esta técnica, así como también realizarla en las condiciones técnicas adecuadas, y junto con el diagnóstico clínico respectivo para evitar interpretaciones erróneas que conlleven a falsos diagnósticos.

2.3 Procesamiento de señales EEG

Las señales EEG, como la mayoría de las señales biológicas, son intrínsecamente difíciles de cuantificar. Estas señales pueden ser caracterizadas como no-estacionarias y a menudo con una relación señal-a-ruido (*signal-to-noise,* SNR) baja. Son de naturaleza irregular y nunca se replican exactamente. Por lo tanto, este tipo de señales no pueden ser razonablemente descritas como señales deterministas y deben ser estudiadas mediante métodos estadísticos (Roessgen, 1996).

La señal EEG ha sido hasta ahora usada para el estudio de las disfunciones cerebrales tales como la actividad epiléptica, detección de convulsiones, detección de husos de sueño, mapeo cerebral, potenciales evocados, trastornos del sueño y niveles de anestesia.

Los métodos disponibles para la predicción y/o detección automática de ataques epilépticos en la literatura se pueden dividir en dos clases principales. La primera clase consta de métodos paramétricos que asumen un modelo para las señales EEG y luego se estima los parámetros del modelo. Varios métodos paramétricos para la detección de las crisis se pueden encontrar en (Wright, Kydd & Sergejew, 1990)

La segunda clase incluye métodos no paramétricos que no requieren un modelo de la señal. Esta clase puede clasificarse además en tres subclases de análisis, como sigue:

(a). Dominio del tiempo
(b). Dominio de la frecuencia
(c). Tiempo-escala (*time-scale*, TS) y tiempo-frecuencia (*time-frequency*, TF)

Los métodos primero y segundo usan las características de las convulsiones del EEG en el dominio del tiempo o frecuencia, respectivamente, como los criterios de detección. Estos métodos se basan en el supuesto de que las señales EEG son cuasi-estacionarias, es decir, el contenido de frecuencia de las señales EEG se supone que es constante durante la ventana de análisis. Sin embargo, trabajos recientes muestran que las señales EEG exhiben características no estacionarias (Boashash, Carson & Mesbah, 2000). Este hecho apoya la idea de utilizar métodos de tiempo-escala y tiempo-frecuencia para la detección de las crisis. Estos métodos tienen en cuenta la naturaleza no estacionaria de las señales EEG y usan simultáneamente las características de la convulsión en el dominio del tiempo y la frecuencia.

2.3.1 Método tiempo-escala y tiempo-frecuencia

La transformada *wavelet* (*wavelet transform*, WT) es una transformación que proporciona la representación tiempo-frecuencia de una señal dada. Esta transformación se desarrolló para superar algunos problemas relacionados con la resolución de las transformaciones tiempo-frecuencia como la transformada de Fourier de corto tiempo (STFT). El análisis WT utiliza un conjunto de funciones para representar la señal. Se puede obtener una representación WT tiempo-escala (TS) de la señal utilizando técnicas de filtro digital. En este caso, se aplican filtros de diferentes frecuencias de corte para analizar una señal con escalas diferentes.

Debido a las siguientes tres características significativas del análisis TS, este es una herramienta adecuada para analizar las señales de EEG:

(a). Multirresolución (para superar el problema de la resolución de tiempo y frecuencia)
(b). Ancho de banda relativo constante (los anchos de tiempo de las ondas *wavelet* son adaptados a sus frecuencias)

(c). Habilidades para presentar la señal si está localizada en el dominio del tiempo o en el dominio de la frecuencia

Está demostrado que varios estados de las señales de EEG presentan algunas características bien definidas en el dominio TS. Esto permite una discriminación eficiente entre los EEG normales y EEG anormales (Yong & Shengxun, 1996).

2.3.2 Transformada *wavelet* discreta

En (Lotte et al., 2007) se analiza como las características extraídas de las señales EEG son no-estacionarias, debido a que dichas señales varían rápidamente en el dominio del tiempo. Estas características deben contener información del tiempo de la señal EEG, ya que los ritmos de actividad cerebral se relacionan con cambios específicos del EEG en el tiempo. Esto hace necesaria una representación que considere dicho esquema.

Una técnica usada comúnmente en estos casos es la transformada *wavelet* discreta (*discrete wavelet transform,* DWT), que permite modelar este tipo de variaciones en el dominio tiempo-escala. Esta transformada permite tener una representación *wavelet* realmente eficiente, debido a que restringe la variación en la traslación y la escala, usualmente a potencias de dos. Por esto en algunos casos la DWT es denominada *wavelet* diádica (Wu, Hsu & Wu, 2009).

La transformada *wavelet* continua (*continuous wavelet transform,* CWT) de una señal *S(t)* es definida como la correlación entre *S(t)* y la función *wavelet* $\psi_{a,b}$ como sigue (Chui, 1992):

$$CWT_{(a,b)} = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} S(t)\psi * \left(\frac{t-b}{a}\right) dt\,, \qquad (1)$$

donde *a*, *b* son llamados parámetros de escala (recíproco de frecuencia) y traslación (localización de tiempo), respectivamente.

Cuando *a* y *b* son tomados como números discretos y definidos en la base de potencias de dos, que son como las siguientes:

$$\begin{cases} a_j = 2^j \\ b_{j,k} = 2^j k, \quad j,k \in Z \end{cases} \qquad (2)$$

Entonces la DWT es obtenida, siendo esta:

$$DWT_{(j,k)} = 2^{-\frac{j}{2}} \int_{-\infty}^{\infty} S(t)\psi * \left(\frac{t - 2^j k}{2^j}\right) dt \qquad (3)$$

La DWT puede implementarse con un esquema simple de filtrado recursivo, hasta alcanzar un número *N* deseado de niveles (resolución). La señal original puede ser reconstruida a través de un filtrado inverso (Xu & Song, 2008).

Mallat (1989) desarrolló una forma eficiente de implementación DWT pasando la señal a través de una serie de filtros pasa bajos y pasa altos. El procedimiento de implementación de DWT está esquematizado y se muestra en la figura 3, donde los filtros *h[n]* y *g[n]* corresponden a los filtros pasa altos y pasa bajos, respectivamente.

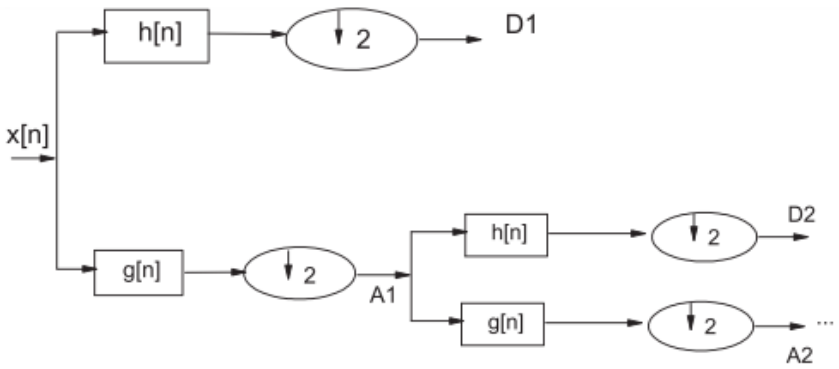


Figura 3. Descomposición de sub-bandas en implementación DWT
Fuente: (Subasi, 2007)

En la primera etapa, la señal pasa simultáneamente a través de los filtros *h[n]* y *g[n]*, siendo la frecuencia de corte la cuarta parte de la frecuencia de muestreo. Las salidas de los filtros *h[n]* y *g[n]* se refieren a los coeficientes de detalle *D1* (altas frecuencias) y de aproximación *A1* (bajas frecuencias) del primer nivel, respectivamente. Se repite el mismo procedimiento con los coeficientes de

aproximación del primer nivel para obtener los coeficientes del segundo nivel. En cada etapa de este proceso de descomposición, la resolución de frecuencia se duplica a través del filtrado y la resolución del tiempo es reducida a la mitad a través del muestreo descendente (submuestreo). Los coeficientes *A1, D1, A2* y *D2* representan el contenido de frecuencia de la señal original dentro de las bandas 0 - *fs/4, fs/4 - fs/2,* 0 - *fs/8,* y *fs/8 - fs/4,* respectivamente, donde *fs* es la frecuencia de muestreo original de la señal *x[n]*.

El proceso anterior es una descomposición iterativa, tal que la señal *S(t)* se expresa de la siguiente manera:

$$S(t) = a_J + \sum_{j \leq J} dj, \qquad (4)$$

donde a_J representa los coeficientes de aproximación y dj los coeficientes de detalle en el *J*-ésimo nivel.

2.3.3 Propiedades transformada *wavelet* discreta

Recientemente, la WT ha sido ampliamente aplicada en muchos campos para resolver varios problemas de la vida real. La transformada de Fourier (*Fourier transform,* FT) de una señal obtiene el contenido de frecuencia de la señal y elimina la información del tiempo. La transformada de Fourier de corto tiempo (*short time Fourier transform,* STFT) es una serie de FT con un tamaño de ventana fijo. Debido a que una gran ventana pierde la resolución de tiempo y una ventana corta pierde resolución de frecuencia, siempre existe una compensación entre las resoluciones de tiempo y frecuencia. Por una parte, una buena resolución temporal requiere una ventana corta con poco tiempo de apoyo; por otro lado una buena resolución de frecuencia requiere una ventana de tiempo prolongado. El tamaño fijo de la ventana (resolución fija tiempo-frecuencia) del STFT resulta ser una restricción en algunas aplicaciones.

Contrariamente a STFT, WT proporciona una forma más flexible de representación tiempo-frecuencia de una señal, permitiendo el uso de ventanas de análisis de tamaño

variable. La característica atractiva de WT es que proporciona información precisa de frecuencia a bajas frecuencias e información precisa de tiempo en altas frecuencias (Adeli et al., 2003). Esta propiedad es importante en aplicaciones biomédicas, porque las señales en este campo siempre contienen información de alta frecuencia con una corta duración y una información de baja frecuencia con larga duración de tiempo. A través de la transformada *wavelet*, las características transitorias son precisamente capturadas y localizadas en ambos dominios, tiempo y frecuencia. WT ha sido comúnmente considerada como una de las más poderosas herramientas para el análisis de señales EEG.

La literatura expone diversas familias de *wavelets.* La calidad y desempeño de cada una de ellas se evalúa de acuerdo en mayor medida a la aplicación, y a las propiedades requeridas por esta, tales como: momentos de desvanecimiento, tamaño del soporte, regularidad, simetría, etc. Entre las familias más empleadas de *wavelets* están: *Haar, Daubechies, Coiflets, Symlets, Biortogonales, Meyer, Mexican hat, Shannon* y *Morlet* (Gómez-Luna, Silva & Aponte, 2013).

Particularmente y por su gran uso en señales EEG, se exploran dos familias *wavelet* como: *Daubechies* y *Haar*.

La *wavelet Haar* tiene un soporte compacto, pero tiene un inconveniente, es discontinua. Mientras que la *wavelet Daubechies* tiene soporte compacto y es continua. Mejor aún, aumentando el nivel es cada vez más suave; es decir, puede tener un número determinado de derivadas continuas. En la figura 4 se presenta la representación gráfica de las dos funciones *wavelet.*

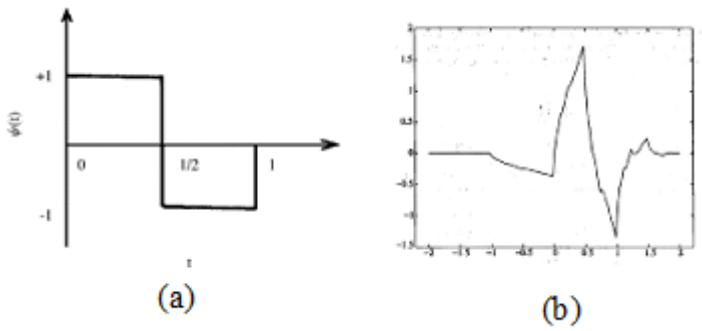


Figura 4. Funciones *wavelet*: (a) Haar y (b) Daubechies
Fuente: (Chui, 1992).

2.4 Extracción de características

Las señales son manifestaciones físicas de procesos naturales o artificiales de naturaleza diferente. Están relacionadas con una o más variables independientes e información sobre el comportamiento de algunos fenómenos físicos. Para caracterizar señales hay algunos parámetros importantes que dan la información necesaria para permitir el análisis de los datos.

Al ser el EEG la gráfica en la que se registra la actividad del cerebro en forma de señales, de esta se puede extraer un sinnúmero de características (*features*). Estas pueden ser lineales y no lineales.

2.4.1 Características lineales

2.4.1.1 Tasa de cruces por cero

La tasa de cruces por cero es la tasa de cambios de signo a lo largo de una señal, es decir, la tasa a la cual la señal cambia de positiva a negativa o viceversa. Esta característica se ha utilizado mucho en el reconocimiento de voz y la recuperación de información musical, siendo una característica clave para clasificar los sonidos de percusión (Gouyon, Pachet & Delerue, 2000).

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{R<0}(s_t s_{t-1}), \qquad (5)$$

donde s es una señal de longitud T y $1_{R<0}$ es una función indicador.

2.4.1.2 Máximo

En el análisis matemático, el máximo de una función, es el valor de mayor amplitud de la misma, ya sea dentro de un rango dado (máximo relativo), o en todo el dominio de una función (máximo global o absoluto) (Larson & Edwards, 2009).

Una función de valor real *f* definida en un dominio *X* tiene un punto máximo global (o absoluto) en *x** si $f(x^*) \geq f(x)$ para todo *x* en *X*.

2.4.1.3 Mínimo

En el análisis matemático, el mínimo de una función, es el valor de menor amplitud de la misma, ya sea dentro de un rango dado, o en todo el dominio de una función (Larson & Edwards, 2009).

Una función de valor real *f* definida en un dominio *X* tiene un punto mínimo global (o absoluto) en *x** si $f(x^*) \leq f(x)$ para todo *x* en *X*.

2.4.1.4 Media

La media aritmética (o media) es la medida más comúnmente utilizada y fácilmente comprendida de la tendencia central. En las estadísticas, el término promedio se refiere a cualquiera de las medidas de tendencia central. La media aritmética se define como la suma de los valores numéricos de cada observación dividida por el número total de observaciones (Medhi, 1992).

Simbólicamente, si tenemos un conjunto de datos que contiene los valores a_1, a_2,…, a_n, entonces la media aritmética *A* se define por la fórmula:

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n} \qquad (6)$$

2.4.1.5 Desviación estándar

La desviación estándar, también representada por la letra griega sigma σ, es una medida que se utiliza para cuantificar la cantidad de variación o dispersión de un conjunto de valores de datos. Una desviación estándar baja indica que los puntos de datos tienden a estar cerca de la media (también llamado el valor esperado) del

conjunto, mientras que una desviación estándar alta indica que los puntos de datos se extienden sobre un rango más amplio de valores (Medhi, 1992).

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|A_i - u|^2}\ , \qquad (7)$$

donde u es la media de A y N el número de valores de la muestra.

2.4.1.6 Varianza

La varianza es la expectativa de la desviación cuadrada de una variable aleatoria de su media y, de manera informal, mide hasta qué punto un conjunto de números (aleatorios) se separan de su media. La varianza tiene un papel central en la estadística. Se utiliza en estadística descriptiva, inferencia estadística, pruebas de hipótesis, bondad de ajuste y muestreo de Monte Carlo, entre muchos otros. Esto lo convierte en una cantidad central en numerosos campos como la física, la biología, la química, la criptografía, la economía y las finanzas. La varianza es el cuadrado de la desviación estándar, el segundo momento central de una distribución (Medhi, 1992).

$$Var = \frac{1}{N-1}\sum_{i=1}^{N}|A_i - u|^2, \qquad (8)$$

donde u es la media de A y N el número de valores de la muestra.

2.4.1.7 Curtosis

Curtosis es una medida estadística que se utiliza para describir la distribución, o asimetría, de los datos observados alrededor de la media. Se utiliza generalmente en el campo estadístico para describir las tendencias en los gráficos y puede estar presente en un gráfico con colas gordas y una distribución baja e igual, así como estar presente en un gráfico con colas flacas y una distribución concentrada hacia la media (Medhi, 1992).

$$k = \frac{E(x-u)^4}{\sigma^4}, \qquad (9)$$

donde u es la media de x, σ es la desviación estándar de x, and $E(t)$ representa el valor esperado de la cantidad t.

2.4.1.8 Asimetría

La asimetría es una medida de simetría, o más precisamente, la falta de simetría. Una distribución o conjunto de datos es simétrica si se ve igual a la izquierda y la derecha del punto central (Medhi, 1992).

$$s = \frac{E(x-u)^3}{\sigma^3}, \qquad (10)$$

donde u es la media de x, σ es la desviación estándar de x, and $E(t)$ representa el valor esperado de la cantidad t.

2.4.1.9 Potencia

La potencia se define como la cantidad de energía consumida por unidad de tiempo. Esta cantidad es útil si la energía de la señal va al infinito. Para las señales, la potencia calculada tomando la instantánea de la señal en un intervalo de tiempo específico, se da como sigue (Sanjay Lall, s.f.):

(a). Tomar una instantánea de la señal durante algún tiempo finito.
(b). Calcular la energía de la señal *Ex*.
(c). Dividir la energía por el número de muestras *N* tomadas para el cálculo.
(d). Extender el límite de número de muestras al infinito ($N \rightarrow \infty$). Esto da la potencia total de la señal.

En dominio discreto, la potencia total de la señal viene dada por:

$$P_x = \lim_{N\rightarrow\infty} \frac{1}{N} \sum_{n=0}^{n=N-1} |x(n)|^2, \qquad (11)$$

donde *N* es el número de valores de la muestra.

2.4.1.10 Energía de sub-banda wavelet

Para una descomposición *wavelet* unidimensional la energía de sub-banda *wavelet* es el porcentaje de energía correspondiente a los coeficientes *wavelet* detallados y aproximados.

En *Matlab* la función *[Ea, Ed] = wenergy (C, L)* devuelve *Ea*, que es el porcentaje de energía correspondiente a la aproximación y *Ed*, que es el vector que contiene los porcentajes de energía correspondientes a los detalles (Misiti, Misiti, Oppenheim & Poggi, 2000).

2.4.2 Característica no-lineal

2.4.2.1 Complejidad de Lempel-Ziv

La complejidad de Lempel-Ziv es una medida de complejidad útil, que puede caracterizar el grado de orden o de desorden de patrones espacio-temporales (Lempel & Ziv, 1976). Es una medida muy sensible a los cambios temporales, por ende es idónea para usarla en señales de actividad cerebral (EEG), ya que el cambio es una de las principales características de estas señales.

Como lo mencionan Zhang, Roy & Jensen (2001) en ciertas patologías el cerebro pierde esa característica de tendencia al cambio, por lo que la actividad cerebral resulta ser más homogénea. Esta tendencia puede ser medida usando la complejidad de Lempel-Ziv de esta forma:

Antes de calcular la medida de complejidad LZ *c(n)*, la señal debe ser transformada en una secuencia de símbolos finitos. Suponiendo que los datos originales son *y(n)*, estos datos son convertidos a una secuencia *P* de “0s” y “1s” como sigue:

$$P = s(1), s(2), \dots, s(n) \qquad (12)$$

donde:

$$s(i) = \begin{cases} 1, & y(i) > Td \\ 0, & si\ no \end{cases} \qquad (13)$$

En donde, *Td* es el valor de la media de la secuencia *P*. La medida de complejidad puede ser descrita como sigue (Aboy, Hornero, Abásolo & Álvarez, 2006):

(a). Dejemos que *S* y *Q* denoten dos subsecuencias de *P* y *SQ* sea la concatenación de *S* y *Q*, mientras que la secuencia $SQ\pi$ se deriva de *SQ* después de que su último caracter sea eliminado (π denota la operación de eliminar el último caracter en la secuencia). Dejemos que $v(SQ\pi)$ denote el vocabulario de todas las diferentes subsecuencias de $SQ\pi$. Inicialmente, $c(n)=1$, $S=s(1)$, $Q=s(2)$, por tanto $SQ\pi=s(1)$.

(b). En general, $S=s(1),s(2),\ldots,s(r)$, $Q=s(r+1)$, entonces $SQ\pi=s(1),s(2),\ldots,s(r)$; si *Q* pertenece a $v(SQ\pi)$, entonces *Q* es una subsecuencia de $SQ\pi$, no una nueva secuencia.

(c). Se renueva *Q*, siendo ahora $s(r+1)$, $s(r+2)$ y juzgar si *Q* pertenece a $v(SQ\pi)$ o no.

(d). Repetir el paso previo hasta que *Q* no pertenezca a $v(SQ\pi)$. Ahora $Q=s(r+1),s(r+2),\ldots,s(r+i)$ no es subsecuencia de $SQ\pi=s(1),s(2),\ldots,s(r+i-1)$, así que $c(n)$ se incrementa en uno.

(e). Por lo tanto, *S* es renovada y pasa a ser $S=s(1),s(2),\ldots,s(r+i)$, y $Q=s(r+i+1)$.

El procedimiento anterior se repite hasta que *Q* sea el último carácter. En ese momento el número de subsecuencias diferentes en *P*, da la medida de complejidad $c(n)$.

El Lempel-Ziv de una secuencia totalmente aleatoria de longitud *n* consistente en dos símbolos diferentes con probabilidades iguales es:

$$b(n) = \frac{n}{log_2(n)}, \qquad (14)$$

y *c(n)* puede ser normalizada por medio de *b(n)*

$$C(n) = \frac{c(n)}{b(n)} = \frac{c(n)log_2(n)}{n}, \qquad (15)$$

donde *C(n)*, es la complejidad LZ normalizada, y refleja la tasa de aparición de nuevos patrones en la secuencia (Aboy et al., 2006).

2.5 Reducción de dimensionalidad

El análisis de componentes principales (PCA) es probablemente la técnica estadística multivariable más popular y es utilizado por casi todas las disciplinas científicas. También es probable que sea la técnica multivariable más antigua.

De hecho, su origen se remonta a Pearson o incluso Cauchy, o Jordan y también Cayley, Silverster y Hamilton, pero su instanciación moderna fue formalizada por Hotelling, que también acuñó el término componente principal. PCA analiza una tabla de datos que representa las observaciones descritas por varias variables dependientes, que son, en general, inter-correlacionadas. Su objetivo es extraer la información importante de los datos. Y expresar esta información como un conjunto de nuevas variables ortogonales llamadas componentes principales. PCA también representa el patrón de similitud de las observaciones y las variables mostrándolas como puntos en los mapas (Hotelling, 1933).

Los objetivos del PCA son:

- Extraer la información más importante de la tabla de datos.
- Comprimir el tamaño del conjunto de datos manteniendo sólo la información importante.
- Simplificar la descripción del conjunto de datos.
- Analizar la estructura de las observaciones y las variables.

Con el fin de lograr estos objetivos, PCA calcula nuevas variables llamadas componentes principales que se obtienen como combinaciones lineales de las variables originales. Se requiere que el primer componente principal tenga la mayor varianza

posible (es decir, inercia y por lo tanto este componente 'explicará' o 'extraerá' la mayor parte de la inercia de la tabla de datos).

El segundo componente se calcula bajo la restricción de ser ortogonal al primer componente y tener la mayor inercia posible. Los otros componentes se calculan de la misma manera. Los valores de estas nuevas variables para las observaciones se denominan puntuación de factores y estas puntuaciones de factores pueden interpretarse geométricamente como las proyecciones de las observaciones sobre los componentes principales (Hotelling, 1933).

2.6 Aprendizaje de máquina

El aprendizaje automático o aprendizaje de máquina (*Machine Learning*) es una rama de la inteligencia artificial que desarrolla algunas técnicas para permitir que las máquinas aprendan. En términos generales, estos sistemas son capaces de analizar el medio ambiente y aprender de él. De manera técnica, el método intenta crear programas que son capaces de generalizar algún comportamiento analizando alguna información. El campo del aprendizaje de máquina está muy relacionado con el de la estadística, ya que ambas técnicas se basan en el análisis de datos. Sin embargo, el aprendizaje de máquina o aprendizaje automático puede considerarse como un intento de automatizar algunas partes del método científico mediante métodos matemáticos como la estadística. Algunos sistemas que forman parte del aprendizaje de máquina son: redes neuronales, máquinas de vectores de soporte y regresión logística (Shai & Shai, 2014).

Las tareas de aprendizaje de máquina suelen clasificarse en tres grandes categorías, dependiendo de la naturaleza de la "señal" o "retroalimentación" de aprendizaje disponible para un sistema de aprendizaje.

Estas son (Fukushima & Miyake, 1982):

- *Aprendizaje supervisado:* el ordenador se presenta con entradas de ejemplo y sus salidas deseadas, dadas por un "maestro", y la meta es aprender una regla general que mapea las entradas a las salidas.

- *Aprendizaje sin supervisión:* no se dan etiquetas al algoritmo de aprendizaje, dejándolo por sí solo para encontrar estructuras en su entrada. El aprendizaje sin supervisión puede ser un objetivo en sí mismo (descubrir patrones ocultos en los datos) o un medio hacia un final.

- *Aprendizaje reforzado:* Un programa informático interactúa con un entorno dinámico en el que debe realizar una determinada meta (como conducir un vehículo), sin que un "maestro" le diga explícitamente si se ha acercado o no a su meta. Otro ejemplo es aprender un juego, jugando contra un oponente.

2.6.1 Regresión logística

La regresión logística es una forma de aprendizaje automático que se utiliza para predecir el resultado de un valor categórico (el que podría adoptar un número de categorías) en función de las variables independientes.

La regresión logística mide la relación entre la variable dependiente categórica y una o más variables independientes, estimando las probabilidades usando una función logística. El objetivo principal es investigar la relación entre la probabilidad relacionada a la variable de salida y a las variables de entrada independientes.

En estas condiciones, la regresión logística se calcula como (Bishop, 2006):

$$y(x) = ln\frac{P(x)}{1 - P(x)} = w^T x + b\ , \qquad (16)$$

donde x es la *data* independiente, P la probabilidad estimada, y es 0 o 1, dependiendo en cada clase y g es la regresión logística.

2.6.2 Análisis discriminante

El análisis discriminante (*Discriminant Analysis*, DA) o análisis de función discriminante se utiliza para determinar qué variables continuas discriminan entre dos o más grupos naturales.

El análisis de la función discriminante es el análisis multivariable de varianza (*Multivariate analysis of variance*, MANOVA) invertido. En MANOVA, las variables independientes son los grupos y las variables dependientes son los predictores. En DA, las variables independientes son los predictores y las variables dependientes son los grupos. Como anteriormente se mencionó, se suele utilizar DA para predecir la pertenencia de un miembro a un grupo. Es decir responde a la pregunta: ¿se puede utilizar una combinación de variables para predecir la pertenencia a un grupo? Por lo general, se incluyen varias variables en un estudio para ver cuáles contribuyen mejor a la discriminación entre grupos (Klecka, 1980).

El DA se divide en un proceso de dos pasos (Klecka, 1980):

(a). Probar la significancia de un conjunto de funciones discriminantes.

(b). Clasificación.

El primer paso es un cómputo idéntico a MANOVA. Existe una matriz de varianzas totales y covarianzas; asimismo, existe una matriz de variaciones dentro del grupo y covarianzas. Las dos matrices se comparan mediante pruebas *F* de multivariables para determinar si existen diferencias significativas (con respecto a todas las variables) entre grupos. Se realiza primero la prueba multivariable, y si es estadísticamente significativa, se procede a ver cuál de las variables tiene medias significativamente diferentes a través de los grupos (Klecka, 1980).

Una vez que la media del grupo es estadísticamente significativa, la clasificación de variables se lleva a cabo. DA determina automáticamente una combinación óptima de las variables, de modo que la primera función proporcione la discriminación más óptima entre grupos, la segunda proporciona la segunda más óptima, y así sucesivamente. Además, las funciones serán independientes u ortogonales, es decir, sus contribuciones a la discriminación entre grupos no se superponen. La primera función recoge la mayor variación; la segunda función selecciona la mayor parte de la variación inexplicada, etc. Computacionalmente, se realiza un análisis de correlación canónica que determinará las sucesivas funciones y raíces canónicas (Klecka, 1980).

La clasificación es entonces posible gracias a las funciones canónicas. Los sujetos se clasifican en los grupos en los que tienen los puntajes de clasificación más altos. El

número máximo de funciones discriminantes será igual a los grados de libertad, o el número de variables en el análisis, lo que sea menor (Klecka, 1980).

2.6.3 *k*-Vecinos más cercanos

El algoritmo *k*-vecinos más cercanos (*k-nearest neighbors*, kNN) es un método no paramétrico utilizado para la clasificación y la regresión. En ambos casos, la entrada consta de los *k* ejemplos de entrenamiento más cercanos en el espacio de características. La salida depende de si kNN se utiliza para la clasificación o la regresión (Altman, 1992):

- En la clasificación kNN, la salida es una pertenencia a una clase. Un objeto es clasificado por un voto mayoritario de sus vecinos, con el objeto siendo asignado a la clase más común entre sus *k* vecinos más cercanos (*k* es un entero positivo, típicamente pequeño). Si *k* = 1, entonces el objeto simplemente se asigna a la clase de ese único vecino más cercano.
- En la regresión kNN, la salida es el valor de propiedad del objeto. Este valor es el promedio de los valores de sus *k* vecinos más cercanos.

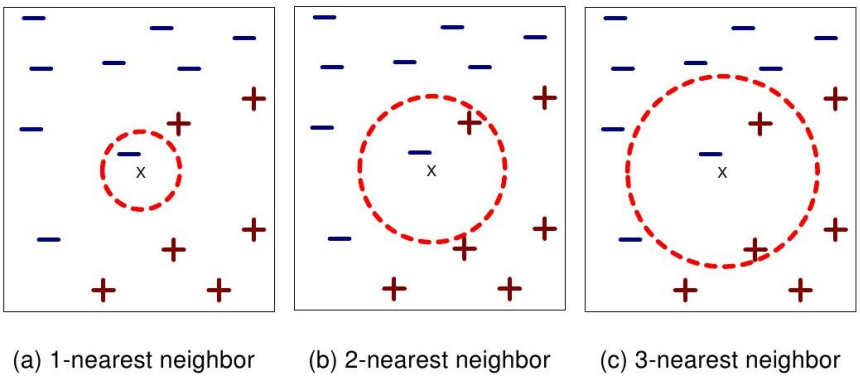


Figura 5. *k*-Vecinos más cercanos de un punto *x* para *k*=1,2,3
Fuente: (Hastie, Tibshirani & Friedman, 2009)

La figura 5 muestra una ilustración de los *k*-vecinos más cercanos a un punto aleatorio *x*, mismos que poseen la menor distancia al punto.

2.6.4 Árboles de decisión

Los árboles de decisión son una forma simple, pero poderosa de análisis de variables múltiples. Estos proporcionan capacidades únicas para complementar, suplementar y sustituir:

- Las formas estadísticas tradicionales (como regresión lineal múltiple)
- Una variedad de técnicas como redes neuronales
- Formas multidimensionales recientemente desarrolladas en el campo de la inteligencia empresarial

Los árboles de decisión son producidos por algoritmos que identifican diferentes formas de dividir un set de datos en segmentos de similares ramas. Estos segmentos forman un árbol de decisiones invertido que se origina con un nodo raíz en la parte superior del árbol.

Para predecir una respuesta, se sigue las decisiones en el árbol desde el nodo raíz (inicio) hasta un nodo hoja. El nodo hoja contiene la respuesta. Los árboles de clasificación dan respuestas que son nominales, tales como "verdadero" o "falso". Los árboles de regresión dan respuestas numéricas. Cada paso en una predicción implica comprobar el valor de un predictor (variable) (Tilbury & Messner, 1998). En la figura 6 se presenta la estructura de un árbol de decisión.

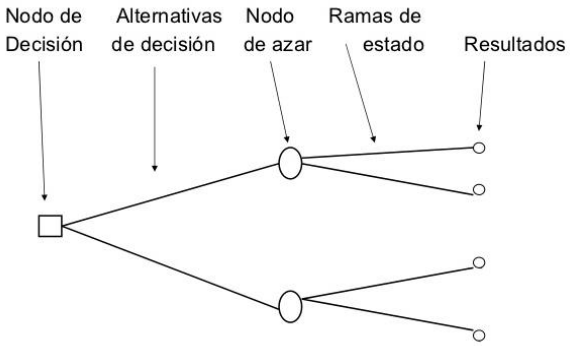


Figura 6. Partes de un árbol de decisión
Fuente: (Safavian & Landgrebe, 1991)

2.6.5 Máquinas de vectores de soporte

Máquinas de vectores de soporte (*Support Vector Machines*, SVM) consiste en un grupo de algoritmos de aprendizaje implementados para resolver problemas relacionados con la clasificación y la regresión. El sistema sigue algunos pasos ordenados. En primer lugar, necesitamos algunas muestras de datos que se clasifican en una categoría o en otra. Entonces, con estos datos, la máquina de vectores de soporte es entrenada para construir un modelo que le permita clasificar algunos datos de prueba en una de ambas clases. Lo que hace exactamente es crear un hiperplano o un grupo de ellos en altas dimensiones, incluso dimensiones infinitas, que podrían utilizarse en la clasificación y problemas de regresión.

Este tipo de algoritmos trata de seleccionar el hiperplano que maximiza la distancia entre este hiperplano y los puntos que están más cerca de este (Bishop, 2006).

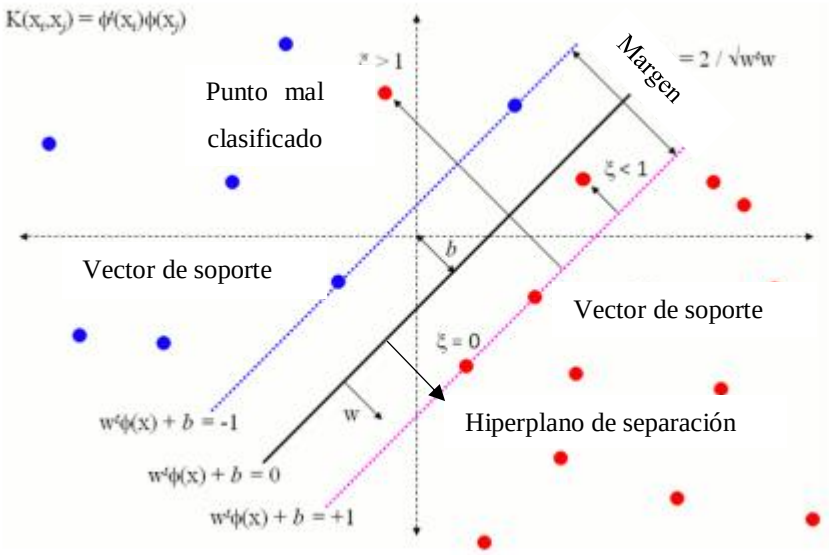


Figura 7. Máquina de vectores de soporte en dos dimensiones
Fuente: (Bishop, 2006)

La figura 7 muestra una ilustración de un hiperplano en dos dimensiones, usando máquina de vectores de soporte (SVM).

2.6.6 Conjunto de clasificadores redundantes

Conjunto de clasificadores redundantes (*ensemble*) es un paradigma de aprendizaje de máquina donde múltiples clasificadores son entrenados para resolver el mismo problema. A diferencia de los métodos ordinarios de aprendizaje automático que intentan aprender una hipótesis de los datos de entrenamiento, *ensemble* trata de construir un conjunto de hipótesis y combinarlas para usarlas (Hansen & Salamon, 1990).

Un *ensemble* contiene un número de clasificadores que usualmente se llaman clasificadores base. La capacidad de generalización de un *ensemble* suele ser mucho más fuerte que la de los clasificadores base. En realidad, el aprendizaje de conjunto es atractivo porque es capaz de impulsar clasificadores débiles, a clasificadores fuertes, que pueden hacer predicciones muy precisas. Por lo tanto, los "clasificadores base" también son referidos como "clasificadores débiles". Cabe señalar, sin embargo, que aunque la mayoría de los análisis teóricos funcionan con clasificadores débiles, los clasificadores base utilizados en la práctica no son necesariamente débiles, ya que el uso de clasificadores base no tan débiles a menudo resulta en un mejor rendimiento (Hansen & Salamon, 1990).

Los clasificadores base suelen ser generados a partir de datos de entrenamiento por un algoritmo de aprendizaje base que puede ser un árbol de decisión, red neuronal u otros tipos de algoritmos de aprendizaje de máquina. La mayoría de los métodos *ensemble* utilizan un único algoritmo de aprendizaje base para producir clasificadores base homogéneos, pero también hay algunos métodos que utilizan múltiples algoritmos de aprendizaje para producir clasificadores heterogéneos. En este último caso no existe un algoritmo de aprendizaje de base único (Hansen & Salamon, 1990).

Se puede observar un ejemplo de implementación del algoritmo *ensemble* en la figura 8, en donde tenemos tres clasificadores base con sus respectivas representaciones de frontera de decisión, de las cuales su combinación, por medio del esquema de combinación elegido, da lugar a la representación final de la frontera de decisión, obtenida con *ensemble*.

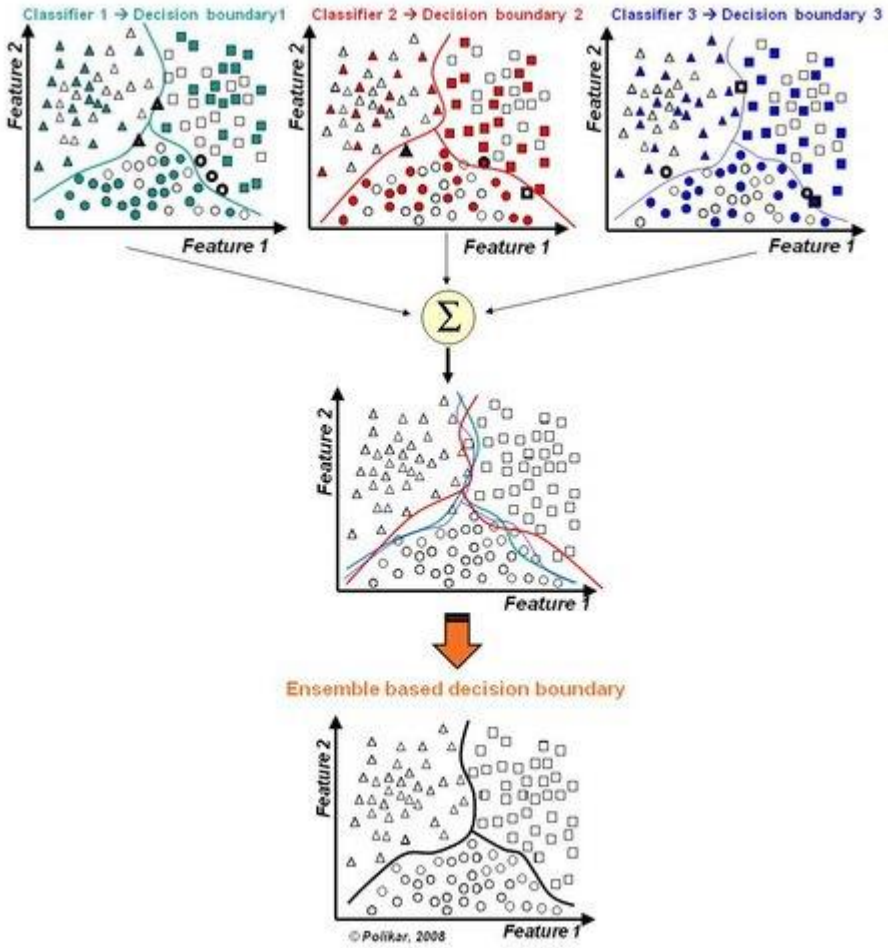


Figura 8. Ejemplo de límite de decisión utilizando *Ensemble*
Fuente: (Fu & Shi, 2002)

2.7 Validación cruzada *k-fold*

La validación cruzada (*Cross-Validation*, CV) es una técnica utilizada para evaluar los resultados de los análisis, para garantizar que estos resultados sean independientes de la partición entre datos de entrenamiento (*train*) y datos de prueba (*test*) (Refaeilzadeh, Tang & Liu, 2009).

En la validación cruzada *k-fold*, los datos de entrenamiento se dividen en *k* subgrupos. Uno de estos subgrupos se utiliza como datos de prueba y el resto de los subgrupos como datos de entrenamiento. Se repite *k* veces cambiando cada vez el subgrupo que es responsable de las pruebas (Refaeilzadeh et al., 2009).

Aunque es un método muy preciso y eficiente, su principal inconveniente es que es un proceso muy lento para la programación y desde el punto de vista computacional.

2.8 Medidas de desempeño

2.8.1 Matriz de confusión

Una matriz de confusión contiene información sobre las predicciones reales y previstas de las clasificaciones realizadas por un sistema de clasificación. El rendimiento de estos sistemas es comúnmente evaluado utilizando los datos de la matriz (Kohavi & Provost, 1998).

Considere un problema de predicción general como el de este proyecto. Donde los resultados se clasifican en positivos (p) o negativos (n). Existen cuatro posibles resultados: si el resultado es positivo y la etiqueta verdadera es positivo, entonces es conocido como verdadero positivo (VP); si el resultado es positivo y la etiqueta verdadera es negativo, es conocido como falso positivo (FP); si el resultado es negativo y la etiqueta verdadera es positivo, es conocido como falso negativo (FN); finalmente, si el resultado es negativo y la etiqueta verdadera es negativo también, es conocido como verdadero negativo (VN). Todos estos cuatro valores se recogen en la matriz de confusión que se muestra en la tabla 1.

Tabla 1
Definición matriz de confusión

		VALOR REAL	
		Positivos	**Negativos**
VALOR DE PREDICCIÓN	**Positivos**	Verdaderos Positivos VP	Falsos Positivos FP
	Negativos	Falsos Negativos FN	Verdaderos Negativos VN

2.8.1.1 Exactitud

La exactitud se utiliza como medida estadística de lo bien que una prueba de clasificación binaria identifica o excluye correctamente una condición. Es decir, la exactitud es la proporción de resultados verdaderos (tanto los verdaderos positivos como los verdaderos negativos) entre el número total de casos examinados.

Se calcula como (Zhu, Zeng & Wang, 2010):

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN} * 100, \qquad (17)$$

donde *VP* representa los Verdaderos Positivos, *FN* los Falsos Negativos, *VN* los Verdaderos Negativos y *FP* los Falsos Positivos.

2.8.1.2 Sensibilidad

La sensibilidad, también denominada tasa de verdaderos positivos (*true positive rate*, TPR), mide la proporción de positivos que se identifican correctamente como tales (por ejemplo el porcentaje de personas enfermas que están correctamente identificadas como tales, es decir que presentan la condición).

Se calcula como (Zhu et al., 2010):

$$Sensibilidad = \frac{VP}{VP + FN} * 100, \qquad (18)$$

donde *VP* representa los Verdaderos Positivos y *FN* los Falsos Negativos.

2.8.1.3 Especificidad

La especificidad, también llamada tasa de verdaderos negativos, mide la proporción de negativos que se identifican correctamente como tales (por ejemplo, el porcentaje de personas sanas que se identifican correctamente como tales, es decir no presentan la condición).

Se calcula como (Zhu et al., 2010):

$$Especificidad = \frac{VN}{VN + FP} * 100, \qquad (19)$$

donde *VN* representa los Verdaderos Negativos y *FP* los Falsos Positivos.

2.8.2 Curva ROC

Con el fin de determinar qué tan bueno es el algoritmo propuesto, surge la necesidad de comparar los datos de la validación cruzada con la etiqueta real de los segmentos de los datos de entrenamiento. Esta determinación se realiza con una curva Característica Operativa del Receptor (*Receiving Operating Characteristics*, ROC), que es una representación de la *sensibilidad* frente a (*1-especificidad*) (Hanley & McNeil, 1982). En base a esto, la mejor predicción posible estará situada en un punto del lado izquierdo, es decir un punto de coordenada (0,1), que representa 100% de sensibilidad y 100% de especificidad (Hanley & McNeil, 1982).

La línea discontinua roja mostrada en la figura 9 representa una clasificación aleatoria, por ejemplo, los resultados de lanzar una moneda.

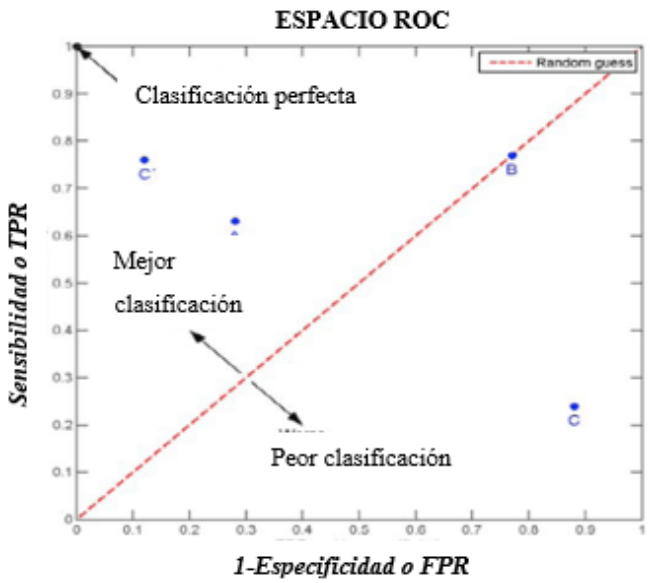


Figura 9. Ejemplo de curva ROC
Fuente: (Hanley & McNeil, 1982)

El parámetro área bajo la curva ROC (*Area Under ROC*, AUR) refleja qué tan bueno es un algoritmo para discriminar entre las dos posibles clases, a lo largo de todo el rango de puntos de corte posibles. Los ejes del gráfico de curva ROC adoptan valores entre 0 y 1 (0% y 100%), delimitando un cuadrado de área = 1,00. Un algoritmo se considera no-discriminativo si su curva ROC coincide con la línea de no-discriminación (AUC = 0,50). A medida que el AUR de un algoritmo se acerca al valor 1,00 (test diagnóstico perfecto), mayor será su capacidad predictiva (Cerda & Cifuentes, 2012).

CAPÍTULO III

3. DISEÑO E IMPLEMENTACIÓN

3.1 *Software* y *hardware*

El *hardware* empleado en la elaboración del proyecto fue el siguiente: un computador portátil (*laptop*) marca TOSHIBA, con sistema operativo Windows 10 de 64-bits, procesador Intel(R) Core(TM) i7 de 2.40 Ghz y memoria RAM de 8 Gb.

El *software* utilizado para la implementación de cada una de las etapas fue la herramienta matemática *Matlab*® (versión R2016b).

En la etapa de procesamiento de la señal, para realizar la descomposición *wavelet* se empleó el *Wavelet Toolbox*, de *Matlab*, que permite analizar cómo el contenido de frecuencia de las señales EEG cambia con el tiempo y revelar posibles patrones variables en el tiempo. Con esta herramienta se puede realizar análisis multirresolución para extraer características de escala fina o de gran escala y detectar eventos que no son visibles en señales sin procesar.

En la etapa de clasificación se utilizó el *Statistics and Machine Learning Toolbox*, de *Matlab*, que proporciona funciones ideales para analizar y modelar datos. Contiene algoritmos de clasificación, tales como: árboles de decisión, máquinas de vectores de soporte, *k*-vecinos más cercanos, etc., necesarios para construir modelos predictivos. Proporciona también funciones de reducción de dimensionalidad, tales como PCA. Una aplicación de este *toolbox* es el *Classification Learner App*, que nos permite entrenar modelos por medio de su configuración, para clasificar datos usando algoritmos de aprendizaje de máquina supervisado, y evaluar su rendimiento, por medio de medidas de desempeño como matriz de confusión y curvas ROC.

Adicionalmente para optimizar el tiempo de cómputo al ejecutar el algoritmo, se hizo necesario utilizar herramientas de cómputo en paralelo, por medio del *Parallel Computing Toolbox*, de *Matlab*, que permite resolver problemas computacionales y de

uso intensivo de datos, utilizando procesadores *multicore*, GPUs y clústeres de computadoras.

3.2 Esquema general del algoritmo

La metodología propuesta en el presente proyecto contempla el diseño, implementación y evaluación de un algoritmo de predicción de crisis epilépticas.

Para tal propósito se emplea señales de electroencefalografía intracraneal (iEEG) de pacientes con epilepsia. Estas son procesadas usando análisis multirresolución, en el dominio tiempo-escala, por medio de la Transformada *Wavelet* Discreta (*Discrete Wavelet Transform*, DWT) y su descomposición frecuencial. La señal EEG descompuesta es caracterizada por medio de la extracción de una serie de características (*features*). La clasificación se lleva a cabo para verificar si una muestra EEG de prueba corresponde a un estado interictal o preictal, empleando los siguientes algoritmos de aprendizaje de máquina supervisado: regresión logística, análisis discriminante, *k*-vecinos más cercanos (kNN), árboles de decisión, máquina de vectores de soporte (SVM) y conjunto de clasificadores redundantes (*ensemble*), y eligiendo el de mejor desempeño. Finalmente, para evaluar el rendimiento del algoritmo se emplea las siguientes medidas de desempeño: exactitud, sensibilidad, especificidad y área bajo la curva ROC (AUR).

La figura 10 muestra la estructura general del algoritmo de predicción de crisis epilépticas.

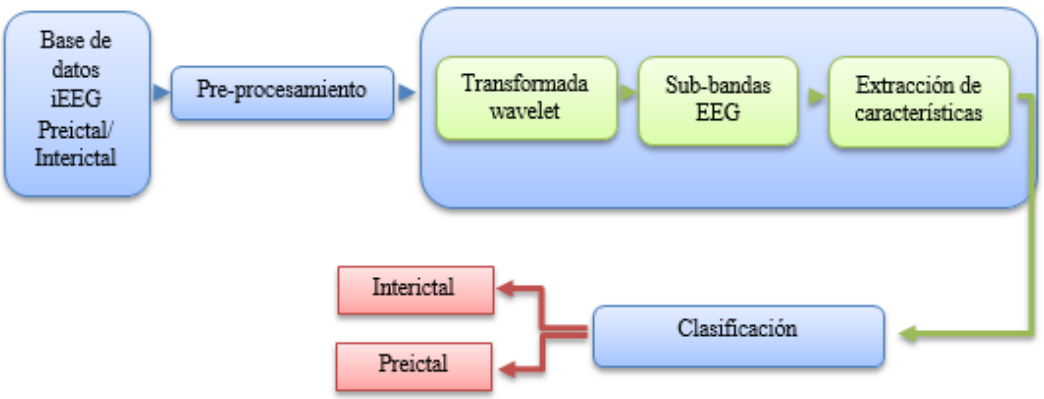


Figura 10. Esquema general del algoritmo de predicción de crisis epilépticas

3.3 Base de datos EEG

El conjunto de datos utilizado en este proyecto está disponible públicamente en el la plataforma *Kaggle - American Epilepsy Society Seizure Prediction Challenge* (Kaggle, 2014). La base de datos contiene registros de actividad cerebral de dos pacientes con epilepsia, registrada en forma de EEG intracraneal (iEEG). La grabación de las señales eléctricas se realizó con un sistema de monitorización ambulatoria. Los registros EEG de 15 electrodos fueron muestreados a 5000 Hz.

Las grabaciones EEG se organizan en registros de diez minutos de duración, con la etiqueta "Preictal" para segmentos de datos pre-convulsión, o "Interictal" para segmentos de datos no convulsivos. Los segmentos de datos se enumeran secuencialmente. Dentro de las carpetas, los segmentos de datos se almacenan en archivos *.mat* de la siguiente manera:

- **preictal_segment_N.mat.-** corresponde al *N*-ésimo segmento de datos pre-convulsivos.
- **interictal_segment_N.mat.-** corresponde al *N*-ésimo segmento de datos no-convulsivos.

Cada archivo *.mat* contiene una estructura de datos con los siguientes campos:

- **data:** matriz de valores de la muestra EEG dispuestos como electrodo-tiempo (fila-columna).
- **data_length_sec:** duración de tiempo de cada fila del campo *data*.
- **sampling_frequency:** número de muestras de datos EEG en un segundo de tiempo.
- **channels:** una lista de nombres de los electrodos, correspondientes a las filas en el campo *data*.
- **sequence:** el índice del segmento de datos dentro de la serie de clips de una hora. Por ejemplo, “preictal_segment_6.mat” tiene un número de secuencia 6, y representa los datos EEG de 50 a 60 minutos dentro de los datos preictales.

Los segmentos de datos preictales se proporcionan cubriendo una hora antes de la convulsión, con un horizonte de convulsiones de cinco minutos, es decir, de 1:05 a 0:05 antes del inicio de la convulsión. Este horizonte previo a la convulsión asegura que:

(a). Las convulsiones podrían predecirse con suficiente advertencia para permitir la administración de medicamentos de acción rápida.

(b). Cualquier actividad convulsiva antes de la anotación del inicio de convulsión que puede haber sido omitida por el epileptólogo no afectará el resultado del algoritmo.

La figura 11 muestra una representación gráfica de la distribución de segmentos EEG de 10 minutos, para todas las secuencias, dentro de los clips preictales de una hora.

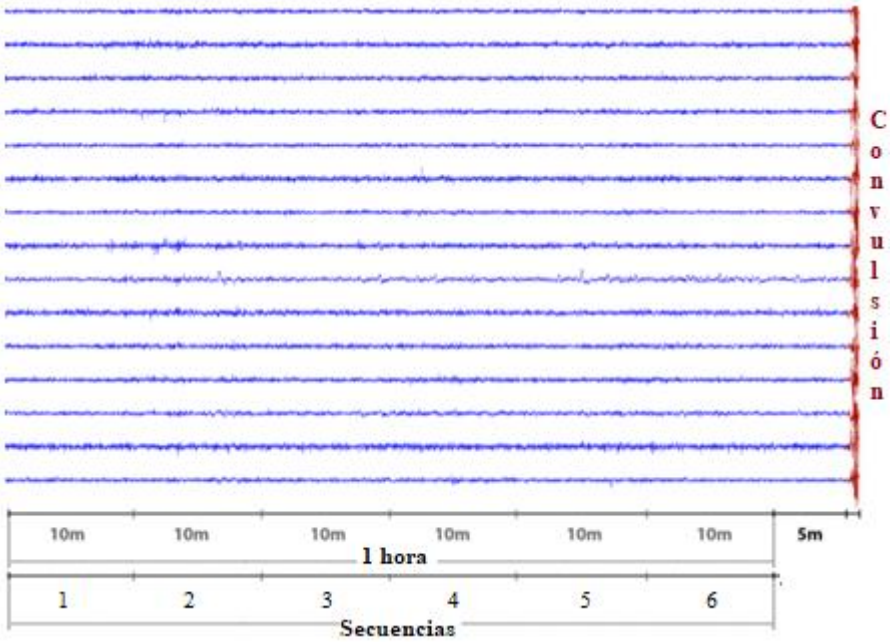


Figura 11. Muestra clip preictal de 1 hora
Fuente: (Kaggle, 2014)

De forma similar, se proporcionan segmentos de datos de 10 minutos interictales, correspondientes a registros de una hora. Los datos interictales fueron elegidos al azar del registro completo de datos, con la restricción de que los segmentos interictales

deben estar al menos a cuatro horas de distancia de cualquier convulsión, para evitar la contaminación con señales preictales o postictales.

Cualquier parte de cualquier segmento de datos de 10 minutos puede contener potencialmente "abandono de datos" (*drops*), donde el implante cerebral intracraneal ha fallado temporalmente para registrar datos. Este abandono de datos corresponde a valores de señal EEG de ceros, a través de todos los canales, en una muestra de tiempo dada. El abandono de datos no proporciona información predictiva en cuanto a si un segmento determinado de 10 minutos es preictal o interictal.

3.4 Pre-procesamiento

Los segmentos EEG de la base de datos, deben ser pre-procesados, con el fin de fragmentarlos en subsegmentos de análisis. Se debe además seleccionar aquellos que tengan un porcentaje de pérdidas de datos (*drops*) bajo, de tal manera que, estas no afecten el desempeño del algoritmo. También es necesario tener un análisis balanceado, es decir, analizar una cantidad de segmentos EEG iguales, tanto para segmentos interictales como para segmentos preictales, y tener también igual número de segmentos analizados, tanto para el paciente uno como para el paciente dos, esto con el objetivo de tener resultados en igualdad de condiciones para los dos estados (interictal y preictal) y los dos pacientes.

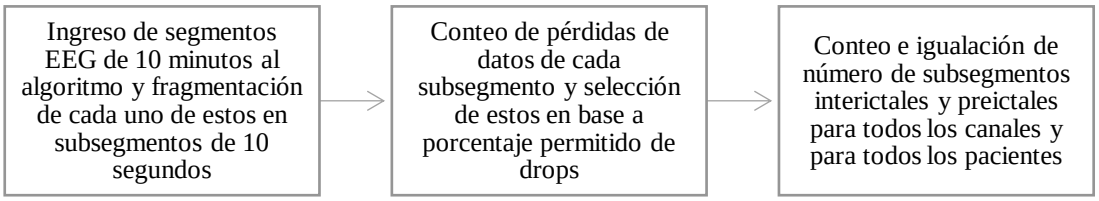


Figura 12. Algoritmo de pre-procesamiento de los segmentos EEG

La figura 12 muestra el algoritmo de procedimiento de la etapa de pre-procesamiento de las señales EEG. A continuación se detalla cada uno de los procesos realizados en esta etapa.

3.4.1 Fragmentación de segmentos EEG en subsegmentos EEG

Para asegurar la posibilidad de una diferenciación entre el estado interictal y el estado preictal, y por ende, una clasificación de estos, se escogió primeramente un tiempo de análisis corto y lo más cercano posible al inicio de la crisis epiléptica (para el caso del estado preictal). De esta forma, inicialmente no se toma en cuenta todas las secuencias de los segmentos de 10 minutos (dentro de los clips EEG de una hora), ni tampoco todas las muestras que conforman un segmento de 10 minutos.

Con este fin, se tomó un tiempo de análisis correspondiente a los últimos 10 segundos anteriores al inicio de la crisis epiléptica, sin tomar en cuenta el horizonte de cinco minutos (véase apartado 3.3 Base de datos iEEG). A este fragmento de señal EEG se lo denomina “subsegmento de los últimos 10 segundos anteriores a la crisis” o simplemente “subsegmento”. El subsegmento es una fracción correspondiente a los últimos 10 segundos (intervalo [590-600] segundos) del segmento EEG de los últimos 10 minutos anteriores a la crisis epiléptica (secuencia 6). Un ejemplo de subsegmento se muestra en la figura 13. El rectángulo de color amarillo representa el tiempo de análisis de este subsegmento.

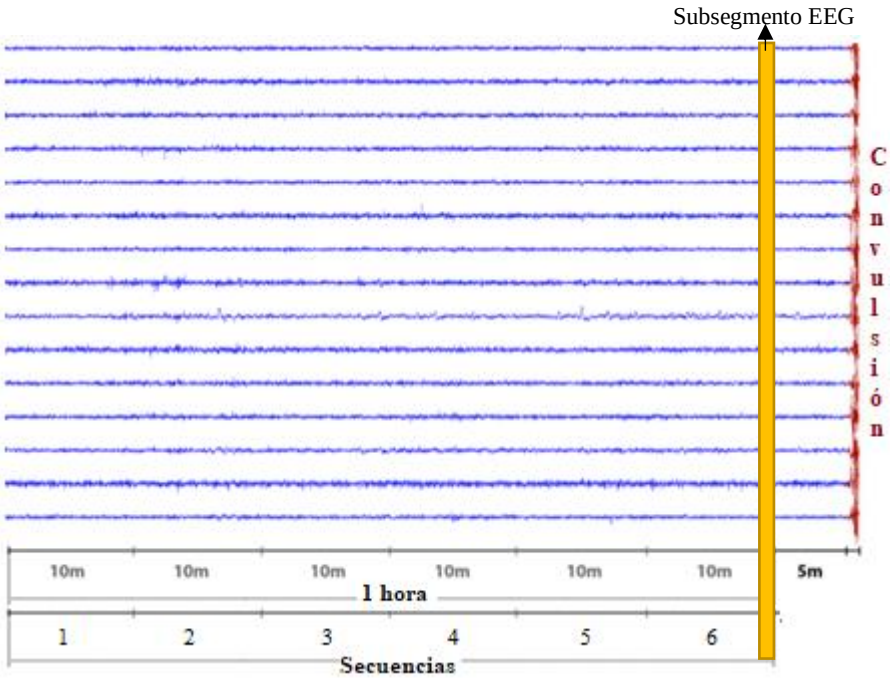


Figura 13. Representación subsegmento EEG
Fuente: (Kaggle, 2014)

Para el estado interictal, se seleccionó el mismo intervalo de tiempo de análisis, tomando en cuenta que los clips EEG de una hora de duración para ambos estados tienen la misma estructura de almacenado.

3.4.2 Selección de subsegmentos EEG por porcentaje de pérdidas

Debido a que existen subsegmentos EEG con pérdidas de datos (*drops*), se debe realizar un estudio del número de *drops* presente en cada subsegmento, con el fin de seleccionar aquellos que contengan un número de pérdidas relativamente bajo. De esta manera, el algoritmo realiza el conteo de *drops* de cada subsegmento y calcula el porcentaje que este representa, respecto al número de muestras totales de dicho subsegmento. Si el porcentaje de pérdidas es menor al permitido, se almacena el subsegmento en un arreglo (*array*), caso contrario, es desechado.

Como se emplea subsegmentos de 10 segundos y una frecuencia de muestreo de 5000 Hz, se tiene por cada subsegmento, un total de 50000 muestras de datos EEG. Si se determina un porcentaje de pérdidas permitidas del 10%, se tiene 5000 muestras perdidas, siendo este un número significativo de muestras, respecto al número total del subsegmento. Al tomar en cuenta que cada *drop* no aporta en nada al entrenamiento del clasificador, se debe elegir un porcentaje permitido de *drops* menos significativo.

En la implementación de este proceso, se puede verificar que la mayoría de subsegmentos presentan un porcentaje de pérdidas menor al 1%, sin embargo existen subsegmentos que presentan un porcentaje mayor de *drops.* Por tal motivo, se decidió establecer un porcentaje permitido de pérdidas del 2%, ya que todos los subsegmentos presentaron un porcentaje de pérdidas menor a este valor, asegurando de esta forma, el uso de todos los subsegmentos correspondientes a los últimos 10 segundos anteriores a la crisis epiléptica, presentes en la base de datos. De esta manera, se garantiza también un porcentaje no significativo de *drops,* representando un número de 1000 muestras perdidas, respecto a las 50000 totales del subsegmento. Recordar que mientras más subsegmentos sean analizados para el entrenamiento de los modelos de predicción, se tendrá un desempeño más aproximado al real, del algoritmo.

3.4.3 Balanceo del número de segmentos y subsegmentos EEG

La base de datos cuenta con 18 segmentos EEG preictales de 10 minutos por paciente, cada uno con su respectiva secuencia, es decir un total de 36 segmentos preictales totales (18 segmentos *x* 2 pacientes). Para tener un análisis balanceado, se ingresa al algoritmo 36 segmentos preictales y 36 segmentos interictales, es decir, un total de 72 segmentos EEG de 10 minutos.

Una vez ingresados estos 72 segmentos al algoritmo, se separa los subsegmentos EEG correspondientes a los últimos 10 segundos anteriores a la crisis epiléptica, y se realiza la selección de subsegmentos por porcentaje permitido de *drops.* Como los subsegmentos preictales e interictales son independientes, cada uno tendrá su propio porcentaje de *drops*, por lo tanto, se tendrá diferente número de subsegmentos interictales y de preictales en una instancia paciente-canal determinada (15 canales *x* 2 pacientes = 30 instancias). Por esta razón, para continuar con un análisis balanceado, se debe igualar el número de subsegmentos de análisis interictales, con el número de subsegmentos preictales.

Los *drops* son pérdidas presentadas en todos los canales al mismo tiempo, es decir, todos los canales presentan aproximadamente igual porcentaje de *drops* (no igual, debido a que un mínimo porcentaje de ceros no son *drops,* sino mediciones específicamente), entonces se tendrá igual número de subsegmentos de análisis en todos los canales del paciente. Sin embargo, este número varía de un paciente a otro, por lo tanto, para terminar de balancear el análisis, se debe proceder a igualar el número de subsegmentos por pacientes, es decir los dos pacientes deben tener igual número de subsegmentos totales analizados (interictales + preictales).

La etapa de pre-procesamiento del algoritmo se realiza para cada instancia paciente-canal (30 instancias), es decir se tendrá un total de 30 listas generadas de subsegmentos EEG seleccionados y balanceados para proceder a su posterior procesamiento.

3.5 Procesamiento

Para obtener el vector de características *m x n* (*m* muestras u observaciones y *n* variables o predictores) que será ingresado en el clasificador, es necesario procesar cada subsegmento EEG, preseleccionado en la etapa de pre-procesamiento. Para ello, es necesario dividir cada subsegmento en un número determinado de ventanas de análisis, cada ventana de análisis debe ser descompuesta en *n* sub-bandas de frecuencia, en el dominio tiempo-escala, por medio de la transformada *wavelet* discreta (DWT), y finalmente se debe extraer una serie de características de cada sub-banda *wavelet* calculada.

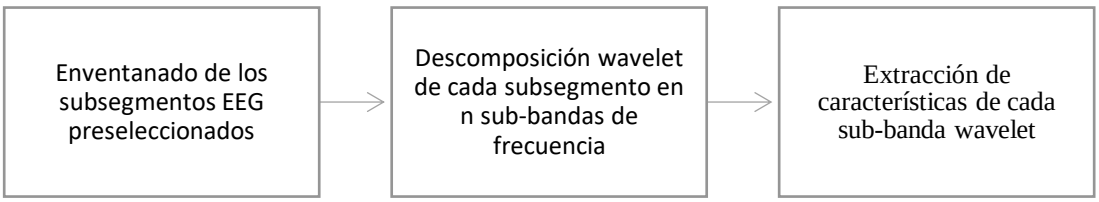


Figura 14. Algoritmo de procesamiento de los subsegmentos seleccionados

La figura 14 muestra el algoritmo de procedimiento de la etapa de procesamiento de los subsegmentos EEG preseleccionados. A continuación se detalla cada uno de los procesos realizados en esta etapa.

3.5.1 Enventanado de los subsegmentos EEG

En este proceso, cada subsegmento EEG, preseleccionado en la etapa de pre-procesamiento, es dividido en *N* ventanas de análisis diferentes.

Con el fin de tener un número adecuado de observaciones en el vector de características, se procedió a fijar un tiempo de duración de ventana de 1 segundo. Se debe tener en cuenta que al dar un tiempo mayor a 1 segundo, se tiene menos observaciones en el vector de características, y por tanto, modelos de predicción menos entrenados. Además, al variar el tiempo de duración de la ventana varía también el número de niveles de descomposición *wavelet* permitidos. Estos niveles se basan en el número de muestras de la ventana de análisis, y con un tiempo de ventana de 1

segundo, el número de niveles está directamente relacionado con la frecuencia de muestreo de las señales EEG.

Por el contrario, con un tiempo de ventana de análisis menor a 1 segundo, se tendrá más observaciones en el vector de características, pero se reducirá la escala de frecuencia de la señal y se verá afectado el número de niveles de descomposición *wavelet* permitidos. Tomando en cuenta que, en aplicaciones biomédicas las señales EEG siempre contienen información importante en sus bandas de baja frecuencia, se necesita una ventana de tiempo relativamente grande. Demostrando así, la idoneidad de una ventana con un valor de tiempo de 1 segundo de duración (5000 muestras de datos EEG). De esta manera, para subsegmentos EEG de 10 segundos se tiene un total de 10 ventanas de análisis resultantes.

Para un entrenamiento mayor de los clasificadores, es necesario aumentar el número de observaciones del vector de características. Por tanto, se procedió a emplear ventanas de análisis con un solapamiento del 50% de la ventana, obteniendo finalmente un total de 19 ventanas de análisis por cada subsegmento EEG.

La figura 15 presenta un ejemplo de enventanado con 3 ventanas de análisis de 1 segundo, frecuencia de muestreo de 5000 Hz y solapamiento del 50%. En verde se encuentra la primera ventana de análisis, en azul la segunda y en anaranjado la tercera.

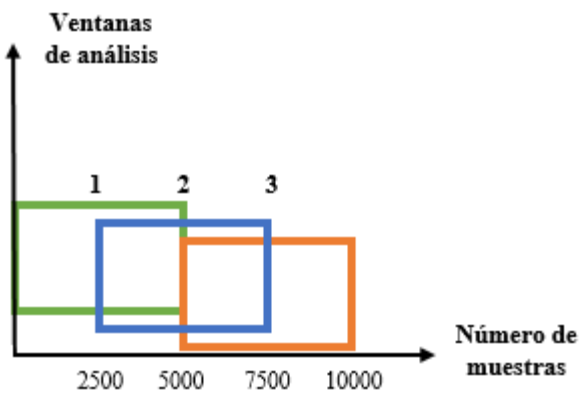


Figura 15. Muestra de ventanas de 1 segundo con solapamiento del 50%

3.5.2 Descomposición *wavelet*

Cada ventana de análisis de los subsegmentos EEG es descompuesta en un número determinado de *n* sub-bandas de frecuencia, por medio de la descomposición *wavelet.* Para tal propósito, se debe afinar los parámetros DWT óptimos, por medio de diferentes pruebas con distintas configuraciones. De esta manera, se encuentra la función *wavelet* y el número de niveles de descomposición más adecuados para un buen desempeño del algoritmo.

La *wavelet Daubechies 4* (*db4*) y la *wavelet Haar*, por su uso en áreas de biomedicina, han sido empleadas en este proyecto.

Para encontrar el número óptimo de niveles de descomposición se usa la función *vmaxlev,* de *Matlab,* la cual retorna el número máximo de niveles *wavelet* permitidos, ingresando el número de muestras analizadas de la señal y el tipo de función *wavelet* utilizada. En general, se toma un valor menor al máximo posible para evitar "efectos de frontera".

La transformada *wavelet* se calcula como el desplazamiento de la función *wavelet* en el tiempo a lo largo de la señal de entrada y el cálculo de la convolución de los mismos. En la mayoría de las aplicaciones prácticas, las señales de interés tienen soporte finito y a medida que la *wavelet* se acerca al borde de la señal, el cálculo de la convolución requiere de valores no existentes, más allá de la frontera de la señal. Esto crea "efectos de frontera", causados por la información incompleta en las regiones límite, haciendo que los resultados de la transformación *wavelet* en estas regiones sean imprecisos (Queiroz, 1992).

De esta manera, usando la función *vmaxlev,* para una ventana de análisis de 1 segundo de duración y una frecuencia de muestreo de 5000 Hz, se obtiene un número máximo de niveles de descomposición de: 12 para *wavelet Haar* y 9 para *wavelet db4.*

Para el caso de *wavelet db4* se decide emplear 8 niveles de descomposición, mientras que, para *wavelet Haar* se selecciona 11 niveles de descomposición. Con esta última, al tener más niveles de descomposición, se realiza un segundo análisis

empleando 9 niveles, con el fin de que las frecuencias de cada nivel concuerden con las bandas EEG (delta, teta, alfa, beta y gamma) de las ondas cerebrales.

Es así que, se tienen 3 posibles configuraciones de prueba para la elección de los parámetros DWT óptimos:

- *Wavelet db4* con 8 niveles de descomposición
- *Wavelet Haar* con 11 niveles de descomposición
- *Wavelet Haar* con 9 niveles de descomposición

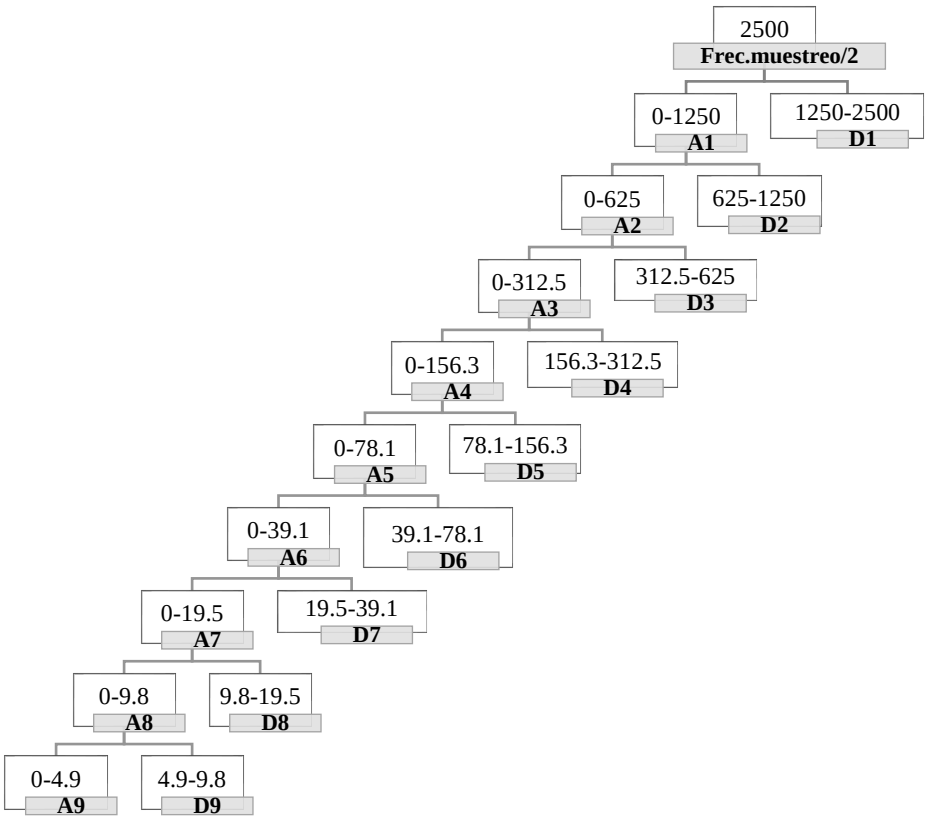
Para las pruebas, se emplea el algoritmo en los subsegmentos EEG, usando la información de todos los canales y pacientes en conjunto, con parámetros de configuración aleatorios, de entre los analizados en este proyecto, resultando en los siguientes: las características tasa de cruces por cero, máximo, mínimo, media y desviación estándar; el método de validación cruzada con k=10 *folds*; y el clasificador árboles de decisión complejo (100 divisiones). Los resultados de las pruebas realizadas son presentados en la tabla 2.

Tabla 2
Elección de parámetros DWT óptimos

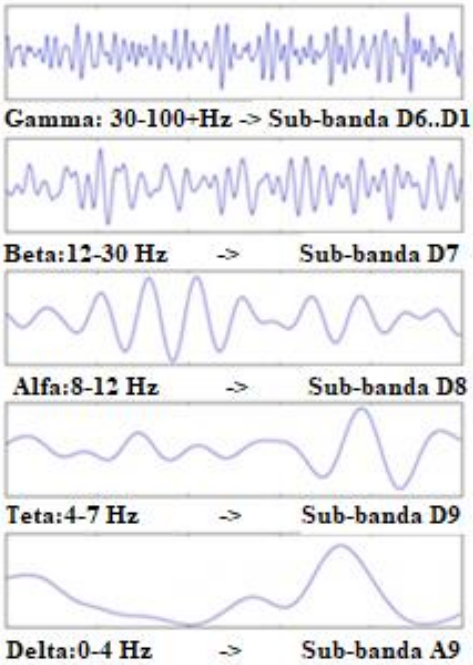
Función *wavelet*, #niveles	**Db4, 8 niveles**	**Haar, 11 niveles**	**Haar, 9 niveles**
Exactitud del algoritmo (%)	99.3	97.4	97.8

Se escoge de esta manera la función *wavelet Daubechies 4* (*db4*) con 8 niveles de descomposición, por ser la configuración que mejor desempeño entrega al algoritmo.

Una representación gráfica de la descomposición *wavelet*, usando *wavelet Haar* con 9 niveles de descomposición, es presentada en la figura 16, donde las sub-bandas de frecuencia *wavelet* A9, D9, D8, D7 y D6-D1 son equivalentes a las bandas EEG delta, teta, alfa, beta y gamma respectivamente. Las frecuencias en esta figura se miden en Hz.



(a)



(b)

Figura 16. (a) Descomposición *DWT* (b) Comparación con bandas EEG

Para la implementación de la descomposición *wavelet*, se emplea la función *wavedec*, de *Matlab*, que ingresando el tipo de función *wavelet* y el número de niveles de descomposición, retorna la descomposición frecuencial de la señal en el vector *[C,L]*, donde *C* contiene los coeficientes *wavelet* de cada sub-banda y *L* es el vector donde cada elemento es la longitud de dichos coeficientes.

Para la reconstrucción de cada sub-banda *wavelet* se emplea la función *wrcoef*, de *Matlab*, misma que reconstruye la señal, ingresando el vector *[C,L]*, la función *wavelet* utilizada, el tipo de coeficiente (aproximado o detallado) y el nivel de descomposición de la sub-banda requerida.

3.5.3 Extracción de características

Una vez descompuesta cada ventana de los subsegmentos EEG en sus respectivas sub-bandas de frecuencia *wavelet*, de cada sub-banda se extrae una serie de características (*features*) para formar el vector de características.

Como un análisis adicional, con propósito únicamente comparativo, y debido a la naturaleza no lineal y no estacionaria de las señales EEG, es necesario realizar un estudio empleando una medida no-lineal. De esta forma, se tiene un análisis de como un conjunto de características tanto lineales como no-lineales son capaces de modelar el comportamiento de las señales cerebrales en pacientes con epilepsia. El primer estudio emplea características exclusivamente lineales, aplicadas a las sub-bandas *wavelet* obtenidas, mientras que, el segundo estudio emplea una característica no-lineal, aplicada directamente a las ventanas de análisis de los subsegmentos, es decir no se emplea descomposición *wavelet.*

3.5.3.1 Análisis lineal

Por su capacidad de rendimiento, bajo costo computacional y fácil implementación, se decidió utilizar las siguientes características lineales: tasa de cruces por cero, máximo, mínimo, media, desviación estándar, asimetría, curtosis, varianza, potencia y energía de sub-banda *wavelet.*

En este análisis, se extrae las características lineales de cada una de las sub-bandas *wavelet* obtenidas a partir de las ventanas de análisis de los subsegmentos EEG preseleccionados.

Las *features* lineales se implementaron en *Matlab* por medio de las siguientes funciones:

- Tasa de cruces por cero.-

H = dsp.ZeroCrossingDetector

ZD=double(step(H,señal))

- Máximo.-

Max=max(señal)

- Mínimo.-

Min=min(señal)

- Media.-

Mean=mean(señal)

- Desviación estándar.-

StdD=std(señal)

- Media.-

Mean = mean(señal)

- Varianza.-

Var = var(señal)

- Curtosis.-

Curt=kurtosis(señal)

- Asimetría.-

Asim=skewness(señal)

- Potencia.-

Pot= sum(señal.^2)/length(señal)

- Energía de sub-banda *wavelet*.-

[Ea,Ed] = wenergy(C,L)

3.5.3.2 Análisis no-lineal

Por su bajo costo computacional, se decidió emplear la Complejidad de Lempel-Ziv (LZ) (usada comúnmente en el procesamiento de señales EEG) como característica no-lineal.

En este análisis, a diferencia del análisis lineal, la medida de complejidad LZ es tomada directamente de las ventanas de análisis de los subsegmentos EEG, sin necesidad de realizar la descomposición *wavelet* de cada ventana previamente. Esto es, debido a que la complejidad LZ mide el grado de orden o de desorden de patrones espacio-temporales de la señal EEG original. Al ser una medida muy sensible a los cambios temporales, es idónea para usarla en señales de actividad cerebral.

La Complejidad de Lempel-Ziv (LZ) se implementa en *Matlab* por medio de la función *lempelziv* (Mora López Tercero, 2013):

[c,l,x]=lempelziv(señal,[vmin vmax n])

LZ=c(2),

donde, *c(2)* es el coeficiente *LZ* normalizado, *l* la longitud de la señal binarizada, *x* la señal binarizada, *vmin* el valor mínimo de la señal EEG, *vmax* el valor máximo de

la señal EEG y *n* el número de niveles para la transformación de la señal (en este caso binarización).

3.6 Etapa de clasificación y evaluación

Una vez obtenidos los vectores de características, es necesario crear los modelos de predicción de crisis epilépticas. Para esto, se emplea varios algoritmos de aprendizaje de maquina supervisado (clasificadores) y se prueba varias configuraciones para cada uno de ellos, con el fin de obtener los parámetros que mejor desempeño entregan a cada clasificador. Una vez obtenidos los clasificadores óptimos, se procede a la creación de los modelos de predicción. Los resultados de cada modelo deben ser validados, para obtener el desempeño más cercano al real de cada uno de ellos. Finalmente, este desempeño debe ser medido por medio de métricas de desempeño, para verificar cuál es el rendimiento que presenta cada modelo de predicción obtenido. Adicionalmente, debido a que existen pruebas con un número de predictores en el vector de características superior a los 200, es necesario emplear una técnica de reducción de dimensionalidad, con el fin de mejorar el costo computacional y de evitar la sobrealimentación del clasificador.

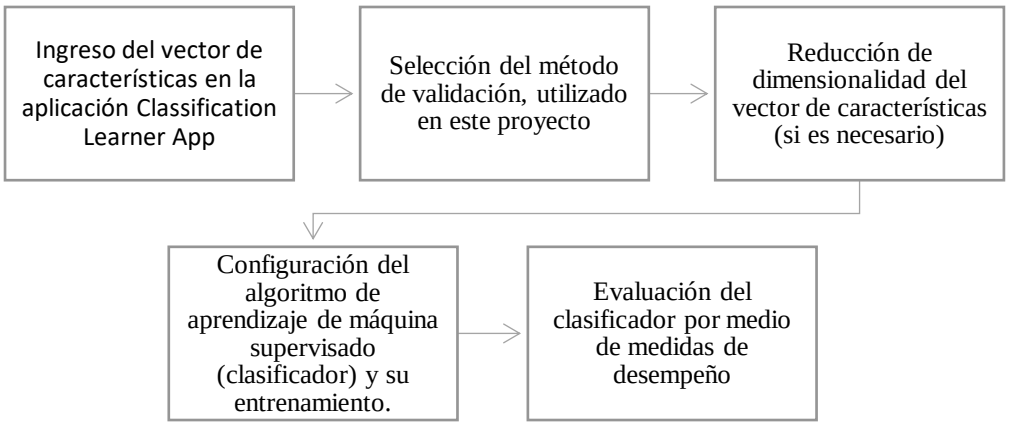


Figura 17. Algoritmo de clasificación y evaluación de modelos

La figura 17 muestra el algoritmo de procedimiento de la etapa de clasificación y evaluación, a partir de los vectores de características obtenidos en la etapa de procesamiento. A continuación se detallan los procesos realizados en esta etapa.

3.6.1 Validación de resultados

Para verificar el desempeño real del algoritmo, se debe emplear un método de validación de resultados. En este proyecto se emplea *Validación cruzada de k-iteraciones* (*k-folds*), que es una técnica bastante acertada a la hora de medir el desempeño real del algoritmo, ya que se basa en la media de varias iteraciones de validación, cada una con diferentes divisiones de los sets de entrenamiento y prueba, es decir, garantiza que los resultados de un modelo de clasificación sean independientes de la partición entre datos de entrenamiento y prueba.

Para tener un desempeño más próximo al real del algoritmo, es necesario utilizar un número óptimo de *k-folds* de validación cruzada, debido a que un exceso en este número puede provocar un excesivo tiempo de cómputo, por el contrario, con un valor pequeño de *k-folds* no se asegura la medición real del desempeño del algoritmo.

En minería de datos y aprendizaje de máquina, validación cruzada con 10-*folds* es el método de validación más común (Refaeilzadeh et al., 2009). De esta manera, se eligió $k = 10$ *folds*, q entrega tiempo de cómputo bajo y valor de desempeño cercano al real del modelo.

3.6.2 Reducción de dimensionalidad

Algunas pruebas realizadas contienen vectores de características con un número de predictores superior a los 200. Para dichas pruebas, es necesario emplear un método de reducción de dimensionalidad del vector de características, con el objetivo de reducir el tiempo de cómputo y evitar una posible sobrealimentación del clasificador. Esto se realiza por medio del Análisis de Componentes Principales (*Principal Component Analysis*, PCA).

PCA analiza el vector de características que representa las observaciones descritas por varias variables dependientes o predictores, que son, en general, inter-correlacionadas, con el objetivo de comprimir el tamaño del vector de características, manteniendo sólo la información importante de los datos, y expresa esta información

como un conjunto de nuevas variables ortogonales llamadas componentes principales (Hotelling, 1933).

Debido que los canales de grabación EEG se encuentran ubicados en posiciones físicas muy cercanas entre sí, sobre la corteza cerebral, los predictores resultantes al emplear dichos canales están inter-correlacionados.

La implementación se realiza por medio de la aplicación *Classification Learner App*, del *Statistics and Machine Learning Toolbox*, de *Matlab*. En esta aplicación, el usuario puede ingresar como criterio de reducción de componente una de las dos posibles configuraciones:

- **Valor de varianza del componente.-** donde la aplicación retorna el número de componentes principales, en base a la cantidad de varianza capturada por estos, es decir, la información que representa cada uno de ellos.
- **Número de componentes.-** donde la aplicación retorna el número de componentes principales, ingresado por el usuario.

De tal forma que, para tener un número variable de componentes principales, con el fin de tener el número óptimo de componentes, se decidió elegir como criterio de reducción de componente: *valor de varianza del componente*, con un valor de varianza del 95%, es decir, el primer componente principal contiene el 95% de la información, el segundo un porcentaje menor y así sucesivamente.

3.6.3 Configuración y entrenamiento de modelos

Para la creación de los modelos de predicción de crisis epilépticas, es necesario emplear distintos algoritmos de aprendizaje de máquina supervisado (clasificadores), y variar sus parámetros de configuración, con el fin de obtener los parámetros y clasificadores óptimos, que mejor desempeño entreguen al algoritmo. En este proyecto, los clasificadores empleados son los siguientes:

- Regresión logística (RL)
- Análisis discriminante (AD)

- *k*-Vecinos más cercanos (kNN)
- Árboles de decisión
- Máquina de vectores de soporte (SVM)
- Conjunto de clasificadores redundantes (*Ensemble*)

Los modelos de predicción son entrenados y probados, en base al etiquetado de cada observación del vector de características. Las observaciones correspondientes a muestras interictales se etiquetan con la clase “0”, mientras que las muestras preictales se etiquetan con la clase “1”.

Para la creación y entrenamiento de los modelos de predicción, se emplea la aplicación *Classification Learner App*. Se realizaron pruebas con distintas configuraciones de parámetros en cada clasificador, para obtener los parámetros óptimos. La tabla 3 resume todos los parámetros de configuración empleados.

Tabla 3
Parámetros de configuración probados en cada clasificador

Clasificador	**Configuraciones**					
RL	Función logística					
AD	Función lineal	Función cuadrática				
kNN	Función fina (1 vecino)	Función media (10 vec.)	Función gruesa (100 vec.)	Función coseno (10 vec.)	Función cúbica (10 vec.)	Función ponderada (10 vec.)
Árboles	Simple (4 divisiones)	Medio (20 divs.)	Complejo (100 divs.)			
SVM	Función lineal	Función cuadrática	Función cúbica	Función gaussiana fina (escala kernel 5.5)	Función gaussiana media (escala kernel 22)	Función gaussiana gruesa (escala kernel 88)
Ensemble	Boosted trees	Bagged trees	Subespacio AD	Subespacio kNN	RUSBoosted trees	

3.6.4 Evaluación de modelos

Finalmente, una vez creados los modelos de predicción, es necesario evaluar el rendimiento de cada uno de ellos. Para tal motivo, se debe utilizar medidas de desempeño, que den un valor del desempeño real de cada modelo en la predicción de crisis epilépticas.

En este proyecto, se emplea las siguientes medidas de desempeño:

- Exactitud
- Sensibilidad
- Especificidad
- Área bajo la curva ROC (AUR)

La implementación de la evaluación se realiza por medio de la aplicación *Classification Learner App*, misma que proporciona al usuario la opción de generar la matriz de confusión y curva ROC respectiva, para cada modelo de predicción. De esta manera, por medio de la matriz de confusión se puede calcular los valores en porcentaje correspondientes a exactitud, sensibilidad y especificidad, y por medio de la curva ROC, se puede calcular el área bajo la curva ROC (AUR).

La medida de desempeño AUR da un estimado de la capacidad predictiva que entrega cada modelo de predicción de crisis epilépticas.

CAPÍTULO IV

4. RESULTADOS Y ANÁLISIS

4.1 Introducción

Una vez obtenidos los parámetros óptimos de diseño del algoritmo de predicción de crisis epilépticas (porcentaje permitido de pérdidas del 2%, ventana de 1 segundo de duración con solapamiento del 50%, función *wavelet db4* con 8 niveles de descomposición y validación cruzada de 10 *folds*) es necesario seleccionar el análisis (lineal o no-lineal) más conveniente para el sistema, es decir aquel que presente mejor desempeño.

En este capítulo, se presenta los resultados obtenidos del estudio comparativo de emplear tanto análisis lineal (características lineales) como análisis no-lineal (característica no-lineal) en el algoritmo de predicción de crisis epilépticas. Estos análisis son implementados en diversas pruebas, ya sea empleando la información individual de cada canal y cada paciente, o en conjunto, de todos los canales y todos los pacientes.

Se expone también: el algoritmo de aprendizaje de máquina supervisado óptimo para la predicción de crisis epilépticas; el canal, paciente y prueba que mejor desempeño presentaron; los resultados de emplear modelos de predicción sin/con el uso de reducción de dimensionalidad y el costo computacional del algoritmo.

Además, se presentan los resultados de emplear el algoritmo en todos los segmentos EEG de 10 minutos, correspondientes a los 10,20, 30, 40, 50 y 60 minutos dentro de los clips EEG de una hora, para probar el desempeño del algoritmo conforme aumenta el tiempo anterior al inicio de la crisis.

Finalmente, se detallan resultados paralelos, empleando el algoritmo de predicción de crisis epilépticas propuesto, en una base de datos iEEG más extensa, que consta de

tres pacientes con epilepsia, 16 canales de grabación EEG por paciente y un número total de 900 segmentos EEG de 10 minutos.

4.1 Casos de estudio

Primeramente, al tener una base de datos iEEG con mucha información (15 canales y 2 pacientes), es necesario delimitar las condiciones de las pruebas necesarias, es decir diseñar pruebas cuyos resultados puedan ser comparativos y se pueda extraer conclusiones a partir de ellos. Para esto, se han planteado pruebas tanto de información individual como de información general, respecto a los canales de grabación EEG y a los pacientes estudiados.

De esta manera, para cada análisis (lineal o no-lineal) implementado en el algoritmo, se ha delimitado una serie de cuatro pruebas respectivas, denominadas en este capítulo como "Casos de estudio". Por lo tanto, como la base de datos empleada cuenta con 2 pacientes y 15 canales de grabación EEG por paciente, los casos de estudio propuestos son los siguientes:

(a). Caso A: estudio de cada paciente y cada canal.- en este estudio se realiza el análisis de cada uno de los canales, que posee cada uno de los pacientes. Formando un total de 30 instancias de análisis diferentes. Esto se muestra en la figura 18.

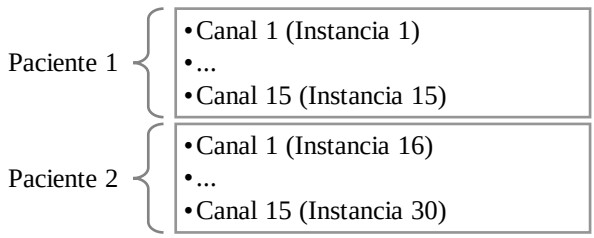


Figura 18. Representación caso de estudio A

(b). Caso B: estudio de cada paciente y todos los canales.- en este estudio se realiza el análisis de todos los canales en conjunto, para cada uno de los

pacientes. Formando un total de 2 instancias de análisis diferentes. Esto se muestra en la figura 19.

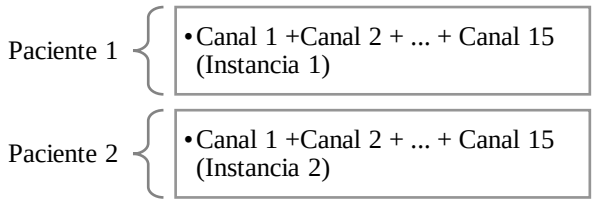


Figura 19. Representación caso de estudio B

(c). Caso C: estudio de todos los pacientes y cada canal.- en este estudio se realiza el análisis de todos los pacientes en conjunto, para cada uno de los canales. Formando un total de 15 instancias de análisis diferentes. Esto se muestra en la figura 20.

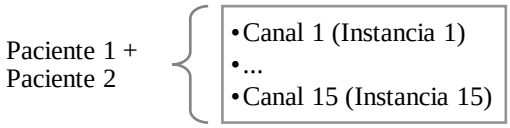


Figura 20. Representación caso de estudio C

(d). Caso D: estudio de todos los pacientes y todos los canales.- en este estudio se realiza el análisis de todos los pacientes en conjunto, para todos los canales en conjunto, es decir se analiza toda la base de datos. Formando una sola instancia de análisis. Esto se muestra en la figura 21.

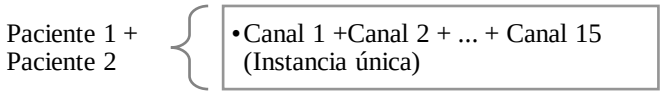


Figura 21. Representación caso de estudio D

4.2 Selección de características (análisis lineal y no-lineal)

Como se menciona en la introducción de este capítulo, es necesario verificar que tipo de análisis (lineal o no-lineal) es el más óptimo para modelar las señales EEG, con el objetivo de predecir crisis epilépticas.

Antes de ir con los resultados de los dos tipos de análisis, es necesario mencionar que, inicialmente se realizó el estudio de estos dos análisis con subsegmentos EEG de los últimos 10 segundos anteriores al inicio de la crisis epiléptica (sin tomar en cuenta el horizonte de cinco minutos), y se obtuvo exactitudes que van de 90% a 100%, esto asegura que, el algoritmo efectivamente logra diferenciar los estados interictal y preictal, usando los últimos 10 segundos anteriores a una crisis. De este análisis, se obtuvo un vector de características total (todos los canales y pacientes en conjunto) de 228 observaciones y 675 predictores. Como se puede notar, el número de observaciones es relativamente pequeño, por tanto, para aumentar dicho número se pueden realizar dos procedimientos: disminuir el valor de tiempo de la ventana o aumentar el tiempo de análisis dentro cada segmento EEG de los últimos 10 minutos anteriores a la crisis.

Para la elección del procedimiento adecuado, se tienen dos consideraciones:

- Al disminuir el tiempo de la ventana de análisis, disminuye la escala de frecuencia de la señal EEG, y por ende se pierde información de baja frecuencia, que en EEG son las bandas con mayor potencia.
- Al aumentar el tiempo de análisis dentro de los segmentos EEG de los últimos 10 minutos anteriores a la crisis, se analiza un número mayor de muestras y por tanto el desempeño obtenido del algoritmo es más próximo al real.

De esta forma, el procedimiento más adecuado es aumentar el tiempo de análisis dentro de cada segmento EEG de los últimos 10 minutos anteriores a la crisis. Para tener resultados de desempeño más cercanos a los reales del algoritmo, se decidió emplear todas las muestras de dicho segmento EEG, es decir utilizar el segmento entero.

Las configuraciones óptimas de los clasificadores fueron elegidas dando prioridad a la sensibilidad, pero se tabularon los valores de exactitud. Se da prioridad a la sensibilidad, ya que se necesita el mayor porcentaje de muestras preictales clasificadas correctamente, debido a que, en caso de presentarse una crisis epiléptica y el paciente no sea informado por el algoritmo (falso negativo) podrían presentarse consecuencias

fatales, mientras que, si no se presenta una crisis epiléptica y el paciente es informado con una alarma de presencia de crisis (falso positivo) esto no conllevaría consecuencias fatales, porque el paciente tomaría las debidas precauciones, a pesar de no existir una crisis venidera.

Las tablas de resultados para cada análisis muestran además de los resultados de exactitud, también el número de observaciones (#obs) y de predictores (#pre) del vector de características de cada prueba. Los algoritmos de clasificación con mejor desempeño se encuentran resaltados.

A continuación se muestra los resultados obtenidos de la implementación en el algoritmo de predicción de crisis epilépticas, de los dos análisis (lineal y no-lineal), empleando los cuatro casos de estudio propuestos anteriormente y un estudio comparativo de ambos tipos de análisis.

4.2.1 Análisis lineal

Es el análisis de enfoque de este proyecto de investigación, y se realiza mediante la extracción de varias características lineales de cada una de las sub-bandas *wavelet*, obtenidas de cada ventana de análisis de las señales EEG.

Inicialmente, se empleó las siguientes características lineales: tasa de cruces por cero, máximo, mínimo, media, desviación estándar, varianza, curtosis, asimetría, potencia y energía de sub-banda *wavelet*. Entre estas características se establecieron varias combinaciones, obteniendo de esta manera, los dos sets de *features* que mejor desempeño presentaron, y estos son:

- Primer set de *features* lineales: tasa de cruces por cero, máximo, mínimo, media y desviación estándar.
- Segundo set de *features* lineales: media, varianza, curtosis, asimetría, potencia y energía de sub-banda *wavelet*.

De entre estos dos sets, el que mejor desempeño obtuvo fue el primero, con una exactitud general (todos los canales y pacientes en conjunto) de alrededor del 99.9%,

mientras que, con el segundo set se obtuvo una exactitud general de alrededor del 92%. Por este motivo, se eligió el primer set de características para los resultados del análisis lineal. A continuación se presentan dichos resultados, empleando todos los casos de estudio definidos anteriormente.

Tabla 4
Exactitud obtenida Análisis Lineal - Caso A - Pacientes 1 y 2

Pac.	#obs, #pred	Canal	Árbol	AD	RL	SVM	KNN	Ensemb.
1	7194,	CH1	97.9	71.9	85.8	93.7	92.4	98.9
	45	CH2	97.9	62.9	82.5	94.4	93.3	98.7
		CH3	98	66.1	84.8	95.3	93.8	99.1
		CH4	99.6	59.3	90.6	94.1	92.3	99.9
		CH5	98.5	67.5	87.8	95.9	95	99.2
		CH6	98.6	64.8	84.7	95.4	94.5	99.6
		CH7	99.3	64.7	85.6	95	94	99.8
		CH8	98.4	69.8	89.1	94.7	93.4	99.1
		CH9	97.1	74.6	87.9	94.4	93	98.2
		CH10	98.4	72.7	87.9	94.4	93.1	98.9
		CH11	98.1	70.6	88.1	95.3	94.1	99.3
		CH12	98.6	64.7	87.7	94.1	91.8	99.3
		CH13	98.7	66	86.7	94.6	92.9	99.5
		CH14	98.4	64.9	85.4	95	93.2	99.1
		CH15	98.7	68.9	88.5	94.6	93.5	99.4
2	7194,	CH1	99.1	96.5	99.4	99.7	98.8	99.7
	45	CH2	96.9	93	98.6	98.9	97.6	98.9
		CH3	94.9	85.1	97.4	97.5	93.7	97.1
		CH4	96.1	91	97.9	97.3	93.8	97.6
		CH5	96.4	94.9	98.8	98.9	96.6	98.5
		CH6	98.9	96.6	99.4	99.8	99.4	99.8
		CH7	99	97.2	99.5	99.9	99.7	99.8
		CH8	98.5	96.4	99.2	99.7	99.1	99.5
		CH9	98.5	94.9	99.2	99.6	98.6	99.3
		CH10	98.5	95.2	99	99.6	99.1	99.4
		CH11	99	94.7	99.4	99.8	99.4	99.8
		CH12	97.1	94.2	99	99.2	97.7	98.6
		CH13	96	92.1	98.9	99.1	97	98.2
		CH14	98.7	95.5	99.3	99.7	99.3	99.7
		CH15	98.7	97	99.3	99.6	99	99.5

Tabla 5
Exactitud obtenida Análisis Lineal - Caso B - Pacientes 1 y 2

Pac.	#obs, #pred	Árbol	AD	RL	SVM	KNN	Ensem.
1	7194, 675	99.5	66.3	84.5	99.8	99.4	99.9
	7194, 3 (con PCA)	83.9	65.7	64.1	66.1	85.9	89.7
2	7194, 675	99.2	94.5	97.2	99.9	99.8	99.9
	7194, 6 (con PCA)	96.5	91.3	96.5	97.7	96.7	97.5

Tabla 6
Exactitud obtenida Análisis Lineal - Caso C

#obs, #pred	Canal	Árbol	AD	RL	SVM	KNN	Ensemble
14388, 45	CH1	98.3	79.9	86.3	95.1	93.5	99.3
	CH2	97.4	73.6	85.3	95.3	93	98.6
	CH3	96	76.6	83	94.5	92	97.7
	CH4	97.4	71.9	85.5	94.1	91.8	98.6
	CH5	97	82.2	87.2	94.5	93	98.6
	CH6	98.6	81.9	84.8	95.8	94.6	99.6
	CH7	99	81.2	86.3	96.4	95.4	99.8
	CH8	98.4	79.2	86.2	94.6	94.1	99.2
	CH9	97.2	76.4	85.5	94.7	92.9	98.6
	CH10	98	81.9	84.9	95.2	94.5	99.2
	CH11	98.2	78.9	85.1	96	95.1	99.3
	CH12	97.5	76.4	85	93.9	92.6	98.9
	CH13	98.2	80.2	85.3	95.9	94.1	99.3
	CH14	98.6	79.2	88.4	95.7	94.3	99.4
	CH15	98.6	79.1	88.1	94.3	94.3	99.4

Tabla 7
Exactitud obtenida Análisis Lineal - Caso D

#obs, #pred	Árbol	AD	RL	SVM	KNN	Ensemb.
14388, 675	99	84.2	87.5	99.7	98.9	99.9
14388, 3 (con PCA)	73.2	57.6	76.1	58.2	73.8	79.7

La tabla 8 y figura 22 presentan la matriz de confusión y curva ROC (AUR redondeado) respectivamente, del desempeño general del sistema al implementar el análisis lineal con la información de todos los canales y pacientes en conjunto (caso de estudio D). El clasificador óptimo para el algoritmo fue *ensemble-bagged trees*, con una exactitud, sensibilidad, especificidad y AUR de 99.9%, 99.8%, 99.9% y 0.99 respectivamente, sin usar reducción de dimensionalidad.

Tabla 8
Matriz de confusión Análisis Lineal - Caso D

MATRIZ DE CONFUSIÓN		VALOR REAL	
		Interictal	**Preictal**
VALOR DE PREDICCIÓN	**Interictal**	7183	11
	Preictal	9	7185

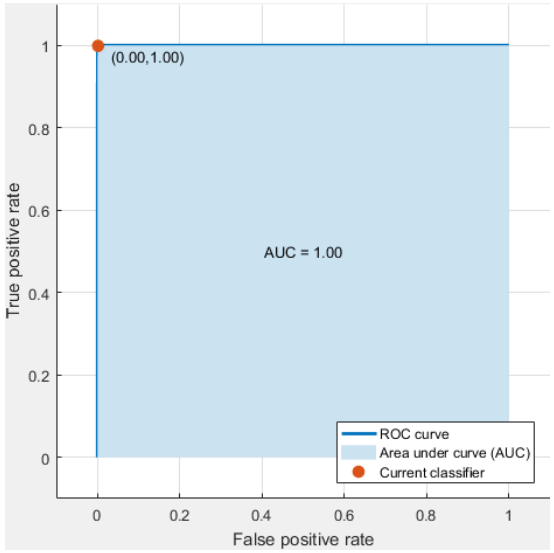


Figura 22. Curva ROC Análisis Lineal - caso D

4.2.2 Análisis no-lineal

Es el análisis de propósito comparativo en este proyecto de investigación, y se realiza mediante la extracción de una característica no-lineal directamente de cada una de las ventanas de análisis de las señales EEG.

Como característica no-lineal, se empleó la siguiente medida de complejidad: Complejidad de Lempel-Ziv.

El desempeño obtenido fue una exactitud general (todos los canales y pacientes en conjunto) de alrededor del 86.1%. A continuación se presentan dichos resultados, empleando todos los casos de estudio definidos anteriormente.

Tabla 9
Exactitud obtenida Análisis No-lineal - Caso A - Pacientes 1 y 2

Pac.	#obs, #pred	Canal	Árbol	AD	RL	SVM	KNN	Ensemb.
1	7194,	CH1	75.2	58.4	57.6	75.9	74.7	75.7
	1	CH2	67.6	54.1	54.1	65.9	64.8	67.9
		CH3	72.7	63.8	63.5	72.8	53	72.6
		CH4	70.3	60.3	59.8	69.8	70.3	70.3
		CH5	77.7	62.9	65.3	78.4	73.4	77.9
		CH6	73.5	67	63.7	73.3	52.5	73.4
		CH7	71.2	61	61	69.7	52.1	71.2
		CH8	68	61.8	60.7	68.8	54.2	68.5
		CH9	67.6	59.9	59.9	67.6	62.2	67.5
		CH10	76.9	69	65.6	76.3	64.4	77.1
		CH11	66.8	64	62.1	67.4	53	66.5
		CH12	61.8	58.9	59	63	53	62.2
		CH13	65.9	62.9	60.6	66.6	52.5	66.4
		CH14	62.4	58.8	56.6	62.4	60.4	62.4
		CH15	73.6	51.9	51.9	74	57	74
2	7194,	CH1	59.1	60.1	54.1	59.9	56.8	59
	1	CH2	58.6	59.6	59.6	59.6	57.1	59.6
		CH3	58.1	55.8	54.8	57.9	56.5	57.8
		CH4	63.7	64.2	61.1	64.2	64.7	62.9
		CH5	63.6	64.7	63.4	64.4	56.5	63.7
		CH6	66.1	66.3	65.8	66.5	55.6	66.5
		CH7	64.5	63.5	63.8	64.8	54.5	64.6
		CH8	63.9	62.7	62.8	63.4	51.3	63.9
		CH9	68.8	66.6	61.3	68.8	67.8	68.4
		CH10	64.4	64.1	58.8	63.5	62.2	64.3
		CH11	64.8	63.9	63.8	64.7	64.2	64.8
		CH12	61.5	61.3	60.1	62.5	60.6	60.8
		CH13	65.3	65.6	64.8	65.4	63.5	65.1
		CH14	64.5	63.7	63.1	64.8	64.4	63.3
		CH15	63.8	62.2	61.4	64	54.6	63.8

Tabla 10
Exactitud obtenida Análisis No-lineal - Caso B - Pacientes 1 y 2

Pac.	#obs, #pred	Árbol	AD	RL	SVM	KNN	Ensem.
1	7194, 15	90	61.5	74	96.9	97.1	96.6
	7194, 10 (con PCA)	86	66.3	68.1	92.1	92.5	94.6
2	7194, 15	72.5	69.7	69.9	79.2	76.8	77.7
	7194, 11 (con PCA)	70.5	70.1	69.9	76.4	73.1	76

Tabla 11
Exactitud obtenida Análisis No-lineal - Caso C

#obs, #pred	Canal	Árbol	AD	RL	SVM	KNN	Ensemble
14388, 1	CH1	67.9	54.3	54.3	67.9	66.2	68
	CH2	60.2	55.3	53.4	59.8	58	60.3
	CH3	60	55.9	53.4	60.2	58.3	59.9
	CH4	65.3	59.3	59.4	65.2	53.2	65.2
	CH5	71.2	63.7	65.4	71.4	57.1	71.3
	CH6	67.5	66.5	65	68	51.6	67.1
	CH7	66.2	63.7	62.1	66	51.5	66.2
	CH8	64.2	61.9	61.8	64.8	52.3	64.3
	CH9	67	58.8	58.3	67.5	60.3	67.2
	CH10	69.1	62.1	61	68.9	53.6	68.8
	CH11	65.9	63.8	62	66.1	55.8	65.9
	CH12	60.7	56.7	57	61.3	53.4	61
	CH13	65.5	61.8	60.5	65.9	54.1	64.6
	CH14	63	62.3	59	63.2	53.1	62.9
	CH15	67.1	56.9	57.2	67.4	53.3	67.7

Tabla 12
Exactitud obtenida Análisis No-lineal - Caso D

#obs, #pred	Árbol	AD	RL	SVM	KNN	Ensem.
14388, 15	79.3	62.2	66.1	86	86	86.1
14388, 10 (con PCA)	75.3	63.8	64.3	81.2	81.8	82.3

La tabla 13 y figura 23 presentan la matriz de confusión y curva ROC respectivamente, del desempeño general del sistema al implementar el análisis no-lineal con la información de todos los canales y pacientes en conjunto (caso de estudio D). El clasificador óptimo para el algoritmo fue *ensemble-bagged trees*, con una exactitud, sensibilidad, especificidad y AUR de 86.1%, 84%, 89% y 0.94 respectivamente, sin usar reducción de dimensionalidad.

Tabla 13
Matriz de confusión Análisis No-lineal - caso D

MATRIZ DE CONFUSIÓN		VALOR REAL	
		Interictal	**Preictal**
VALOR DE PREDICCIÓN	**Interictal**	6259	935
	Preictal	1079	6115

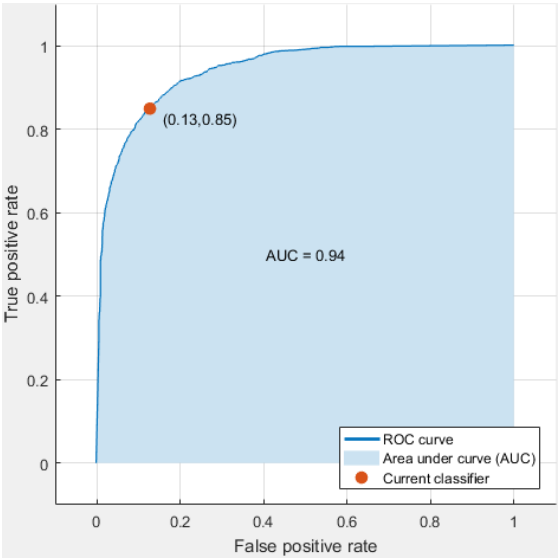


Figura 23. Curva ROC Análisis No-lineal - caso D

4.2.3 Resumen y análisis de resultados

A continuación se presenta un estudio comparativo de los análisis lineal y no-lineal, implementados en el algoritmo de predicción de crisis epilépticas propuesto. La tabla 14 muestra el resumen de resultados usando la información de todos los canales y pacientes en conjunto. Y se realiza el análisis respectivo, presentando los canales, clasificadores, pacientes con mejor desempeño y el tiempo de cómputo empleado por el algoritmo.

Tabla 14
Resumen de resultados Análisis Lineal y No-Lineal

Análisis	Caso	Pac.	Exact	Sens	Espec	AUR	Canal	Clasific.
Lineal	**A**	1	99.9	99.8	99.9	0.99	4	Ensemb.
		2	99.9	99.9	99.9	0.99	7	SVM
	B	1	99.9	99.9	99.8	0.99	Todos	Ensemb.
		2	99.9	99.9	99.9	0.99	Todos	Ensemb.
	C	Todos	99.8	99.7	99.8	0.99	7	Ensemb.
	D	Todos	99.9	99.8	99.9	0.99	Todos	Ensemb.
No Lineal	**A**	1	78.4	80	77	0.83	5	SVM
		2	68.8	68	70	0.74	9	Árbol
	B	1	97.1	95	99	0.97	Todos	kNN
		2	79.2	78	80	0.85	Todos	SVM
	C	Todos	71.4	70	72	0.77	5	SVM
	D	Todos	86.1	84	89	0.94	Todos	Ensemb.

Se puede verificar que el análisis lineal tiene un desempeño notablemente mejor que el análisis no-lineal, con exactitud, sensibilidad, especificidad y AUR de 99.9%, 99.8%, 99.9% y 0.99 respectivamente, para el primer caso, y exactitud, sensibilidad, especificidad y AUR de 86.1%, 84%, 89% y 0.94 respectivamente, para el segundo caso.

Tomando en cuenta que el análisis lineal emplea 675 predictores (9 sub-bandas *wavelet* x 5 características x 15 canales) y el análisis no-lineal solamente 15 predictores (1 característica x 15 canales), los resultados de este último son bastante aceptables. Recordar que los dos análisis utilizan el mismo número de observaciones (14388 observaciones).

De los dos análisis, los clasificadores con mejor desempeño son los siguientes: kNN, árboles de decisión, SVM y e*nsemble-bagged trees.* De los cuatro el de menor tiempo computacional es árboles de decisión y el de mejor desempeño entregado es e*nsemble-bagged trees*. De tal forma que, el clasificador óptimo para el algoritmo de predicción de crisis epilépticas es *ensemble-bagged trees*, que combina la rapidez y desempeño entregados por un conjunto de clasificadores base del tipo "árboles de decisión" con agregación *bootstrap*.

En ambos análisis, el desempeño del algoritmo aumenta notablemente al utilizar la información de todos los canales de grabación EEG. El paciente con mejor desempeño es el paciente 1. Para el análisis lineal, los canales con mejor desempeño entregado son

los canales 4 y 7, mientras que, para el análisis no-lineal, los canales mejor desempeñados son los canales 5 y 9. El uso de PCA (análisis de componentes principales) reduce notablemente la dimensión del vector de características, pero empeora el desempeño del algoritmo.

Respecto al tiempo de cómputo empleado, el análisis lineal conlleva un tiempo total de 2951 segundos (2781 en procesamiento y 170 en clasificación), mientras que, el análisis no-lineal emplea un tiempo de 680 segundos (625 en procesamiento y 55 en clasificación). Por otro lado, al implementar funciones del *Parallel Computing Toolbox*, el tiempo de cómputo disminuye en gran porcentaje, obteniendo así, 1100 segundos (930 en procesamiento y 170 en clasificación) para el análisis lineal y 215 segundos (160 en procesamiento y 55 en clasificación) para el análisis no-lineal.

Se puede verificar que, el análisis óptimo para la implementación del algoritmo de predicción de crisis epilépticas es el Análisis Lineal, empleando la función *wavelet db4* con 8 niveles de descomposición *wavelet*, y las siguientes características lineales: tasa de cruces por cero, máximo, mínimo, media y desviación estándar.

4.3 Análisis en registros EEG de una hora de duración

Una vez realizado el análisis en los últimos 10 minutos anteriores al inicio de una crisis epiléptica (sin tomar en cuenta los cinco minutos de horizonte), es necesario verificar el desempeño del algoritmo conforme el tiempo de análisis anterior al inicio de la crisis va aumentando, con el fin de, comprobar cómo mientras más alejado sea el tiempo anterior al inicio de la crisis, el desempeño del algoritmo debe ir disminuyendo, debido a que, al incrementar de manera significativa este tiempo, el estado preictal tendría interferencia con el estado interictal, por su cada vez más pequeña distancia en tiempo de separación entre estos dos estados.

Para esto, es necesario implementar el algoritmo en los segmentos EEG de 10 minutos, correspondientes a los últimos 60,50,40,30, 20 y 10 minutos anteriores al inicio de la crisis (secuencias 1,2,3,4,5 y 6), que se encuentran dentro de los clips EEG preictales de una hora de duración. Para los clips interictales de una hora se emplean igualmente todas las secuencias de 10 minutos dentro de estos.

Este análisis nos permite verificar el deterioro en el desempeño del algoritmo (exactitud, sensibilidad, especificidad y AUR) conforme aumenta el tiempo anterior al inicio de la crisis. El desempeño general del algoritmo está dado por el promedio de los desempeños individuales obtenidos para cada segmento EEG de 10 minutos, correspondiente a los últimos 60,50,40,30, 20 y 10 minutos anteriores al inicio de la crisis.

Usando el número de registros totales presentados en la tabla 15, los resultados son mostrados en la tabla 16, y se obtuvieron empleando la información de todos los canales y pacientes en conjunto, junto con el clasificador óptimo *ensemble-bagged trees,* y el análisis lineal (*wavelet db4,* 8 niveles y 5 *features* lineales).

Tabla 15
Número de segmentos EEG de 10 minutos totales analizados

Secuencia	Tiempo antes de crisis (minutos)	#Segmentos EEG	#Interictal	#Preictal
1	60	6	3	3
2	50	6	3	3
3	40	6	3	3
4	30	6	3	3
5	20	6	3	3
6	10	6	3	3
TOTAL	1 hora	36	18	18

Tabla 16
Resultados totales para clips EEG de una hora

Secuencia	Tiempo antes de crisis (minutos)	Exact	Sens	Espec	AUR	Clasificador
1	60	99	99	99.1	0.98	Bagged Trees
2	50	98.5	98	99	0.98	Bagged Trees
3	40	98.3	98.4	98.2	0.98	Bagged Trees
4	30	99.2	99.1	99.2	0.99	Bagged Trees
5	20	99.7	99.8	99.6	0.99	Bagged Trees
6	10	99.9	99.9	99.8	0.99	Bagged Trees
MEDIA		99.1	99	99.1	0.98	

Se puede verificar en la tabla 16 como a medida que el tiempo anterior al inicio de la crisis epiléptica (sin tomar en cuenta el horizonte de 5 minutos) va aumentando hasta llegar a una hora, también va disminuyendo el desempeño del algoritmo, sin embargo, su reducción no es significativa.

De la tabla 15 se obtiene un tiempo total de análisis del algoritmo de 36 segmentos EEG x 600 segundos = 21600 segundos, equivalente a 6 horas de registros EEG analizados.

Respecto al tiempo de cómputo total, empleado al utilizar los registros EEG de una hora, se tiene un tiempo total de 17706 segundos (2951 por segmento EEG de 10 minutos *x* 6 secuencias). Utilizando el *Parallel Computing Toolbox*, el tiempo de cómputo disminuye a 6600 segundos (1100 por segmento EEG de 10 minutos *x* 6 secuencias).

Finalmente, el desempeño total del algoritmo utilizando los clips EEG de una hora de duración, tanto interictales como preictales, es del 99.1% de exactitud, 99% de sensibilidad, 99.1% de especificidad y 0.98 de AUR. Comprobando de esta manera, la posibilidad de predecir el inicio de una crisis epiléptica, por medio de la diferenciación entre el estado interictal y el estado preictal, usando registros EEG de una hora de duración.

4.4 Algoritmo empleado en una base de datos alternativa

Inicialmente, se implementó el algoritmo en una base de datos distinta y más extensa. Debido a que dicha base de datos ya no se encuentra disponible públicamente, presentando derechos de autor, entonces los análisis obtenidos en base a ella, no fueron presentados en esta investigación, sin embargo, es conveniente mencionar los resultados más importantes hallados con el uso de esta base de datos, con propósito únicamente comparativo.

Esta base de datos se encontraba en la plataforma *Kaggle-Melbourne University AES/MathWorks/NIH Seizure Prediction* (Kaggle, 2016), y contaba con registros de actividad cerebral de tres pacientes con epilepsia, registrados en forma de EEG intracraneal (iEEG), con 16 electrodos (canales de grabación EEG) por paciente, y una frecuencia de muestreo de 400 Hz. De esta base de datos, se empleó 150 segmentos EEG de 10 minutos preictales y 150 segmentos EEG de 10 minutos interictales, para

cada paciente, teniendo un total de 900 segmentos EEG de 10 minutos analizados (300 segmentos *x* 3 pacientes).

Los parámetros de diseño que mejor desempeño entregaron al algoritmo fueron los siguientes: porcentaje permitido de pérdidas del 1%; ventanas de 1 segundo de duración y 50% de solapamiento; función *wavelet db4* con 5 niveles de descomposición; características lineales (tasa de cruces por cero, máximo, mínimo, media y desviación estándar); validación *holdout* con 75% de entrenamiento y 25% de prueba (debido al excesivo tiempo de cómputo no se utilizó validación cruzada con 10 *folds*); y el clasificador *ensemble-bagged trees.*

Los resultados presentados a continuación, fueron obtenidos analizando los clips EEG de una hora anterior al inicio de la crisis epiléptica, por medio de los segmentos EEG de 10 minutos (utilizando todas las muestras de datos de dichos segmentos), correspondientes a los últimos 10, 20, 30, 40, 50 y 60 minutos anteriores al inicio de la crisis

Una vez realizado el pre-procesamiento se seleccionó un total de 414 segmentos EEG de 10 minutos, con un tiempo total de análisis del algoritmo de 248400 segundos (414 segmentos EEG x 600 segundos), equivalente a 69 horas de registros EEG analizados. Esto se detalla en la tabla 17.

Tabla 17
Número de segmentos EEG analizados en base de datos alternativa

Secuencia	Tiempo antes de crisis (minutos)	#Segmentos EEG	#Interictal	#Preictal
1	60	42	21	21
2	50	72	36	36
3	40	66	33	33
4	30	78	39	39
5	20	84	42	42
6	10	72	36	36
TOTAL	1 hora	414	207	207

En la tabla 18 se muestran los resultados finales de desempeño al implementar el algoritmo propuesto en la base de datos mencionada.

Tabla 18
Resultados para clips EEG de una hora en base de datos alternativa

Secuencia	Tiempo antes de crisis(min.)	Exact	Sens	Espec	AUR	Clasificador
1	60	92.3	89	95	0.98	Bagged Trees
2	50	90.3	88	93	0.96	Bagged Trees
3	40	93	89	97	0.98	Bagged Trees
4	30	93.5	91	96	0.98	Bagged Trees
5	20	89	84	94	0.96	Bagged Trees
6	10	92.2	87	97	0.98	Bagged Trees
MEDIA		91.77	88	95.33	0.973	

Finalmente, el desempeño obtenido por el algoritmo de crisis epilépticas propuesto en esta investigación, con esta nueva base de datos fue de exactitud 91.77%, sensibilidad 88%, especificidad 95.33% y AUR 0.973.

Tomando en cuenta que la predicción de crisis epilépticas a nivel mundial está todavía en etapa de investigación, y en base a los resultados obtenidos empleando las dos bases de datos anteriores, se puede verificar que el desempeño del algoritmo propuesto, en la predicción de crisis epilépticas, por medio de la diferenciación, y por ende, la clasificación del estado interictal y el estado preictal, es bastante aceptable.

CAPÍTULO V

5. CONCLUSIONES

5.1 Conclusiones

Se concluyó que, haciendo uso de parámetros de diseño óptimos y técnicas de procesamiento de señales EEG adecuadas, se puede identificar, y por tanto clasificar los estados interictal y preictal, para llevar a cabo la predicción de una crisis epiléptica, por medio de la detección del estado preictal.

Se comprobó que el análisis lineal, por medio de la transformada *wavelet* discreta y características lineales (tasa de cruces por cero, máximo, mínimo, media y desviación estándar), es el más adecuado para la predicción de crisis epilépticas, con un desempeño obtenido de 99.1% de exactitud, 99% de sensibilidad, 99.1% de especificidad y 0.98 de AUR. Por otro lado, el análisis no-lineal es una buena alternativa a la hora de modelar el comportamiento no-estacionario de las señales EEG.

La transformada *wavelet* discreta (DWT), por medio del análisis multirresolución, es una herramienta realmente útil en el procesamiento de señales no-estacionarias, como es el caso de las señales electroencefalográficas (EEG), gracias a la información precisa de frecuencia a bajas frecuencias y de tiempo en altas frecuencias que esta proporciona. Para procesamiento de señales EEG, el uso de la función *wavelet Daubechies 4* brinda mejor desempeño que la función *wavelet Haar*, ambas empleadas comúnmente en biomedicina.

El clasificador óptimo para predicción de crisis epilépticas, por su rapidez y desempeño, es *Conjunto de clasificadores redundantes* (*ensemble*) con agregación de *bootstrap* de árboles de decisión (*bagged trees*).

La utilización de la información de todos los canales de grabación EEG, de determinado paciente, entrega un mejor desempeño al algoritmo de predicción de crisis epilépticas, que al utilizar la información de un solo canal de grabación EEG.

5.2 Recomendaciones

Investigar más técnicas de pre-procesamiento, procesamiento y clasificación de señales electroencefalográficas (EEG), empleadas en la detección y/o predicción de crisis epilépticas.

Dar prioridad a la sensibilidad del algoritmo predictor, debido a que en un caso hipotético de falla en la detección de registros preictales, el paciente con epilepsia no tomaría las medidas adecuadas para evitar una convulsión inminente, pudiendo resultar en consecuencias graves.

Se recomienda emplear el *Parallel Computing Toolbox*, con un ordenador de características más potentes, tales como: mejor procesador, más memoria RAM, mayor espacio de almacenamiento de disco duro, uso de GPU, etc., para mejorar el tiempo de cómputo empleado por el algoritmo predictor.

Para la optimización del código de implementación del algoritmo predictor en *Matlab*, se recomienda el uso de funciones óptimas, clases, objetos, estructuras y tipos de archivos de escritura rápida, tales como: csv, dat, mat, etc.

5.3 Trabajos Futuros

Posteriormente, se podría probar el desempeño y robustez del algoritmo de predicción de crisis epilépticas, utilizando una base de datos EEG con grabación de tipo superficial, es decir EEG de cuero cabelludo (sEEG), y comparar los resultados, con los obtenidos en registros de EEG intracraneal (iEEG), empleados en esta investigación.

Un análisis mucho más amplio y complejo sería añadir otras señales bioeléctricas al estudio, tales como el Electrocardiograma (ECG), del cual existen indicios que su morfología cambia cuando se presenta una crisis epiléptica.

Se podría implementar el algoritmo predictor en lenguajes de programación más potentes y rápidos, tales como: Python, R, etc., y comparar el tiempo de cómputo y resultados obtenidos.

Una idea realmente fascinante, y que está en investigación actualmente, sería llevar a cabo la implementación del algoritmo en un dispositivo de monitoreo de convulsiones. En un futuro dispositivos como este, podrían ser utilizados para el monitoreo y posterior tratamiento de pacientes con epilepsia, ya sea en ambientes hospitalarios o caseros.

REFERENCIA

Aboy, M., Hornero, R., Abásolo, D., & Álvarez, D. (2006). Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis. *IEEE Transactions on Biomedical Engineering*, *53*(11), 2282-2288.

Adeli, H., Zhou, Z., & Dadmehr, N. (2003). Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of neuroscience methods*, *123*(1), 69-87.

Adewuya, A. O., & Ola, B. A. (2005). Prevalence of and risk factors for anxiety and depressive disorders in Nigerian adolescents with epilepsy. *Epilepsy & behavior*, *6*(3), 342-347.

Alam, S. S., & Bhuiyan, M. I. H. (2013). Detection of seizure and epilepsy using higher order statistics in the EMD domain. *IEEE journal of biomedical and health informatics*, *17*(2), 312-318.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

Annegers, J. F., Grabow, J. D., Groover, R. V., Laws, E. R., Elveback, L. R., & Kurland, L. T. (1980). Seizures after head trauma A population study. *Neurology*, *30*(7), 683-683.

Annett, M. (1985). *Left, right, hand and brain: The right shift theory*. Psychology Press.

Başar, E., Başar-Eroglu, C., Karakaş, S., & Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International journal of psychophysiology*, *39*(2), 241-248.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Boashash, B., Carson, H., & Mesbah, M. (2000). Detection of seizures in newborns using time-frequency analysis of EEG signals. In *Statistical Signal and Array*

Processing, 2000. Proceedings of the Tenth IEEE Workshop on (pp. 564-568). IEEE.

Brown, S., Betts, T., Crawford, P., Hall, B., Shorvon, S., & Wallace, S. (1998). Epilepsy needs revisited: a revised epilepsy needs document for the UK. *Seizure*, *7*(6), 435-446.

Calle, M. L., & Sánchez-Espigares, J. A. (2007). Árboles de clasificación y regresión en la investigación biomédica. *Medicina clínica*, *129*(18), 702-706.

Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología*, *29*(2), 138-141.

Chaovalitwongse, W. A., Pottenger, R. S., Wang, S., Fan, Y. J., & Iasemidis, L. D. (2011). Pattern-and network-based classification techniques for multichannel medical data signals to improve brain diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *41*(5), 977-988.

Chaovalitwongse, W. A., Pottenger, R. S., Wang, S., Fan, Y. J., & Iasemidis, L. D. (2011). Pattern-and network-based classification techniques for multichannel medical data signals to improve brain diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *41*(5), 977-988.

Chen, L. L., Zhang, J., Zou, J. Z., Zhao, C. J., & Wang, G. S. (2014). A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection. *Biomedical Signal Processing and Control*, *10*, 1-10.

Chui, C. K. (1992). An Introduction to Wavelets, Wavelet Analysis and its Application Vol. 1. *1, Academic Press, Boston*.

Crowe, J. A., Gibson, N. M., Woolfson, M. S., & Somekh, M. G. (1992). Wavelet transform as a potential tool for ECG analysis and compression. *Journal of biomedical engineering*, *14*(3), 268-272.

Ditterrich, T. G. (1997). Machine learning research: four current direction. *Artificial Intelligence Magzine*, *4*, 97-136.

Dua, T., De Boer, H. M., Prilipko, L. L., & Saxena, S. (2006). Epilepsy care in the world: results of an ILAE/IBE/WHO global campaign against epilepsy survey. *Epilepsia*, *47*(7), 1225-1231.

Elgohary, S., Eldawlatly, S., & Khalil, M. I. (2016, October). Epileptic seizure prediction using zero-crossings analysis of EEG wavelet detail coefficients. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2016 IEEE Conference on* (pp. 1-6). IEEE.

Epilepsy Australia: Epilepsy Information. (2015). Available: http://www.epilepsyaustralia.net/Epilepsy_Information/Epilepsy_explained/Epilepsy_explained.aspx

Faure, P., & Korn, H. (2001). Is there chaos in the brain? I. Concepts of nonlinear dynamics and methods of investigation. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, *324*(9), 773-793.

French, C. (1996). *Data Processing and Information Technology*. Cengage Learning EMEA.

Fu, Z. Z. H. C. S., & Shi, Z. H. C. (2002). Neural Network Ensemble [J]. *Chinese Journal of Computers*, *1*, 000.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267-285). Springer, Berlin, Heidelberg.

Gadhoumi, K., Lina, J. M., Mormann, F., & Gotman, J. (2016). Seizure prediction for therapeutic devices: A review. *Journal of neuroscience methods*, *260*, 270-282.

Gevins A., Remond A.: Handbook of EEG and clinical neurophysiology. (1987) Elsevier, Amsterdam

Gómez-Luna, E., Silva, D., & Aponte, G. (2013). Selección de una wavelet madre para el análisis frecuencial de señales eléctricas transitorias usando WPD. *Ingeniare. Revista chilena de ingeniería*, *21*(2), 262-270.

Gouyon, F., Pachet, F., & Delerue, O. (2000, December). Classifying percussive sounds: a matter of zero-crossing rate. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*.

Guger, C., Schlogl, A., Neuper, C., Walterspacher, D., Strein, T., & Pfurtscheller, G. (2001). Rapid prototyping of an EEG-based brain-computer interface (BCI). *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *9*(1), 49-58.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, *12*(10), 993-1001.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41). Springer New York.

Hauser, W. A., & Hesdorffer, D. C. (1990). *Epilepsy: frequency, causes and consequences* (Vol. 376). Demos Medical Pub.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Jirsch, J. D., Urrestarazu, E., LeVan, P., Olivier, A., Dubeau, F., & Gotman, J. (2006). High-frequency oscillations during human focal seizures. *Brain*, *129*(6), 1593-1608.

Kaggle: American Epilepsy Society Seizure Prediction Challenge. (2014). Available: https://www.kaggle.com/c/seizure-prediction/data

Kaggle: Melbourne University AES/MathWorks/NIH Seizure Prediction. (2016). Available: https://www.kaggle.com/c/melbourne-university-seizure-prediction

Kantz, H., & Schreiber, T. (2004). *Nonlinear time series analysis* (Vol. 7). Cambridge university press.

Klecka, W. R. (1980). *Discriminant analysis* (Vol. 19). Sage.

Kohavi, R., & Provost, F. (1998). Confusion matrix. *Machine learning*, *30*(2-3), 271-274.

Kossoff, E. H. (2011, September). Intractable childhood epilepsy: choosing between the treatments. In *Seminars in pediatric neurology* (Vol. 18, No. 3, pp. 145-149). WB Saunders.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems* (pp. 231-238).

Larson, R., & Edwards, B. H. (2009). *Calculus*. Cengage Learning.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on information theory*, *22*(1), 75-81.

Liu, B., Yan, L., Li, L., & Wang, W. (2010, June). Comparing study of nonlinear model for epileptic preictal prediction. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on* (pp. 1-4). IEEE.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, *4*(2), R1.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, *11*(7), 674-693.

Medhi, J. (1992). *Statistical methods: an introductory text*. New Age International.

MedicWiz. (2016). *9 Types of EEG tests - Everything about Brainwave Monitoring.* Recuperado el 3 de Abril de 2016, de:

https://www.medicwiz.com/medtech/diagnostics/9-types-of-eeg-tests-everything-about-brainwave-monitoring

Misiti, M., Misiti, Y., Oppenheim, G., & Poggi, J. M. (2000). *Wavelet toolbox for use with MATLAB: User's Guide*. MathWorks.

Montgomery, D. C., & Runger, G. C. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons.

Mora López Tercero, I. (2013). *Detección de crisis epilépticas a partir de señales EEG mediante índices basados en el algoritmo de Lempel-Ziv* (Bachelor's thesis, Universitat Politècnica de Catalunya).

Mormann, F., Andrzejak, R. G., Elger, C. E., & Lehnertz, K. (2007). Seizure prediction: the long and winding road. *Brain*, *130*(2), 314-333.

Nasehi, S., & Pourghassem, H. (2011, December). Automatic prediction of epileptic seizure using kernel fisher discriminant classifiers. In *Intelligent Computation and Bio-Medical Instrumentation (ICBMI), 2011 International Conference on* (pp. 200-203). IEEE.

National Institute of Neurological Disorders, & Stroke (US). (2004). *Seizures and epilepsy: Hope through research* (No. 4). National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Pacia, S. V., & Ebersole, J. S. (1997). Intracranial EEG substrates of scalp ictal patterns from temporal lobe foci. *Epilepsia*, *38*(6), 642-654.

Park, Y., Luo, L., Parhi, K. K., & Netoff, T. (2011). Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. *Epilepsia*, *52*(10), 1761-1770.

Poulsen, J., & French, A. (2008). Discriminant function analysis. *Retrieved from*.

Press, W. H. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Queiroz, R. L. (1992, March). Subband processing of finite length signals without border distortions. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (Vol. 4, pp. 613-616). IEEE.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems* (pp. 532-538). Springer US.

Roessgen, M. A. (1996). *Analysis and modelling of eeg data with application to seizure detection in the newborn* (Doctoral dissertation, Queensland University of Technology).

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

Sanjay Lall (s.f.). *Norm and Vector spaces*. Information Systems Laboratory, Stanford University.

Shai, S. S., & Shai, B. D. (2014). Understanding machine learning: from theory to algorithms.

Shorvon, S. D. (2011). The etiologic classification of epilepsy. *Epilepsia*, *52*(6), 1052-1057.

Tilbury, D., & Messner, W. (1998). Control Tutorials for Matlab: Software Instruction Over the World-Wide Web. *NASA*, (19980237506).

Viglione, S. S., & Walsh, G. O. (1975). Proceedings: Epileptic seizure prediction. *Electroencephalography and clinical neurophysiology*, *39*(4), 435.

Winterhalder, M., Maiwald, T., Voss, H. U., Aschenbrenner-Scheibe, R., Timmer, J., & Schulze-Bonhage, A. (2003). The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods. *Epilepsy & Behavior*, *4*(3), 318-325.

Wright, J. J., Kydd, R. R., & Sergejew, A. A. (1990). Autoregression models of EEG. *Biological cybernetics*, *62*(3), 201-210.

Wu, J. D., Hsu, C. C., & Wu, G. Z. (2009). Fault gear identification and classification using discrete wavelet transform and adaptive neuro-fuzzy inference. *Expert Systems with Applications*, *36*(3), 6244-6255.

Xu, B., & Song, A. (2008). Pattern recognition of motor imagery EEG using wavelet transform. *Journal of Biomedical Science and Engineering*, *1*(01), 64.

Yong, L., & Shengxun, Z. (1996, January). Apply wavelet transform to analyse EEG signal. In *Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE* (Vol. 3, pp. 1007-1008). IEEE.

Zeljković, V., Druzgalski, C., Yequn, L., Jian, C., Xin, Y., Bojic, M., & Mayorga, P. (2013, April). Automated pre-ictal phase detection algorithm from eeg signals. In *Health Care Exchanges (PAHCE), 2013 Pan American* (pp. 1-5). IEEE.

Zhang, X. S., Roy, R. J., & Jensen, E. W. (2001). EEG complexity as a measure of depth of anesthesia for patients. *IEEE Transactions on Biomedical Engineering*, *48*(12), 1424-1433.

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, *19*.