Selection of Relevant Features to Support Automatic Detection of Epileptiform Events

MARITZA FERNANDA MERA GAONA

Appendix
PhD Thesis in Telematics Engineering

Supervisors:
PhD. Diego Mauricio López
PhD. Rubiel Vargas Cañas
PhD. Maria Eugenia Miño

Universidad del Cauca
Faculty of Electronics and Telecommunications Engineering
Department of Telematics
eHealth
Popayán, June 2021

MARITZA FERNANDA MERA GAONA

Selection of Relevant Features to Support Automatic Detection of Epileptiform Events

Appendix

Dissertation submitted to the Faculty of Electronics and Telecommunications Engineering of the Universidad del Cauca, Colombia
for granting the academic degree of

Doctora en:
Ingeniería Telemática

Supervisors
PhD. Diego Mauricio López
PhD. Rubiel Vargas Cañas
PhD. Maria Eugenia Miño

Popayán
2021

APPENDIX A

PROPOSAL: Selection of Relevant Features to Support Automatic Detection of Epileptiform Events

Selección de Características Relevantes en Señales EEG para la Detección Automática de Eventos Epileptiformes

Anteproyecto de Tesis de Doctorado

Maritza Fernanda Mera Gaona

Director: PhD. Diego Mauricio López
Co-Director: PhD Rubiel Vargas Cañas

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Doctorado en Ingeniería Telemática
Línea: e-Salud
Popayán, Junio de 2018

TABLA DE CONTENIDO

1. PLANTEAMIENTO DEL PROBLEMA

1.1. Contexto General

La tasa de incidencia de enfermedades neurológicas a nivel mundial reporta que en la actualidad existen cerca de 450 millones de personas que padecen algún tipo de trastorno mental [1]. Una de las dificultades para tratar a los pacientes que presentan este tipo de trastornos, es la falta de personal especializado y los altos costos que implica el diagnóstico [2]. En países de bajos o medianos ingresos como Colombia, la atención brindada a los pacientes que padecen este tipo de trastornos no es de buena calidad, debido a la falta de recursos que invierte el gobierno nacional en atención a problemas de Salud Mental.

Adicionalmente, los especialistas encargados de diagnosticar y llevar los tratamientos de estas enfermedades por lo general se encuentran en centros médicos especializados, los cuales son de difícil acceso para las poblaciones que se encuentran ubicadas sobre zonas rurales. Esta situación es uno de las razones por las que se dificulta la detección temprana de trastornos neurológicos en países de bajos y medios ingresos. Según las estadísticas, en Colombia el número total de neurólogos en el 2011 era de 231, lo que representaba una distribución de un neurólogo por cada 199.327 habitantes [3]. Esta cifra es demasiado baja si la comparamos con la disponibilidad de neurólogos en países desarrollados como Estados Unidos y España, en donde la relación es de un neurólogo por cada 22.880 habitantes y un especialista en neurología por cada 23.500 habitantes, respectivamente [4]. Por otro lado, en ciudades pequeñas y medianas el problema se incrementa debido a que la cantidad de especialistas y centros clínicos disminuye, por ejemplo, en Popayán se cuenta con solo cuatro centros de especialistas en neurología para atender una población de 267.976 habitantes [5]. Teniendo en cuenta que, en el resto del departamento del Cauca, no se cuenta con centros de estas características, los 4 centros de la ciudad de Popayán deben prestar sus servicios a toda la población del departamento del Cauca, los cuales son cerca de 1.051.007 habitantes [5]. Por lo anterior, y considerando la importancia de diagnosticar a los pacientes en una edad temprana para obtener mejores resultados en los tratamientos aplicados, este proyecto se enfocará en soportar el diagnóstico de eventos epileptiformes en EEGs de pacientes menores de edad.

Por otro lado, la poca disponibilidad de profesionales especializados en neurología genera largos tiempos para obtener un resultado de la lectura de Electroencefalogramas (EEG), además de los altos costos en la prestación de estos servicios. Esto debido a que la inspección visual de una señal EEG puede ser demasiado compleja, consumir bastante tiempo por la duración y tener diferentes interpretaciones de los neurólogos que la analizan. Por lo anterior, herramientas que detecten automáticamente episodios epilépticos en una señal EEG ayudarían a disminuir el tiempo empleado para la inspección visual de la señal EEG, especialmente cuando se analicen EEGs de larga duración (24, 48 o 72 horas). La posibilidad de reducir el tiempo de inspección de un especialista podría representar la oportunidad de atender a más pacientes, algo muy útil en países como Colombia debido a la baja disponibilidad de especialistas.

Durante los últimos años la investigación sobre el desarrollo de sistemas de captura y análisis de señales biomédicas se ha incrementado con el fin de encontrar nuevos mecanismos de diagnóstico clínico sobre determinadas patologías. A través del procesamiento de señales EEG, se puede monitorear la actividad neuronal del cerebro y extraer información que describe información útil para la detección de patologías neurológicas. Sin embargo, el reto de la caracterización de señales EEG, se asocia a la dificultad de extraer la información relevante capaz de describir la presencia o ausencia de una enfermedad. Esta dificultad se debe a la complejidad de conocer a priori cual es la información contenida en la señal que se debe considerar como relevante para la detección de los patrones que caracterizan cada patología [6].

Durante el procesamiento y análisis de señales EEG, se extrae de cada señal un conjunto de características, representado en un vector, que describe la mayor cantidad de información posible con el fin de obtener una representación completa de la señal. Sin embargo, cuando se realiza el proceso de extracción de características de una señal multicanal como es el caso de un EEG, se deben emplear una gran cantidad de descriptores que generan cientos de datos por cada canal, de los cuales algunos pueden contener ruido o información redundante que al descartarlos no eliminarían información relevante de la señal [7]. Teniendo en cuenta lo anterior, la selección de características se hace necesaria en la implementación de un sistema de este tipo, con el fin de identificar las características que contienen información útil para la detección de patologías al analizar la señal y reducir la complejidad y el costo computacional generado por el cálculo de características redundantes, ruido y manejo de vectores con tamaño n muy grande [8].

Durante los últimos años, el diagnóstico de enfermedades como la Epilepsia a través de análisis digital de señales EEG ha sido una de las principales áreas de investigación en neurociencias. A través de diferentes mecanismos de extracción de características se ha logrado obtener información que posteriormente permite clasificar una señal como normal o anormal [6]. Así mismo, se han realizado otros estudios basados en el análisis de señales EEG para analizar la actividad cerebral durante estado de sueño, análisis de señales sobre recién nacidos, entre otros [7] [8], con el fin de soportar procesos de diagnóstico clínico. Algunos mecanismos de clasificación de señales han empleado redes neuronales, análisis de los datos estadísticos extraídos de las señales, reglas basadas en conocimiento del dominio y mecanismos de clustering para clasificar una nueva señal encontrando la clase del grupo de señales más cercano [9] [10]. Sin embargo, aunque en la literatura se reportan algunos trabajos que han implementado métodos de selección de características para identificar aquellas con mayor poder de diferenciación en un proceso de clasificación o detección de señales epilépticas, la gran mayoría de los trabajos revisados en el estado del arte, se han enfocado en la identificación de patrones específicos definiendo a prori un conjunto de características sin considerar la relevancia real de cada una de ellas [9].

En este sentido, en la literatura se encuentran numerosos mecanismos de caracterización de señales EEG para la detección y/o clasificación de eventos asociados a la Epilepsia, sin embargo, el gran reto de investigación en esta área se ha enfocado en mejorar el desempeño de la clasificación en términos de precisión, exactitud y recall, con el objetivo de brindar herramientas confiables que soporten y/o ayuden a los especialistas durante el proceso de diagnóstico de Epilepsia. Considerando lo anterior, una de las principales estrategias para mejorar los modelos de clasificación en Aprendizaje de Máquina o Minería de Datos es entrenar los modelos encargados de determinar la clase de una instancia con características relevantes, es decir, aquellas características que no representen ruido para el aprendizaje y por el contrario tengan un alto poder de diferenciación entre las clases. Generalmente, la identificación de características relevantes se convierte en un proceso bastante útil en escenarios en los cuales se tiene un dataset con alta dimensionalidad, es decir, un número de características o instancias considerablemente alto [10]. Así mismo, escenarios en los que los datasets de entrenamiento tienen más características que instancias, como la clasificación de microarreglos de datos, la selección de características se hace una tarea obligatoria para reducir la dimensionalidad de los datasets y soportar una clasificación robusta [11]. Este último escenario coincide con la clasificación de anormalidades en EEG si se consideran la gran cantidad de descriptores que han sido reportados en la literatura y la baja disponibilidad de datasets con un número elevado de instancias que describan anormalidades o eventos epileptiformes.

Por otro lado, existen diferentes propuestas en la literatura para realizar el proceso de selección de características en conjuntos de datos diversos y de gran tamaño, sin embargo, la implementación de algunos de estos mecanismos, como los envolventes y embebidos, es compleja debido a que para su funcionamiento se requiere el desarrollo de algoritmos de clasificación que calculen el índice de relevancia de las características, que a su vez pueden afectar la consistencia de los conjuntos de características seleccionados cuando se agregan o eliminan instancias del dataset [12]. Así mismo, en la revisión de la literatura de esta propuesta se pudo identificar algoritmos de selección de características basados en filtros, los cuales pueden ser computacionalmente más sencillos que los otros enfoques (envolventes y embebidos). Sin embargo, este tipo de algoritmos generalmente analizan la relevancia de las características de forma individual, de tal manera que dos características que tengan correlación entre sí podrían ayudar a clasificar mejor, aunque estas mismas características por separado podrían haber sido consideradas como irrelevantes.

Con el fin de mejorar el funcionamiento de los algoritmos de selección de características se han planteado soluciones mediante la identificación de correlaciones entre características y clases, de tal manera que se pueda mejorar su efectividad y mantener un bajo costo computacional en la selección de características [12]. Así como también, la incorporación de técnicas como Bootstrap para seleccionar características empleando muestras del dataset original e integrando los diferentes subconjuntos de características generados [13] [14]. Aunque algunos trabajos presentan resultados satisfactorios, estos métodos pueden ser sensibles al balanceo de los datasets y el tratamiento de datos continuos. Por lo tanto, al aplicar este tipo de algoritmos no siempre se lograría seleccionar todas las características relevantes. Considerando lo anterior, algunos autores han propuesto el ensamble de algoritmos de selección de características para mejorar la identificación de características relevantes a través del consenso de algoritmos con diferentes enfoques [15].

Teniendo en cuenta lo anterior, la presente propuesta de doctorado plantea la siguiente pregunta de investigación **¿Cómo mejorar la efectividad en la selección de características relevantes en procesos de clasificación de Eventos Epileptiformes en señales EEG?**

Para responder a esta pregunta de investigación se propone la siguiente hipótesis: **el ensamble de métodos de selección de características puede mejorar la efectividad en la selección de caracteristicas relevantes en procesos de clasificación de Eventos Epileptiformes en señales EEG.**

Esta aproximación se basa en la premisa de los multiclasificadores: “varios clasificadores clasifican mejor que uno”, la cual sería aplicada a la selección de características, donde se pretende demostrar que “varios selectores de características seleccionan mejor que uno”. Se pretende emplear los resultados de múltiples algoritmos de selección para proponer un conjunto de características relevantes definido a partir de un mecanismo (s) de consenso de todos los conjuntos de características generados.

2. ESTADO DEL ARTE

2.1. Marco Teórico

En esta sección se describen conceptos teóricos relacionados tanto con el contexto clínico de la Epilepsia, así como con el proceso de Selección de Características.

2.1.1. Contexto Clínico

A continuación, se explican algunos conceptos relacionados a la Epilepsia y la toma de electroencefalogramas.

2.1.1.1. Electroencefalograma

Un Electroencefalograma (EEG) es une herramienta empleada por neurólogos para medir la actividad eléctrica del cerebro humano generada por las corrientes que fluyen durante las excitaciones

sinápticas. Cuando las neuronas son activadas, las corrientes sinápticas se activan en las dendritas y generan un campo eléctrico sobre el cuero cabelludo [16].
A través de la inspección visual de la actividad eléctrica del cerebro los especialistas pueden analizar las funciones cerebrales a través del tiempo y detectar desórdenes neurológicos. Los EEGs están representados en grabaciones de señales multicanal tomadas a través de la digitalización de los datos capturados desde electrodos ubicados en el cuero cabello o intracranealmente [17].
Los electroencefalogramas son empleados principalmente para soportar el diagnóstico de enfermedades neurológicas que afectan la actividad cerebral como, por ejemplo:

- Epilepsia,
- Tumores cerebrales,
- Lesiones de la cabeza,
- Desordenes de sueño,
- Demencia,
- Monitoreo de anestesia durante cirugías.

Durante la toma de un EEG, los electrodos capturan pequeñas descargas eléctricas que se generan a partir de la actividad de las células en el cerebro. Estas descargas son amplificadas y digitalizadas con el fin de obtener una representación gráfica en forma de ondas. Para la toma de un EEG se pueden emplear diferente número de electrodos, generalmente, cada par de electrodos genera la información de un canal (Ver Figura 2). Dependiendo de la cantidad de electrodos empleados se obtendrán más canales en el EEG y en consecuencia, más información.

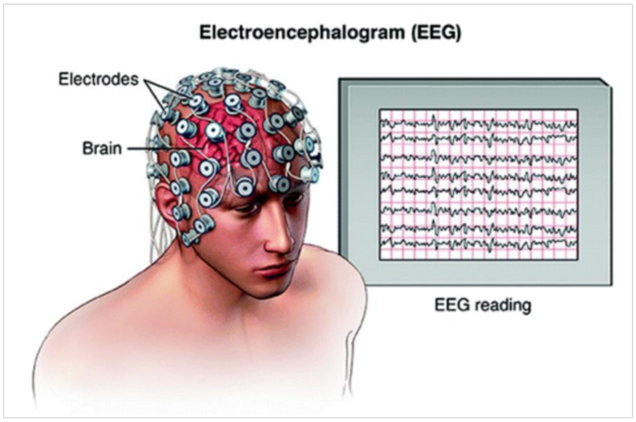


Figura 1.Esquema de Adquisición de EEG, tomado de [18]

2.1.1.2. Sistema 10-20

La ubicación de los electrodos durante la toma de un EEG indica la información que se va a capturar desde los diferentes lóbulos del cerebro, cada lóbulo está asociado a algunas actividades que desarrolla el ser humano. Generalmente, el método que se emplea para ubicar los electrodos en el cuero cabelludo es el sistema de electrodos internacional 10-20 [19].

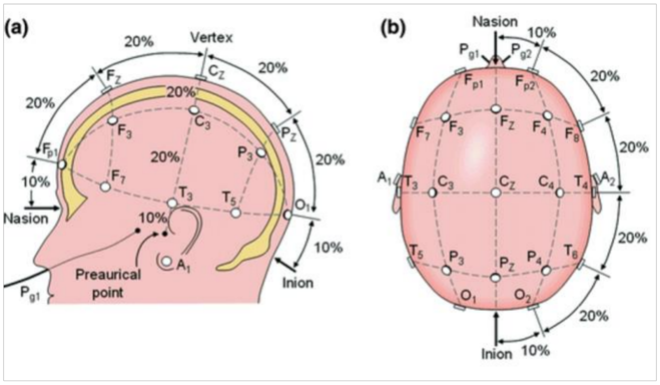


Figura 2.Ubicación de electrodos según esquema 10-20. Tomado de [18]

En la figura 3, se puede observar el montaje estándar que se emplea para posicionar cada electrodo. La ubicación de cada electrodo está determinada por las distancias entre electrodos vecinos, la cual puede ser entre el 10% y 20% de la distancia total medida desde la parte frontal hasta la parte trasera del cráneo o la distancia total medida desde la parte derecha hasta la parte izquierda del cerebro. Los puntos de referencia empleados para medir las distancias son: *Nasion,* el punto de intersección entre la nariz y la frente, e *Inion*, el punto definido por el hueso de la base del cráneo.

El montaje estándar también emplea letras para identificar la ubicación del hemisferio en el que se encuentra el electrodo [20]. Las letras empeladas son:

- F: Frontal
- T: Temporal
- C: Central
- P: Parietal
- O: Occipital

Adicionalmente, cuando se emplea la letra “z”, esta indica que el electrodo está ubicado en la línea media. Para diferenciar los electrodos ubicados en el hemisferio derecho del hemisferio izquierdo, se emplean números pares o números impares respectivamente. Una vez ubicados los electrodos en el cuero cabelludo del paciente, se miden las diferencias de voltaje entre un par de electrodos y mostradas al especialista como la información un canal del EEG.

Teniendo en cuenta lo anterior, el estudio de EEGs se convierte en una herramienta de gran utilidad para el diagnóstico de diferentes desórdenes neurológicos y otras anormalidades que alteran la actividad eléctrica.

2.1.1.3. Epilepsia

Es un desorden neurológico generado por descargas eléctricas en una o varias zonas del cerebro. El diagnóstico es dado a través de la inspección visual de EEGs practicados a los pacientes junto con datos clínicos que describan antecedentes asociados a posibles lesiones cerebrales.

Existen diferentes síntomas que presentan los pacientes que padecen esta enfermedad, el más común, las convulsiones epilépticas. Estas generan en el EEG cambios abruptos de la frecuencia y la amplitud. La desincronización de la actividad eléctrica se ve reflejada en la aparición de ondas desordenadas que después de un tiempo, vuelven a ordenarse. Algunos de los cambios que se pueden observar en las ondas del EEG durante una convulsión epiléptica son picos afilados o aparición de ondas lentas [21].

En algunos casos, se observa picos en la señal EEG debido a artefactos producidos por el movimiento muscular, apertura y cierre de los ojos, movimientos del cuello, entre otros, estos artefactos pueden ser eliminados fácilmente gracias a que los EEGs generalmente incluyen información en canales adicionales, usando electroencefalografía o electromiografía, que permiten conocer si el paciente se movió o si cerró los ojos.

2.1.2. Selección de Características

El proceso de selección de características está relacionado al análisis de relevancia y reducción de dimensionalidad. En múltiples trabajos los tres términos son empleados para describir el proceso reducir el conjunto original de datos generado por el proceso de extracción características en un conjunto mínimo que represente la mayor cantidad de información útil en el proceso de clasificación [22]. Comúnmente, la selección de características puede realizarse haciendo análisis de relevancia para medir la importancia de las características en la clasificación. Por otro lado, la selección de características incluye un proceso de reducción de dimensionalidad, pero únicamente a nivel de columnas. Por lo anterior, en el contexto de esta tesis doctoral, los términos de selección de características, análisis de relevancia y reducción de dimensionalidad fueron considerados como sinónimos debido a que a través del análisis de relevancia de características y la reducción de la dimensionalidad se puede soportar el proceso de selección de características relevantes en un dataset.

Durante el proceso de análisis de relevancia se analizan tres tipos datos: relevantes, no relevantes (ruido) y redundantes. A través de métodos de eliminación y transformación de datos, el conjunto original de características es transformado en un conjunto de menor tamaño, el cual solo contiene las características identificadas como relevantes. A continuación, se describe las categorías en las que típicamente se dividen los métodos de selección de características.

2.1.2.1. Filtros

Los métodos de selección de características basados en filtros evalúan una métrica para evaluar la relevancia de las características [12]. En este sentido, la función del filtro no se retroalimenta de un predictor o clasificador para tomar una decisión en cuanto a la relevancia de las características, sino por un índice de relevancia que determina el poder de diferenciación que tiene la característica. Dicha relevancia puede ser calculada a través de una función sencilla como la correlación o incluso se puede usar otro algoritmo de selección de características para estimar un índice de relevancia.

Los índices de relevancia pueden ser calculados por cada característica para posteriormente organizar todas las características en un ranking calculado a partir del valor del índice individual calculado. Una vez se obtiene el ranking se establece un criterio o un umbral para eliminar las características que están ubicadas en la parte inferior del ranking de relevancia. El establecer el criterio que permite descartar las características peor ubicadas en el ranking se puede realizar a partir del uso de algoritmos de clasificación que permita evaluar la exactitud en función de las características mejor ubicadas que fueron empleadas para lograrla.

Aunque los algoritmos de selección de características basados en filtros son generalmente más fáciles de implementar, enfrentan las dificultares asociadas a la identificación de características redundantes, la correlación entre características y la correlación entre clases.

2.1.2.2. Embebidos

Los métodos embebidos involucran en el proceso de selección de características el mecanismo de clasificación, de tal manera que la estructura bajo la cual se clasifica está ligada al proceso de análisis de relevancia de las características [23]. En este sentido, el algoritmo de clasificación siempre va a depender del proceso de selección de características.

2.1.2.3. Envolventes

Este tipo de algoritmos de selección de características incorporan un algoritmo de aprendizaje para evaluar la relevancia de los subconjuntos de características seleccionadas sin tener relación con el proceso de clasificación. Los métodos envolventes generalmente requieren mayor capacidad de recursos computacionales, y aunque existen estrategias para mejorar la eficiencia de la búsqueda de las características relevantes, estas suelen ser más complejas de implementar.

Considerando los diferentes enfoques de los algoritmos de selección de características a continuación se presenta una tabla con sus principales características:

Tabla 1. Caracterización de algoritmos.

	Filtros	Embebidos	Envolventes
Alta Velocidad	x		
Alto costo computacional		x	x
Complejos de implementar		x	x
Análisis individual de características	x		
Redundancia de características	x		
Alta Velocidad	x		
Análisis no individual de características		x	x
Mayor efectividad		x	x
Incluye algoritmo de aprendizaje/clasificación		x	x
Aprendizaje y selección son una sola tarea		x	
Adición y/o eliminación de características de forma iterativa	x	x	x
Incorporan la estructura de clasificación en la selección		x	
Incluyen Validación Cruzada			x

Mejor rendimiento con pocos datos	x		
Mejor rendimiento muchos datos		x	x

2.1.3. Métodos de consenso

En el desarrollo de multiclasificadores se han implementado diferentes mecanismos para soportar el consenso de los diferentes ‘expertos’ (clasificadores) empleados durante la clasificación. Debido a que el papel de cada experto es representado por un clasificador que puede funcionar de forma independiente a los demás, y en consecuencia se obtienen múltiples respuestas, las cuales deben ser fusionadas para obtener una única clase como respuesta que represente a decisión final de clasificación. Teniendo en cuenta lo anterior, a continuación, se presentan algunos de los principales mecanismos de consenso reportados en la literatura.

2.1.3.1. Voto mayoritario simple

En esta técnica cada experto tiene un voto con la misma importancia que los demás expertos. Bajo este esquema, la clasificación final depende de la respuesta más popular en el conjunto de respuestas, es decir, la respuesta con más votos es la respuesta final [24].

2.1.3.2. Voto mayoritario por peso

En este método se simula la manera en que se realizan las votaciones en las empresas por parte de los accionistas, de tal manera que cada votante puede tener mayor o menor influencia en la decisión final. En este sentido, se busca que cada experto de un voto y mediante un esquema basado en pesos se le brinda una importancia a cada voto [25]. Teniendo en cuenta que algunos expertos pueden tener mejor criterio que otros, es importante darles a éstos mayor grado de relevancia en el resultado final, de tal manera, que los cade con menor criterio no tengan la misma importancia que los mejores expertos, el cual es el principal problema que el voto mayoritario simple.

2.1.3.3. Reglas basadas en el enfoque de Bayes

Son un conjunto de reglas que definen la fusión de un conjunto de respuestas, sin embargo, es uno de los mecanismos menos popular para la combinación de decisiones de conjuntos de clasificadores, debido a que la estadística bayesiana es empleada para determinar clasificaciones teóricamente óptimas.

2.1.3.4. Regla del promedio simple

Es una de las reglas más básicas para promediar las repuestas de un conjunto de expertos. En un multiclasificador, se define un problema con K clases y un vector de probabilidades v(k) en el que la suma de los componentes debe ser igual a 1. Para cada clasificador se unifica en vector V las probabilidades de todos los clasificadores. El vector V es empleado para determinar cuál es clase con mayor probabilidad.

2.1.3.5. Regla del producto

Este mecanismo multiplica las probabilidades de una misma clase en todos los clasificadores para determinar la decisión final.

2.2. Trabajos Relacionados

En esta sección, se presenta un resumen de los principales trabajos identificados en la revisión de la literatura realizada. Los trabajos fueron seleccionados a partir de dos mapeos sistemáticos realizados de manera independiente sobre diferentes bases de datos bibliográficas (PubMed, IEEE, Science Direct y SCOPUS) respecto a la detección automática de Epilepsia y el proceso de selección de características, respectivamente. Para realizar cada uno de los mapeos sistemáticos se emplearon las siguientes cadenas de búsqueda:

- (Automated EEG analysis AND epileptic) AND (classification OR seizures OR EEG records OR detection) AND (feature selection OR feature relevance OR normal OR dimensionality)
- (Feature selection OR feature relevance OR dimensionality reduction)
- (Feature selection OR feature relevance OR dimensionality reduction) AND (voting or consensus OR combination rule OR Weighted Majority OR Belief Functions OR multiple combination OR voting by majority)

Teniendo en cuenta que en la literatura se pueden encontrar diferentes conceptos que en la práctica están relacionados al diagnóstico automático de Epilepsia a través de EEGs y la selección de características, a continuación, se describe en la tabla 1 los conceptos que fueron considerados para este proyecto como equivalentes.

Tabla 2. Sinónimos de palabras clave.

Concepto	Sinónimo
Automatic analysis	Automatic detection Automatic classification
Feature selection	Feature relevance Dimensionality reduction
Seizure	Abnormality

Considerando que las cadenas de búsqueda definidas retornaron gran cantidad de trabajos en cada una de las bases de datos bibliográficas exploradas, a continuación, se describen los criterios de inclusión y exclusión que fueron considerados para la revisión de los trabajos encontrados de acuerdo a cada una de las cadenas de búsqueda.

Criterios de Inclusión para la cadena 1:

- Estudios que describen el diagnostico automático de alguna enfermedad a través del análisis EEG.
- Estudios que describen la construcción de sistemas de diagnóstico automático de señales EEG.
- Estudios que describan como clasificar señales EEG de forma automática.
- Estudios que describan como detectar anormalidades en señales EEG.
- Estudios que describan el proceso de selección de características

Criterios de Inclusión para la cadena 2:

- Estudios que describan técnicas de reducción de dimensionalidad, análisis de relevancia o selección de características.
- Estudios que describen los resultados obtenidos con la solución propuesta.

Criterios de Inclusión para la cadena 3:

- Estudios que describan técnicas de consenso de diferentes soluciones en trabajos que hayan incluido selección de características relevantes.

Criterios de Exclusión para las cadenas 1, 2 y 3:

- Estudios que no describan qué técnicas y algoritmos que fueron empleados.

Los resultados obtenidos empleando cada cadena de búsqueda fueron analizados de acuerdo a los criterios de inclusión y exclusión definidos. A continuación, en la tabla 2 se puede observar un breve resumen de los trabajos recuperados por cada base de datos para la cadena de búsqueda 1, así como los artículos seleccionados después de aplicar los criterios de inclusión y exclusión.

Tabla 3. Resultados Cadena de Búsqueda 1.

Fuente	Artículos recuperados	Artículos seleccionados
Science Direct	1150	20
IEEE	30	11
PubMed	930	26

A partir de los trabajos seleccionados en el primer mapeo sistemático, se identificaron las principales técnicas empleadas para el análisis automático de señales EEG en las etapas de extracción y clasificación, los resultados se pueden observar en las figuras 1 y 2 respectivamente.

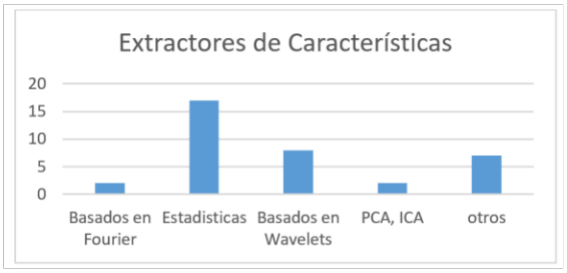


Figura 3.Extractores de Características.

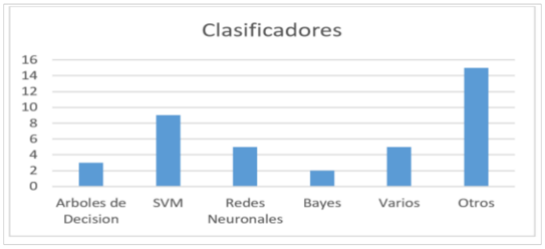


Figura 4.Clasificadores.

En la tabla 3 se pueden observar la lista de las 39 técnicas que fueron empleadas en los trabajos recuperados para el proceso de selección de características en el segundo mapeo sistemático. Esta búsqueda se realizó independiente del dominio de aplicación. En algunos trabajos se probó más de una técnica, sin embargo, las técnicas que más fueron usadas en diferentes trabajos son: Minimum redundancy-maximal relevance en 11 trabajos, PCA en 9, algoritmos genéticos en 5, random forest en 5 y sequential Floating Search en 5.

Tabla 3. Distribución de técnicas empleadas para soportar la selección de características

Técnica	Cantidad
minimum redundancy-maximal relevance	11
PCA	9
genetic algorithm	5
random forest	5
Sequential Floating Search	5
Correlation-based Feature Selection	4
incremental feature selection	4
RELIEFF	4
Linear discriminant analysis	3

Mutual Information	3
SVM-RFE	3
ANOVA	2
Chi-square	2
Information Gain	2
Information theory	2
Monte-Carlo Feature Selection	2
weight	2
rank of features	2
Cohort Intelligence	1
Consistency-Based Filter	1
eye-inspection	1
factor analysis	1
Fast Correlation-Based Feature	1
Feature extraction via partial least squares	1
Feature selection via concave minimization	1
feature/variable selection algorithm	1
filtering	1
Fisher linear discriminant	1
Gram-schmidt Orthogonalization	1
graph-based features	1
harmony search (HS) algorithm	1
ICA	1
KSVM	1
maximal information coefficient	1
neighborhood rough set	1
particle swarm optimization	1
relaxed linear separability	1
Tf-idf (del inglés Term frequency – Inverse document frequency)	1
wrapper for feature-subset selection	1

En la tabla 4 se puede observar el resumen de las técnicas de consenso empleadas en los trabajos recuperados de la base de datos de SCOPUS con la tercera cadena de búsqueda. En los resultados encontrados se identificaron trabajos en los que se realiza consenso para el ensamble de métodos de selección de características o clasificación:

Tabla 4. Distribución de técnicas de consenso empleadas

Técnica	Cantidad
Voting majoritary	15
Weighted Majority Voting	2
Consentimiento común	1
Layered Majority Voting	1
Layered Weighted Voting	1

Los principales trabajos identificados en los tres mapeos sistemáticos se describen a continuación:

2.2.1. Trabajos sobre diagnóstico automático de epilepsia y selección de carácterísticas

Automated EEG analysis of epilepsy: A review [26]

La investigación aporta un análisis de los métodos heurísticos y descriptivos como el dominio del tiempo, frecuencia y tiempo-frecuencia para la extracción de características de una señal EEG. Además, provee una revisión de técnicas de clasificación para la detección de epilepsia automatizada mediante el desarrollo de una herramienta informática (Computer Aided Diagnostic tool - CAD). En los resultados, lo más destacado es el valor de 99% obtenido en la detección de desórdenes epilépticos utilizando las técnicas de análisis no lineal. Los resultados obtenidos en la evaluación son muy buenos, sin embargo, no se evidencia una revisión de la relevancia de las características extraídas.

Epileptic seizure detection by analyzing EEG signals using different transformation techniques [27]

En este trabajo se implementa un método para detectar en señales EEG periodos durante la convulsión y periodos entre convulsiones. Esto sería de gran ayuda para el tratamiento e identificación de convulsiones epilépticas. En el estudio se proponen nuevas formas para la extracción de características estadísticas de las señales empleando coeficientes de alta frecuencia de las señales transformadas. Para el proceso de clasificación se emplearon máquinas de vectores de soporte y la evaluación del método arrojó un 91.36% de sensibilidad. La evaluación arroja resultados positivos en cuanto a la clasificación, sin embargo no se evidencia un mecanismo o evaluaciòn de la relevancia de las características extraídas.

EEG analysis of seizure patterns using visibility graphs for detection of generalized seizures

En esta investigación se recolectaron registros EEG de 29 pacientes epilépticos empleando 24 electrodos y el sistema de posicionamiento 10-20 con una tasa de muestreo de 100 Hz. Los datos fueron empleados para caracterizar patrones epilépticos sobre EEGs con convulsiones generalizadas. Cada patrón fue caracterizado y posteriormente a través de una máquina de soporte vectorial (SVM) se hizo el proceso de clasificación. En este trabajo los investigadores logran caracterizar los EEGs tomando como base el comportamiento caótico de estos y proponen un conjunto de características caóticas para la detección de patrones epilépticos en pacientes adultos. Sin embargo, no se evidencia una revisión de la relevancia de las características extraídas.

Epileptic seizure detection in EEGs signals based on the weighted visibility graph entropy [28]

En este estudio se proponen un conjunto de nuevas características basadas en la representación visual de la entropía para la detección automática de convulsiones epilépticas. El análisis de la entropía es analizado debido a que las señales producidas en los EEGs son caóticas y no estacionarias. El proceso de clasificación se hace a través de características extraídas de WVGEs mapeados de cada canal (weighted visibility graph entropy) y los algoritmos de clasificación SVM, KNN, Decision Tree y

Naive Bayes. Las señales son clasificadas en tres clases: normal, libre de convulsión(interictal) y convulsión (ictal). Sin embargo, no se evidencia una revisión de la relevancia de las características extraídas.

EEG signal classification for epilepsy diagnosis via optimum path forest – A systematic assessment [29]

La investigación describe la construcción de un clasificador de señales EEG para el diagnóstico de epilepsia. A través de 4 tipos de funciones de Wavelets se realizó el proceso de extracción de características y empleando tres métodos de filtrado bien conocidos se realizó el análisis de relevancia de características que permitió reducir el conjunto de características inicial. Por otro lado, empleando Maquinas de Vectores de Soporte – SVM (Support Vector Machine), redes neuronales y clasificadores bayesianos se probó la clasificación con el objetivo de encontrar el mecanismo de mejor comportamiento en términos de eficiencia y eficacia. La evaluación arrojó buenos resultados en términos de tiempo cuando la clasificación se realizó con máquinas de soporte de vectores y redes neuronales. Sin embargo, la solución propuesta está diseñada para clasificar segmentos de señal y no para detectar eventos epileptiformes en una señal de larga duración. Tampoco se propone una comparación de la efectividad de los mètodos de selecciòn de carácterísticas.

2.2.2.Trabajos Relacionados con el proceso de Selección de Características

En esta sección se resumen los trabajos sobre selección de características que describen técnicas de consenso.

Application of non-linear and wavelet based features for the automated identification of epileptic EEG signals [30]**.**

En este trabajo se desarrolla una técnica de diagnóstico asistido por computador que permite clasificar una señal como normal, ictal e interictal a partir de un número reducido de número de características relevantes (no-lineales) extraídas de cada señal. Las características extraídas están basados en los espectros de orden mayor (Higher Order Spectra), AmpEn, SampEn, dimensión fractal y exponente de Hurst. Se realizó un proceso de reducción de dimensionalidad empleando el test ANOVA. Las características relevantes fueron empleadas con diferentes clasificadores para seleccionar el de mejor desempeño. Para la evaluación de la técnica se emplearon 100 segmentos de señal de cada clase. Empleando diferentes combinaciones de características y clasificador se evidenció que el clasificador difuso obtuvo una exactitud promedio del 99.7%. Está técnica desarrollada puede ser empleada para el desarrollo de nuevos sistemas que soporten el monitoreo automático de pacientes. La evaluación realizada evidencia buenos resultados, en este sentido, las características empleadas representan un punto de partida para la selección de un conjunto mínimo de características que permitan identificar descargas epilépticas en un EEG.

Statistical features based epileptic seizure EEG detection - an efficacy evaluation [31]

La investigación describe la construcción de un clasificador de señales EEG en ictal, inter-ictal y normal. Las características son extraídas de 500 señales de las tres clases (100 ictal, 200 inter-ictal, 200 normal) para ser procesadas por un conjunto de clasificadores. Así mismo, se empleó dos algoritmos de análisis de relevancia de características: relief y rank the features. Los dos algoritmos rankearon a la entropía como la característica más relevante. Los clasificadores empleados para las pruebas fueron: SVM, Fuzzy-KNN. KNN y Naïve Bayes. En este trabajo se propone la entropía como una característica con alto poder de diferenciación, sin embargo, no se realiza un estudio profundo sobre una base amplia de descriptores ni se prueba la efectividad con selectores de características más robustos.

Feature weighting as a tool for unsupervised feature selection [32]

En este trabajo se propone emplear el peso como una herramienta para la eliminación de características que no son relevantes en el conjunto de datos inicial. Para esto proponen dos algoritmos no supervisados para la selección de características. Para cada conjunto de datos dado, los algoritmos calculan los pesos para cada columna y eliminan aquellas que tengan un peso pequeño. Este enfoque, evalúa el análisis de características de forma individual, es decir, analiza la relevancia de una característica sin tener en cuenta si ésta tiene una relación con otras.

Interaction-based feature selection using Factorial Design [33]

Este trabajo aborda la selección de características no solo como una tarea de eliminación de columnas en un dataset, sino también como el análisis de la relación existente entre las columnas. Para esto, los autores proponen dos etapas para identificar y ordenar las interacciones según su impacto y la segunda para identificar las interacciones relevantes de los subconjuntos de características. Este enfoque podría ser empleado para mejorar la selección de características en algoritmos basados en filtros, ya que dichos métodos, por lo general realizan el análisis de relevancia de forma individual.

Improved multiclass feature selection via list combination [34]

En esta investigación los autores proponen realizar el proceso de selección de características empleando el algoritmo Maquina de Soporte Vectorial-Eliminación recursiva de características. Normalmente, los problemas de selección de características son diseñados para problemas de clasificación binaria, sin embargo, en este trabajo los autores proponen aplicar la Eliminación Recursiva de Características en un problema de clasificación multiclase.

Minimum redundancy maximum relevance feature selection approach for temporal gene expression data [35]

Se desarrolló un método de selección de características basado en filtros a partir del cálculo de mínima redundancia y máxima relevancia, a través del cual se puede determinar la relevancia de cada característica. El método fue empleado para seleccionar características en datos temporales de expresiones genéticas, en el cual incorporaron información temporal en el proceso de selección de características con el objetivo de obtener características más discriminativas.

A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining [36]

En este trabajo se propone un algoritmo basado en el coeficiente de máxima información y la ortogonalización Gram-Schmidt. Esta aproximación tiene como objetivo solucionar el problema que se presenta en la eliminación de redundancia irrelevante en métodos como mínima redundancia-máxima relevancia. Este método busca mejorar el análisis de relevancia en métodos basados en filtros, ya que, generalmente son métodos que realizan análisis de relevancia de forma individual y no consideran las relaciones entre características y clases.

Differential evolution for filter feature selection based on information theory and feature ranking [37]

En este trabajo se propone un nuevo criterio para filtrar las características a partir la combinación de información Mutua, ReliefF y Puntuación Fisher. El objetivo es combinar los métodos para seleccionar las características mejor ranqueadas determinadas por ReliefF y Puntuación Fisher al calcular la relevancia mutua entre características y clases. Los resultados obtenidos en este trabajo muestran que combinar estas tres técnicas podría compensar las falencias de cada técnica cuando se usa por separado. Con lo anterior, se podría trabajar en el futuro en la construcción de un esquema de multi-selección de características, de tal manera que a partir de un consenso entre varios métodos se pueda proponer un algoritmo de selección de características más robusto.

2.2.3.Métodos de ensamble de algoritmos de selección de características

En esta sección se resumen los trabajos que describen métodos de ensamble de algoritmos de selección de características en diferentes dominios.

Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection [38]

En este trabajo se plantea el ensamble de las respuestas dadas por un clasificador que ha sido entrenado con *n* diferentes conjuntos de características relevantes generados por diferentes algoritmos de selección de características. El proceso se repite para C clasificadores y a través de voto mayoritario se llega un consenso para definir la clase de la instancia. Los autores proponen el ensamble a partir de las respuestas de los clasificadores, pero no de los métodos de selección de características.

A novel information theory-based ensemble feature selection framework for high-dimensional microarray data [39]

Los autores proponen un método basado en ensamble a partir del entrenamiento de un conjunto de clasificadores, el entrenamiento de cada clasificador está basado en diferentes conjuntos de características relevantes que han sido generados previamente. En la etapa final las respuestas dadas por cada clasificador son integradas a través de un clasificador final que funciona empleando el voto mayoritario. Aunque los resultados de la propuesta superan los resultados obtenidos en la literatura con la que se comparan, el método de ensamble funciona para la clasificación y no para la selección de características.

Indicator selection with committee decision of filter methods for stock market price trend in ISE [40]

En esta investigación los autores proponen a través de un comité de decisión mezclar los conjuntos de características relevantes seleccionados con algoritmos de selección de características basados en filtros. Con este enfoque se busca considerar enfoques de filtros de características basados en diferentes métricas, con el objetivo de obtener un conjunto de características relevantes más robusto sin importar los cambios que pueda sufrir el dataset de prueba. En esta aproximación se busca realizar un consenso de diferentes conjuntos de selección de características generados a partir de algoritmos basados en filtros, sin embargo, también se deberían tener en cuenta los métodos envolventes, ya que generalmente, estos tienen un rendimiento superior que los basados en filtros.

Detection of Aβ plaque deposition in MR images based on pixel feature selection and class information in image level [41]

En esta investigación se propone generar varios conjuntos de características relevantes a partir del uso de diferentes métodos de selección de características sobre el dataset original, posteriormente a través del voto mayoritario se construye el conjunto final de características relevantes y se prueba con diferentes clasificadores para obtener el mejor rendimiento. Las pruebas fueron realizadas empleando características extraídas de imágenes digitales. El proceso de clasificación construido arrojó una precisión del 80 %. Este mecanismo empleó diferentes algoritmos basados en filtros y envolventes, lo cual podría garantizar que el conjunto de características seleccionado incluye las ventajas de cada uno de estos enfoques.

Unsupervised ensemble feature selection for underwater acoustic target recognition [13]

En esta investigación se emplea una técnica de muestreo sobre el dataset original con el fin de generar un conjunto de características relevantes por cada muestra. Finalmente se establece a través de voto mayoritario la fusión de los diferentes conjuntos de características generados para obtener un conjunto final. Este esquema funciona bien para la tratar de controlar las inconsistencias del algoritmo de selección de características cuando el dataset de entrenamiento cambia, sin embargo, no se logran evitar problemas como la detección de características redundantes.

2.3. Brechas Existentes

Considerando la literatura revisada, se han identificado proyectos de detección automática de convulsiones epilépticas que arrojan resultados muy positivos en su evaluación respecto a la precisión y exactitud de la clasificación realizada. Sin embargo, la gran mayoría de proyectos revisados proponen una solución basada en un número reducido de características para evitar el alto costo computacional del cálculo de éstas, la complejidad que se le puede agregar al clasificador para generalizar y la ausencia de datasets con suficientes instancias que permitan soportar un proceso de aprendizaje y/o clasificación analizando diferentes descriptores. En este sentido, no se evidencia un estudio completo en el que se comparen los diferentes enfoques de caracterización (descriptores) que tienen este tipo de señales, y en consecuencia no se realiza la clasificación a partir de un análisis robusto de relevancia de características. Adicionalmente, para hacer un análisis de relevancia de las características que permiten clasificar eventos epileptiformes, se encuentra un escenario con un gran número de descriptores para soportar la extracción de características, pero datasets con un número reducido de instancias, lo cual obliga a realizar un proceso de selección de características.

Teniendo en cuenta lo anterior, es necesario identificar las características que representan la información más relevante de la señal y extraer únicamente ésta para el proceso de clasificación. Debido a que incluir información no relevante implicaría adicionar ruido al clasificador y reducir la exactitud y precisión del clasificador. En este sentido, para la construcción de un sistema para la identificación automática de anomalías epileptiformes en una señal, se debe considerar hacer un proceso de análisis de relevancia de características con el fin de identificar las características relevantes que permiten describir y clasificar correctamente una señal.

Por otro lado, los algoritmos de análisis de relevancia son frecuentemente utilizados en problemas de recuperación de información con el fin de reducir la complejidad de la gestión de datasets de gran volumen, como por ejemplo en datasets de EEGs y bioinformática dónde se cuentan con datasets de gran dimensionalidad. Sin embargo, la literatura revisada evidencia que no han sido ampliamente considerados en problemas de procesamiento digital de señales EEG. Teniendo en cuenta esto, en la presente propuesta tiene como objetivo describir una solución para la detección automática de anomalías epileptiformes en señales EEG a partir de características relevantes, para lo cual se plantea la implementación de un método de ensamble de algoritmos de selección de características, que determine las características a partir de un consenso de varios algoritmos de selección de características, de tal manera que se puedan combinar las diferentes aproximaciones y ventajas de los selectores en una solución más completa.

Como se pudo observar en el estado del arte, los enfoques que se han empleado para el ensamble de selección de características han sido empleados en dominios diferentes al análisis de señales EEG, y con el objetivo principal de mezclar las ventajas de los diferentes tipos de algoritmos y reducir el impacto de sus debilidades, por ejemplo, la sensibilidad de la detección de características relevantes a partir de la relación existente entre ellas o la inconsistencia de la selección de características cuando se agregan o eliminan datos del dataset. Sin embargo, el método más empleado para la tarea de ensamble de métodos de selección de características es el voto mayoritario, técnica que funcionaria correctamente si todos los selectores de características tuvieran el mismo nivel de desempeño, caso que no ocurre en el dominio del análisis de señales EEG.

A continuación, se resumen las principales brechas identificadas.

- Por la naturaleza de la señal el costo computacional del análisis de EEGs es alto y las investigaciones se han enfocado en tareas específicas de clasificación.
- El análisis de relevancia no ha sido empleado ampliamente durante la caracterización de señales EEG debido a que se han empleado métodos específicos para caracterizar algunas alteraciones sobre la señal.
- En la clasificación de señales EEG se encuentra que existen una gran cantidad de descriptores, sin embargo, los datasets de datos brutos de EEGs tienen pocas instancias. En consecuencia, al extraer características de un dataset de EEGs se obtiene un dataset de entrenamiento con un número n de características superior al número de instancias.
- Los algoritmos de selección de características tienen ventajas y desventajas dependiendo su aproximación. Sin embargo, el ensamble de diferentes enfoques de algoritmos de selección de características podría ser una solución más robusta en el análisis de EEGs.

3. OBJETIVOS

General

Proponer un mecanismo para la selección de características relevantes en la detección automática de eventos epileptiformes en señales EEG.

Específicos

1. Construir un dataset de señales EEG y datos clínicos de pacientes con epilepsia.
2. Construir un método de ensamble se selección de características relevantes para la clasificación de eventos epileptiformes.
3. Evaluar el desempeño del mecanismo para la selección de características relevantes en un proceso de clasificación de eventos epileptiformes en señales EEG pediátricas.

4. ACTIVIDADES Y CRONOGRAMA

A continuación, se describen las actividades más importantes que se deben realizar para dar cumplimiento a cada objetivo.

Objetivo 1.

Empleando la metodología las 3 primeras fases del proceso para minería de datos CRISP-DM se construirá el Dataset empleado para realizar las pruebas de la propuesta.

- Fase 1: Comprensión del negocio: revisión en la literatura sobre aspectos técnicos y teóricos asociados a las señales EEG.
- Fase 2: Compresión de los datos: Identificar el formato y representación de las señales EEG.
- Fase 3: Preparación de los datos: extraer las características de las señales EEG.

Objetivo 2.

Siguiendo la fase de Modelado de CRISP-DM se prueban los algoritmos de extracción de características, se diseña el esquema de análisis de relevancia basado en un ensamble de algoritmos de selección de características, y a partir de las características más relevantes se implementa el proceso de clasificación de eventos epileptiformes.

- Revisión de la literatura y selección de algoritmos de extracción de características de señales EEG.
- Diseño e implementación de un ensamble de algoritmos de selección de características para la selección de características en señales EEG.

Objetivo 3

Para la evaluación del construido se diseñará un experimento de acuerdo al proceso experimental definido en [42]. Teniendo en cuenta que la propuesta se describe en este documento y el objetivo número 3 describe la evaluación del método construido en el objetivo 2 en función del desempeño de la clasificación que se logre con las características seleccionadas.

5. RECURSOS, PRESUPUESTO Y FUENTES DE FINANCIACIÓN

De acuerdo a la tabla sobre criterios de referencia para la elaboración del presupuesto en anteproyectos, establecida por la FIET y en la que se determinan diferentes aspectos financieros a ser tenidos en cuenta al momento de desarrollarlo se tiene en la Tabla 2:

	FUENTE			**TOTAL**
RUBROS	**Estudiante**	**FIET**	Colciencias	
Recursos Humanos	0	23’040.000	144’000.000	**147’040.000**
Recursos Hardware	2.000.000	0		**2.000.000**
Recursos Software	0	0		**0**
Recursos Varios	0	0		**0**
Publicaciones y Pasantía	0	4.000.000	**4’000.000**	**8.000.000**
TOTAL	**2.000.000**	**23’040.000**	**148’000.000**	**157.040.000**

Tabla 2. Presupuesto del proyecto.

6. CONDICIONES DE ENTREGA

Los resultados de la propuesta de doctorado serán entregados en:

- Una monografía que contenga la información recopilada a través del estado actual del conocimiento y los resultados obtenidos mediante la ejecución del presente proyecto.
- Un prototipo que valide las características propuestas.
- Un disco compacto que contenga toda la información en formato digital utilizada o generada en el transcurso del proyecto.
- Mínimo (1) artículo en una revista indexada en Colciencias en categoría A en el que se presenten los resultados del proceso investigativo llevado a cabo en la ejecución del proyecto.

REFERENCIAS

[1] Organización Mundial de la Salud., «WHO | Media Centre,» 2010. [En línea]. Available: http://www.who.int/mediacentre/factsheets/fs220/en. [Último acceso: 2015 06 06].

[2] Organización Mundial de la Salud., «The world health report 2002 - Reducing Risks, Promoting Healthy Life,» *World Health Report ,* vol. 1, 2002.

[3] P. U. Javeriana, «ESTUDIO DE DISPONIBILIDAD Y DISTRIBUCIÓN DE LA OFERTA DE MÉDICOS ESPECIALISTAS, EN SERVICIOS DE ALTA Y MEDIANA COMPLEJIDAD EN COLOMBIA,» Pontificia Universidad Javeriana, Bogotá, 2013.

[4] J. F. Ceron, «Encuesta colombiana de neurología - 2011,» *Acta Neurológica Colombiana,* vol. 28, nº 4, pp. 181-186, 2012.

[5] DANE, «Censo 2005- Cauca,» 2011B. [En línea]. Available: http://www.dane.gov.co/files/censo2005/PERFIL_PDF_CG2005/19000T7T000.PDF. [Último acceso: 2015 04 2014].

[6] S. Motamedi-Fakh, M. Moshrefi-Torbati, M. Hill, C. M.Hill y P. R. White, «Signal processing techniques applied to human sleep EEG signals — A review,» *Medical Engineering & Physics,* vol. 32, p. 679–689, 2010.

[7] B. Boashash , G. Azemi y N. Ali, «Principles oftime–frequency feature extraction for change detection in non-stationary signals: Applications to newborn EEG abnormality detection,» *PatternRecognition,* 2014.

[8] T. M. Nunes , A. L. Coelho , C. Lima, J. P. Papa y V. H. de Albuquerque, «EEG signal classification for epilepsy diagnosis via optimum,» *Neurocomputing,* vol. 136, pp. 103-123, 2014.

[9] P. Karoly, D. Freestone, R. Boston, D. Grayden, D. Himes, K. Leyde, U. Seniratne, S. Berkovic, T. O'brien y M. Cook, «Interictal spikes and epileptic seizures: their relationship and underlying rhythmicity,» *Brain a journal of neurology,* pp. 1-13, 2016.

[10] L. Gao, J. Song, X. Liu, J. Shao, J. Liu y J. Shao, «Learning in high-dimensional multimedia data: the state of the art,» *Multimedia Systems, Special Issue Paper,* vol. 3, pp. 303-313 , 2015.

[11] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. Benítez y F. Herrera, «A review of microarray datasets and applied feature selection methods,» *Information Sciences,* vol. 282, pp. 111-135, 2014.

[12] W. Duch, «Filter Methods,» de *Feature Extraction*, Berlin, Springer, 2009, p. 89.

[13] H. Yang, A. Gan, S. Shen, Y. Pan, J. Tang y J. Li, «Unsupervised ensemble feature selection for underwater acoustic target recognition,» de *Proceedings of the INTER-NOISE 2016 - 45th International Congress and Exposition on Noise Control Engineering: Towards a Quieter Future*, Hamburg, 2016.

[14] J. Meng, H. Hao y Y. Luan, «Classifier ensemble selection based on affinity propagation clustering,» *Journal of Biomedical Informatics,* vol. 60, pp. 234-242, 2016.

[15] Y. Saeys, T. Abeel y Y. Van de Peer, «Robust Feature Selection Using Ensemble Feature Selection Techniques,» *Machine Learning and Knowledge Discovery in Databases,* vol. 5212, 2008.

[16] S. Sanei y J. Chambers, EEG Singal Processing, Cardiff: John Wiley & Sons, Ltd, 2007, p. 6.

[17] M. TEplan, «FUNDAMENTALS OF EEG MEASUREMENT,» *MEASUREMENT SCIENCE REVIEW,* vol. 2, p. 1, 2002.

[18] S. Siuly, Y. Li y Y. Zhang, EEG Signal Analysis and Classification: Techniques and Applications, Melbourne: Springer, 2017.

[19] H. H. Jasper, «The ten-twenty electrode system of the International Federation,» *Electroencephalogram. CLinica Neurophyosiology,* vol. 10, pp. 367-380, 1958.

[20] K. E. Misulis, Atlas od EEG, Seizure Semiology and Management, 2nd Edition ed., Oxford University Press, 2014, p. 30.

[21] A. T. Tzallas, M. G. Tsipouras, D. G. Tsalikakis, E. C. Karvounis, L. Astrakas, S. Konitsiotis y Margaret Tzaphlidou , «Automated Epileptic Seizure Detection Methods: A Review Study,» de *Epilepsy-histological, electroencephalographic and psychological aspect*, Intech Open, 2012, pp. 76-98.

[22] I. Guyon y A. Elisseeff, «An Introduction to Variable and Feature Selection,» *Journal of Machine Learning Research ,* vol. 3, pp. 1157-1182, 2003.

[23] T. Navin Lal, O. Chapelle, J. Weston y A. Elisseeff, «Embedded Methods,» de *Feature Extraction: Foundations and Applications*, Berlin , Springer Berlin Heidelberg, 2008, pp. 131-161.

[24] E. Bauer y R. Kohavi, «An empirical comparison of voting classification algorithms: Bagging, boosting and variants.,» *Machine Learning,* vol. 36, nº 1, p. 105–139, 1999.

[25] N. Littlestone y M. K. Warmuth, «The weighted majority algorithm,» *Information and Computation,* vol. 108, pp. 212-261, 1994.

[26] U. R. Acharya y et al, «Automated EEG analysis of epilepsy: A review,» *Knowledge-Based Systems,* vol. 45, pp. 147-165, Junio 2013.

[27] M. Z. Parvez y M. Paul, «Epileptic seizure detection by analyzing EEG signals using different transformation techniques,» *Neurocomputing,* vol. 145, pp. 190-200, Diciembre 2014.

[28] Z. Mohammadpoory, M. Nasrolahzadeh y J. Haddadnia, «Epileptic seizure detection in EEGs signals based on the weighted visibility graph entropy,» *Seizure,* vol. 50, pp. 202-208, 2017.

[29] T. M. Nunes, A. L. Coelho, C. A. Lima, J. P. Papa y V. H. C. de Alburquerque, «EEG signal classification for epilepsy diagnosis via optimum path forest – A systematic assessment,» *Neurocomputing,* vol. 136, pp. 103-123, 2014.

[30] U. RAJENDRA ACHARYA, S. VINITHA SREE, P. C. ALVING ANF y J. S. SUIRI, «Application of non-linear and wavelet based features for the automated identification of epileptic EEG signals,» *International Journal of Neural Systems ,* vol. 22, nº 2, 2012.

[31] G. Gopan K , N. Sinha y D. Babu J, «Statistical features based epileptic seizure EEG detection - an efficacy evaluation,» *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on,* 2015.

[32] D. Panday, R. Cordeiro de Amorin y P. Lane, «Feature weighting as a tool for unsupervised feature selection,» *Information Processing Letters,* vol. 129, pp. 44-52, Enero 2008.

[33] X. Tang, Y. Dai y S. Meng, «Interaction-based feature selection using Factorial Design,» *Neurocomputing,* Diciembre 2017.

[34] J. Izetta, P. Verdes y P. Granitto, «Improved multiclass feature selection via list combination,» *Expert Systems with Applications,* vol. 88, pp. 205-216, Diciembre 2017.

[35] M. Radovic, M. Ghalwash, N. Filipovic y Z. Obradovic, «Minimum redundancy maximum relevance feature selection approach for temporal gene expression data,» *BMC Bioinformatics,* vol. 18, nº 9, 2017.

[36] H. Lyu, M. Wan y J. Han, «A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining,» *Computers in Biology and Medicine,* vol. 89, nº 1, pp. 264-274, 2017.

[37] E. Hancer, B. Xue y M. Zhang, «Differential evolution for filter feature selection based on information theory and feature ranking,» *Knowledge-Based Systems,* vol. 140, nº 15, pp. 103-119, 2018.

[38] C.-J. Qin y X.-P. Wang, «Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection,» *Biomedical Engineering - Applications, Basis and Communications,* vol. 29, nº 06, 2017.

[39] J. Cai, J. Luo, C. Liang y S. Yang, «A Novel Information Theory-Based Ensemble Feature Selection Framework for High-Dimensional Microarray Data,» *Advances in Intelligent Systems and Computing,* vol. 13, nº 5, pp. 742-753, 2017.

[40] A. Çakmak Pehlivanlı, B. Aşıkgil y G. Gülay, «Indicator selection with committee decision of filter methods for stock market price trend in ISE,» *Applied Soft Computing Journal,* vol. 49, pp. 792-800, 2016.

[41] Y. Li, X. Zhu, J. Wang, S. Liu, F. Li y M. Qiu, «Detection of Aβ plaque deposition in MR images based on pixel feature selection and class information in image level,» *BioMedical Engineering Online,* vol. 15, nº 18, 2016.

[42] C. Wohlin, P. Runeson, M. Hóst, M. Ohlsson, B. Regnell y A. Wesslén, Experimentation in software engineerig - An Introduction, Boston: Kluwer Academic Publishers, 2000.

[43] A. Garcés Correa, L. Orosco, P. Diez y E. Laciar, «Automatic detection of epileptic seizures in long-term EEG records,» *Computers in Biology and Medicine,* vol. 57, pp. 66-73, February 2015.

[44] Y. Song, J. Crowcroft y J. Zhang, «Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine,» *Journal of Neuroscience Methods,* vol. 210, nº 2, pp. 132-146, 2012.

[45] T. Sunil Kumar, V. Kanhangad y R. Bilas Pachori, «Classification of seizure and seizure-free EEG signals using local binary patterns,» *Biomedical Signal Processing and Control,* vol. 15, pp. 33-40, Enero 2015.

[46] F. Provost y T. Fawcett, «Robust Classification for Imprecise Environments,» *Machine Learning,* vol. 42, nº 3, pp. 203-231, 2001.

[47] C. Seale, Real-Time Processing of EEG Signals for Mobile Detection of Seizures, Irlanda: Dept. Electron. and Comp. Eng., Nui Galway Univ., Galway , 2012.

[48] G. Chen, «Automatic EEG seizure detection using dual-tree complex wavelet-Fourier features,» *Expert Systems with Applications,* vol. 41, nº 5, pp. 2391-2394, Abril 2014.

[49] S.-H. Lee, J. S. Lim, J.-K. Kim, J. Yang y Y. Lee, «Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance,» *Computer Methods and Programs in Biomedicine,* vol. 116, nº 1, pp. 10-25, Agosto 2014.

[50] L. Wang, X. Long, J. Arends y R. Aarts, «EEG analysis of seizure patterns using visibility graphs for detection of generalized seizures,» *Journal of Neuroscience Methods,* vol. 290, nº 1, pp. 85-94, Octubre 2017.

[51] A. Shahidi Zandi, G. A. Dumont, M. Javidan y R. Tafreshi , «Detection of Epileptic Seizures in Scalp Electroencephalogram: An Automated Real-Time Wavelet-Based Approach,» *Journal of Clinical Neurophysiology,* vol. 29, nº 1, pp. 1-16, 2012.

[52] W. Kerr, A. Anderson, E. LAu, A. Cho, J. Bramen, P. Douglas, E. Braun , J. Stern y M. Cohen, «Automated diagnosis of epilepsy using EEG power spectrum.,» *Epilepsia,* vol. 53, nº 11, pp. 189-192, 2012.

[53] H. Khamis, A. Mohamed y S. Simpson, «Frequency-moment signatures: a method for automated seizure detection from scalp EEG,» *Clin Neurophysiol,* vol. 124, nº 12, pp. 2317-2327, 2013.

[54] J. Jing, J. Dauwels y C. Sydney, «Automated localization of the seizure focus using interictal intracranial EEG,» *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE ,* 2014.

[55] K. Kale y J. P. Gawande, «Automated feature extraction of epileptic EEG using Approximate Entropy,» *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on ,* 2012.

[56] P. Bizopoulos, D. Tsalikakis, A. Tzallas, D. Koutsouris y D. Fotiadis, «EEG epileptic seizure detection using k-means clustering and marginal spectrum based on ensemble empirical mode decomposition,» *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on,* 2012.

[57] R. Yadav, A. Shah, J. Loeb y R. Agarwal, «Morphology-based automatic seizure detector for intracerebral EEG recordings.,» *IEEE Trans Biomed Eng,* vol. 59, nº 7, pp. 1871-1881, 2012.

[58] S. Santaniello, D. Sherman, M. Mirski, N. Thakor y S. Sarma, «A Bayesian framework for analyzing iEEG data from a rat model of epilepsy.,» *Conf Proc IEEE Eng Med Biol Soc,* p. 2011, 2011.

[59] K. Gopika, A. Harsha, L. A. Joseph y E. Kollialil, «Adaptive neuro-fuzzy classifier for 'Petit Mal' epilepsy detection using Mean Teager Energy,» *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on ,* 2013.

[60] «Classification of Normal and Epileptic EEG Signal Using Time-Frequency Domain Features through Artificial Neural Network,» *Advances in Computing and Communications (ICACC), 2012 International Conference on ,* 2012.

ACTA DE PROPIEDAD INTELECTUAL

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
ACTA DE ACUERDO SOBRE LA PROPIEDAD INTELECTUAL DEL TRABAJO DE GRADO

En atención al acuerdo del Honorable Consejo Superior de la Universidad del Cauca, número 008 del 23 de Febrero de 1999, donde se estipula todo lo concerniente a la producción intelectual en la institución, los abajo firmantes, reunidos el día ___ del mes de __________ de _________ en el salón del Consejo de Facultad, acordamos las siguientes condiciones para el desarrollo y posible usufructo del siguiente proyecto.
Materia del acuerdo: Tesis de Maestría para optar al título de Magíster en Computación.

Título de la Tesis: Sistema para la Detección Automática de Anomalías Epileptiformes en Señales EEG.

Objetivo de la Tesis: Proponer un modelo que guie el diseño de sistemas de soporte a la toma de decisión inteligente en la planeación de programas de Salud Pública

Duración de la Tesis: 48 meses.

Cronograma de actividades: ---
Término de vinculación de cada partícipe en el mismo: ----------------

Organismo financiador: -----------------, naturaleza y cuantía de sus aportes -------------, porcentaje de los costos del trabajo ---------------.

Los participantes de la Tesis, el (los) señor(es) estudiante(s) de Maestría / Doctorado Maritza Fernanda Mera Gaona, identificado(s) con la cédula de ciudadanía número 1061726592, a quien(es) en adelante se le(s) llamara "estudiante(s)", el ingeniero en calidad de Director de Tesis de Doctorado, Diego Mauricio López identificado con la cédula de ciudadanía _________, a quien en adelante se le llamará "docente", y la Universidad del Cauca, representada por el Decano de la FIET, manifiestan que:
1.- La idea original del proyecto es de Carolina González quien la propuso y presentó al Grupo de investigación respectivo Inteligencia Computacional, que la aceptó como tema para el proyecto de grado en referencia.

2.- La idea mencionada fue acogida por el estudiante como proyecto para obtener el grado de Magíster en Computación, quien la desarrollará bajo la dirección del docente.

3.- Los derechos intelectuales y morales corresponden al docente y a los estudiantes.

4.- Los derechos patrimoniales corresponden al docente, a los estudiantes y a la Universidad del Cauca por partes iguales y continuarán vigentes, aún depués de la desvinculación de alguna de las partes de la Universidad.

5.- Los participantes se comprometen a cumplir con todas las condiciones de tiempo, recursos, infraestructura, dirección, asesoría, establecidas en el anteproyecto, a estudiar, analizar, documentar y hacer acta de cambios aprobados por el Consejo de Facultad, durante el desarrollo del proyecto, los cuales entran a formar parte de las condiciones generales.

6.- El estudiante se compromete a restituir en efectivo y de manera inmediata a la Universidad los aportes recibidos y los pagos hechos por la Institución a terceros por servicios o equipos, si el comité de Postgrados, previo concepto del Comité de Maestría/Doctorado respectivo declara suspendido el proyecto por incumplimiento del cronograma o de las demás obligaciones contraídas por los estudiantes; y en cualquier caso de suspensión, la obligación de devolver en el estado en que les fueron proporcionados y de manera inmediata, los equipos de laboratorio, de cómputo y demás bienes suministrados por la Universidad para la realización del proyecto.

7.- El docente y los estudiantes se comprometen a dar crédito a la Universidad y de hacer mención del Fondo de Fomento de Investigación en caso de existir, en los informes de avance y de resultados, y en registro de éstos, cuando ha habido financiación de la Universidad o del Fondo.

8.- Cuando por razones de incumplimiento, legalmente comprobadas, de las condiciones de desarrollo planteadas en el anteproyecto y sus modificaciones, el participante deba ser excluido del proyecto, los derechos aquí establecidos concluyen para él. Además se tendrán en cuenta los principios establecidos en el reglamento del programa y el acuerdo 035 de 1992 vigente de la Universidad del Cauca en lo concerniente a la cancelación y la pérdida del derecho a continuar estudios.

9.- El documento del anteproyecto y las actas de modificaciones si las hubiere, forman parte integral de la presente acta.

10.- Los aspectos no contemplados en la presente acta serán definidos en los términos del acuerdo 008 del 23 de febrero de 1999 expedido por el Consejo Superior de la Universidad del Cauca, del cual los participantes del acuerdo aseguran tener pleno conocimiento.

Director: ____
PhD. Diego Mauricio López

Estudiante: ______________________________________
Ing. Maritza Fernanda Mera Gaona
C.C. 1.061.726.592 de Popayán (Cauca)

Decano Facultad: _________________________________
PhD. Francisco Pino

APPENDIX B

PAPER: Towards a Selection Mechanism of Relevant Features for Automatic Epileptic Seizures Detection

Towards a Selection Mechanism of Relevant Features for Automatic Epileptic Seizures Detection

Maritza MERA-GAONA[1a] Rubiel VARGAS-CANAS [b] and Diego M. LOPEZ [a]
[a] *Telematics Engineering Research Group, Universidad del Cauca, Colombia*
[b] *Dynamics systems, Instrumentation and Control Research Group, Universidad del Cauca, Colombia*

Abstract. Background: Epilepsy diagnosis is frequently confirmed using electroencephalogram (EEG) along with clinical data. The main difficulty in the diagnosis is associated with the large amount of data generated by EEG, which must be analyzed by neurologists for identifying abnormalities. One of the main research challenges in this area is the identification of relevant EEG features that allow automatic detection of epileptic seizures, especially when a large number of EEG features are analyzed. Objective: The aim of this paper is to analize the accuracy of algorithms typically used in feature selection processes, in order to propose a mechanism to identify a set of relevant features to support automatic epileptic seizures detection. Results: This paper presents a set of 161 features extracted from EEG signals and the relevance analysis of these features in order to identify a reduced set for efficiently classifying EEG signals in two categories: normal o epileptic seizure (abnormal). A public EEG database was used to assess the relevance of the selected features. The results show that the number of features used for classification were reduced by 97.51%. Conclusions: The paper provided an analysis of the accuracy of three algorithms, typically used in feature selection processes, in the selection of a set of relevant features to support the automatic epileptic seizures detection. The Forward Selection algorithm (FSA) produced the best results in the classification process, with an accuracy of 80.77%.

Keywords. Epilepsy, Epileptic seizure, EEG signal processing, Feature selection.

1. Introduction

Epilepsy is a disorder which affects approximately one in every 100 people worldwide [1]. The diagnosis is frequently confirmed through electroencephalogram (EEG), which is a noninvasive and low-cost method used to examine electric activity of the brain [2]. The data captured during an EEG is represented as waveforms signals where a specialist can identify abnormal neuronal activity in the acquired signal. Inspection of EEG is a long-lasting task, because duration of typical Epilepsy EEG recordings are between 20 and 30 minutes and in some cases reaches up to 48 hours [3]. This represents one of the main reasons for the high cost of diagnosis and treatment of Epilepsy. In addition, in Low- Middle Income countries (LMIC) the difficulties and cost of neurological diseases arise due to the lack of trained physicians. In Colombia, for example, there is approximately one neurologist per 200.000 inhabitants, which is almost nine times lower

[1] Corresponding Author. Maritza Mera-Gaona, PhD Student, Telematics Department, University of Cauca, Calle 5 No 4-70, Popayán, Colombia; email: maritzag@unicauca.edu.co

than European Countries [4] [5]. The situation is even worst in rural areas. In recent years, a large number of research has been developed on the automatic diagnosis of Epilepsy, e.g, [6]. Some solutions automatically detect epileptic seizures and even propose to classify the type of Epilepsy. Furthermore, the evaluations performed showed that most of these solutions have a sensitivity or specificity higher than 95%. Nonetheless, these studies do not provide any analysis of the computational performance, the classification results presented are preliminary [7][8], and the data used for evaluation are collected in non-clinical settings.

The aim of this paper is to analyze the accuracy of algorithms typically used in feature selection process, in order to propose a mechanism to identify a set of relevant features to support the automatic epileptic seizures detection.

2. Methods

An automatic epileptic seizures detection system was built according to the scheme described in Figure 1. It is wort mentioning that the stages of signal acquisition and signal processing components are generally included into the software that captures the EEG signals, therefore, they were not addressed in this study. Feature Extraction describes the process used to calculate a representation of the EEG signals while Classification describes how to determine whether the signal is normal or abnormal through Machine Learning and Artificial Intelligence algorithms. In addition, if an abnormality is identified, it is possible classify the signal according to the type of Epilepsy suffered by the patient.

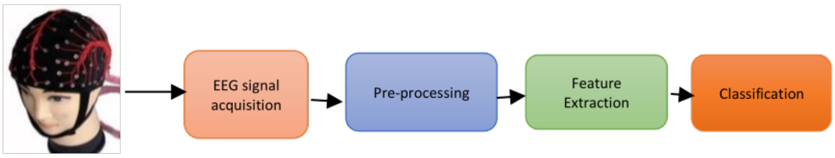


Figure 1. Scheme of the automatic epileptic seizures detection process.

2.1. Feature Extraction

The feature extraction process performs a series of mathematical operations on the EEG signals to obtain a set of descriptors which represent information in EEG signals. Some approaches propose an additional task in the feature extraction stage in order to validate the relevance of the obtained descriptors. This is known as Feature Selection (FS) [9] and is used to eliminate redundant information or noise from descriptors. Consequently, it is an optimization task, since it avoids performing operations for calculating irrelevant descriptors reducing the computational load and increasing the accuracy of the classification.

2.1.1. Experimental data

In this work, the CHB-MIT database recorded at Children's Hospital Boston was used [10]. The database contains recordings from 23 pediatric patients using the international 10-20 system of EEG electrode positions and nomenclature with a Sampling rate of 256 samples per second. Each EEG record contains information of 23 channels following the standard European Data Format (.edf) used to exchange and store multi-channel signals of biological and physical origin. In the database, each patient has several EEG records; some of them have epileptic seizures and other describe normal brain activity.

2.1.2. Feature Extractors

Feature extraction describes the process of applying a number of operators to obtain a set of descriptors. The descriptors represent the information contained in the EEG signal. According to the literature, the following descriptors are identified and calculated: Entropy [11], Maximum Amplitude, Minimum amplitude, Mean, Variance, Maximum Power and Mean Power [12]. The abovementioned features were computed for each EEG channel ending up in 161 descriptors, 23 EEG channels plus 138 computed features. Moreover, the Forward Selection, Optimize Selection and Backward Elimination algorithms were used for the Feature Selection process.

2.2. Classification

The classification process uses the features extracted in the previous phase to determine whether the signal is normal or abnormal. In this work, the algorithms Naive Bayes, Rule Induction, Decision Tree and KNN were evaluated.

3. Results

The obtained precision of the classification stage as a function of the FS process is described in Table 1.

Table 1. Accuracy of the Algorithms used in the Feature Selection process.

	Forward Selection	**Optimize selection**	**Backward Elimination**
Naïve Bayes	57.69%	42.30%	48.07%
Rule Induction	70.00%	**63.46**%	40.38%
KNN	71.11%	42.30%	40.38%
Decision Tree	**80.77**%	34.61%	48.07

The Forward Selection algorithm (FSA) produced the best results in the classification process with the subset of selected features.

The subsets of features generated by the FSA according to each classification algorithm tested are described in Table 2. The FSA reduced 161 features to: (i) 4 using Naive Bayes, (ii) 8 using Rule Induction, (iii) 4 using Decision Tree and (iv) 5 using KNN. However, the best accuracy was obtained using the subset of features generated using Decision Tree as classifier (Table 1).

The features selected were a6, a27, a94 and a121. According to the encoding scheme used to name the features, it was observed that the Mean Power of the first channel, Power Media of the second channel, the Minimum Amplitude of the fourteenth channel and the Maximum Amplitude of the eighteenth channel are the features that determine whether a signal is normal or abnormal.

The obtained results became in a contribution for the automatic detection of epileptic seizure due to it describes a new set of features that can be extracted from an EEG signal for detecting some abnormalities in the cerebral activity of a person. In addition, our proposal was tested with data from a real database, i.e., data collected in a clinical setting, obtaining an accuracy of 80.77% using only 5 features whereas in the literature, others proposals achieved better results using much more than five features and using experimental data [12][13][14], i.e., data collected in non-clinical settings.

Table 2. Features selected using the FS Algorithm.

Feature	Naive Bayes	Rule Induction	Decision Tree	KNN
a6			x	
a26	x			x
a27			x	
a32		x		
a42				x
a45	x			
a51	x			
a32		x		
a69		x		
a81		x		x
a82				x
a83		x		
a94			x	
a121			x	
a129		x		
a136		x		
a137		x		
a160	x			x

4. Discussion and Conclusions

This paper provided an analysis of the accuracy of three algorithms: Forward Selection, Optimize Selection and Backward Elimination, typically used in feature selection processes, in the selection of a set of relevant features to support the automatic epileptic seizures detection. The Forward Selection algorithm (FSA) produced the best results in the classification process, with an accuracy of 80.77%.

The results in this study represent a contribution to the process of feature extraction of EEG signals. The Feature selection process reduced the initial set of features extracted in a 97.51%. This shows that, although it is possible to obtain different data from an EEG, not all of them are relevant to support Classification.

Regarding the improvement of the accuracy in the Classification process, the Decision Tree algorithm showed better results compared to the other classification algorithms used to evaluate subsets of features generated by the Forward Selection algorithm (Naive Bayes, Rule Induction, and KNN). Therefore, it can also be concluded that the calculation of the Mean Power of the first channel, Power Media of the second channel, the Minimum Amplitude of the fourteenth channel and the Maximum Amplitude of the eighteenth channel in an EEG signal supports the automatic detection of epileptic seizures with an accuracy of 80.77%.

To the best of our knowledge, no single study has been conducted to provide an analysis of the computational performance and accuracy of relevant features selection in EEG signals, as the one presented in this study. The main limitation of this work is that, despite the proposed features have shown positive results to classify EEG signals as normal or abnormal, the results have not yet been evaluated to classify the abnormal signals according to the type of Epilepsy. This due to the EEG database used does not have this information.

The main contribution of this work to the medical field is the identification of the features that detect automatically an epileptic seizure in an EEG signal. This can decrease time of reading of an EEG signal and facilitate the diagnosis of the Epilepsy.

The accuracy analysis presented is relevant for the design of mechanisms to automatically identify a relevant features to support the automatic epileptic seizures detection, as well as for the proposal of new thechniques for automatic relevance analysis in EEG signals. As a further study, we propose to calculate and evaluate the relevance of new features in order to improve the accuracy of the results. In this context, it is very important to consider in the analysis new features, such as those extracted from electronic health records, in order to provide a more accurate diagnosis of Epilepsy. In addition, a data set combining EEG data and clinical information, as well as including information about the type of Epilepsy, have to be provided.

Acknowledgements

The work is funded by a grant from the Colombian Agency for Science, Technology and Innovation Colciencias - Call 647- 2015 Doctorados Nacionales.

References

[1] Y. Song, J. Crowcroft y J. Zhang, Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine, Journal of Neuroscience Methods 210, (1993). 131-146.

[2] C. Stam , j. Pijin, P. Suffczynski and F. Lopez da Silva, Dynamics of the human alpha rhythm: evidence for non-linearity., Clin Neurophysiol 110 (1999), 1801-1813.

[3] M. . E. Menshawy, A. Benharref y M. Serhani, « automatic mobile-health based approach for EEG epileptic seizures detection, *Expert Systems with Applications* **42** (2015), 7157–7174.

[4] P. U. Javeriana, Estudio de Disponibilidad y Distribución de la Oferta de Médicos Especialistas, en Servicios de Alta y Mediana Complejidad en Colombia, Pontificia Universidad Javeriana, Bogotá, 2013.

[5] WHO. Mental Health Atlas. Available http://www.who.int/mental_health/evidence/atlas/mental_health_atlas_2014(Last accessed: 8-feb-2016).

[6] J. Jin, J. Dauwels and S. Cash, Automated localization of the seizure focus using interictal intracranial EEG., Conf Proc IEEE Eng Med Biol Soc (2014), 4439-4442.

[7] G. Chen, Automatic EEG seizure detection using dual-tree complex wavelet-Fourier features, Expert Systems with Applications 41 (2014), 2391-2394.

[8] T. Sunil Kumar, V. Kanhangad y R. Bilas Pachori, Classification of seizure and seizure-free EEG signals using local binary patterns, Biomedical Signal Processing and Control 15 (2015), 33-40.

[9] H. Liu, H. Motoda and L. Yu, A selective sampling approach to active feature selection, Artificial Intelligence, 159 (2004), 49-74.

[10] N. Ahammad, T. Fathima and P. Joseph, Detection of Epileptic Seizure Event and Onset Using EEG,»BioMed Research International, 2014 (2014), 1-7.

[11] Mathworks - Entropy, 03 March 2016. Available: http://www.mathworks.com/help/wavelet/ref/wentropy.html (Last accessed: 03-Mar-2016).

[12] M.A. Naderi and H. Mahdavi-Nasab, Analysis and classification of EEG signals using spectral analysis and recurrent neural networks, Biomedical Engineering (ICBME) 17th Iranian Conference of (2010), 1-4.

[13] G. Gopan, N. Sinha and D. Babu, Statistical Features based Epileptic Seizure EEG Statistical Features based Epileptic Seizure EEG, Advances in Computing, Communications and Informatics (ICACCI), (2015), 1394-1398.

[14] A.M. Aldabbagh, Low computational complexity EEG epilepsy data classification algorithm for patients with intractable seizures, Biomedical Engineering (ICoBE) (2015), 1-4.

APPENDIX C

Paper: Automatic Detection of Epileptic Spike in EEGs of Children using Matchted Filter

Automatic Detection of Epileptic Spike in EEGs of Children using Matchted Filter

Maritza Mera[a,1], Diego M. López[a], Rubiel Vargas[a] and Maria Miño[a]
[a] *University of Cauca, Colombia*

Abstract. Problem: Electroencephalogram (EEG) is one of the most used tools for the diagnosis of Epilepsy. By analyzing the EEG, neurologists can identify alterations in brain activity associated with this disease. However, this task is not always easy to perform, because of the duration of the EEGs or simply the subjectivity of the specialist in detecting alterations. **Goal:** To present an epileptic spike detector based on matched filter for supporting diagnosis of Epilepsy through a tool able to automatically detect spikes in EEG of children. **Results:** The results of the evaluation showed that the developed detector achieved a sensitivity of 89.28 % which is within the range of what has been reported in the literature (82.68% and 94.4%), and a specificity of 99.96%, improving the specificity of the best reviewed work. **Conclusions:** Taking into account the results obtained in the evaluation, the solution becomes an alternative to support the automatic identification of epileptic spikes by neurologists.

Keywords. Matched Filter, Spike detection, Epilepsy, Seizure.

1. Introduction

Reading EEGs by specialists is a task which can consume a lot of effort and time due to the duration of EEG signal recordings. In general, EEG records have durations between 20 and 30 minutes and in some cases, the records are even longer (48 or 72 hours) [1] representing one of the main causes of the high cost of diagnosing neurological diseases such as epilepsy. In the same manner, the difficulty of diagnosing this kind of disease increases in developing countries, due to the lack of medical personnel; in countries like Colombia, for example, there is a rate of one neurologist per 200,000 inhabitants [2], as a result, it is difficult to guarantee diagnosis and timely attention to patients. The situation is more worrisome in the case of patients residing in rural areas because the specialists are located in the clinical centers of main cities.

Taking into account the above mentioned, automatic detection of different abnormal events present in EEG signals arises as an alternative to reduce the reading times of an EEG signal and increase the opportunity of EEG reading services, because once the abnormalities on

[1] Corresponding Author. Maritza Mera-Gaona, PhD Student, Telematics Department, University of Cauca, Calle 5 No 4-70, Popayán, Colombia; email: maritzag@unicauca.edu.co.

the signal are identified, the specialist would only have to confirm or denied them.

The automatic reading of EEGs is a field of research in which different approaches have been developed in order to offer tools that facilitate the reading of EEG records, especially for those which are of long duration. In [3], the authors proposed to classify epileptiform events using time-frequency analysis and a random forest-based classifier, achieving an accuracy of 83%. Likewise, in [4] features extracted from wavelet coefficients are used to classify the EEG segments with a 93% sensitivity and specificity. In [5], a tool based on neural networks for the detection of epileptic seizures was developed, accomplishing an accuracy, specificity and sensitivity of 88.67%, 90 %% and 95% respectively.

Considering the above, the main challenge of the works that have been developed so far is to improve the percentages of effectiveness and reliability of the detection or classification of epileptic seizures. Consequently, some investigations have been developed for the identification of specific patterns in order to increase the reliability of the reading. Due to epileptic discharges do not occur under the same pattern, thus, characterize and classify them under the same model can reduce the effectiveness of detection or classification.

The objective of this research is to propose an epileptic spike detector based on matched filter for supporting diagnosis of Epilepsy through a tool able to automatically detect spikes in EEG of children.

This paper has been organized as follows: section 2 describes the dataset used to support the development and evaluation of the proposal, theorical description of the Matched Filter and the development of the detector to automatically identify epileptic spikes. In section 3, the evaluation and the results are presented. In section 4 the discussion of the results obtained. Finally, some conclusions are described in section 5.

2. Materials y Methods

In this section, a description of the main materials, methods and concepts considered for the implementation of the automatic detection of epileptic spikes in an EEG signal is made.

2.1. *Database*

For this research, 100 electroencephalograms from children with suspected epilepsy were collected. This collection was made as part of the Neuromotic project whose general objective is the development of a TeleEEG system to support the diagnosis of epilepsy in rural areas in Colombia [6]. as part of this project, we seek to develop a component to support the reading of EEG by a neurology professional.

In the construction of the dataset and in accordance with bioethics standards, an informed consent was obtained for each EEG record, the aforementioned consent was approved by the Ethics Committee of the Universidad of Cauca, Colombia. Each EEG record was acquired using the BWII EEG device and the BW Analysis software, both developed by Neurovirtual, which has FDA certification.

Each EEG record was acquired under the electrode positioning system 10-20 [7], considering a sampling rate of 200 samples per second, and an approximate duration of 30 minutes. Some EEG records were taken in patients in the waking state (46 records) and others in sleep (54 records).

Once the records were digitized, they were evaluated by a neuropaediatrician who established the diagnosis. The EEGs diagnosed as abnormal went through an annotation process, in which segments with epileptic alterations were documented describing in detail the beginning and end of an epileptic abnormality.

2.2. *Matched Filter*

Matched filters are basic signal analysis tools used to extract known waveforms from a signal that has been contaminated with noise [8]. The model used for the extraction or detection of the wave can be seen in Figure 1.

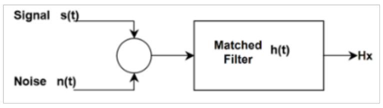


Figure 1. Detection scheme.

The scheme defined in figure 1 describes the implementation of a filter h(t) to extract the signal s(t) contaminated with noise n(t), as a result of applying h(t) it is obtained the hypothesis H_x. In this scheme, the null (H_0) and alternative (H_1) hypotheses are considered in equations (1) and (2). If the waveform that is sought is present in the signal, hypothesis H_1 is confirmed, otherwise H_0 hypothesis is confirmed. In the context of the detection of epileptic spikes, the noise n(t) represents the brain activity of the patient, the signal s(t) the epileptic spike to be found, H_0 patient’s normal activity and H_1 the presence of the epileptic spike.

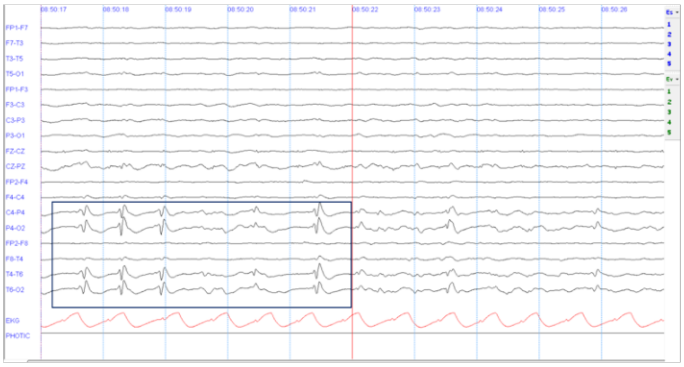
$$\boxed{H_0 : x(t) = n(t)} \qquad \textit{present signal} \tag{1}$$

$$\boxed{H_1 : x(t) = s(t) + n(t)} \qquad \textit{absent signal} \tag{2}$$

This mechanism works very well in practice when a known pattern or waveform is sought, because the filter allows to maximize the SNR (signal noise ratio) of the filtered signal and reduce the effect of noise on the original signal [9]. However, when waveforms are not known, the method does not work efficiently.

In this work, the development of a tool that supports the diagnosis of epilepsy through the identification of epileptiform events is foreseen. For this purpose, a review has been made in the literature on characteristic patterns that describe the presence of an epileptic discharge. In this sense, it could be observed that epileptic seizures generate electric shocks on some areas of the brain generating unexpected changes in the waveform of EEGs. In some cases, the appearance of waveforms is identified periodically or semiperiodically or simply the disorganization of the patient's electrical activity. Some of the most wanted patterns by neurologists during the inspection of EEGs correspond to peaks (narrow and broad). Taking into account the above, a tool that allows the automatic detection of peaks from an EEG waveform that functions as a template is proposed. This template was constructed by averaging 25 segments diagnosed as spikes by a neuropediatric expert in reading EEGs. Figure 2 shows an example of the appearance of epileptic spikes in the base rhythm of the EEG wave on channels 17, 18, 22 and 23 of the

EEG.



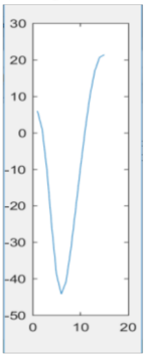


Figure 2. Epileptic Spikes.

Figure 3. Epileptic spike pattern.

From the epileptic spikes detected, the epileptic spike pattern of Figure 3 was constructed. Considering the visual analysis performed by the neurologist, it was defined that the size of the segments of epileptic spikes extracted should contain data of 15 samples, 75 ms, in order to capture the data from the beginning of the spike until the end of it.

2.3. Epileptic Spike Detector

Considering the wave pattern which describes an epileptic spike, a spike detector algorithm was constructed using matched filter and sliding windows over an EEG channel. The algorithm is defined below:

Algorithm 1. Spikes detector.

```
Void SpikesDetector (windowSize, slidingSize, pattern, EEGChannel, spikesBeginnings, spikesEnds)
    startIndex = 0
    maxIndex = Lenght (EEGChannel)
    while (startIndex< maxIndex) do
        x = EEGChannel(:, startIndex: maxIndex + windowSize)
        matches = matchedFilter(x, pattern)
        if (isNotEmpty(matches)) do
            spikesBegginings.Add (startIndex)
            spikesEnds.Add (startIndex+windowSize)
            startIndex = startIndex + slidingSize
        else
            startIndex = startIndex + windowSize
        end if
```

```
End SpikesDetector
```

The algorithm receives 6 as arguments, the size of window, size of sliding, pattern, EEG channel, Beginnings and ends of detected segments. The size of the window allows determining the start and end of the segment to be analyzed, as well as the size of the sliding allows knowing how many samples move to the right of the beginning of the segment that has been analyzed. Figure 4 illustrates the afore mentioned process.

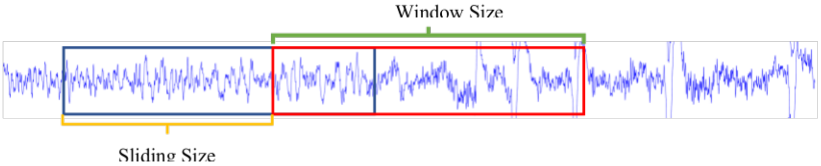


Figure 4. Analysis scheme by window.

The *pattern* corresponds to the template constructed from the epileptic spikes, *EEGChannel* corresponds to a channel extracted from the EEG in which the pattern will be searched. *spikesBeginnings* and *spikesEnds* correspond to the arrangements where the beginnings and ends of the segments that have presence of the pattern of epileptic spikes are stored. The algorithm analyzes the entire EEG channel while segments can be extracted through the sliding window, and for each window extracted a review is made with the Matched Filter to determine if this window has the presence of the epileptic spike pattern.

The algorithm that describes the Matched Filter is described below:

Algorithm 2. Matched Filter.

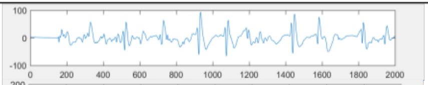
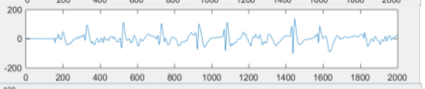
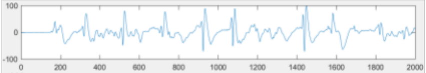
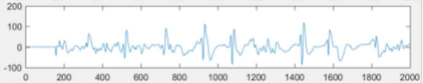
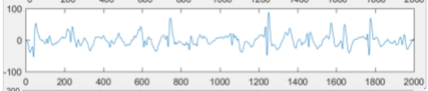
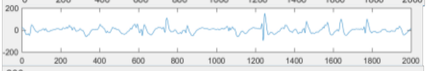
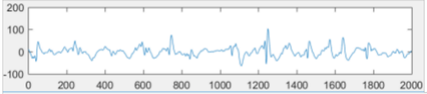
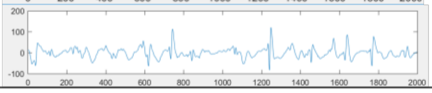
```
Var matches MatchedFilter (segment, template, thresh)
    b = createMatchedFilter(template)
    y = FilterSignal(b,segment)
    u = template.'*template
    matches = ReviewThreshold (y,thresh,u)
    return matches
End MatchedFilter
```

Where *segment* describes segment to evaluate, *template* represents de epileptic spike pattern, *thresh* sets a detection threshold, which was defined empirically, the function *createMatchedFilter* creates a matched filter based on template, *y* contains the segment filtered with the matched filter, *u* stores the autocorrelation matrix of the template and function *ReviewThreshold* establishes if *y* exceeds the threshold.

3. Results

For the evaluation of the epileptic spike detector, 8 segments of EEG records extracted from the dataset of 100 patients were used. Taking into account the annotations made on the dataset, beginnings and ends of 56 segments in which epileptic discharges occur in the form of a spike are known. In this sense, the spike detector was used for each EEG segment and the correctly identified, badly identified and unidentified segments were counted to determine the sensitivity and specificity of the detection. Table 1 describes the number of tips contained by each extracted segment.

Table 1. Description of each segment.

Numer of Segmnet	Number of spikes	Segment
S1	7	
S2	7	
S3	6	
S4	8	
S5	6	
S6	7	
S7	6	
S8	8	

Each segment described in Table 1 was reviewed by the built-in spikes detector. The results can be seen in Table 2.

Table 2. Results of Evaluation.

Segment	Real spikes	Spikes detected
S1	7	9
S2	7	29
S3	6	10
S4	8	16
S5	6	6
S6	7	19
S7	6	7
S8	8	14

In the results obtained it can be seen that the number of spikes detected in each segment is greater than the actual number of spikes. With this in mind, each spike identified by the detector was reviewed to analyze the reason for the error. It was possible to identify that in some cases, a real spike was being identified twice or three times by the detector, due to the reduced size of the sliding window and in other cases, the abrupt fall of the slow waves that occur just after the spike occurrence is also considered by the detector. It is also important to mention that slow waves are also considered an abnormality by the neurologists annotating the EEG. Thus, the spikes detected with close beginnings (difference between beginnings less than 20 samples) were considered as a single one.

Taking into account the above, Table 3 presents the results of the evaluation eliminating repeated spikes, the detection of slow waves and the number of spikes not detected.

Table 3. Results of Evaluation

Segment	Real segments	Detected spikes	Slow waves	Spikes not detected	Wrong detected spikes
S1	7	7	1	0	0
S2	7	7	12	0	1
S3	6	5	6	1	0
S4	8	7	6	1	0
S5	7	5	1	2	0
S6	7	7	9	0	0
S7	6	4	2	2	0
S8	8	8	5	0	0
TOTAL	**56**	**50**	**42**	**6**	**1**

Considering the results obtained in Table 3, it can be concluded that the built-in epileptic spike detector achieved a sensitivity of 89.28%. To calculate the sensitivity, the size of the segments (2000 samples) and the number of windows that were generated through the sliding window implemented in the detector were taken into account, which generated for each segment analyzed a total of 369 windows that had to be evaluated. Considering that there were 8 segments analyzed, 2,952 windows were revised, which allows obtaining a specificity of 99.96%.

4. Discussion

In this paper, the development of a new mechanism for the automatic detection of epileptic spikes based on the implementation of a matched filter and a template representing the waveform of an epileptic spike is presented. The tool developed reached a sensitivity of 89.28% and specificity of 99.96% in the identification of epileptic spikes on a dataset with EEG records of children.

The construction of the dataset arose as a need to have a set of training data which describes in detail the beginning and end of an epileptic abnormality, due to in the literature there are different datasets that only describe periods of time in which the appearance of an abnormality can be observed and then disorganization or new appearances of the abnormality. However, they do not describe the exact segments of the start and end of specific abnormalities [10].

The main contribution of this work for the field of neurology is the implementation of a method that automatically detects epileptic spikes with high reliability with respect to the values found in the literature. This could decrease the reading time of EEGs and facilitate the diagnosis of Epilepsy by neurologists. Additionally, the proposed method was tested using real data from a Dataset built by the authors and annotated with the help of a neuropediatrician to document the exact segments where the epileptic abnormalities occur in electroencephalograms.

In previous investigations, many tools have been designed to detect points in EEG signals. The main objective of these in the majority is to reduce the reading time of the specialists, since normally they face large volumes of data [11]. In [12] the authors describe the development of a tool for the detection of epileptic spikes using neural networks, in which a PPV (positive prediction) of 72.67% and a sensitivity of 82.68% were obtained. In [13] it is proposed to analyze the EEG record following a Markov paradigm in order to increase the sensitivity of the detection, however, the result becomes a solution with high computational complexity. In [14] a spike detector developed using analysis of energy and frequency changes is described, for this a SNEO (smoothed nonlinear energy operator) is used testing different window functions, however, the results were performed using a dataset with animal records and the objective of the tool is to support real-time evaluation of EEGs. In [15] a detector of single spikes and spikes with slow waves is proposed. The results of the evaluation show that the built model improves the accuracy of the classification when the single spikes and spikes with slow waves are considered as different classes, however, the detection is done in two stages, the first to detect a possible spike and the second one to extract features of the window and classify it as a spike, a spike with a slow wave or not spike, this could imply a greater computational load. In this study, the authors performed several configurations, obtaining a sensitivity between 87.9% and 94.4%, as well as a specificity between 86.7% and 92.3%.

Taking into account the works reviewed, the solution developed in this study obtained a sensitivity (89.28 %) within the range of what has been reported in the literature (82.68% and 94.4%) and improving the specificity of the best reviewed work. Furthermore, it is expected that this proposal reduces the computational load due to it perform fewer stages.

As future work the characterization of the greatest number of abnormalities associated with epilepsy in order to develop an epileptic event detector that includes abnormalities other than epileptic spikes is proposed. Considering that not all the abnormalities associated with epilepsy can be easily represented in a wave pattern, it is also recommended to include a classification process based on a process of character extraction through signal processing to support the classification of the segments that cannot be represented through a wave pattern. Finally, the spike detector implemented in this project was tested using EEG records of children, however, this mechanism could be used to detect epileptic spikes in adult patients, since the waveform does not change.

5. Conclussions

This paper described the implementation of an epileptic spike detector through the development of a sliding windowing mechanism that allows to screen an EEG signal window by window and determine whether these correspond to epileptic spikes by comparing of a template with each window using a matched filter. The template was constructed from the calculation of the average of 25 segments corresponding to 25 epileptic spikes and the Matched Filter method implemented achieved a sensitivity of 89.28% and a specificity of 99.96%. The main contribution of this work for the field of neurology is the implementation of a method that automatically detects epileptic spikes with high reliability with respect to the values found in the literature. This could potentially decrease the reading time of EEGs and facilitate the diagnosis of Epilepsy by neurologists.

Acknowledgements

The work is funded by a grant from the Colombian Agency for Science, Technology, and Innovation Colciencias – under Calls 715-2015, project: “NeuroMoTIC: Sistema móvil para el Apoyo Diagnóstico de la Epilepsia”, Contract number FP44842-154-2016, and Call 647- 2015.

References

[1] A. Garcés, L. Orosco, P. Diez, and E. Laciar, “Automatic detection of epileptic seizures in long-term EEG records,” *Comput. Biol. Med.*, vol. 57, pp. 66–73, 2015.

[2] J. Liliana, A. Lara, M. Alexandra, M. Gómez, and F. R. Gómez, “Informe Final Proyecto Estudio de disponibilidad y distribución de la oferta de médicos especialistas , en servicios de alta y mediana María Alexandra Matallana Gómez Autores,” 2013.

[3] L. Fraiwan, K. Lweesy, N. Khasawneh, and H. Wenz, “Automated sleep stage identification system based on time – frequency analysis of a single EEG channel and random forest classifier,” *Comput. Methods Programs Biomed.*, vol. 108, no. 1, pp. 10–19, 2011.

[4] D. G. Tsalikakis and M. G. Tsipouras, “Epileptic Seizures Classification Based on Long-Term EEG Signal Wavelet Analysis,” *Precis. Med. Powered by pHealth Connect. Heal.*, pp. 165–169, 2018.

[5] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, “Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals,” *Comput. Biol. Med.*, 2017.

[6] L. D. Molina E,. Salazar E., “NeuroEHR: Open Source Telehealth System for the Management of Clinical Data, EEG and Remote Diagnosis of Epilepsy,” 2018, pp. 915–916.

[7] M. QUIGG, “ACQUISITION OF THE ELECTROENCEPHALOGRAM 2,” in *EEG Pearls*, Mosby, 2006, pp. 17–35.

[8] J. C. Bancroft, “Introduction to Matched Filters,” 2002.

[9] J. Hermand and W. I. Roderick, “Acoustic Model-Based Matched Filter Processing for Fading Time-Dispersive Ocean Channels : Theory and Experiment,” vol. 18, no. 4, 1993.

[10] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals,” *Components a New Res. Resour. Complex Physiol. Signals*, vol. 101, no. 23, pp. 215–220, 2000.

[11] N. Gaspard, R. Alkawadri, P. Farooque, I. I. Goncharova, and H. P. Zaveri, “Clinical Neurophysiology Automatic detection of prominent interictal spikes in intracranial EEG : Validation of an algorithm and relationsip to the seizure onset zone,” *Clin. Neurophysiol.*, 2013.

[12] H. J. Carey, M. Manic, and P. Arsenovic, “Epileptic Spike Detection with EEG using Artificial Neural Networks,” pp. 89–95, 2016.

[13] H. Kumar, G. Amit, and K. Kohli, “EEG Spike Detection Technique Using Output Correlation Method : A Kalman Filtering Approach,” 2015.

[14] H. K. Garg and A. K. Kohli, “Nonstationary-Epileptic-Spike Detection Algorithm in EEG Signal using SNEO,” pp. 80–86, 2013.

[15] Y. Liu, “Model-Based Spike Detection of Epileptic EEG Data,” pp. 12536–12547, 2013.

APPENDIX D

Paper: Evaluating the impact of multivariate imputation by MICE in feature selection

Evaluating the impact of multivariate imputation by MICE in feature selection

Maritza Mera-Gaona[a] , Ursula Neumann[b], Rubiel Vargas-Canas[a] and Diego M. López[a]

[a]University of Cauca, Colombia
[b]Fraunhofer Center for Applied Research on Supply Chain Services SCS, Nuremberg, Germany

* Corresponding author
maritzag@unicauca.edu.co

Abstract

Handling missing values is a crucial step in preprocessing data in Machine Learning. Most available algorithms for analyzing datasets in the feature selection process and classification or estimation process analyze complete datasets. Consequently, in many cases, the strategy for dealing with missing values is to use only instances with full data or to replace missing values with a mean, mode, median, or a constant value. Usually, discarding missing samples or replacing missing values by means of fundamental techniques causes bias in subsequent analyzes on datasets. **Aim**: Demonstrate the positive impact of multivariate imputation in the feature selection process on datasets with missing values. **Results**: We compared the effects of the feature selection process using complete datasets, incomplete datasets with missingness rates between 5 and 50%, and imputed datasets by basic techniques and multivariate imputation. The feature selection algorithms used are well-known methods. The results showed that the datasets imputed by multivariate imputation obtained the best results in feature selection compared to datasets imputed by basic techniques or non-imputed incomplete datasets. **Conclusions**: Considering the results obtained in the evaluation, applying multivariate imputation by MICE reduces bias in the feature selection process.

Keywords. Missing values, listwise deletion, variable deletion, Multiple Imputation, Multivariate Imputation, Multivariate Imputation by Chained Equations.

Introduction

Missing data is a common problem in real-world datasets. Even if the researchers work hard to avoid them, missing values frequently occur for different reasons. Consequently, missingness can lead to issues in analyzing the data because most statistical methods and packages exclude subjects with any missing value. The result is that analyzes are made only with complete cases, affecting precision and leading to biased results. Although removing incomplete data is a fast and straightforward technique, it is also a risky solution since in applying it we must assume that discarded data does not influence the dataset. As a result of discarding cases with missing values, datasets could lose many instances of interest [1].

Considering the above, before deciding how to handle missing values in a dataset, the researchers must determine what the missing values depend on. The choice of a correct strategy will ensure an appropriate dataset to support subsequent analyzes such as Feature Selection and Classification.

According to Rubin [2] [3] there are three types of mechanisms of missing values: (i) Missing Completely At Random (MCAR), (ii) Missing At Random, and (iii) Missing Not At Random (MNAR). Missingness is MCAR if the probability of having missing data does not depend on the observed data or missing variables. For example, when a sensor's battery runs out, the sensor stops sending data to servers. Missing data is called MAR when the missing values (values can be missing or not) are related to other available information but not on unobserved data, which means that some variables depend on others. An example is that women usually avoid revealing their age in surveys (gender is related to missingness in the age variable). MNAR occurs if the probability of missingness depends on the values of unobserved variables. For example, people with high salaries avoid revealing their incomes in surveys. For some researchers, the mechanisms of MAR and MNAR are similar and indistinguishable [4].

Many studies have been carried out in order to explore mechanisms for handling missing values in different fields [5][6][7][8][9][10][11][12][13]. Although choosing the method may be difficult, most studies conclude that imputation is better than removing data due to the fact that deleting data could bias datasets as well as subsequent analyzes on these [14]. Consequently, data imputation is an important preprocessing task in Machine Learning.

An additional problem in the last few years is the proliferation of datasets with hundreds or even tens of thousands of variables. Thus, feature selection (FS) has become an option for reducing high dimensionality, redundant features, or noise from datasets [15]. Nevertheless, in real scenarios it is necessary to deal with missing values in the datasets and the most common FS techniques consider only datasets with complete data in the independent variables.

According to [16], missing values could be present in the target variable in the classification context. For example, when a classification or estimation model is evaluated, missing values are imputed in the test data's target variable and the model predicts values for the target variable. However, when a dataset has missing values in the features, we must find a way to handle the missing values and perform preprocessing tasks to get a dataset with complete data. Commonly, the missing data problem is solved by removing the instances or features with missing values or replacing the missing values using basic mechanisms such as mean, mode, etc. Although these strategies are easy to implement, they change the distribution of the datasets and may bias subsequent Machine Learning analyzes, for instance the feature selection or classification processes. On one hand, the methods to handle missing values could eliminate from the dataset: (i) relevant features or (ii) instances that reveal the importance of the relevant features. On the other hand, the machine learning models could be trained using only a part of the original datapoints.

Some studies have proposed new techniques to carry out FS on datasets with missing values [17][18][19]. Although these studies showed promising results, the authors' experiments did not evaluate the effect of data imputation on datasets to analyze whether or not the imputation methods bias the FS process. Moreover, the experiments in [17] and [19] were carried out using only rates of missing values less than or equal to 10%.

In previous studies, we evaluated how feature selection improved the performance of the classification of epileptic events and normal brain activity in Electroencephalograms [20][21]. The experiments were carried out using datasets with high dimensionality in a scenario with the need of reducing the computational complexity. The results indicated that the best subset of relevant features was selected by an approach based on Ensemble Feature Selection (EFS).

We thus proposed a Framework of Ensemble Feature Selection to improve the selection of relevant features in datasets with high dimensionality [22]. Nonetheless, one of the weakness of the original proposal was the handling of datasets with missing values. In the real world, datasets have a high probability of having incomplete data, which means that handling missing values is necessary before selecting relevant features. This renders the results of FS uncertain when the dataset has incomplete data.

This research aims to describe how data imputation can improve feature selection on datasets with missing data and avoid biasing the dataset. For this, we showed the impact of missing values in the FS process by implementing a data imputation algorithm and evaluating it with different datasets to compare the FS process using datasets without handling missing values versus imputed datasets. In light of this, this paper is organized as follows: Section 2 presents the datasets used to evaluate our proposal and theoretical descriptions about basic mechanisms for handling Missing Values, Multivariate Imputation, Multiple Imputation, and Feature Selection. In Section 3, the evaluation and results are presented. Section 4 describes the discussion of results. Finally, the main conclusions are laid out in Section 5.

Materials and Methods

Systematic mapping studies in software engineering

To review works related to FS and data imputation, we carried out two systematic mappings focused on identifying studies related to imputation and the assembly of feature selection algorithms following the guidelines described by Petersen [23]. We used two search strings, one for each topic:

- Imputation data: (imputation data) and (missing values or missingness rates or incomplete data or incomplete dataset)
- Feature selection: ("framework" and "ensemble") and ("dimensionality reduction" or "feature selection") and ("EEG" and "automatic") and ("detector" or "reading" or "recognition" or “analysis”).

The searches guided by the previous keywords, were used to find relevant papers from IEEE, PubMed, and Science Direct databases. The analysis of the papers was led following review criteria based on the quality of their contributions, particularly the proposal of imputation and assembly of feature selection algorithms.

Datasets

This research uses 4 datasets [24], [25], [26] [27], *Breast-cancer, letter-recognition, Statlog – Heart and Spambase,* from UCI Machine Learning Repository [28] to evaluate our proposal. These collections include categorical and numerical features and contain data from different fields.

The Breast-Cancer dataset contains data provided by the Oncology Institute [24]. Each instance is described by 9 attributes and represents information from a patient.

Letter-recognition is a dataset that represents 26 capital letters in the English alphabet [25]. The dataset was built considering the black-and-white pixel representation on 20 different fonts. Each representation was randomly distorted to get 20.000 instances. Each instance was converted into 16 numerical features.

The *Statlog – (Heart)* dataset contains information about heart diseases. This dataset is a modified version of the *Heart Disease* dataset [26].

The *Spambase* dataset is a collection of spam and non-spam emails [27]. It is described by 57 attributes representing emails from emails classified as spam, work or personal emails.

Table 1 describes the number of categorical and numerical features and the number of instances in each dataset.

Table 1. Datasets

Dataset	Categorical	Numerical	Instances
Breast-Cancer	9	0	286
Letter-recognition	0	16	20000
Statlog - (Heart)	7	6	269
Spambase	0	57	4601

Removing Data

The most basic method for handling missing values in datasets is removing data. However, this option could delete all class instances, remove relevant variables, unbalance the dataset, and generate biases in classification or prediction.

Listwise

Listwise deletion removes all data for a case with at least one missing value. If the dataset contains a small number of instances, this strategy can remove all samples from one or more classes. Besides, when we remove the dataset cases, the result unbalances the dataset in most cases.

Dropping Variables

Dropping variables is a good option when the variables with missing values are insignificant. Nonetheless, it is difficult to know the relevant features without making a feature selection analysis. Considering the above, imputation is usually better than dropping variables.

Imputation

Imputation allows replacing missing values with substitute or replacement values. There is a wide variety of imputation methods, and their main differences are associated with the process used to calculate the new values. It is relevant to mention that imputation does not necessarily give better results because a suitable imputation method cannot always be found.

Mean, Median and Mode replacement

A primary imputation method is to replace missing values with the overall mean, median, or mode. Although it is a fast strategy, this method presents clear disadvantages such as the mean, median, or mode imputation to reduce variance in the dataset.

Multivariate Imputation by Chained Equations

Multivariate imputation by chained equations (MICE) is an imputation method based on Fully Conditional Specification, where different models impute incomplete attributes. Hence, MICE can impute missing values in datasets with continuous, binary, and categorical attributes by using a different model for each attribute. Thus, each attribute is modeled according to its distribution; for example, binary or categorical variables are modeled using logistic regression and continuous variables using linear regression. In the regression models, the modeled attribute represents the dependent variable, and the remaining attributes represent the independent variables. MICE algorithm considers the assumption that missing values are MAR, which means that its use in a dataset where the missing values are not MAR could generate biased imputations.

The MICE algorithm is described below.

1. Build a basic imputation for every missing value in the dataset.
2. Set back missing values for one feature (F_x).
3. The observed values of F_x are used to train a prediction model in which F_x is a dependent variable, and the other features are independent.
4. The missing values for F_x are replaced with the predictions calculated by the model built in step 3.
5. For each feature with missing values, steps 2-4 are repeated. When a prediction model has imputed all features with missing values, one cycle or iteration is finished.
6. Steps 2-5 are repeated for n iterations, and the imputations are updated at each cycle. The objective is to use the number of iterations to achieve a stable imputation. The imputed dataset is obtained in the last iteration.

The researcher determines the number of iterations *n*. Many iterations can improve imputation or promote overfitting. The stable number of iterations must be found by testing different values and depends on the data and missing values.

According to the MICE algorithm, we obtain one imputed dataset when the algorithm performs *n* iterations. Additionally, if the previous process is repeated *m* times, we get multiple imputed datasets.

Multiple Imputation

Multiple imputation is a mechanism for creating multiple complete datasets in which for each missing value we calculate *m* predictions [29]. The goal of multiple imputation is predicting or estimating the missing values and considering the uncertainty about missing values and the imputation model. This approach is not meant for generating new values only because a single unique value could be calculated using more straightforward means [30].

Fig. 1 shows the main steps of Multiple Imputation.

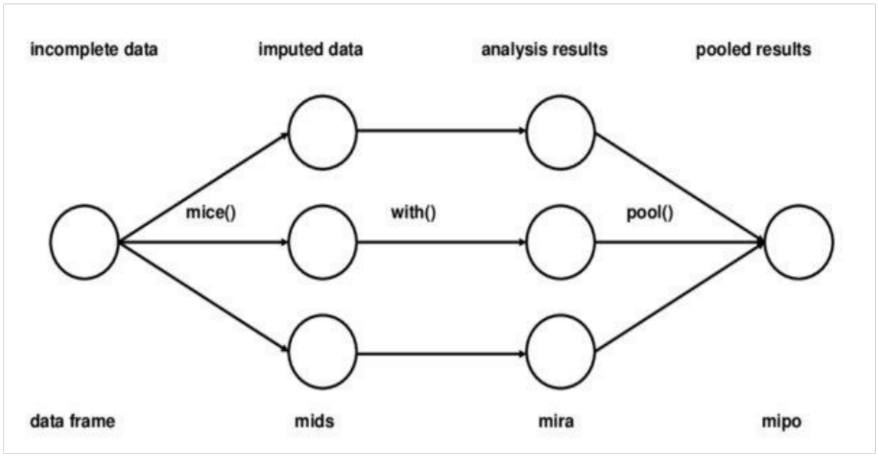


Fig. 1. Main steps used in multiple imputation [31].

MICE is a technique used to produce multiple imputations and pool them into one imputed dataset [32]. The standard strategy in Multiple Imputation is building a large joint model to predict all attributes with missing values. However, this approach is challenging to implement when there are hundreds of variables of different types. In these cases, MICE is an excellent option for handling the types [33], since the algorithm establishes a series of regression models according to the distribution and type of each attribute.

The setting of Multiple Imputation by MICE

- Number of Imputations

A critical task in Multiple Imputation is defining the number of datasets that we must impute. All imputed datasets contain the same data according to the original observed data; the differences appear initially with only the missing values. The literature recommends the number of imputed datasets ought to be between 5 and 10 [29].

- Data to train the prediction models.

A relevant aspect to consider in setting up MICE is selecting the variables or attributes included in the imputation process. Usually, we use all available variables, especially those used in subsequent analyses such as feature selection and classification/estimation. In [29], the authors consider three important points in selecting variables and their values: (i) the imputation model must be more general than the analysis model; then, if it is possible, including "auxiliary" variables (in the imputation regression model of a variable) that will not be used in the analysis process but offer information to improve the imputations; (ii) Defining whether the imputations are calculated at the item level or the summary level; for example, when there are variables constructed from other variables, it is necessary to decide if it is better to impute the original variables or the resulting variables; and, (iii) determining if the imputations will be calculated to reflect raw scores or standardized scores.

In some cases, researchers have proposed using outcome-dependent variables in the imputation model to include all possible relationships in the imputation regression model [34]. This assumption is based on the fact that the outcome depends on variables to impute. If outcomes are excluded from the imputation process, imputations will be calculated assuming that these are independent of the outcome.

- Pooling

The *m* imputed datasets generated by multiple imputation are pooled considering the types of attributes with missing values in the dataset. For instance, binary or categorical attributes are usually pooled, finding the mode of predictions and numerical attributes, calculating the mean of predictions [31].

Feature selection

Select K Best

Select K Best (SKB) is an FS algorithm for selecting a set of features according to the *k* highest scores. Scores are calculated using a test between each feature and the target. Some of the most widely used tests are described below.

Chi-squared

Chi-squared is a statistical test to evaluate features and determine whether these are dependent or independent of the target. If a feature is independent, it is considered irrelevant to the classification. Equation 1 describes the Chi-squared test.

$$X^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency} \quad (1)$$

Where *observed frequency* is the number of class observations and *expected frequency* the number of expected class observations if there was no relationship between feature and target.

F-test and ANOVA F-test

These are statistical tests to evaluate features and obtain the significance of each feature to improve a classification or regression model. The result of these measures is a subset of features with the *k* most informative features.

Recursive feature elimination

The RFE algorithm uses an external estimator to evaluate the importance of the features. Recursively, it removes features and evaluates the remaining subset by building a model with the current subset of features. The accuracy of the model is used to identify which features contribute to improving the prediction. The algorithm thus eliminates the worst-performing features on a model until the best subset is found.

Feature importance measures for tree models

The importance of a feature is calculated using Decision Trees, or the ensemble methods built upon them. One of the most common measures is Gini importance [35], based on the impurity reduction of splits. This counts when a feature is used to split a node, weighted by the number of samples it divides. When a tree model is trained using scikit-learn [36], a vector with the importance of each feature is calculated. The sum of the vector values is 1. Vector values can be used as scores to select the k most essential features, where the feature with the highest score is the most important.

Metrics to evaluate imputation methods.

We calculated the mean absolute error (MAE) and root mean square error (RMSE) between imputed values and original values in numerical variables and accuracy in categorical variables to evaluate the imputation quality.

- MAE and RMSE

The mean absolute error and the root mean square error are the standard statistical metrics used to evaluate models [37].

MAE and RMSE are described by equations 2 and 3,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \quad (3)$$

where e_i represents n samples of model errors $(e_i, i = 1, 2, \dots, n)$. To evaluate the quality of imputations, we considered equations 4 and 5. Where $\hat{Y}_i$ represents the values predicted by imputation and Y_i real values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{5}$$

- Accuracy

Accuracy is an error-rate used to evaluate the performance of classification models. It estimates the overall probability of correct classification of a test sample [38]. Accuracy is described by equation 6,

$$error = \frac{FN + FP}{N} \tag{6}$$

where N is the total of instances, FN the number of false negatives, and FP the number of false negatives.

Results

In this section, we present the evaluation results for analyzing the quality of imputation and the behavior of the process of feature selection on datasets imputed by MICE and mean/mode replacement.

Evaluating the quality of imputation

The described datasets were used to create simulated realistic datasets with missing values. Each original dataset was transformed considering 10 levels of missing data (% missingness = 5, 10, 15…,45, and 50), and for each level, the transformation was repeated 100 times. Hereafter we refer to datasets with randomly removed missing values as *simulated datasets*. Besides, each simulated dataset was imputed using MICE and mean/mode replacement.

Once the imputed datasets were generated and processed, we compared them with the original datasets to evaluate the quality of the imputations.

Outcomes

The MICE algorithm was evaluated comparing the imputed values with real values in the original dataset. We further compared the imputation calculated by MICE with the imputation calculated by mean/mode replacement. The latter is the most common and basic solution implemented to impute missing values. For this, the simulated datasets were imputed 100 times with the two methods mentioned for each missingness rate. To evaluate if the imputed values were correct, we calculated MAE and RMSE for imputations in numerical variables and accuracy for categorical variables.

- Evaluation: Breast-cancer

Fig. 2 describes the overall accuracy of imputations calculated by the MICE algorithm and mode imputation.

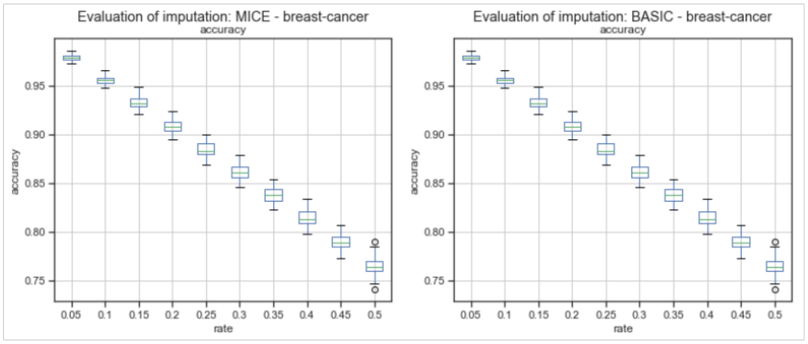


Fig. 2. Accuracy of imputations by MICE and Mode.

Fig. 3 describes the accuracy by the feature of imputations calculated using the MICE algorithm.

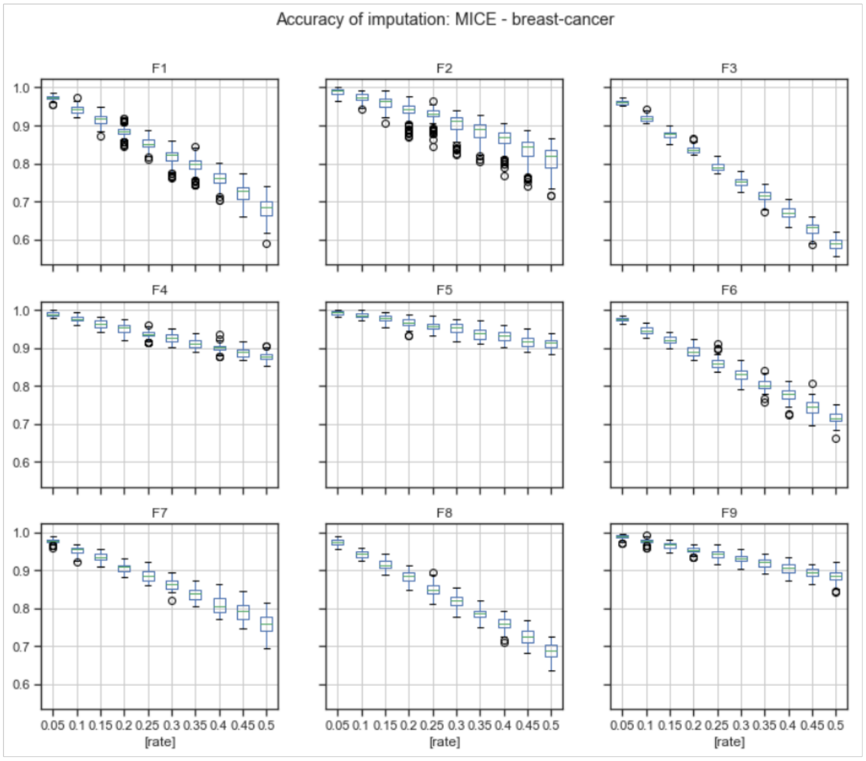


Fig. 3. Accuracy of imputations of MICE by feature.

Fig. 4 describes the accuracy by the feature of imputations calculated using mode replacement.

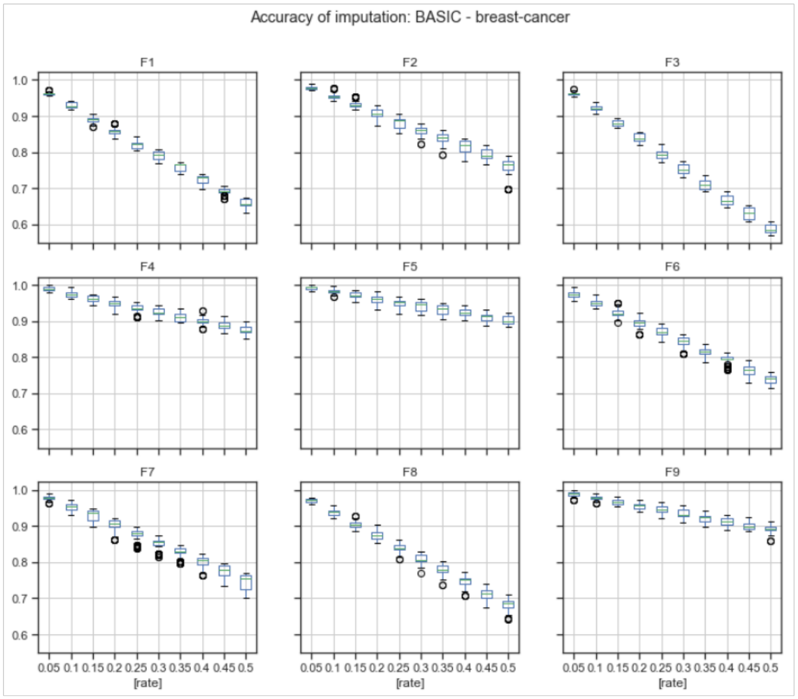


Fig. 4. Accuracy of mode imputations by feature.

Table 2 describes the overall accuracy of imputations calculated using MICE and mode replacement. According to the results, the overall accuracy achieved by MICE was better than the overall accuracy achieved by mode replacement in 100% of the missingness rates.

Table 2. The overall accuracy of MICE and Mode.

RATE	MICE	MODE
0.05	0.979	0.970
0.1	0.957	0.950
0.15	0.936	0.921
0.2	0.912	0.900
0.25	0.889	0.869
0.3	0.865	0.846
0.35	0.843	0.812
0.4	0.819	0.801
0.45	0.793	0.781
0.5	0.768	0.750

According to the results given in Appendix A, Table 15 and Table 16, the accuracy of the MICE imputation outperformed the accuracy of mode replacement in 97.59% of missingness rates by feature. Mode replacement obtained the best performance only for missingness rates of 35% and 40% in feature F3.

- Evaluation: Letter-recognition

Table 3 describes the overall MAE and RMSE of imputations calculated using MICE and mean replacement.

Table 3. The overall MAE and RMSE.

RATE	MICE		MEAN	
	MAE	RMSE	MAE	RMSE
0.05	0.0773	0.1063	0.1171	0.1542
0.1	0.0798	0.1094	0.1171	0.1543
0.15	0.081	0.1109	0.1173	0.1544
0.2	0.0834	0.1138	0.1171	0.1542
0.25	0.0872	0.1187	0.1172	0.1544
0.3	0.0924	0.1256	0.1171	0.1541
0.35	0.0929	0.1263	0.1172	0.1542
0.4	0.0938	0.1271	0.1172	0.1542
0.45	0.0948	0.1279	0.1176	0.1544
0.5	0.0952	0.1283	0.1176	0.1544

According to the results, the overall MAE and RMSE achieved by MICE were better than the overall MAE and RMSE achieved by mean replacement in 100% of missingness rates.

The MAE and RMSE of imputation calculated using MICE outperformed the MAE and RMSE of imputation calculated by mean replacement in 99.62% and 96.87% of the missingness rates by feature. Considering the number of variables of the letter-recognition dataset, we calculated these percentages but did not show each feature's results and each missingness rate.

- Evaluation: Statlog (heart)

Considering that the *statlog* dataset has categorical and numerical variables, we showed MAE and RMSE for numerical variables and accuracy for categorical variables. .

Table 4 describes the overall accuracy of imputations calculated using MICE and mode replacement. According to the results, the overall accuracy achieved by MICE was better than the overall accuracy achieved by mode replacement in 100% of missingness rates.

Table 4. The overall accuracy of MICE and Mode.

RATE	MICE	MODE
0.05	0.984	0.982
0.1	0.966	0.962
0.15	0.949	0.943
0.2	0.931	0.923
0.25	0.914	0.904
0.3	0.895	0.885
0.35	0.877	0.866
0.4	0.858	0.848
0.45	0.838	0.83
0.5	0.819	0.812

According to the results given in Appendix B, **Table 17**and **Table 18**, the accuracy of MICE's imputation outperformed the accuracy of mode replacement in 75% of the missingness rates by feature.

Table 5 describes the overall MAE and RMSE of imputations calculated using MICE and mean replacement.

Table 5. The overall MAE and RMSE.

RATE	MICE		MEAN	
	MAE	RMSE	MAE	RMSE
0.05	0.141	0.189	0.173	0.217
0.1	0.142	0.191	0.175	0.22
0.15	0.145	0.195	0.174	0.22
0.2	0.146	0.198	0.174	0.22
0.25	0.152	0.205	0.174	0.22
0.3	0.156	0.212	0.174	0.22
0.35	0.162	0.22	0.174	0.22
0.4	0.167	0.226	0.175	0.221
0.45	0.168	0.226	0.174	0.221
0.5	0.168	0.225	0.174	0.221

According to the results, the overall MAE and RMSE achieved by MICE were better than MAE and RMSE achieved by mean replacement in 100% of the missingness rates.

In accordance with the results given in Appendix B, **Table 19** and **Table 20**, the MAE of the imputation of MICE outperformed the MAE of mean replacement in 81.42% of the missingness rates by feature. Also, **Table 21** and **Table 22** show that the RMSE of the imputation of MICE outperformed the RMSE of the mode replacement in a 68.85% of the missingness rates by feature.

- Evaluation: Spambase

Table 6 describes the overall MAE and RMSE of imputations calculated using MICE and mean replacement.

Table 6. The overall MAE and RMSE.

RATE	MICE		MEAN	
	MAE	RMSE	MAE	RMSE
0.05	0.0185	0.0508	0.0229	0.0568
0.1	0.0187	0.0509	0.023	0.0569
0.15	0.0189	0.0511	0.0231	0.0565
0.2	0.0195	0.0522	0.0234	0.057
0.25	0.02	0.0531	0.0234	0.0566
0.3	0.0215	0.0553	0.0237	0.0565
0.35	0.0234	**0.0579**	0.0241	**0.0568**
0.4	0.0233	**0.0572**	0.0241	**0.0565**
0.45	0.0239	**0.0579**	0.0247	**0.0571**
0.5	0.0241	**0.0575**	0.0249	**0.0569**

According to the results, the overall MAE and RMSE achieved by MICE outperformed the overall MAE and RMSE achieved by mean replacement in 100% and 60% of missingness rates, respectively.

The MAE and RMSE of imputation calculated using MICE outperformed the MAE and RMSE of imputation calculated by mode replacement in 77.36% and 70% of the missingness rates by feature, respectively. Considering the number of variables of the *spambase* dataset, we calculated these percentages but did not show each feature's results and each missingness rate.

Densities

Fig. 5, Fig. 6, and Fig. 7 describe each variable's probability density function of the complete *breast-cancer* dataset and datasets imputed using MICE and mode replacement. According to the figures, the imputation calculated using MICE has densities similar to the complete dataset ones. However, most densities of datasets imputed using mode replacement did not only change in their shape but also increased the probabilities for some values compared to the complete dataset.

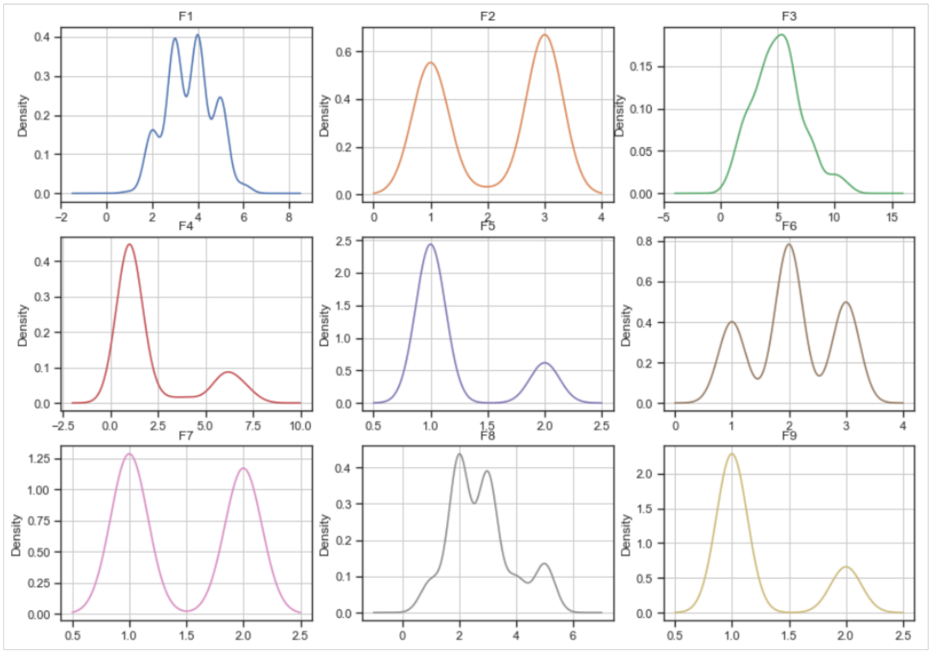


Fig. 5. Distribution of complete breast-cancer dataset.

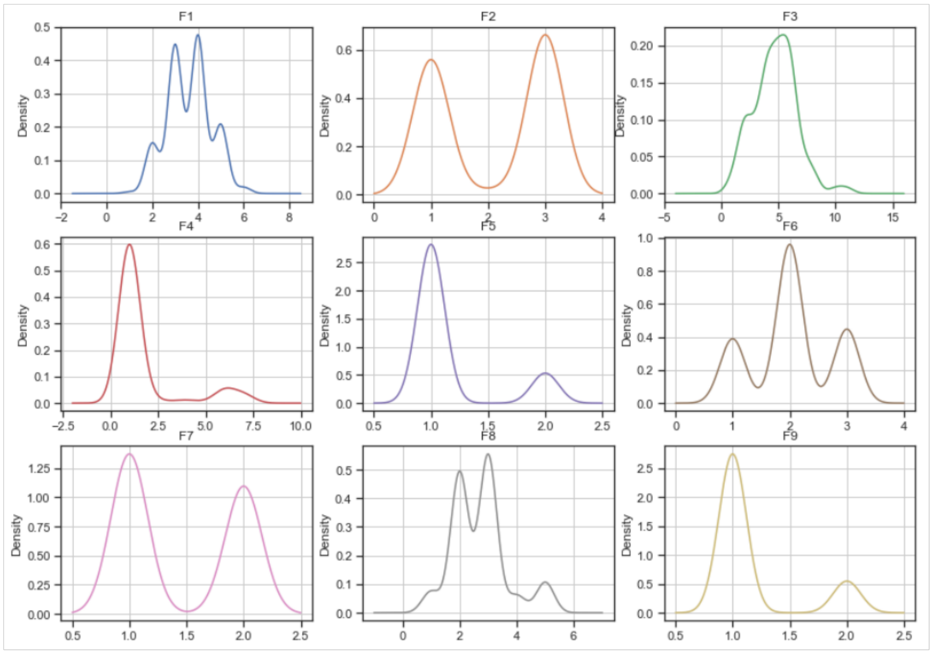


Fig. 6. Distribution of breast-cancer dataset imputed by MICE.

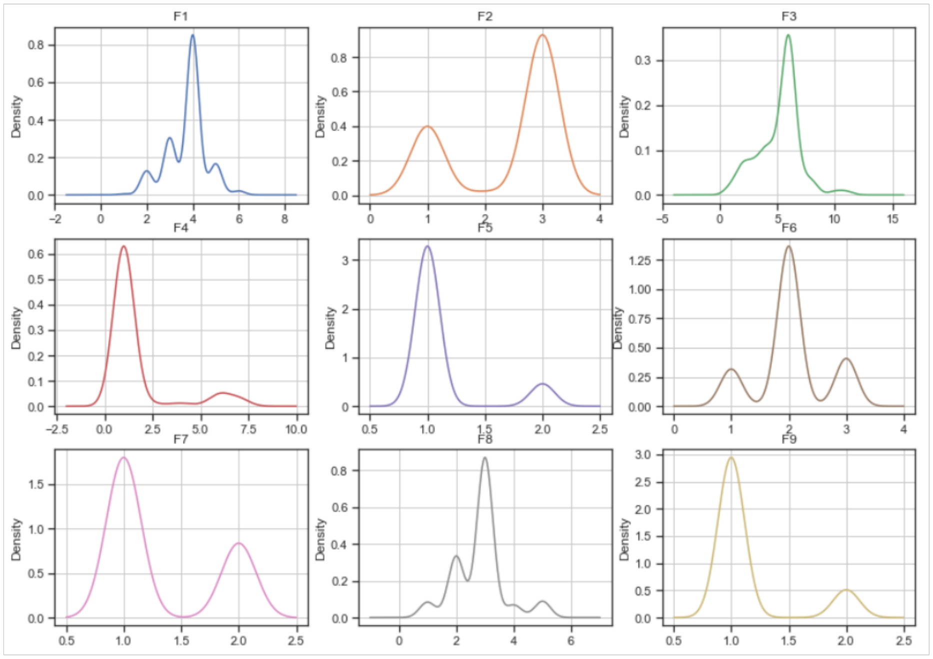


Fig. 7. Distribution of breast-cancer dataset imputed by mode.

Considering the number of variables of the *startlog* (heart), *spambase*, and *letter-recognition* datasets, the densities of their variables are not shown in this paper. However, they were plotted and analyzed. As a result of this analysis, the imputations calculated by MICE maintain their densities close to the densities of the complete dataset, while densities of the imputed dataset using mode/mean replacement changed their shapes and probabilities.

Evaluating feature selection

To evaluate the impact of missing values on feature selection, we simulated realistic datasets using the datasets described in **Table 1**. For each dataset, we generated three datasets with three different missingness rates: 25%, 30%, and 35%. Considering the simulated datasets, five FS algorithms were used to select relevant features on the complete dataset, on the dataset imputed using MICE, the dataset imputed using basic methods (Mean/Mode replacement), the dataset without missing values in instances (listwise elimination), and the dataset without missing values on variables (dropping variables).

Letter-recognition

Table 7 describes the *letter-recognition* dataset's relevant features that were selected using five algorithms of feature selection.

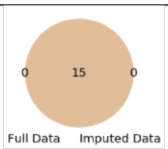
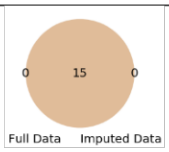
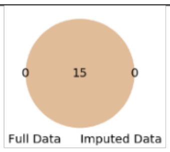
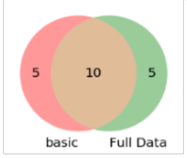
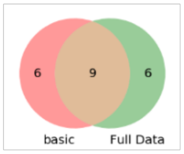
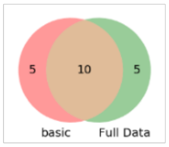
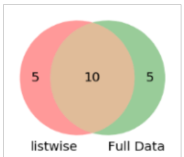
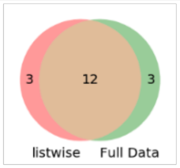
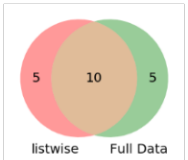
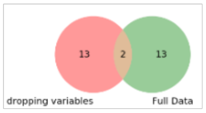
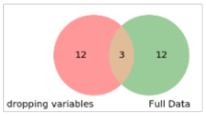
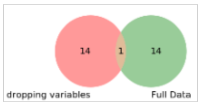
Table 7. Results of feature selection of the letter-recognition dataset.

dataset	Algorithm	
Full	Select K Best (Chi-squared)	F11, F13, F15
	Select K Best (F-value)	F7, F11, F14
	Select K Best (ANOVA F-value)	F7, F11, F13
	Feature Recursive Elimination	F12, F13, F14
	Feature Importance	F9, F13, F15

The results of applying five feature selection algorithms on datasets generated from simulations of missing values in the *letter-recognitio*n dataset are described in Appendix C, **Table 23**. Each simulated dataset handled missing values with imputation by MICE and mean/mode replacement, listwise deletion, and dropping variables.

Table 8 describes the intersection between the set of relevant features of *the letter-recognition* dataset and each simulated dataset's relevant features.

Table 8. Intersections of sets of relevant features of the letter-recognition dataset and its simulated datasets.

dataset	25%	30%	35%
Imputed			
Basic			
listwise			
Dropping variables			

According to the results in Table 7 and Table 8, the datasets imputed using MICE obtained the same set of relevant features as the complete dataset. The results also showed that datasets that were imputed using basic methods or removing instances of variables with missing values were influenced by dataset changes and produced different sets of relevant features.

Statlog (heart)

Table 9 describes the relevant features of the *statlog* dataset that were selected using five algorithms of feature selection.

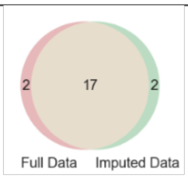
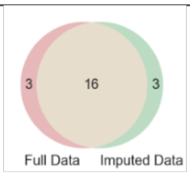
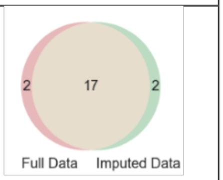
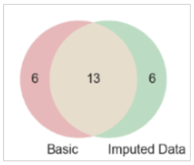
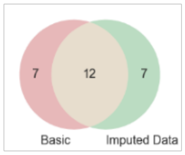
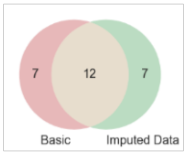
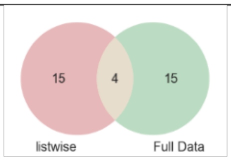
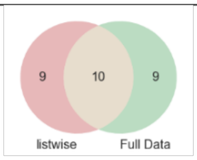
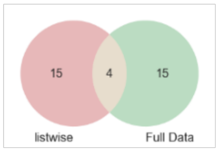
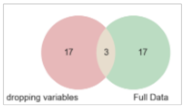
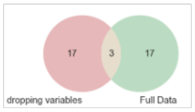
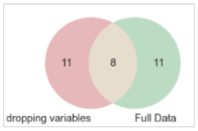
Table 9. Results of feature selection of the statlog dataset.

dataset	Algorithm	
Full	Select K Best (Chi-squared)	F3, F9, F12, F13
	Select K Best (F-value)	F3, F9, F12, F13
	Select K Best (ANOVA F-value)	F3, F9, F12, F13
	Feature Recursive Elimination	F8, F10, F12
	Feature Importance	F3, F9, F12, F13

The results of applying five feature selection algorithms on datasets generated from simulations of missing values in the *statlog* dataset are presented in Appendix D, **Table 24**. Each simulated dataset handled missing values with imputation by MICE and mean/mode replacement, listwise deletion, and dropping variables.

Table 10 describes the intersection between the set of relevant features of *the statlog* dataset and each simulated dataset's relevant features.

Table 10. Intersections of sets of relevant features of the statlog dataset and its simulated datasets.

dataset	25%	30%	35%
Imputed			
Basic			
Listwise			
Dropping variables			

Spambase

Table 11 describes the relevant features of the *spambase* dataset that were selected using five algorithms of feature selection.

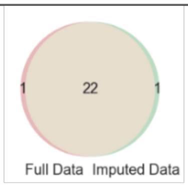
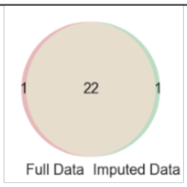
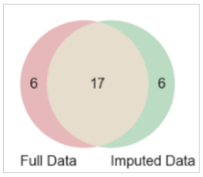
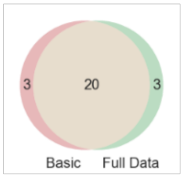
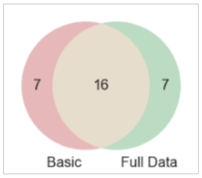
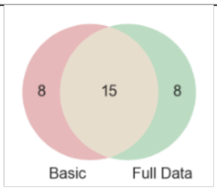
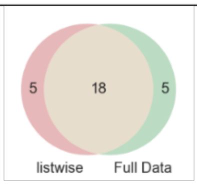
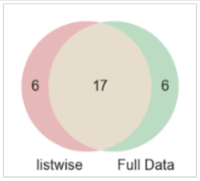
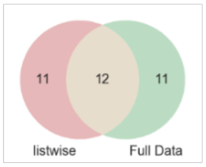
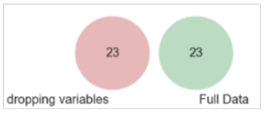
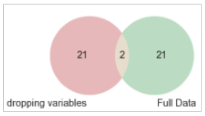
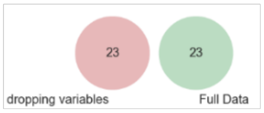
Table 11. Results of feature selection of the *spambase* dataset.

dataset	Algorithm	
Full	Select K Best (Chi-squared)	F25, F27, F55, F56, F57
	Select K Best (F-value)	F7, F19, F21, F23, F53
	Select K Best (ANOVA F-value)	F7, F19, F21, F23, F53
	Feature Recursive Elimination	F7, F27, F53
	Feature Importance	F7, F16, F21, F52, F53

The application of five feature selection algorithms on datasets generated from simulations of missing values in the *spambase* dataset is shown in Appendix E, **Table 25**. Each simulated dataset handled missing values with imputation by MICE and mean/mode replacement, listwise deletion, and dropping variables.

Table 12 describes the intersection between the set of relevant features of the *spambase* dataset and the relevant features of each simulated dataset.

Table 12. Intersections of sets of relevant features of the *spambase* dataset and its simulated datasets.

dataset	25%	30%	35%
Imputed			
Basic			
Listwise			
Dropping variables			

Breast-cancer

Table 13 describes the relevant features of the *breast-cancer* dataset selected using five feature selection algorithms.

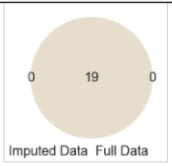
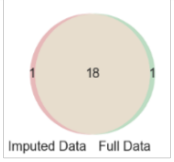
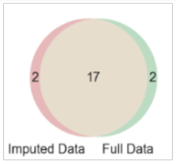
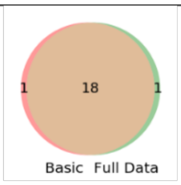
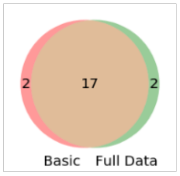
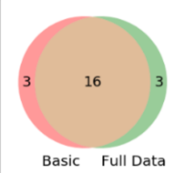
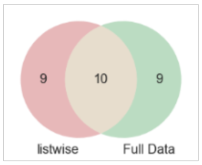
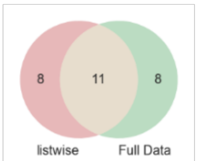
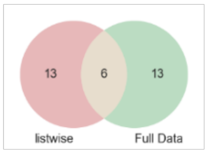
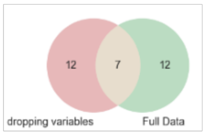
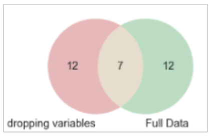
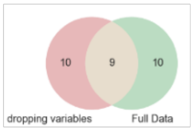
Table 13. Results of feature selection of the breast-cancer dataset.

dataset	Algorithm	
Full	Select K Best (Chi-squared)	F3, F4, F5, F6
	Select K Best (F-value)	F4, F5, F6, F9
	Select K Best (ANOVA F-value)	F4, F5, F6, F9
	Feature Recursive Elimination	F5, F6, F7
	Feature Importance	F1, F3, F6, F8

The results of applying five feature selection algorithms on datasets generated from simulations of missing values in the *breast-cancer* dataset are given in Appendix F, **Table 26**. Each simulated dataset handled missing values with imputation by MICE and mean/mode replacement, listwise deletion, and dropping variables.

Table **14** describes the intersection between the set of relevant features of the *breast-cancer* dataset and the set of relevant features of each simulated dataset.

Table 14. Intersections of sets of relevant features of the breast-cancer dataset and its simulated datasets

dataset	25%	30%	35%
Imputed			
Basic			
Listwise			
Dropping variables			

Discussion

In this work, we built an implementation of the MICE algorithm to evaluate the impact of multivariate and multiple imputation in datasets with categorical, numerical, and mixed categorical and numerical variables. The algorithm was assessed using datasets with different rates of missing values, which were generated randomly. The results were compared with the results of simple methods to handle missing values. The evaluation measured the quality of imputation, the distribution of imputed variables, and the impact in feature selection on imputed datasets.

To set up our MICE algorithm for each dataset, we took into account some aspects discussed in previous studies. For instance, Graham [39] suggests increasing the number of imputations to as many as 40 to improve imputation power when datasets have a high percentage of missing values. In practice, Graham also describes that many imputations could be inappropriate due to the dataset size, the models used to impute it, the amount of missingness in the data, and the available computer resources. In this sense, the imputation of a single dataset can take minutes, hours, or days. Thus, for datasets with hundreds or thousands of attributes and instances and a high rate of missingness, it would be impractical to calculate 40 imputed datasets as this could take hours or days. Consequently, we used many imputations for datasets with small sizes and smaller imputations for datasets of larger dimensions.

In accordance with the evaluation, the RMSE described in the previous section showed a good performance of all imputations calculated using MICE for all missingness rates. According to [40], a good result must be low (<0.3), and all results of RMSE of the MICE algorithm are less than 0.3 in overall results and results by feature.

The evaluation conducted in this paper was divided into two stages: reviewing of quality of imputation and analyzing results of FS on imputed datasets. For *the breast-cancer* dataset, the overall accuracy achieved by MICE was better than the overall accuracy of mode replacement in 100% of missingness rates, **Fig. 2** and **Table 2** . The accuracy calculated by feature showed that some features obtained better accuracy than others, **Fig. 3** and **Fig. 4**. For feature F3, the accuracy achieved using mode replacement was better than the MICE imputations. When FS was carried out, feature F3 was not considered relevant, which meant that this feature could represent noise. Besides, analyzing the accuracies calculated for missingness rates by feature, the imputation of MICE outperformed mode replacement in 97.54% of cases.

For the *letter-recognition* dataset, the overall RMSE and MAE achieved by MICE were better than the overall RMSE and MAE of mean replacement in 100% of the overall errors, **Table 3**. In the analysis by the feature of missingness rates, the MAE and RMSE achieved by MICE were better than the MAE and RMSE of mean replacement in 99.62% and 96.87% of cases, respectively.

The *statlog (heart)* contained mixed numerical and categorical variables. For this dataset, the overall accuracy achieved by MICE was better than the overall accuracy of mode replacement in 100% of missingness rates analyzed, **Table 4**. The accuracies calculated of missingness rates by feature showed that MICE was better than the accuracy of mode replacement in 75% of cases. The overall RMSE and MAE achieved by MICE were better than the overall RMSE and MAE of mean replacement in 100% of missingness rates, **Table 5**. Moreover, the RMSE and MAE calculated for missingness rates by the MICE feature outperformed the RMSE and MAE of mean replacement in 68.85% and 81.42% of cases, respectively. Features for which imputation calculated by mode/mean replacement was better than the corresponding MICE imputation were F2, F4, F5, and F6. These were not selected as relevant features in the FS process carried out on a complete *letter-recognition* dataset.

In the *spambase* dataset, **Table 6**, the overall RMSE and MAE achieved by MICE were better than the overall RMSE and MAE of mean replacement in 60% and 100% of missingness rates, respectively. However, in the analysis by feature, the RMSE and MAE of MICE were better than RMSE and MAE of mean replacement in 70% and 77.36%, respectively. In the results by feature, the percentages of RMSE and MAE decreased because *the spambase* dataset has a high number of features, and several of them are irrelevant and considered as noise.

In addition, **Fig. 5**, **Fig. 6** and **Fig. 7** show how the distribution of the *breast-cancer* dataset changed when the method of imputing data was mode replacement while the imputation performed by MICE algorithm achieved a similar distribution to the original dataset. Likewise, the *startlog* (heart), *spambase*, and *letter-recognition* datasets had changes in their distributions when the mode replacement method was employed.

For evaluating the impact of missing values in the FS process, three simulated datasets were built for each complete dataset (*breast-cancer*, *letter-recognition*, *statlog*, and *spambase*) using different missingness rate percentages (25%, 30%, and 35%) and four techniques to handle missing values were applied on each simulated dataset. The results showed the differences among the sets of relevant features of the datasets processed with techniques to handle missing values. For *letter-recognition,* the datasets imputed by MICE and complete dataset obtained the same set of relevant features, see **Table 7** and **Table 8**. However, the datasets imputed by basic replacement and dropped datasets changed their sets of relevant features regarding the complete dataset set of relevant features. In the *statlog* dataset, the set of relevant features of datasets imputed by MICE had two or three elements different to those of the complete dataset, see **Table 9** and **Table 10**. The other sets of relevant features changed in 6, 15, and 17 elements regarding the complete dataset set of relevant features. The results of FS on the *spambase* dataset showed that the most similar set to the set of relevant features of the complete dataset was the set of relevant features of the dataset imputed by MICE, **Table 11** and **Table 12**. For *breast-cancer,* the set of relevant features of the complete dataset and the datasets imputed by MICE differ in very few elements. The sets of relevant features of datasets imputed by basic replacement changed slightly, **Table 13** and

Table 14. The sets of relevant features of datasets imputed by listwise and dropping variables have many different elements.

In general, FS results showed that the datasets imputed by using MICE obtained sets of relevant features similar to the sets of relevant features calculated using the complete datasets. Likewise, the biggest differences were found between the sets of relevant features of the complete datasets and the datasets imputed by listwise and dropping variables.

Researchers have compared methods to impute data in previous work to determine how to improve the quality of imputation or to establish which method is better for a specific mechanism of missing values, type of variables, or dataset. Nonetheless, most studies did not evaluate the impact of imputation or removing data in the feature selection process. For instance, a comparison of imputation methods was carried out in [41]. The study used a complete dataset about smoking habits to simulate datasets with missingness rates of 5% and 15%. Although the authors showed imputation results for different missingness simulations, they only considered two missingness rates, and the dataset contained only categorical variables. Another work compared basic imputation and deletion methods. The results showed that pairwise deletion was the best technique for the dataset used in the evaluation [42]. The study evaluated missingness rates of 5%, 10%, 15%, 20% and 30%. However, the study considered neither imputation in numerical variables nor analysis of feature selection. The comparison of six methods for missing data was carried out in [43]. For the evaluation, simulated datasets were built using different missingness rate percentages (from 5% to 45%). Although the evaluation showed a detailed and reliable process to evaluate the quality of imputations calculated by the most popular methods, this did not show the impact of imputation in the feature selection process. The comparison of imputation methods in [44] also evaluated some of the most common techniques to impute data. However, the results only showed the limitations of the algorithms to impute data in any dataset. In general, most studies showed the evaluation of imputation quality but did not present the impact of missing values in subsequent analyzes. Some researchers have studied the influence of missing values in classification. However, they did not review the effect caused for missing values or imputed values in the FS process [45][46][47][48].

This study has several limitations, and the results of the quality of imputation for each method are limited to the datasets used. Hence, researchers should study their datasets to decide which method applies. In this sense, the main contribution of our research is not providing a universal solution to handle missing values or to select relevant features. Rather it involves presenting evidence about the need to consider the impact of missing values in the feature selection process.

As future work, we are considering improving the implementation of the MICE algorithm to use regression models and other methods to predict or estimate missing values. Another enhancement to ponder is evaluating whether or not the imputations improve when the target variable is included as an independent variable in predicting missing values. Besides, it is important to mention that although we designed an experiment to evaluate the impact of missing values in the feature selection process, we did not experiment simulating the three different mechanisms of missing values. For future work, we consider that the evaluation and results should be analyzed treating the mechanism of missing values separately.

Conclusions

In this paper, the implementation and evaluation of the MICE algorithm are described. MICE was developed to handle missing data, a commonly occurring problem in real datasets. Our implementation was evaluated by calculating imputed datasets from simulated datasets with different missingness rates. The evaluation compared the imputation quality of the MICE algorithm and basic methods, and the results of feature selection on complete datasets and imputed datasets (by MICE and basic methods).

According to the overall results of accuracy, MAE and RMSE shown in the evaluation, the MICE algorithm was better than the basic methods in all missingness rates used to simulate missing values in the *breast-cancer*, *letter-recognition*, and *statlog* (heart) datasets. For the *spambase* dataset, although the MICE algorithm achieved an overall MAE in all missingness rates better than the overall MAE of the basic imputations, the RMSE of the MICE algorithm only outperformed the RMSE of the basic method in 60% of all missingness rates.

The analysis of accuracy, MAE, and RMSE by feature showed that the basic method of imputation outperformed the imputation of the MICE algorithm for some features. According to the feature selection process applied to the complete datasets, these features were not relevant.

The evaluation results showed that for missingness rates greater than 5% and less than 50%, the complete datasets and imputed datasets calculated using MICE obtained similar distributions of their variables and similar results in the analyzes of feature selection.

Moreover, the datasets imputed using basic methods showed better results in the feature selection process than the simulated datasets handled by dropping missing variables or missing cases. However, the distribution of the variables of imputed datasets changed, meaning that the basic methods bias the datasets and accordingly that learning models could be biased.

Furthermore, selecting an appropriate method to handle missing values depends on the dataset, the mechanism of missing values, and the missingness rate. This paper showed evidence about the impact of missing values in common subsequent analyzes, such as the feature selection process.

Finally, as with any study, this work has limitations, and we cannot conclude that the MICE algorithm is the best method to handle missing values in all situations. However, the evidence presented in this paper shows that imputation could potentially be better for the avoidance of bias in subsequent analyzes than simply removing data in datasets with missing values.

Appendixes

Appendix A: Results of Breast-cancer

Table 15. Accuracy of MICE by feature.

RATE	F1	F2	F3	F4	F5	F6	F7	F8	F9
0.05	0.973	0.988	0.962	0.988	0.991	0.975	0.977	0.974	0.988
0.1	0.943	0.972	0.922	0.975	0.985	0.948	0.953	0.942	0.978
0.15	0.915	0.958	0.881	0.962	0.976	0.924	0.934	0.914	0.966
0.2	0.883	0.937	0.838	0.951	0.967	0.896	0.908	0.883	0.955
0.25	0.853	0.927	0.793	0.936	0.958	0.87	0.885	0.848	0.944
0.3	0.816	0.902	0.752	0.926	0.951	0.843	0.862	0.818	0.934
0.35	0.795	0.884	**0.712**	0.91	0.938	0.816	0.836	0.786	0.921
0.4	0.758	0.861	**0.668**	0.9	0.93	0.794	0.809	0.759	0.911
0.45	0.722	0.834	0.63	0.887	0.917	0.763	0.791	0.725	0.899

Table 16. Accuracy of Mode by feature.

RATE	F1	F2	F3	F4	F5	F6	F7	F8	F9
0.05	0.962	0.978	0.96	0.988	0.99	0.973	0.977	0.969	0.988
0.1	0.929	0.954	0.919	0.973	0.981	0.945	0.953	0.936	0.978
0.15	0.891	0.93	0.875	0.961	0.971	0.92	0.929	0.904	0.966
0.2	0.859	0.906	0.835	0.949	0.96	0.89	0.903	0.874	0.953
0.25	0.821	0.882	0.791	0.933	0.949	0.86	0.875	0.839	0.942
0.3	0.791	0.859	0.752	0.924	0.941	0.831	0.85	0.808	0.931
0.35	0.758	0.837	**0.714**	0.909	0.932	0.803	0.828	0.779	0.919
0.4	0.722	0.815	**0.669**	0.899	0.922	0.776	0.8	0.748	0.906
0.45	0.692	0.794	0.626	0.888	0.911	0.744	0.775	0.711	0.893

Appendix B: Results of statlog (heart)

Table 17. Accuracy of MICE by feature.

RATE	F2	F3	F6	F7	F9	F13
0.05	0.984	0.978	0.991	0.977	0.988	0.986
0.1	**0.967**	0.952	0.982	0.954	0.973	0.967
0.15	**0.951**	0.928	0.976	0.931	0.961	0.949
0.2	0.932	0.903	**0.968**	0.907	0.943	0.931
0.25	**0.916**	0.878	**0.961**	0.884	0.93	0.914
0.3	**0.901**	0.854	**0.954**	0.855	0.91	0.897
0.35	**0.879**	0.828	**0.947**	0.833	0.895	0.881
0.4	**0.867**	0.803	**0.939**	0.801	0.878	0.858
0.45	**0.849**	0.772	**0.93**	0.78	0.863	0.837
0.5	**0.833**	0.75	**0.922**	0.756	0.843	0.813

Table 18. Accuracy of mode replacement by feature.

RATE	F2	F3	F6	F7	F9	F13
0.05	0.984	0.976	0.991	0.974	0.985	0.98
0.1	**0.968**	0.95	0.982	0.946	0.967	0.957
0.15	**0.952**	0.924	0.976	0.922	0.951	0.935
0.2	0.932	0.897	**0.969**	0.895	0.934	0.912

0.25	**0.918**	0.872	**0.962**	0.865	0.921	0.888
0.3	**0.902**	0.848	**0.956**	0.838	0.902	0.863
0.35	**0.885**	0.82	**0.95**	0.81	0.886	0.848
0.4	**0.871**	0.797	**0.942**	0.788	0.867	0.821
0.45	**0.856**	0.771	**0.934**	0.764	0.851	0.802
0.5	**0.841**	0.748	**0.926**	0.741	0.837	0.78

Table 19. MAE of MICE by feature.

RATE	**F1**	**F4**	**F5**	**F8**	**F10**	**F11**	**F12**
0.05	0.133	0.123	0.089	0.112	0.12	0.199	0.212
0.1	0.134	0.125	0.09	0.119	0.112	0.195	0.22
0.15	0.133	0.128	0.088	0.12	0.119	0.203	0.223
0.2	0.136	0.129	**0.09**	0.12	0.118	0.202	0.225
0.25	0.144	**0.132**	**0.093**	0.125	0.119	0.214	0.236
0.3	0.145	**0.14**	**0.094**	0.13	0.125	0.215	0.242
0.35	0.154	**0.147**	**0.097**	0.14	0.129	0.225	0.242
0.4	0.157	**0.151**	**0.102**	0.143	0.132	0.233	0.251
0.45	0.157	**0.15**	**0.105**	0.141	0.133	0.236	0.253
0.5	0.156	**0.148**	**0.1**	0.141	0.134	0.241	0.253

Table 20. MAE of mode replacement by feature.

RATE	**F1**	**F4**	**F5**	**F8**	**F10**	**F11**	**F12**
0.05	0.165	0.126	0.091	0.134	0.158	0.286	0.264
0.1	0.159	0.129	0.094	0.143	0.147	0.28	0.26
0.15	0.156	0.13	0.09	0.143	0.149	0.282	0.267
0.2	0.156	0.131	**0.089**	0.142	0.147	0.282	0.268
0.25	0.159	**0.13**	**0.089**	0.143	0.146	0.284	0.271
0.3	0.156	**0.13**	**0.087**	0.144	0.149	0.281	0.27
0.35	0.159	**0.131**	**0.089**	0.144	0.147	0.281	0.267
0.4	0.157	**0.13**	**0.088**	0.146	0.147	0.283	0.267
0.45	0.157	**0.13**	**0.088**	0.145	0.148	0.283	0.266
0.5	0.157	**0.131**	**0.087**	0.144	0.147	0.283	0.268

Table 21. RMSE of MICE by feature.

RATE	**F1**	**F4**	**F5**	**F8**	**F10**	**F11**	**F12**
0.05	0.163	0.154	0.115	0.138	0.155	0.25	0.272
0.1	0.166	0.16	0.123	0.149	0.147	0.245	0.283
0.15	0.165	0.163	0.12	0.149	0.157	0.256	0.289
0.2	0.166	0.165	**0.121**	0.152	0.156	0.256	0.298
0.25	0.178	**0.169**	**0.121**	0.157	0.159	0.27	0.31
0.3	0.179	**0.179**	**0.124**	0.164	0.167	0.274	0.321
0.35	0.19	**0.187**	**0.128**	0.178	0.173	0.286	**0.323**
0.4	**0.195**	**0.193**	**0.135**	**0.184**	0.175	0.294	**0.334**
0.45	**0.194**	**0.191**	**0.139**	0.18	0.177	0.297	**0.334**
0.5	**0.195**	**0.189**	**0.132**	**0.179**	0.176	0.299	**0.33**

Table 22. RMSE of mode replacement by feature.

RATE	F1	F4	F5	F8	F10	F11	F12
0.05	0.197	0.157	0.116	0.163	0.191	0.31	0.307
0.1	0.191	0.163	0.125	0.177	0.179	0.301	0.305
0.15	0.189	0.166	0.121	0.174	0.186	0.307	0.314
0.2	0.189	0.166	**0.12**	0.174	0.183	0.306	0.317
0.25	0.191	**0.164**	**0.117**	0.176	0.182	0.309	0.322
0.3	0.188	**0.167**	**0.116**	0.176	0.187	0.304	0.321
0.35	0.191	**0.168**	**0.117**	0.178	0.183	0.305	**0.316**
0.4	**0.189**	**0.166**	**0.116**	**0.179**	0.184	0.309	**0.317**
0.45	**0.189**	**0.168**	**0.116**	0.18	0.187	0.309	**0.316**
0.5	**0.189**	**0.168**	**0.114**	**0.178**	0.184	0.309	**0.32**

Appendix C: Results of feature selection on letter-recognition

Table 23. Results of feature selection of simulated datasets.

dataset	Algorithm	25%	30%	35%
Imputed	Select K Best (Chi-squared)	F11, F13, F15	F11, F13, F15	F11, F13, F15
	Select K Best (F-value)	F7, F11, F14	F7, F11, F14	F7, F11, F14
	Select K Best (ANOVA F-value)	F7, F11, F13	F7, F11, F13	F7, F11, F13
	Feature Recursive Elimination	F12, F13, F14	F12, F13, F14	F12, F13, F14
	Feature Importance	F9, F13, F15	F9, F13, F15	F9, F13, F15
Basic	Select K Best (Chi-squared)	**F8, F9**, F13	**F8, F9**, F13	**F9**, F13, F15
	Select K Best (F-value)	F7, F11, F14	F7, F11, F14	F7, F11, F14
	Select K Best (ANOVA F-value)	**F2, F13**, F14	**F2, F13**, F14	**F2, F13**, F14
	Feature Recursive Elimination	F12, F13, F14	F12, F13, F14	F12, F13, F14
	Feature Importance	**F8**, F9, F13	**F8**, F9, **F13**	F9, **F12, F14**
Listwise	Select K Best (Chi-squared)	**F8**, F13, F15	**F2,** F13, F15	**F5**, F13, F15
	Select K Best (F-value)	F7, F11, F14	**F9**, F11, F14	F7, F11, F14
	Select K Best (ANOVA F-value)	F7, F11, **F12**	F7, F11, F13	F11, **F12, F14**
	Feature Recursive Elimination	**F3, F5,** F13	F12, F14, **F15**	F12, F13, F14
	Feature Importance	F9, **F12**, F13	F9, F13, F15	**F12**, F13, **F16**
Dropping variables	Select K Best (Chi-squared)	**F6, F7, F10**	**F7, F8**, F11	**F6, F7, F9**
	Select K Best (F-value)	**F9**, F11, **F12**	F7, **F10, F12**	**F8, F10,** F11
	Select K Best (ANOVA F-value)	**F9, F10**, F11	**F10**, F11, **F12**	**F8, F9, F10**
	Feature Recursive Elimination	**F9, F10**, **F11**	**F10, F11, F12**	**F8, F9, F10**
	Feature Importance	**F6, F7, F10**	**F7, F8, F11**	**F6, F7, F9**

Appendix D: Result of feature selection on Statlog(heart)

Table 24. Results of feature selection of simulated datasets.

dataset	Algorithm	25%	30%	35%
Imputed	Select K Best (Chi-squared)	F3, F9, F12, F13	F3, F9, F12, F13	F3, F9, F12, F13
	Select K Best (F-value)	F3**, F10,** F12, F13	F9, **F10**,F12, F13	F3, F9, **F10**, F13
	Select K Best (ANOVA F-value)	F3**, F10**, F12, F13	F9, **F10**,F12, F13	F3, F9, **F10**, F13
	Feature Recursive Elimination	F8, F10, F12	F8, F10, F12	F8, F10, F12
	Feature Importance	F3, F9, F12, F13	F3**, F10**, F12, F13	F3, F9, F12, F13
Basic	Select K Best (Chi-squared)	F3,**F10**,F12, F13	F3,**F10**,F12, F13	F3,**F10**,F12, F13
	Select K Best (F-value)	**F8,F10**,F12, F13	F3**, F8,F10**,F12	**F8,F10**,F12, F13
	Select K Best (ANOVA F-value)	**F8,F10**,F12, F13	F3**, F8,F10**,F12	**F8,F10**,F12, F13

	Feature Recursive Elimination	F8, F10, F12	F8, F10, F12	F8, F10, F12
	Feature Importance	F3, **F10**,F12, F13	F3, **F8,F10**,F12	**F8,F10**,F12, F13
listwise	Select K Best (Chi-squared)	**F2**, F3, **F11**, F12	**F2**, F3, F9, F13	**F2, F7, F8**, F9
	Select K Best (F-value)	**F2, F5, F7,** F12	**F4, F5, F8,** F12	**F2, F6, F7,** F9
	Select K Best (ANOVA F-value)	**F2, F5, F7,** F12	F3, **F8**, F9, F13	**F2, F6, F7**, F9
	Feature Recursive Elimination	**F2, F3, F11**	**F3, F7, F13**	**F6, F7, F9**
	Feature Importance	**F2, F3, F5, F7**	F3, F8, **F9**, F13	**F2, F6, F7**, F9
Dropping variables	Select K Best (Chi-squared)	**F2, F8**, F9, **F10**	**F2, F8,** F9, **F10**	F3, **F7, F8**, F9
	Select K Best (F-value)	**F7, F8,** F9, **F10**	F3, **F7, F8, F10**	F3, **F6, F7,** F9
	Select K Best (ANOVA F-value)	**F7, F8,** F9, **F10**	F3, **F7, F8, F10**	F3, **F6, F7,** F9
	Feature Recursive Elimination	**F4, F7, F8**	**F7, F8,** F10	**F6, F7, F9**
	Feature Importance	**F3, F7, F8, F10**	**F3, F7, F8, F10**	F3, **F6, F7,** F9

Appendix E: Result of feature selection on Spambase

Table 25. Results of feature selection of simulated datasets.

dataset	Algorithm	25%	30%	35%
Imputed	SKB (Chi-squared)	F25, F27, F55, F56, F57	F25, F27, F55, F56, F57	F25, F27, F55, F56, F57
	SKB (F-value)	F7, F19, F21, F23, F53	F7, F19, F21, F23, F53	F7, **F17**, F21, F23, **F56**
	SKB (ANOVA F-value)	F7, F19, F21, F23, F53	F7, F19, F21, F23, F53	F7, **F17**, F21, F23, **F56**
	FRE	F7, **F23**, F53	F7, **F23**, F53	F7, **F41**, F53
	Feature Importance	F7, F16, F21, F52, F53	F7, F16, F21, F52, F53	F7, F21, F52, F53, **F56**
Basic	SKB (Chi-squared)	F25, F27, F55, F56, F57	**F16**, F27, F55, F56, F57	**F16**, F27, F55, F56, F57
	SKB (F-value)	F7, **F16**, F21, F23, F53	**F7, F16**, F21, F23, F53	F7, **F16, F17**, F21, F53
	SKB (ANOVA F-value)	F7, **F16**, F21, F23, F53	**F7, F16**, F21, F23, F53	F7, **F16, F17**, F21, F53
	FRE	F7, **F23**, F53	F7**, F23**, F53	F7, **F23**, F53
	Feature Importance	F7, F16, F21, F52, F53	F16, F21, **F23**, F52, F53	F16, F21, F52, **F56, F55**
listwise	SKB (Chi-squared)	**F16**, F27, F55, F56, F57	F25, F27, F55, F56, F57	**F22**, F27, F55, F56, F57
	SKB (F-value)	F7, **F16**, F21, F23, **F57**	**F16**, F21, F23, F53, **F56**	**F8, F17**, F21, **F52**, F53
	SKB (ANOVA F-value)	F7, **F16**, F21, F23, **F57**	**F16**, F21, F23, F53, **F56**	**F8, F17**, F21, **F52**, F53
	FRE	F7, F24, F53	**F16, F23**, F53	**F16, F21,** F27
	Feature Importance	F7, F16, F21, F52, F53	**F5**, F16, F21, F52, F53	F16, **F17**, F21, F52, **F56**
Dropping variables	SKB (Chi-squared)	**F18, F20, F21, F42, F43**	**F12, F16, F38, F39, F40**	**F9, F12, F34, F37, F38**
	SKB (F-value)	**F14, F16, F18, F19, F40**	**F6, F12, F13, F16, F18**	**F9, F10, F12, F14, F35**
	SKB (ANOVA F-value)	**F14, F16, F18, F19, F40**	**F6, F12, F13, F16, F18**	**F9, F10, F12, F14, F35**
	FRE	**F19, F21, F40**	**F16, F18, F25**	**F14, F19, F35**
	Feature Importance	**F13, F18, F39, F40, F42**	F6, **F12**, F16, **F18, F37**	**F9, F12, F34, F35, F37**

Appendix F: Result of feature selection Breast-cancer

Table 26. Results of feature selection of simulated datasets.

dataset	Algorithm	25%	30%	35%
Imputed	Select K Best (Chi-squared)	F3, F4, F5, F6	F3, F4, F5, F6	F3, F4, F5, F6
	Select K Best (F-value)	F4, F5, F6, F9	F4, F5, F6, F9	F4, F5, F6, F9
	Select K Best (ANOVA F-value)	F4, F5, F6, F9	F4, F5, F6, F9	F4, F5, F6, F9
	Feature Recursive Elimination	F5, F6, F7	**F1**, F5, F6	**F1**, F5, F6
	Feature Importance	F1, F3, F6, F8	F1, F3, F6, F8	F1, F3, **F4**, F8
Basic	Select K Best (Chi-squared)	F3, F4, F5, F6	F3, F4, F5, F6	F3, F4, F6, **F9**
	Select K Best (F-value)	F4, F5, F6, F9	F4, F5, F6, F9	F4, F5, F6, F9
	Select K Best (ANOVA F-value)	F4, F5, F6, F9	F4, F5, F6, F9	F4, F5, F6, F9

	Feature Recursive Elimination	**F1**, F6, F7	**F1**, **F4**, F6	F5, F7, **F9**
	Feature Importance	F1, F3, F6, F8	F1, F3, F6, F8	F1, F3, **F4**, F8
listwise	Select K Best (Chi-squared)	**F1, F2**, F3, F4	F3, F4, F5, **F8**	**F1, F2,** F4, **F7**
	Select K Best (F-value)	**F3**, F5, **F7**, F9	**F3**, F4, F5, F6	**F1, F2,** F4, **F7**
	Select K Best (ANOVA F-value)	**F3**, F5, **F7**, F9	**F3**, F4, F5, F6	**F1, F2,** F4, **F7**
	Feature Recursive Elimination	F5, F7, **F9**	**F2, F4**, F7	**F2**, F7, **F8**
	Feature Importance	F1, F3, **F4, F7**	**F4, F5**, F6, **F7**	**F2**, F3, **F7**, F8
pairwise	Select K Best (Chi-squared)	**F2, F3**, F4, **F7**	**F2**, F3, F4, **F7**	**F1**, F3, F4, F6
	Select K Best (F-value)	**F2, F3**, F4, **F7**	**F2**, **F3**, F4, **F7**	**F1, F3**, F4, F6
	Select K Best (ANOVA F-value)	**F2, F3**, F4, **F7**	**F2**, **F3**, F4, **F7**	**F1, F3**, F4, F6
	Feature Recursive Elimination	**F4**, F5, F7	**F3, F4**, F5	**F1, F3, F4**
	Feature Importance	**F2**, F3, **F4**, F6	F1, **F2**, **F4**, F6	F1, F3, **F4, F5**

Funding

The work was funded by a grant from the Colombian Agency of Science, Technology, and Innovation Colciencias – under Funding call 647- 2015, project: "Mechanism of selection of relevant features for the automatic detection of epileptic seizures"; the funder provided support in the form of a scholarship for MMG but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Additionally, Fraunhofer Center for Applied Research on Supply Chain Services SCS and University of Cauca provided support in the form of salaries for DML and RVC, and UN, respectively. However, the employers did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Acknowledgements

The authors are grateful too to Mr Colin McLachlan for suggestions related to the English text.

Authors' contributions

MMG developed and evaluated the proposal and wrote the original draft of the manuscript. UN conceptualized the idea, proposed the methodology, and supported the writing of the manuscript, DML and RVC reviewed and edited the manuscript and supervised the research. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests. The academic and commercial affiliations of the authors, Colciencias, Fraunhofer Center for Applied Research on Supply Chain Services SCS and University of Cauca, do not alter their adherence to PLOS ONE policies on sharing data and materials.

References

[1] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[2] D. B. Rubin, "Biometrika Trust Inference and Missing Data Author (s): Donald B .

Rubin Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : https://www.jstor.org/stable/2335739," *BiometrikaTrust*, vol. 63, no. 3, pp. 581–592, 1976.

[3] D. . Rubin, *Multiple imputation for nonresponse in surveys*, vol. 31, no. 1. New York: Wiley, 1990.

[4] N. J. Perkins *et al.*, "Principled Approaches to Missing Data in Epidemiologic Studies," *Am. J. Epidemiol.*, vol. 187, no. 3, pp. 568–575, 2018.

[5] M. E. Quinteros *et al.*, *Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile*. Elsevier Ltd, 2019.

[6] M. Cheliotis, C. Gkerekos, I. Lazakis, and G. Theotokatos, "A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems," *Ocean Eng.*, vol. 188, no. June, p. 106220, 2019.

[7] D. A. Williams, B. Nelsen, C. Berrett, G. P. Williams, and T. K. Moon, "A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data," *Environ. Model. Softw.*, vol. 102, pp. 172–184, 2018.

[8] Q. Lan, X. Xu, H. Ma, and G. Li, "Multivariable Data Imputation for the Analysis of Incomplete Credit Data," *Expert Syst. Appl.*, vol. 141, p. 112926, 2019.

[9] D. F. Young-Saver, J. Gornbein, S. Starkman, and J. L. Saver, "Handling of Missing Outcome Data in Acute Stroke Trials: Advantages of Multiple Imputation Using Baseline and Postbaseline Variables," *J. Stroke Cerebrovasc. Dis.*, vol. 27, no. 12, pp. 3662–3669, 2018.

[10] G. Delaporte, M. Cladière, and V. Camel, "Missing value imputation and data cleaning in untargeted food chemical safety assessment by LC-HRMS," *Chemom. Intell. Lab. Syst.*, vol. 188, no. February, pp. 54–62, 2019.

[11] P. Chittora *et al.*, "Prediction of Chronic Kidney Disease -A Machine Learning perspective," *IEEE Access*, vol. 9, 2021.

[12] L. Ali *et al.*, "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019.

[13] M. Raihan-Al-Masud and M. Rubaiyat Hossain Mondal, "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms," *PLoS One*, vol. 15, no. 2, pp. 1–21, 2020.

[14] P. McKnight, K. McKnight, S. Sidani, and A. Figueredo, *Missing data: A gentle introduction*. The Guildord Press, 2007.

[15] I. Guyon, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[16] S. Seaman, J. Galati, D. Jackson, and J. Carlin, "What is meant by 'missing at random'?," *Stat. Sci.*, vol. 28, no. 2, pp. 257–268, 2013.

[17] A. Aussem and S. Rodrigues de Morais, "A conservative feature subset selection algorithm with missing data," *Neurocomputing*, vol. 73, no. 4–6, pp. 585–590, 2010.

[18] G. Doquire and M. Verleysen, "Feature selection with missing data using mutual information estimators," *Neurocomputing*, vol. 90, pp. 3–11, 2012.

[19] W. Qian and W. Shu, "Mutual information criterion for feature selection from incomplete data," *Neurocomputing*, vol. 168, pp. 210–220, 2015.

[20] M. Mera-Gaona, R. Vargas-Canas, and D. M. Lopez, "Towards a Selection Mechanism of Relevant Features for Automatic Epileptic Seizures Detection.," *Stud. Health Technol. Inform.*, vol. 228, no. 4, pp. 722–6, 2016.

[21] M. Mera, D. M. Lopez, and R. Vargas-Canas, “Feature Selection in EEG Signals to Support Automatic Detection of Epileptiform Events,” University of Cauca, 2021.

[22] M. Mera-Gaona, D. M. Lopez, and R. Vargas-Canas, “Framework for the Ensemble of Feature Selection Methods,” 2021.

[23] K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” in *Information and Software Technology*, 2015, vol. 64, pp. 1–18.

[24] M. Tan and J. Schilimmer, “Breast cancer dataset.” Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia, 1988.

[25] P. W. Frey and D. J. Slate, “Letter Recognition Using Holland-Style Adaptive Classifiers,” *Mach. Learn.*, vol. 6, no. 2, pp. 161–182, 1991.

[26] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, “Heart Disease Databases.” 1988.

[27] M. Hopkins, E. Reeber, G. Forman, and J. Suermondt, “SAMP E-mail Database.” Hewlett-Packard Labs, 1999.

[28] D. Dheeru and E. Karra Taniskidou, “UCI Machine Learning Repository,” *University of California, Irvine, School of Information and Computer Sciences*. School of Information and Computer Science, Irvine, CA, 2017.

[29] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple Imputation by Chained Equations What is it and how does it work?,” *Int J Methods Psychiatr Res*, vol. 20, no. 1, pp. 40–49, 2012.

[30] J. S. Murray, “Multiple Imputation : A Review of Practical and Theoretical Findings,” *Stat. Sci.*, vol. 33, no. 2, pp. 142–159, 2018.

[31] S. van Buuren and K. Groothuis-oudshoorn, “MICE: Multivariate Imputation by Chained,” *JSS J. Stat. Softw.*, vol. 45, no. 3, 2011.

[32] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, and J. van Hoewyk, “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models Key Words: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations,” 2001.

[33] Y. He, M. B. Landrum, D. P. Harrington, and P. Catalano, “Multiple imputation in a large-scale complex survey: a practical guide *,” *Stat. Methods Med. Res.*, vol. 19, pp. 653–670, 2010.

[34] K. G. M. Moons, R. A. R. T. Donders, T. Stijnen, and F. E. Harrell, “Using the outcome for imputation of missing predictor values was preferred,” *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1092–1101, Oct. 2006.

[35] B. H. Menze *et al.*, “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” *BMC Bioinformatics*, vol. 10, no. August, 2009.

[36] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2825–2830, 2011.

[37] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[38] T. C. W. Landgrebe, P. Paclik, R. P. W. Duin, and A. P. Bradley, “Precision-Recall Operating Characteristic (P-ROC) curves in imprecise environments,” *Proc. - Int. Conf. Pattern Recognit.*, vol. 4, no. July 2014, pp. 123–127, 2006.

[39] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, “How many imputations are really needed? Some practical clarifications of multiple imputation theory,” *Prev. Sci.*, vol.

8, no. 3, pp. 206–213, Sep. 2007.

[40] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C. P. Varghese, and R. K. Agrawal, "Validation of QSAR Models -Strategies and Importance," *Int. J. Drug Des. Discov.*, vol. 2, no. 3, pp. 511–519, 2011.

[41] J. A. Torres Munguía, "Comparison of imputation methods for handling missing categorical data with univariate pattern," *Rev. Metod. Cuantitativos para la Econ. y la Empres.*, vol. 17, no. 1, pp. 101–120, 2014.

[42] A. Lotsi, L. Asiedu, and J. Katsekpor, "Comparison of Imputation Methods for Missing Values in Longitudinal Data Under Missing Completely at Random (mcar) mechanism," *African J. Appl. Stat.*, vol. 4, no. 1, pp. 241–258, 2017.

[43] P. Schmitt, J. Mandel, and M. Guedj, "A Comparison of Six Methods for Missing Data Imputation," *J. Biom. Biostat.*, vol. 06, no. 01, pp. 1–6, 2015.

[44] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones, "A Comparison of Imputation Techniques for Handling Missing Data," *West. J. Nurs. Res.*, vol. 24, no. 7, p. 815, 2002.

[45] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognit.*, vol. 41, no. 12, pp. 3692–3705, 2008.

[46] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," *Classif. Clust. Data Min. Appl.*, no. 1995, pp. 639–647, 2004.

[47] T. Orczyk and P. Porwik, "Influence of missing data imputation method on the classification accuracy of the medical data," *J. Med. Informatics Technol.*, vol. 22, pp. 111–116, 2013.

[48] D. Mundfrom and A. Whitcomb, "Imputing missing values: The effect on the accuracy of classification," *Mult. Linear Regres. Viewpoints*, vol. 25, pp. 13–19, 1998.