Universidad Juárez Autónoma de Tabasco

Tesis Doctoral

Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento

Que presenta

Elías Beltrán Naturi

Para obtener el grado de

Doctor en Ciencias de la Computación

Director

Dr. Francisco Diego Acosta Escalante

Dr. José Adán Hernández Nolasco

Línea de generación y aplicación del conocimiento:

Tecnologías emergentes en ingeniería de software

Institución sede:

Universidad Juárez Autónoma de Tabasco

Cunduacán, Tabasco, México **Marzo 2020**

Universidad Juárez Autónoma de Tabasco

Tesis Doctoral

Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento

Que presenta

Elías Beltrán Naturi

Para obtener el grado de

Doctor en Ciencias de la Computación

Comité Tutoral: **Dr. Francisco D. Acosta Escalante**
Dr. José Adán Hernández Nolasco

Jurado: **Dr. Óscar Alberto Chávez Bosquez**
Dr. Francisco Javier Álvarez Rodríguez
Dra. Betania Hernández Ocaña
Dr. José Adán Hernández Nolasco
Dr. Francisco Diego Acosta Escalante
Dr. José Hernández Torruco
Dr. Pablo Payró Campos

Cunduacán, Tabasco, México **Marzo 2020**

Oficio No. 402/2020/DACYTI/D
02 de marzo de 2020

C. Elías Beltrán Naturi
Matrícula 162H9001

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudio de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional "**Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento**", para presentar examen y obtener el Grado de Doctor en Ciencias de la Computación bajo la modalidad de Tesis.

Sin otro particular, aprovecho la oportunidad para saludarle.

Atentamente

UNIVERSIDAD JUAREZ AUTONOMA DE TABASCO

MTE. Óscar Alberto González González
Director

DIVISION ACADEMICA DE INFORMATICA Y SISTEMAS

C.c.p. MASI. Arturo Corona Ferreira.- Encargado del Despacho de la Coordinación de Posgrado.
Archivo.
Consecutivo.

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx
http://ujat.mx/dacyti

Oficio No. 436/2020/DACYTI/D
28 de febrero de 2020

Dr. Francisco Diego Acosta Escalante
Profesor-Investigador
Presente

De acuerdo al artículo 46 fracción III del Reglamento General de Estudios de Posgrado Vigente, de la Universidad Juárez Autónoma de Tabasco, me permito informarle a Usted, que ha sido asignado director del trabajo de tesis titulado **"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento"**, a realizar por el **C. Elías Beltrán Naturi**, para obtener el grado de Doctor en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

UNIVERSIDAD JUAREZ AUTONOMA DE TABASCO

MTE. Óscar Alberto González Gonzalez
Director

DIVISION ACADEMICA DE INFORMATICA Y SISTEMAS

C.c.p. MASI. Arturo Corona Ferreira.-Encargado del Despacho de la Coordinación de Posgrado.
Archivo.
Consecutivo.

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx
http://ujat.mx/dacyti

Oficio No. 392/2020/DACYTI/D
28 de febrero de 2020

Dr. José Adán Hernández Nolasco
Profesor-Investigador
Presente.

De acuerdo al artículo 46 fracción III del Reglamento General de Estudios de Posgrado Vigente, de la Universidad Juárez Autónoma de Tabasco, me permito informarle a Usted, que ha sido asignado Director del Trabajo de Tesis titulado **"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington ataxias hereditarias basados en sensores de movimiento",** a realizar con el **C. Elías Beltrán Naturi**, para obtener el grado de Doctor en Ciencia de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

MTE. Oscar Alberto González González
Director

C.c.p. MASI. Arturo Corona Ferreira.- Encargado del Despacho de la Coordinación de Posgrado.
Archivo.
Consecutivo.

Cunduacán, Tabasco. 28 de febrero de 2020.

MTE. Óscar Alberto González González.
Director de la DACyTI.
Universidad Juárez Autónoma de Tabasco.
Presente

Por medio de la presente nos permitimos comunicarle que después de haber realizado las asesorías correspondientes al Proyecto de Titulación: **"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento",** elaborada por el C. **Elías Beltrán Naturi**, del Doctorado Interinstitucional en Ciencias de la Computación, consideramos que la ha concluido satisfactoriamente, por lo que puede continuar con los trámites para la obtención del grado.

Sin otro particular, aprovecho la oportunidad para saludarle y desearle el mayor de los éxitos en este importante proyecto

Atentamente

Dr. Francisco Diego Acosta Escalante
Director de Tesis

Dr. José Adán Hernández Nolasco
Co-director de Tesis

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690. Cunduacán, Tabasco, México.
E-mail: direccion.dais@ujat.mx
Teléfonos: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870

Cunduacán, Tabasco. 28 de febrero de 2020.

En la Universidad Juárez Autónoma de Tabasco, de acuerdo con el Reglamento de Estudios de Posgrado vigente, se revisó el trabajo de investigación titulado **"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento"**, realizado por el C. **Elías Beltrán Naturi estudiante**, estudiante del Doctorado Interinstitucional de Ciencias de la Computación para obtener el Grado de Doctor en Ciencias de la Computación bajo la modalidad de tesis.

Los integrantes del jurado, después de revisar el trabajo escrito, y en virtud de que se han atendido satisfactoriamente las observaciones y recomendaciones, otorgamos nuestra aprobación para la impresión de la tesis de tal forma que se pueda continuar con los trámites correspondientes para la obtención del grado.

Dr. Oscar Chávez Bosquez
Profesor-Investigador
Universidad Juárez Autónoma de Tabasco

Dra. Betania Hernández Ocaña
Profesor-Investigador
Universidad Juárez Autónoma de Tabasco

Dr. Pablo Payró Campos
Profesor-Investigador
Universidad Autónoma de Ciudad Juárez

Dr. José Hernández Torruco
Profesor-Investigador
Universidad Autónoma de Aguascalientes

Dr. Francisco Javier Álvarez Rodríguez
Profesor Externo
Universidad Autónoma de Aguascalientes

Carretera
acán, Tabasco, México.
E-mail: direccion.dais@ujat.mx
Teléfonos: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870

Cunduacán, Tabasco, a 28 de febrero de 2020.

Asunto: Cesión de Derechos

A QUIEN CORRESPONDA:

Los abajo firmantes, declaramos que el trabajo de tesis doctoral titulado, **"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento"**, es de nuestra autoría intelectual y por lo tanto cedemos los derechos de comunicación pública, reproducción, distribución, difusión en general y puesta a disposición electrónica de la citada tesis doctoral, de forma gratuita y no exclusiva, a la Universidad Juárez Autónoma de Tabasco, a la cual relevamos de cualquier sanción y asumimos responder a cualquier reclamo de derechos de autor ante las autoridades competentes.

Atentamente

Autores:

Nombre	Dirección	Firma
C. Elías Beltrán Naturi.	Av. Alcatraces 229B, Fraccionamiento Real del Bosque, Tuxtla Gutiérrez, Chiapas, CP. 29055.	
Dr. Francisco Diego Acosta Escalante.	Universidad Juárez Autónoma de Tabasco, División académica de Ciencias y Tecnologías de la Información; Carretera Cunduacán-Jalpa KM. 1 Col. La Esmeralda CP. 86690, Cunduacán, Tabasco, CP. 86690.	
Dr. José Adán Hernández Nolasco.	Universidad Juárez Autónoma de Tabasco, División académica de Ciencias y Tecnologías de la Información; Carretera Cunduacán-Jalpa KM. 1 Col. La Esmeralda CP. 86690, Cunduacán, Tabasco, CP. 86690.	

c.c.p. **MTE. Óscar Alberto González González.**- Director de la DACyTI
Mtro. Arturo Corona Ferreira. Encargado de Despacho de Posgrado DACyTI-UJAT
Directores de Tesis
Estudiante

El que suscribe, autoriza por medio del presente escrito a la Universidad Juárez Autónoma de Tabasco para que utilice tanto física como digitalmente la Tesis de Doctorado **"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento",** de la cual soy autor y titular de los Derechos de Autor.

La finalidad del uso por parte de la Universidad Juárez Autónoma de Tabasco de la tesis antes mencionada será única y exclusivamente para la difusión, educación, sin fines de lucro; autorización que hace de manera enunciativa más no limitativa para subirla a la Red Abierta de Bibliotecas Digitales (RABID) y cualquier otra Red Académica con las que la Universidad Tenga relación institucional.

Po lo antes mencionado, libero a la universidad Juárez Autónoma de Tabasco de cualquier reclamación legal que pudiera ejercer respecto al uso y manipulación de la Tesis mencionada y para los fines estipulados en este documento.

Se firma la presente autorización en la Ciudad de Cunduacán Tabasco a los 28 días del mes de febrero del año 2020.

Autorizo

C. Elías Beltrán Naturi

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690. Cunduacán, Tabasco, México.
E-mail: direccion.dais@ujat.mx
Teléfonos: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870

A mi madre y mis hermanos por su valioso apoyo y compresión en todo el proceso de mi formación …

Agradecimientos

Este trabajo fue financiado por Consejo de Ciencia y Tecnología (CONACYT). Gracias por todo el apoyo económico y técnico para el desarrollo de esta investigación, se extiende el más sincero agradecimiento.

Quiero externar mi agradecimiento al programa de posgrado Doctorado Interinstitucional en Ciencias de la Computación(PNPC) de la Universidad Juárez Autónoma de Tabasco (UJAT) por permitirme ingresar como estudiante en dicho programa educativo.

A mi director el Dr. Francisco Acosta Escalante y mi codirector Dr. José Adán Hernández Nolasco por aceptarme como tesista, por su entrega, tiempo y dedicación al guiarme en el camino de la investigación. Los doctores que formaron parte de revisión del presente trabajo mi gratitud por los consejos y sugerencias para el mejoramiento de este trabajo.

Agradezco al Instituto Nacional de Neurocirugía y neurología “Manuel Velasco Suarez” (INN-MVS) por su apoyo en la realización de las mediciones para este proyecto. Estoy muy agradecidos al personal y a los empleados de INNN-MVS por su colaboración, a todos los pacientes que participaron y a los familiares que los acompañaron, y en especial a la Dr. Marie Catherine Boll por su apoyo y asesoría en el desarrollo de todo el proyecto de investigación.

Abstract

"Gait pattern recognition in patients with Huntington's disease and Hereditary Ataxias based on movement sensors"

by Beltrán-Naturi ELÍAS

Gait pattern recognition techniques have been studied to discriminate between healthy and sick patients according to their common features. The classification algorithms have been able to recognize patients with Huntington's disease (HD) with 88.2 % accuracy, while patients with Hereditary Ataxias (HA) have only been correctly classified at 78.78 %. These results have been obtained in recent work with various automatic learning techniques and gait features extracted from various data sources including movement sensors. This research work focused on proposing a method to obtain greater accuracy in the classification of patients with these diseases than that obtained from published works, with the particularity of focusing on implementing a minimum amount of gait characteristics based on data collected with smartphone movement sensors. The data was collected using the smartphone movement sensors on patients with these diseases and healthy people as controls (HC). A method of data preparation was implemented and a segmentation algorithm of the gait cycle was proposed. The gait features were obtained from a data segment equivalent to 10 strides, from which a total of 56 features were extracted. Different attribute selection algorithms showed that 11 features of Huntington's disease and 15 of Hereditary Ataxias were sufficient to achieve excellent performance. The algorithms *LogitBoost & RandomForest* allowed to reach an accuracy of 92.85 % when classifying HD vs HC; while searching for a small number of characteristics, *KNN* achieved 96 % accuracy and *MLP* obtained 100 % accuracy, using 2 and 3 features respectively, using data from HA patients.

Resumen

"Reconocimiento de patrones de la marcha en pacientes con las enfermedades de Huntington y Ataxias Hereditarias basado en sensores de movimiento"

por Beltrán-Naturi ELÍAS

Las técnicas del reconocimiento de patrones de marcha han sido objeto de estudio para discriminar a los pacientes enfermos de sanos según sus características comunes. Los algoritmos clasificación han sido capaces de reconocer a los pacientes con la enfermedad de Huntington (HD) con un 88.2 % de exactitud, mientras que los enfermos con Ataxias Hereditarias (HA) sólo han sido clasificados correctamente en un 78.78 %. Estos resultados se han obtenido en trabajos recientes con diversas técnicas de aprendizaje automático y características de la marcha extraída de diversas fuentes de datos incluyendo sensores de movimiento. Este trabajo de investigación se enfocó en proponer un método para obtener una mayor exactitud en la clasificación de pacientes con estas enfermedades que la obtenida de los trabajos publicados, con la particularidad de enfocarnos en implementar una cantidad mínima de características de la marcha basada en datos recolectados con sensores de movimiento de teléfonos inteligentes. Los datos fueron recolectados usando los sensores de movimiento de *teléfonos inteligentes* en pacientes con estas enfermedades y personas sanas como controles (HC). Se implementó un método de preparación de los datos y se propuso un algoritmo de segmentación del ciclo de la marcha. Las características de la marcha fueran obtenidas de un segmento de datos equivalente a 10 zancadas, de los cuales se extrajeron un total de 56 características. Diversos algoritmos de selección de atributos mostraron que 11 características de enfermos Huntington y 15 de Ataxias Hereditarias eran suficientes para alcanzar un excelente rendimiento. Los algoritmos *LogitBoost & RandomForest* permitieron obtener una exactitud del 92.85 % al clasificar HD vs HC; mientras que, al buscar un reducido número de características, KNN alcanzo una precisión de 96 % y MLP obtuvo exactitud del 100 %, usando 2 y 3 características respectivamente, usando datos de pacientes con HA.

Publicaciones

Autores	Título	Foro	Año
Beltran-Naturi, E.; Acosta-Escalante, F. D. & Boll C.	Detección y seguimiento de enfermedades de Huntington y Ataxia Espinocerebelosa, mediante la caracterización de la marcha con sensores de movimiento.	Avances en las Tecnologías de la Información (2016th ed., pp. 610–619). Toluca, MX: alfa-omega.	2016
Acosta-Escalante F. D.; Beltran-Naturi, E.; Boll M. C.; Hernandez-Nolasco J. A. & Pancardo P.	Meta-classifiers in Huntington's disease patients classification, using iPhone's movement sensors placed at the ankles.	IEEE Access, Volume 6, Issue 1 (pp. 30942–30957)	2018
Beltran-Naturi, E.; Acosta-Escalante F. D.; Boll M. C. & Hernandez-Nolasco J. A.	Reconocimiento de la marcha en pacientes con ataxias hereditarias basados en sensores de movimiento del iPhone.	Libro Digital "Colección de avances en Tecnologías de la Información", Volúmen 5, en el Congreso Nacional e Internacional de Informática y Sistemas 2019, Villahermosa, MX.	2019
Beltran-Naturi, E.; Acosta-Escalante F. D.; Boll M. C. & Hernandez-Nolasco J. A.	Un enfoque del reconocimiento de patrones de marcha en pacientes con Huntington y Ataxias Hereditarias usando datos de acelerómetros del iPhone.	XXXII Congreso Nacional y XXVIII Congreso Internacional de Informática y Computación (ANIEI 2019), Puebla, MX.	2019

Índice general

Índice de figuras

Índice de Tablas

Abreviaturas

TPR	**T**rue **P**ositive **R**ate o la Tasas de Verdaderos Postivos.
FPR	**F**alse **P**ositive **R**ate o la tasa de Falsos Positivos.
HA	**H**ereditary **A**taxia o Ataxia Hereditaria.
HC	**H**ealthy **C**ontrol o sujetos de control.
HD	**H**untington **D**isease o Enfermos con Huntington.
PD	**P**arkinson **D**isease (Enfermedad de Parkinson).
ALS	**A**myotrophic **L**ateral **S**clerosis (Esclerosis Lateral Amiotrófica).
ND	**N**eurodegenerative **D**iseases (Enfermedades Neurodegenerativas).
ARA	**A**utosomal **R**ecessive **A**taxia (Ataxia Autosómica Recesiva).
ADA	**A**utosomal **D**ominant **A**taxia (Ataxia Autosómica Dominante).
SCA	**S**pino**C**erebellar **A**taxias (Ataxias Espinocerebelares).
EA	**E**pisodic **A**taxias (Ataxias Episódicas).
ExtraTrees	**Ext**remely **Ra**ndomized **Trees** (Árboles Extremadamente Aleatorios).
CART	**C**lassification **A**nd **R**egression **T**rees (Árboles de clasificación y regresión).
LMT	**L**ogistic **M**odel **T**rees (Árboles de Regresión Logística).
KNN	**K**-**N**earest **N**eighbours (Los K-Vecinos más cercanos).
MLP	**M**ulti **L**ayer **P**erceptron (Perceptron Multicapa).
SVM	**S**upport **V**ector **M**achine (Máquinas de Soporte Vectorial).

Capítulo 1

Introducción

Las enfermedades neurodegenerativas son trastornos neurológicos degenerativos caracterizados por un desgaste en el desempeño motor de los pacientes, provocando movimientos corporales anormales y alterando los patrones de la marcha debido a una falta de coordinación del control motor del movimiento.

La identificación de estas enfermedades a partir del análisis de alteraciones en la marcha con métodos computacionales ha sido objeto de diferentes estudios. En particular en el estudio de la diferenciación de grupos de personas sanas y enfermos con patologías motoras como las enfermedades de Huntington (HD), Parkinson (PD) y Ataxias hereditarias (HA)[1] se han utilizado métodos basados en algoritmos probabilísticos, máquinas de soportes vectorial y minería de datos, capaces de reconocer las alteraciones en los patrones de la marcha que causan estas patologías [1]-[6].

Para el reconocimiento de estas alteraciones se utilizan sensores de movimiento que registran grandes volúmenes de datos. Aplicaciones como *Google Fit*, *Apple Health* y *Samsung Health* recolectan los datos de estos sensores a través de dispositivos como teléfonos o pulseras inteligentes [7]-[9]. Sin embargo, la identificación de características de la marcha que permitan encontrar las diferencias sutiles de las afectaciones que producen las enfermedades neurodegenerativas en la forma de caminar es un tema que actualmente cobra mucho interés en la comunidad científica.

Esta tesis tiene como objetivo la propuesta de métodos de inteligencia artificial y el estudio de algoritmos para el reconocimiento de patrones y la implementación de técnicas de minería de datos para la identificación de los grupos patológicos HD y HA basado en información de la marcha recolectada con sensores de movimiento embebidos dentro de dispositivos de uso cotidiano como son los teléfonos inteligentes (SmartPhones).

[1] En esta tesis se usan las siglas en inglés en las abreviaturas debido a su amplia difusión.

Los resultados de esta investigación buscan sentar las bases para desarrollar herramientas tecnológicas que apoyen a los especialistas en el diagnóstico y seguimiento de la progresión de estas enfermedades a largo plazo.

El presente trabajo se encuentra estructurado de la siguiente manera: el capítulo 1 expone los antecedentes, la problemática y los objetivos que se buscan alcanzar; el capítulo 2 presenta el marco teórico con la finalidad de proporcionar un marco de referencia para la comprensión de los aspectos teóricos que requiere esta investigación; el capítulo 3 hace una revisión de los trabajos relacionados para comprender hasta donde los temas propuestos han sido investigados y como se diferencia esta investigación de los trabajos recientes; el capítulo 4 muestra la propuesta metodológica, herramientas y sujetos involucrados en la investigación; el capítulo 5 expone los resultados obtenidos al aplicar nuestra metodología y los hallazgos relevantes de la investigación; finalmente, el capítulo 6 exhibe las conclusiones relacionadas a los resultados encontrados en la investigación y los trabajos futuros.

1.1. Antecedentes

Las enfermedades neuro-degenerativas son progresivas e incurables; los tratamientos médicos permiten disminuir las síntomas, pero hasta ahora, no son útiles para detener su progreso [10], [11]. Los pacientes que padecen las enfermedades de Huntington y Ataxias hereditarias presentan alteraciones psiquiátricas, psicológicas y motoras. Estos problemas del sistema motor se hacen visibles en actividades de la vida diaria al perder progresivamente la precisión y estabilidad en actividades básicas como el caminar; las alteraciones en los patrones de la marcha se vuelven evidentes conforme al avance de la enfermedad [12]-[18].

Los estudios para establecer un grado de avance en la enfermedad mediante la identificación de alteraciones significativas en el caminar incluyen estudios retrospectivos y longitudinales que permiten establecer un grado del deterioro de la marcha mediante la observación clínica y el uso de escalas clínicas como UHDRS (Unified Huntington's Disease Rating Scale), SARA (Application of a Scale for the Assessment and Rating of Ataxia), FARS ("Friedreich's ataxia impact scale) y TCF (Escala de la Capacidad Funcional) [19]-[21].

Los estudios recientes buscan fortalecer las medidas subjetivas de la observación clínica implementando diversas tecnologías para recolectar información del caminar de los enfermos para determinar el deterioro de la marcha; lo que ha permitido establecer diferencias entre diversos grupos patológicos basados en parámetros espacio-temporales y características de la marcha extraídos de información de sensores de presión y movimiento [22]-[26].

Las tecnologías de sensores de movimiento son capaces de recolectar información de aceleraciones pequeñas en el desplazamiento de las personas, la cual procesada con algoritmos de reconocimiento de patrones logra identificar diferencias sutiles en los movimientos. Los sensores de movimiento se han vuelto útiles para recopilar datos de la marcha por su bajo costo, facilidad de uso y poco intrusivo [27]. Sin embargo, son pocos los estudios que reportan el uso de estos dispositivos en pacientes con enfermedades neurodegenerativas como Huntington (HD) y Ataxias Hereditarias (HA); además no se han reportado estudios para determinar la cantidad necesaria de sensores para caracterizar correctamente la marcha de estos enfermos, el tipo de sensores que recolectan datos de los movimientos eficientemente, los ejes que permitan identificar patrones anormales característicos de estas enfermedades y la ubicación de los dispositivos en cuerpo para obtener datos totalmente confiables.

Los sensores de movimiento, por su reducido tamaño puede usarse para registrar información de pacientes mientras caminan, que analizada con algoritmos de reconocimiento de patrones podría determinar afectaciones en los patrones de la marcha de individuos con enfermedades neurodegenerativas; lo anterior, como herramienta de apoyo al diagnóstico y seguimiento de la progresión de las alteraciones motoras en este tipo de enfermedades debido son portables y poco intrusivos.

1.2. Relevancia y justificación de la investigación

HD y HA son enfermedades neurodegenerativas y hereditarias por lo que los descendientes de enfermos tiene una alta posibilidad de presentar la enfermedad; sin embargo, hasta el momento es imposible saber cuándo la enfermedad aparecerá (en la mayoría de los casos en la vejez). El aumento de la expectativa de vida en nuestro país y el consecuente incremento de la población adulta, favorecerá la aparición de un

mayor número de pacientes con enfermedades neurodegenerativas [28]. Por otra parte, investigaciones recientes han revelado que en las zonas montañosas del estado de Veracruz los casos de pacientes con Ataxias Hereditarias se han cuadriplicado en los últimos años [29], [30]. Los especialistas sugieren que existe una fase inicial que ocurre meses o años antes de la manifestación completa de la enfermedad en donde las alteraciones en los movimientos no son perceptibles completamente para el ojo humano. Identificar esta fase inicial y realizar el tratamiento adecuado pueden ayudar a aplazar la manifestación completa de estas patologías [15], [31], [32].

Las tecnologías de sensores de movimientos permiten recolectar información relacionada con la forma de caminar de las personas basado en la aceleración (acelerométros) y la rotación (giroscopios) del cuerpo [27], [33]. Diversos estudios han comprobado la capacidad de estos sensores para obtener información de la marcha en pacientes con estas enfermedades de manera confiable [22], [23], [25], [34]-[37]. Estos dispositivos han sido objeto de investigación por ser de bajo costo, portables, livianos y de fácil uso comparados con otras herramientas para el mismo propósito, por lo que resulta ideal para obtener datos de los movimientos de personas con alteraciones en los patrones de la marcha.

Algunos trabajos recientes publicados se enfocan en el reconocimiento de patrones de la marcha con datos de sensores de presión y muy pocos que usan sensores de movimiento. El reconocimiento de patrones de la marcha en enfermos con HD y HA no supera el 90 % de precisión para ninguna de las dos patologías mencionadas [3], [38]; y no se han reportado investigaciones que involucren ambas patologías. Una de las motivaciones de este trabajo es la búsqueda de métodos que puedan incidir en la mejora de los resultados hasta ahora publicados. El desafío que se debe resolver es el hecho de que ambas enfermedades afectan a la marcha de maneras muy similares.

1.3. Línea de Investigación

Este proyecto de investigación está centrado en el uso de tecnologías de la vida diaria para la recopilación de datos e información con el propósito de ser analizados usando métodos y tecnologías computacionales para resolver un problema que afecta una población o un grupo de personas.

Los sensores de movimiento son dispositivos que permiten recopilar datos en diferentes ambientes y se pueden encontrar en dispositivos de uso común como los teléfonos inteligentes. Por lo tanto, este trabajo se encuentra alineado con la línea de investigación *Temas emergentes de Ingeniería de Software*; la cual busca encontrar nuevos ámbitos de aplicación de las tecnologías existentes que aún no han sido explotadas, el área de *Inteligencia Ambiental* explora el uso de estos dispositivos para incrementar su rentabilidad, uso y productividad.

1.4. Alcances de la investigación

La presente investigación busca sentar las bases para el desarrollo de tecnologías que permitan contribuir al monitoreo y seguimiento de la progresión de la enfermedad de pacientes con Huntington y Ataxias hereditarias. Sin embargo, el campo de estudio es muy extenso, por lo que esta investigación se centra en:

- El Diseño de un método para procesar la información recolectada con sensores de movimiento en los tobillos de los pacientes, usando la menor cantidad de sensores de movimiento.
- El Diseño de un método automático o semi-automático para la extracción de características de la marcha representativas de cada patología.
- La identificación de algoritmos que permitan la selección de características de la marcha y el reconocimiento de patrones de la marcha de cada patología.
- La selección de aquellas características que permitan representar a ambos grupos patológicos, así como también, los elementos de los sensores de movimiento que permiten obtener información relevante a cada enfermedad.
- La obtención de una exactitud en los resultados de clasificación superior a los reportados en investigaciones relacionadas.

En este trabajo se realizó con un número reducido de pacientes debido a su poca disponibilidad y acceso. Las pruebas de validación de resultados se realizan de manera analítica, técnica común en este tipo de estudios.

1.5. Problema de la investigación

Las tecnologías de sensores permiten recolectar información en diferentes ámbitos e investigaciones; los sensores de presión y movimiento han sido utilizados para recolectar información de la marcha con diversos fines. Los volúmenes de información obtenidos con estas tecnologías han sido analizadas y procesadas con técnicas de reconocimiento de patrones, tales como algoritmos de minería de datos y aprendizaje automático.

Esta investigación está motivada por diversos factores encontrados en el análisis del estado del arte, entre los que podemos mencionar los siguientes:

- Es necesario el desarrollo de tecnologías computacionales que vayan acorde a los avances tecnológicos en el análisis y cuantificación de información de la marcha; lo anterior debido a que los trabajos relacionados se han basado en pruebas estandarizados con la observación clínica de los especialistas; otros trabajos en reducido número, fueron desarrollados en laboratorio de la marcha usando videometría o esteras de la marcha con sensores de presión, los cuales son de alto costo y requieren grandes infraestructuras para implementarse.
- Las técnicas computacionales de reconocimiento de patrones permiten encontrar patrones que diferencien entre diversos grupos. Los algoritmos de clasificación son útiles para diferenciar información de grupos muy similares; como es el caso de las enfermedades de HD y HA, las cuales comparten ciertas similitudes patológicas en la afectación de la marcha; lo anterior constituyen una complejidad al momento de reconocer las patologías basados en las características de la marcha.
- Los métodos computacionales permiten procesar grandes volúmenes de información; lo que resulta útil para analizar la información de sensores en diversas partes del cuerpo, con la finalidad de identificar los sensores que mejor permiten caracterizar la marcha de pacientes con las enfermedades mencionadas; debido a que no se reportan trabajos que establezcan una configuración de sensores en el cuerpo para adquirir información de los movimientos relacionados a las alteraciones de la marcha en pacientes con enfermedades neuro-degenerativas.
- Las técnicas de minería de datos y aprendizaje automático para diferenciar estos grupos patológicos basados en las características de la marcha aun no es completa,

por lo que se necesitan desarrollar nuevos métodos computacionales para el procesamiento, extracción de características y reconocimiento de patrones basados en información de la marcha recolectada con sensores portátiles de movimiento.

- Se necesitan desarrollar nuevas tecnologías computacionales que permitan el seguimiento y monitoreo de la progresión de las alteraciones en la marcha a largo plazo para estas enfermedades. Por lo tanto, se deben desarrollar investigaciones orientadas a contribuir en este ámbito de aplicación.

1.6. Pregunta de investigación

Los elementos encontrados en el planteamiento del problema nos llevan adentrarnos en las siguientes preguntas de investigación:

- ¿Cuál es la configuración mínima de sensores de movimiento que permiten capturar las alteraciones de la marcha en pacientes con HD y HA?
- ¿Cuáles son los métodos y técnicas para el tratamiento de la información obtenida de los dispositivos (sensores de movimiento)
- ¿Cuáles son los métodos y técnicas para extraer la información que caracterice la marcha de los pacientes según su patología?
- ¿Cuál sería la configuración mínima de características de la marcha y del número de sensores de movimiento que permitan a más de un algoritmo la discriminación de enfermos y sanos con alto porcentaje de exactitud?
- ¿ Cuáles son los algoritmos que permite clasificar correctamente las características de la marcha de pacientes según su patología?

1.7. Objetivo general

La investigación gira entorno al siguiente objetivo principal:

Desarrollar un método para la extracción y selección de características de la marcha de enfermos con Huntington y Ataxias hereditarias, que permita diferenciar eficientemente la categoría a la que pertenecen cada grupo basado en datos de sensores de movimiento de teléfonos inteligentes.

1.8. Objetivos específicos

Los objetivos secundarios que derivan de nuestro objetivo principal y que nos permiten el cumplimiento de lo planteado son los siguientes:

- Determinar la cantidad mínima de sensores y su ubicación para obtener la información representativa de la marcha alterada en los enfermos.
- Plantear un conjunto de técnicas para el procesamiento, segmentación y extracción de características de la marcha.
- Evaluar algoritmos que permitan seleccionar las características que mejor describen la marcha en los enfermos.
- Identificar los algoritmos que permitan discriminar las características entre los diversos grupos de forma eficiente.

1.9. Hipótesis de investigación

Para el desarrollo de esta investigación se plantea la siguiente hipótesis:

"Las técnicas de clasificación permiten diferenciar enfermos con Huntington y Ataxias hereditarias con una alta precisión (entre el 90 % y el 100 %) usando un número reducido de características de la marcha extraídas de datos obtenidos con un número reducido de sensores de movimiento de teléfonos inteligentes.

1.10. Metodología de la investigación

El objetivo de la investigación científica es generar nuevos conocimientos que nos permitan conocer la realidad y describir un conjunto de fenómenos que ocurren en nuestro

alrededor. Para lograr lo anterior, se usan diversos métodos con el fin de un nuevo conocimiento científico; los métodos se apoyan de diversas técnicas para lograr o alcanzar los objetivos. Esta investigación involucra métodos de investigación implementados en la informática según *Barchini, G.* [39]; se aplica un conjunto de investigación de tipo empírica usando métodos cuantitativos experimentales.

Los métodos de investigación cuantitativos según Sampieri R.[40] involucran los pasos descritos en la figura 1.1. El investigador *plantea un problema* de estudio delimitado y concreto sobre el fenómeno que lo llevan a plantear *preguntas de investigación* sobre cuestiones específicas. Con el problema planteado el investigador busca *investigaciones relacionadas al problema* y construye una guía de estudios denominado *marco teórico*; de lo anterior deriva una o varias *hipótesis* (objeto de estudio) y las somete a pruebas con diseños de investigación apropiados para corroborarlas o descartarlas. Aunque las investigación difieren en muchos aspectos por su naturaleza, las investigaciones cualitativas involucran al menos los siguientes pasos: definición y selección de la muestra de los sujetos de estudios, recolección de los datos usando diversos instrumentos, análisis de los datos usando un conjunto de técnicas propias a los métodos de cada área del conocimiento con el que se trabaja, y finalmente el análisis de los resultados encontrados para interpretar los hallazgos para aceptar o descartar la hipótesis. La última fase involucra intrínsecamente la elaboración de los reportes necesarios para el cierre de la investigación.

El presente trabajo se basa en dos herramientas principales:

- La recolección de información de pacientes con Sensores de Movimiento para capturar datos relacionado a los patrones de la marcha de enfermos con HD y HA.
- El análisis de los datos e interpretación de los resultados usando Algoritmos computacionales (clasificadores) que basados en información de la marcha de cada pacientes puedan reconocer las características representativas de cada grupo (HD o HA).

El desarrollo del proceso de la investigación se realizó acorde a los pasos definidos en la figura 1.1, de la siguiente manera:

- Identificación de la problemática relacionado a los problemas de la marcha de los pacientes con las enfermedades de HD y HA, y su delimitación del problema como se muestra en este capítulo.
- Revisión de conceptos más significativos relacionados al tema de investigación.

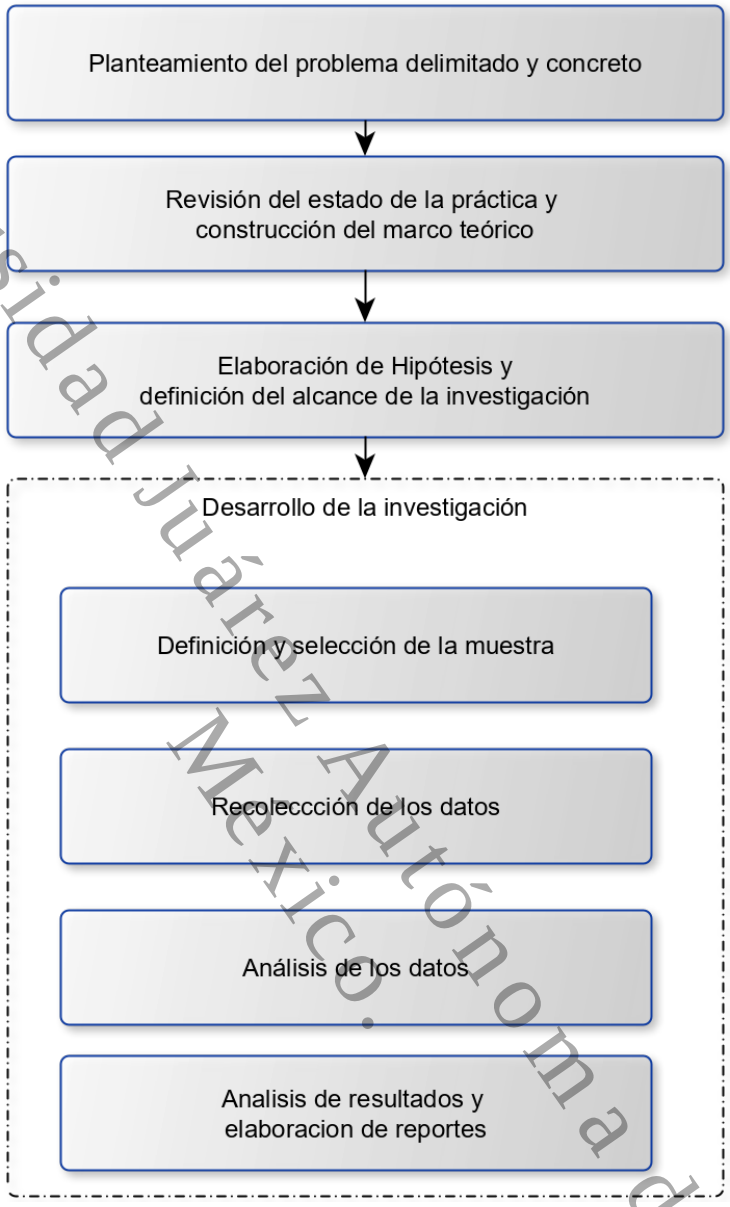


FIGURA 1.1: Metodología de la investigación adoptada.

- Revisión del estado del arte del uso de los sensores de movimiento y el reconocimiento de patrones de la marcha con pacientes con enfermedades neurodegenerativas.
- Formulación de la hipótesis y definición de los alcances de la investigación.
- Adquisición, organización y limpieza de los datos de la marcha de pacientes con HD y HA.

- Generación de un método para el pre-procesamiento de los datos extraídos de los sensores de movimiento.
- Generación de un método para la segmentación de la marcha, con la finalidad de extraer la información útil de los datos recolectados por los sensores.
- Identificación de métodos y técnicas para cuantificar y extracción de características de los patrones de la marcha según la patología de los pacientes con HD y HA.
- Identificación de los algoritmos que permitan seleccionar las características de la marcha que mejor representen a cada grupo con el menor número de sensores de movimiento.
- Implementar algoritmos que permitan diferenciar los patrones de marcha de ambos grupos, elaborar ensayos de prueba y seleccionar los que muestran mejor rendimiento.

1.11. Resultados

Los resultados obtenidos son:

- Un marco de trabajo que permita reconocer los patrones de la marcha de pacientes con las enfermedades de HD y HA.
- Un método para el tratamiento de datos de sensores de movimiento que permita identificar con precisión el final de cada ciclo de la marcha.
- Un método para reducir el número de características de la marcha y sensores de movimiento que permitan la diferenciación de la marcha atáxica de la marcha normal con una alta precisión.
- Conclusión de la investigación en los tiempos establecidos. Dar a conocer los resultados de este trabajo en publicaciones científicas de prestigio y la obtención del grado académico.

Capítulo 2

Marco teórico

En este capítulo se introducen los conceptos básicos que sustentan esta investigación como: la marcha y su descripción, las enfermedades de Huntington, las Ataxias Hereditarias, así como las explicaciones básicas de los algoritmos que fueron usados en el desarrollo de los experimentos.

2.1. La marcha humana

La marcha es definida por *Daza J.* en [41] *como el paso bípedo que utiliza la raza humana para desplazarse de un lugar a otro empleando un mínimo consumo energético y con mínimo esfuerzo*, se caracteriza por el contacto permanente del individuo con el suelo, con uno o ambos pies, requiere la integración del sistema musculo-esquelético y los reflejos posturales. Aunque existen pequeñas diferencias en la forma de la marcha de un individuo, estas diferencias caen dentro de pequeños límites, lo que permite identificar a una persona por su forma de caminar.

La descomposición de la marcha en el denominado *ciclo de la marcha (GC)* permite a los especialistas identificar movimientos normales o patológicos (Fig. 2.1). El ciclo de marcha comienza cuando el pie contacta con el suelo y termina con el siguiente contacto con el suelo del mismo pie. Los problemas de la marcha provocan cambios en la velocidad y estabilidad en el desplazamiento lo que se refleja en un aumento del relativo del tiempo de desplazamiento la velocidad y un mayor consumo energético.

El objetivo del estudio de la marcha es determinar la presencia o ausencia de alteraciones motoras. La marcha se describe en [43] por parámetros *espacio-temporales*, tales como:

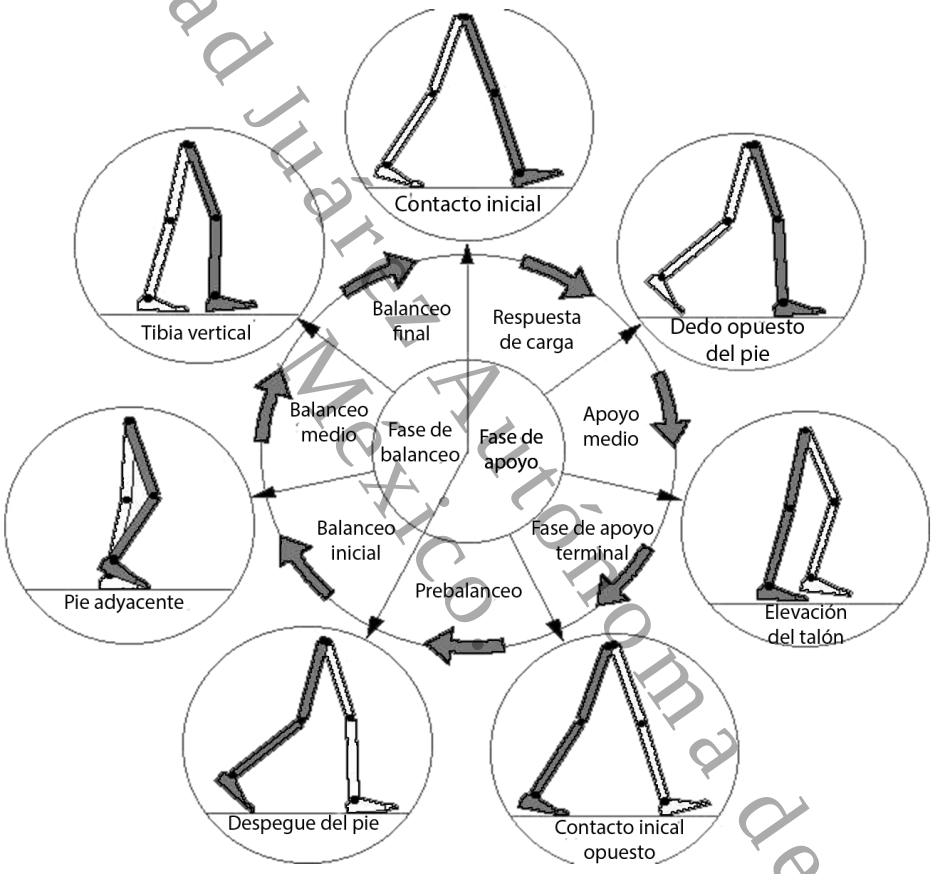


FIGURA 2.1: Fases y descomposición del ciclo de la marcha [42].

- **Ciclo de marcha**: recorrido entre dos apoyos sucesivos de un mismo talón.
- **Longitud de zancada**: secuencia de acontecimientos que tiene lugar entre dos choques de talón consecutivos del mismo pie. La distancia media entre dos apoyos consecutivos del mismo pie se denomina longitud de zancada y es, en definitiva, la suma de las longitudes del paso izquierdo y del derecho.
- **Velocidad de marcha**: distancia que recorre el cuerpo hacia delante en la unidad de tiempo (por ejemplo, 1.5 m/s).
- **Cadencia de marcha**: ciclos o pasos por unidad de tiempo (120 pasos/min o 1 ciclo/s).
- **Longitud de ciclo**: distancia entre dos choques consecutivos de talón de un mismo pie.
- **Longitud de paso**: distancia entre ambos pies cuando contactan con el suelo.
- **Amplitud de paso**: distancia entre los centros de las huellas plantares.
- **Ángulo de la marcha**: formado por el eje longitudinal del pie y la línea media de la progresión de la marcha.
- Cadencia: corresponde al promedio de pasos por minuto que un individuo es capaz de realizar.

Estos parámetros varían entre sujetos y en el mismo sujeto; sin embargo, resultan ser representativos de una persona cuando las condiciones y los factores que afectan la marcha se mantienen constantes [44]. Evaluar el movimiento corporal en el desplazamiento permiten identificar deficiencias motoras que inciden en la marcha que limitan las actividades de la vida diaria.

En el GC se puede analizar tres tipos de variables [45]:

1. *Las variables cinemáticas* que describen el desplazamiento del cuerpo mediante el registro de las variaciones angulares de las articulaciones y los movimientos relativos de los segmentos corporales en el espacio.
2. *Las variables cinéticas* que cuantifican las relaciones entre las fuerzas de acción-reacción, los momentos y las potencias medidas para cada uno de los segmentos corporales.

3. *Las variables de activación muscular* que evalúan la actividad eléctrica de la acción muscular durante el ciclo de marcha. Estas variables son registradas por señales de EMG (electromiografía dinámica superficial).

Durante la marcha, el movimiento que imprime el centro de gravedad es sinuoso y no rectilíneo, lo cual exige ciertos intercambios de energía: conversiones entre energía cinética y potencial y transferencias de energía entre segmentos. Durante la fase de apoyo bipodal el centro de gravedad del tronco se encuentra en su posición más baja y presenta su máxima velocidad hacia delante, es decir, su energía potencial es mínima y su energía cinética máxima.

Para el análisis de la marcha se usan diversos métodos [41]:

- *Acelerometría:* permite medir la aceleración en cualquier segmento o articulación del cuerpo.
- *Goniometría digital:* mide la posición angular en cualquier instante de tiempo.
- *Sistemas de análisis en 2D y 3D (cámaras normales, infrarrojas)*: registran en video el movimiento y permiten digitalizar el cuerpo como un sistema de segmentos unido por puntos.
- *Electromiografía (EMG):* mide la actividad muscular.
- *Baropodometría:* mide la presión ejercida sobre el piso.
- *Plataformas de fuerza:* registran la fuerza durante el apoyo del pie en el ciclo de marcha.

Las causas que pueden originar alteraciones en la marcha son [46]:

- *Anormalidades frecuentes:* Acortamiento de miembro inferior, anquilosis o limitación de la amplitud articular, inestabilidad articular o marcha antiálgica.
- D*éficit neurológico de origen central* causado por enfermedades, tales como: Hemiplejía espastica, Ataxias Hereditarias, Parkinson, Huntington, etc.
- *Lesiones neurológicas periféricas:* parálisis de extensores de cadera, de glúteo medio, de cuádriceps, isquiotibiales, flexores dorsales del pie o del tríceps sural.

2.2. Las enfermedades neurodegenerativas y su afectación a la marcha

Las enfermedades neurodegenerativas (ND) provocan un deterioro neurológico debido a una inevitable pérdida de la función cerebral; suelen ser de origen genético y progresivas; hasta el momento, la mayoría de ellas no tienen cura; los tratamiento disponibles ayudan a atenuar los síntomas para mejorar la calidad de vida de los pacientes [47]-[49].

Las enfermedades neuro-degenerativas afectan diversas funcionalidades corporales como el equilibrio, movimiento, hablar, respirar. La afectación de las capacidades motrices provocan alteraciones en los patrones de marcha tales como como la falta de coordinación y control cuando se desplazan [10], [50]-[54]. **La enfermedad de Huntington y las Ataxias Hereditarias** son enfermedades ND que comparten mecanismos patológicos similares en la afectación del control del movimiento; por lo tanto resulta de interés estudiar y analizar los patrones de la marcha de ambos grupos para establecer diferencias significativas que aporten al diagnóstico y seguimiento médico [55]-[58].

2.2.1. La enfermedad de Huntington

La enfermedad de Huntington es un trastorno hereditario autosómico dominante caracterizado clínicamente por una combinación de síntomas motores, cognitivos y psiquiátricos. Los trastornos motores son, por lo general, los más frecuentes y los más notorios [12], [59], son progresivos y empeoran con la gravedad de la enfermedad; afectan la precisión y la velocidad de movimiento; causando pérdida del equilibrio y de la mecánica normal de la marcha provocando accidentes y caídas en los pacientes [60], [61].

En la fase inicial de la enfermedad se presentan movimientos involuntarios, corea (movimientos espasmódicos) y bradicinesia (lentitud de movimiento). Con la evolución de la enfermedad surgen otros signos motores como la rigidez, distonía (contracciones sostenidas del músculo), acinesia (déficit en la iniciación del movimiento), inestabilidad postural, hipocinesia (actividad motora anormal disminuida) [12], [18], [60]. Las alteraciones de marcha en pacientes con HD se caracterizan por el desequilibrio en la postura y desplazamiento, tendencia del deslizamiento en zigzag, ampliación del polígono de soporte, pérdida de movimientos asociados a los brazos, dificultades en los

giros en U y una velocidad variable [61], [62]. Los movimientos irregulares pueden ser excesivos, espontáneos, distribuidos al azar y abruptos. La gravedad de los trastornos puede variar desde la agitación con una exageración leve e intermitente del gesto y la expresión; movimientos agitados de las manos, marcha inestable y en forma de danza hasta un flujo continuo de movimientos violentos e incapacitantes (corea) [63].

La corea es generalmente la anormalidad más temprana del movimiento visible en adultos; se confunde con otras dolencias cuando los movimientos son aislados e incipientes. No obstante, la corea en la HD no afecta de forma apreciable al centro de gravedad durante la deambulación, y la consistencia de los perfiles de marcha en el golpe de talón muestra que el objetivo final se alcanza en cada paso a pesar de la variabilidad aleatoria y frecuente durante el ciclo de marcha [64]. Los pacientes con enfermedad más avanzada tienen un equilibrio más pobre y una mecánica de andar disminuida, lo que resulta en un estado de ausencia deambulación [15].

La edad media de aparición de la enfermedad varía entre los 30 y los 50 años (cuarta década de vida) [15]. Después del diagnóstico, se observa un empeoramiento progresivo de los síntomas en un período de 15 a 30 años hasta la muerte [31]. No existe ninguna terapia o intervención disponible que demuestre un inicio retrasado o que ralentice la progresión de la enfermedad [11]. La duración media de la enfermedad hasta la muerte se estima entre 15 y 20 años después del inicio de la corea. La duración real es probablemente mucho más larga, basada en biomarcadores y observaciones clínicas [65].

Las pruebas médicas para evaluar el control motor incluyen análisis de la marcha con largas pasarelas que revelan una postura más larga, tendencia a recostarse sobre los talones, disminución de la velocidad y de la variabilidad de la longitud de la zancada; las alteraciones en la marcha se utilizan como predictor de la progresión de la enfermedad. Se han aplicado herramientas tecnológicas para evaluar alteraciones tales como: procesamiento de imágenes, bandas de marcha con sensores, sensores de presión y sensores portátiles basados en modelos musculoesqueléticos [43], [45].

2.2.2. Ataxias Hereditarias

Las ataxias hereditarias son un grupo heterogéneo de trastornos caracterizados fenotípicamente por una ataxia progresiva de la marcha; los trastornos motores son el resultado de la degeneración de la corteza cerebelosa y la médula espinal; movimientos

descontrolados de las manos, descoordinación de los movimientos oculares y generalmente asociados con atrofia del cerebelo [14].

Los trastornos motores tienen como síntoma principal un caminar inestable, tambaleante, descoordinado, con una base ancha y los pies tirados, descendiendo primero sobre el talón y luego sobre los dedos de los pies con un doble golpe a la que se denomina como marcha atáxica [66], [67]; la inestabilidad al caminar, la reducción de la longitud de zancada, disminución de la velocidad de la marcha, el aumento de la variabilidad de andar y trastornos del equilibrio son características del progresión de la enfermedad [68], [69]. La alteración de la marcha y otros trastornos del movimiento son claramente visibles en etapas avanzadas de la enfermedad.

Las Ataxias Hereditarias que más prevalecen en la población son la Ataxia Autosómica Recesiva (ARA) y la Ataxia Autosómica Dominante (ADA), que pueden subdividirse en Ataxias Espinocerebelares Progresivas (SCA) y Ataxias Episódicas (EA) caracterizadas por la aparición Paroxística de Ataxia [14], [70]. Actualmente se han descrito unas 36 variedades de SCA en los últimos 20 años; la SCA de tipo 3 (enfermedad de Machado-Joseph) es la más extendida y de mayor frecuencia [71] representa el 21 % de los enfermos; seguido por la SCA2 y el SCA6, con un 15 %; finalmente la SCA1, cuenta con el 6 %; y la SCA7, con un 5 % de enfermos [71], [72]. En 2014, *Boll et al.* iniciaron un registro institucional de ataxias degenerativas [21] e informaron que la SCA2 es la más prevalente en México; con un número creciente desde 2012, 817 pacientes por cada 10,000 habitantes; en la región montañosa de Veracruz se estima que en cuatro años se ha duplicado estas cifras [29].

El examen de los trastornos del movimiento a través de la observación clínica ayuda a los especialistas a establecer un grado de afectación motor para determinar la progresión de la enfermedad; sin embargo, la diversidad de síntomas y su lenta progresión es muy difícil encontrar un instrumento fiable para medir los cambios en alteraciones motoras a lo largo del tiempo [73]. *García P. et al.* [74] determinaron que la Ataxia se puede presentar en otras enfermedades como la enfermedad de Huntington y pueden se confundidos porque comparten mecanismos patogénicos comunes en trastornos del movimiento, oculomotores y demencia. En ambos grupos, la variabilidad de la marcha es el principal un indicador del deterioro de la marcha en ambos grupos.

2.3. Sensores de movimiento

El término *Sensores de Movimiento (MS) o inerciales* se refiere a una familia de sensores representada esencialmente por sensores de aceleración lineal (acelerómetros) y sensores de velocidad angular (giroscopios). El término "inercial" es usado debido a que estos sensores miden su propio movimiento y, en consecuencia, el movimiento del cuerpo rígido al que están sujetados, utilizando el principio de inercia: *la aceleración puede estar relacionada con la inercia del movimiento de una masa libre contenida en el sensor, cuando es acelerada por una fuerza externa o por un par de torsión (para el acelerómetro o el giroscopio, respectivamente).*

Los acelerómetros y giroscopios pueden utilizarse para cuantificar directamente el movimiento de un cuerpo rígido en términos de aceleración lineal o velocidad angular, para estimar otras magnitudes mecánicas (como el desplazamiento angular y lineal) y/o para extraer otro tipo de información del análisis de señales en el tiempo y en el dominio de la frecuencia.

2.3.1. Acelerómetros

Los acelerómetros son sensores que miden las aceleraciones de objetos en movimiento a lo largo de ejes de referencia. Es preferible medir la actividad física utilizando la acelerometría porque la aceleración es proporcional a la fuerza externa y, por lo tanto, puede reflejar la intensidad y la frecuencia del movimiento humano. Los datos de aceleración se pueden utilizar para obtener información de velocidad y desplazamiento mediante la integración de datos de aceleración con respecto al tiempo [75].

El principio de funcionamiento de los acelerómetros se basa en un elemento sensor microelectromecánico (MEMS), el cual consiste en una masa de apoyo (o masa sistólica) unida a un sistema de suspensión con respecto a un marco de referencia. La fuerza inercial debida a la aceleración o a la gravedad hará que la masa de apoyo se desvíe de acuerdo con la Segunda Ley de Newton. La aceleración puede medirse eléctricamente con los cambios físicos en el desplazamiento de la masa de apoyo con respecto al marco de referencia. Los acelerómetros piezoresistivos, piezoeléctricos y capacitivos diferenciales son los más comunes [76], [77].

Algunos acelerómetros pueden responder a la gravedad para proporcionar detección de inclinación con respecto a los planos de referencia cuando los acelerómetros se mueven con respecto a otros objetos. Los datos de inclinación resultantes pueden utilizarse para clasificar las posturas del cuerpo (orientaciones). Con estas características, la acelerometría es capaz de proporcionar información suficiente para medir los movimientos y una serie de actividades humanas. Los acelerómetros han sido ampliamente aceptados como sensores útiles y prácticos en dispositivos portátiles para medir y evaluar la actividades comunes en entornos clínicos/laboratorios o de vida diaria [78], [79].

2.3.2. Giroscopios

Un giroscopio es un dispositivo que utiliza la gravedad de la Tierra para ayudar a determinar la orientación. Su diseño consiste en un disco de rotación libre llamado rotor, montado sobre un eje de rotación en el centro de una rueda más grande y estable. A medida que el eje gira, el rotor permanece inmóvil para indicar la atracción gravitacional central y, por lo tanto, puede determinar cuando está "hacia abajo".

Un giroscopio utiliza el principio del momento angular basado en el principio de rigidez del espacio para indiciar la orientación y posición angular; a diferencia del acelerómetro que mide la aceleración lineal basada en la vibración. Los giroscopios más comunes son: mecánicos, ópticos y electromecánicos.

El giroscopio mantiene su nivel de eficacia al poder medir la velocidad de rotación alrededor de un eje en particular. Cuando se mide la velocidad de rotación alrededor del eje de balanceo de una aeronave, se identifica un valor real hasta que el objeto se estabiliza. Un giroscopio se utiliza en la navegación en vehículos aéreos no tripulados, brújulas y grandes embarcaciones [80].

2.3.3. Sensores de movimiento del iPhone 5S

Los dispositivos inteligentes que portan el sistema operativo *IOS*, aunque no todos los modelos (iPhone, iPad o iPod touch) tienen casi el mismo conjunto de sensores. Los sensores están relacionados con la versión del sistema operativo y el *Hardware que este contiene*. Los sensores más comunes son: sensor de proximidad, sensor de movimiento/acelerómetro, sensor de luz ambiental, sensor de humedad, giroscopio, brújula, barómetro, identificador táctil e identificador facial.

El acelerómetro del iPhone mide los cambios de velocidad a lo largo de un eje. Todos los dispositivos iOS tienen un acelerómetro de tres ejes, que proporciona valores de aceleración en cada uno de los tres ejes mostrados en la Figura 2.2. Los valores reportados por los acelerómetros se miden en incrementos de la aceleración gravitacional, representando el valor 1.0 una aceleración de 9.8 metros por segundo (por segundo) en la dirección dada. Los valores de aceleración pueden ser positivos o negativos dependiendo de la dirección de la aceleración [81], [82].

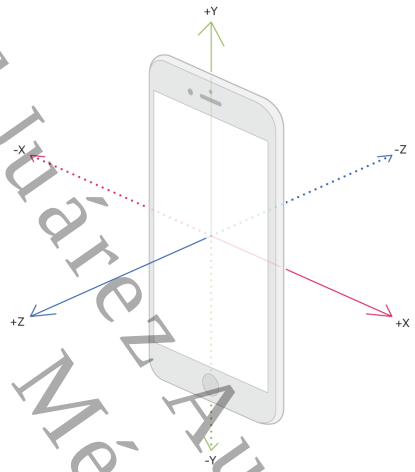


FIGURA 2.2: Los acelerómetros miden los cambios de velocidad a lo largo de los ejes X, Y y Z [82].

El giroscopio mide la velocidad a la que gira un dispositivo alrededor de un eje espacial. Muchos dispositivos *iOS* tienen un giroscopio de tres ejes, que proporciona valores de rotación en cada uno de los tres ejes que se muestran en la Figura 2.3. Los valores de rotación se miden en radianes por segundo alrededor del eje dado. Los valores de rotación pueden ser positivos o negativos dependiendo del sentido de rotación.

Los datos brutos de la velocidad de rotación suministrados por las interfaces del giróscopo pueden estar sesgados por otros factores, como la aceleración del dispositivo, por lo tanto, si se requieren valores de rotaciones reales se deben usar algoritmos especiales para eliminar cualquier sesgo.

El sistema operativo *iOS* proporciona el framework *"Core Motion"* que permiten obtener los datos relacionados a los sensores de movimiento. Los servicios de los sensores de movimiento permiten obtener los datos para implementarse en diversas aplicaciones. El *framework* permite realizar el procesamiento necesario para obtener datos refinados y ser usados directamente en las aplicaciones. El servicio *"Core Motion"* utiliza todo

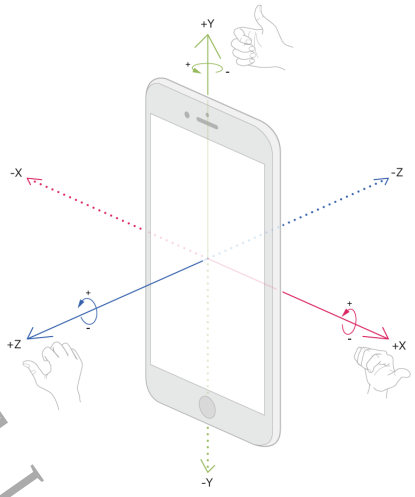


FIGURA 2.3: Los giróscopos miden la velocidad de rotación alrededor de los ejes X, Y y Z [82].

el hardware relevante para generar una secuencia que contiene la siguiente información: la orientación (o actitud) del dispositivo en un espacio tridimensional en relación con un marco de referencia, la velocidad de rotación no sesgada, el vector de gravedad actual, el vector de aceleración generado por el usuario (sin gravedad) y el vector del campo magnético actual [81], [82].

Los sensores de movimiento de los iPhones han revelado que son lo suficientemente fiables y precisos para evaluar e identificar los patrones cinemáticos de marcha en [83], [84]. Estudios relacionados con la evaluación de la marcha y el cuidado de la salud han demostrado la capacidad de los sensores del iPhone para cuantificar los parámetros de la marcha con un nivel suficiente de consistencia, específicamente en la posición del tobillo, de forma cómoda, portátil y manejable. [85]-[87].

2.4. Reconocimiento de patrones

El uso del reconocimiento de patrones para el análisis del conjunto de datos adquiridos con sensores de movimiento permite el descubrimiento de información y el establecimiento de propiedades entre un diferentes objetos (instancias); es decir, analizar un conjunto dado de datos, tomando cada instancia y los asigna a una clase particular (predicción) [88].

El reconocimiento de patrones es el proceso de reconocimiento o agrupamiento de similitudes en un conjunto de información (dataset) mediante el uso de algoritmos de aprendizaje automático. El reconocimiento de patrones puede definirse como la clasificación de datos basada en conocimientos ya adquiridos o en información estadística extraída de patrones y/o su representación. El reconocimiento de patrones debe: encontrar patrones familiares en los datos de manera confiable, reconocer y clasificar instancias de objetos desconocidos con precisión, identificar patrones incluso cuando se encuentran parcialmente ocultos y debe realizarse de manera fácil y automática [89].

El reconocimiento de patrones implica *clasificación y/o agrupación de patrones*. En la *clasificación*, se asigna una *etiqueta de clase* apropiada a un patrón basado en una abstracción que se genera utilizando un conjunto de información previamente conocida para ser comparada con un dato desconocido; la clasificación es utilizada en el *aprendizaje supervisado*. La agrupación basada en particiones de los datos ayuda a la toma de decisiones específicas de interés para una problemática; el proceso de *agrupamiento o clustering* se denomina *aprendizaje no supervisado* [90].

Las tareas de reconocimiento de patrones están basadas en *características*; las cuales pueden ser representadas como variables binarias, continuas o discretas. Una *característica* es una función de una o más mediciones, calculada de manera que cuantifica datos representativos de los objetos o instancias que participan en el proceso [91]. En una aplicación típica de reconocimiento de patrones, los datos brutos se procesan y se convierten en un conjunto de características o atributos que puede ser utilizados por un algoritmo de aprendizaje automático. Los atributos que se toman en conjunto, forman un *vector de características* [92], [93].

El aprendizaje es la fase en la que un sistema forma y adapta la información para obtener un resultado preciso. El aprendizaje es la fase más importante, ya que el rendimiento del modelo aprendido con los datos proporcionados dependerá de los algoritmos utilizados en los datos. Los conjuntos de datos (características) se dividen en un subconjunto de formación o entrenamiento (utilizado para construir un modelo) y otro de prueba, para verificar el desempeño del modelo aprendido [94], [95].

Las reglas de entrenamiento y los algoritmos utilizados proporcionan información relevante sobre cómo asociar los datos de entrada con la decisión de salida. El sistema se entrena aplicando estos algoritmos en el conjunto de datos, se extrae toda la información relevante de los datos y se obtienen los resultados. Generalmente se usan un 80 %

de datos para entrenar el modelo y el 20 % restante permiten verificar el desempeño de ese modelo [96]-[98].

2.5. Meta-clasificadores

La etapa de clasificación en el reconocimiento de patrones puede incluir el uso de uno o más algoritmos a la vez para mejorar la precisión del reconocimiento de datos. El uso de meta-clasificadores, múltiples o ensambladores permite mejorar los resultados de los algoritmos de clasificación combinando predicciones de clasificadores individuales en iteraciones. Un meta-clasificador utiliza un clasificador base para realizar la clasificación de tareas, añade otro paso de procesamiento que se realiza antes de que el clasificador base implemente los datos. En esta sección se describen brevemente los meta-clasificadores y árboles de clasificación tomados en cuenta en esta investigación. Se incluyen también las descripciones algorítmicas con el fin de facilitar la lectura.

2.5.1. LogitBoost

El término *"regresión"* se refiere comúnmente a un tipo particular de modelo paramétrico (o proceso) para estimar una variable objetivo (numérica): mientras que en los problemas de clasificación se producen estimaciones para una variable objetivo categórica, llamada clase.

La transformación de una tarea de clasificación en un problema de regresión puede realizarse utilizando un modelo de *regresión lineal* estándar denominado *regresión logística*. La *Regresión Logística Lineal* (ecuación 2.1) modela las probabilidades de la clase j mediante funciones lineales en x y se asegura de que los resultados permanezcan entre los valores $[0, 1]$. El proceso de regresión ajusta un vector de parámetros β a una variable numérica destino tomando el modelo $f(x) = \beta^T x$; donde x es el vector de los valores de atributo para la instancia (asumiendo un componente constante en el vector de entrada para la interceptación).

$$Pr(G = j|X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}, \quad \text{donde } F_j(x) = \beta_j^T \cdot x \tag{2.1}$$

El procedimiento estándar en estadística es buscar la estimación de máxima fiabilidad, es decir, elegir los parámetros que maximizan la probabilidad de los puntos de datos observados. Para el modelo de regresión logística, no existen soluciones de forma cerrada para estas estimaciones; en su lugar, los algoritmos de optimización numérica tienen que ser usados para acercarse a la solución de máxima probabilidad de manera iterativa y alcanzarla hasta el límite.

Uno de estos métodos iterativos es el algoritmo *LogitBoost* [99], el cual permite ajustar los modelos aplicando el coste funcional de regresión logística a un modelo aditivo generalizado por probabilidad máxima, parecido al algoritmo *AdaBoost*. En la clasificación, realiza una regresión logística aditiva utilizando un esquema de regresión como algoritmo base de aprendizaje.

El algoritmo **Logitboost** 1, en cada iteración, ajusta un regresor de mínimos cuadrados a una versión ponderada de los datos de entrada con una variable objetivo transformada. Aquí, y^*_{ij} (ecuación 2.2) son las variables binarias de pseudo-respuesta que indican la pertenencia de un grupo a una observación como,

$$y^*_{ij} = \begin{cases} 1 & \text{if } y_i = j \quad \text{donde } y_i \text{ es la clase observada} \\ 0 & \text{if } y_i \neq j \quad \text{para la instancia } x_i \end{cases} \tag{2.2}$$

Al restringir f_{mj} para ser lineal en *x*, la regresión logística lineal conseguirá que el algoritmo se ejecute hasta la convergencia. Cuando se restringe aún más f_{mj} da como resultado un error cuadrático más bajo, entonces el algoritmo realiza una selección automática de atributos. Al utilizar la validación cruzada para determinar el mejor número de iteraciones de LogitBoost *M*, sólo se incluyen aquellos atributos que mejoran el rendimiento en instancias no vistas. En lugar de minimizar el error con respecto a y (como AdaBoost), se eligen los modelos débiles para minimizar el error (de mínimos cuadrados ponderados) de $f_t(x)$ con respecto a;

$$z_t = \frac{y^* - p_t(x)}{p_t(x)(1 - p_t(x))} \tag{2.3}$$

donde $p_t(x)$ es regresión logística, $p_t(x)(1-p_t(x))$ es el peso (w_t), $y^* = (y+1)/2$ son modelos aprendidos débiles, y z_t es la aproximación de Newton-Raphson sobre el reductor del error de probabilidad de registro en la etapa t, y $f_t(x)$ es elegido como el aprendiz que mejor se aproxima a z_t por los mínimos cuadrados ponderados. En las iteraciones,

el ajuste de la función $f_t(x)$ pondera una regresión de mínimos cuadrados de z_t a x con peso w_t. La salida de los clasificadores será $argmax f(x)$.

Algoritmo 1 Meta-clasificador LogitBoost (Jclasses) [99].

Entrada: $w_{ij} = \frac{1}{n}$, en donde $i = 1, ..., n$, $j = 1, ..., J$
Salida: $\underset{j}{\text{argmax}}\, F_j(x)$

1: $F_j = 0$
2: $p_j(x) = \frac{1}{J}$; $\forall j$
3: Repetir para $m = 1, ..., M$:
4: Repetir para $j = 1, ..., J$:
Calcular las fuerzas de respuesta y los pesos para la jth clase
5: $z_{ij} = \frac{y^*_{ij} - p_j(x_i)}{p_j(x_i)(1 - p_j(x_i))}$
6: $w_{ij} = p_j(x_i)(1 - p_j(x_i))$
Ajustar la función $f_{mj}(x)$ para una regresión ponderada de los mínimos cuadrados de z_{ij} hasta x_i con los pesos w_{ij}
7: $f_{mj}(x) = \frac{J-1}{J}(f_{mj}(x) - \frac{1}{J}\sum_{k=1}^{j} f_{mk}(x))$
8: $F_j(x) = F_j(x) + f_{mj}(x)$
9: $p_j(x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}$

2.5.2. RandomCommittee

El meta-clasificador *RandomCommittee* construye un conjunto de clasificadores base de manera aleatoria a partir de datos de entrenamiento S para las iteraciones M. Se generan M copias del algoritmo de aprendizaje que conformarán el conjunto de modelos aprendidos T. Cada clasificador base se construye usando un número aleatorio de semillas s_w diferente dado por [100],

$$(s[0] \times 31^{(n-1)} + s[1] \times 31^{(n-2)} + ... + s[n-1]) + semilla$$

pero basado en los mismos datos s_w. Cuando el aprendiz es incapaz de manejar los pesos de las instancias de datos, se remuestrean con una distribución discreta usando el método Walker dado por,

$$P(X = i) = \frac{Q(i)}{M} + \sum_{j:A(j)=i} \frac{[1 - Q(j)]}{M} \tag{2.4}$$

El algoritmo 2 describe la estructura general de funcionamiento de *RandomCommittee*.

Algoritmo 2 Pseudo-código del meta-clasificador *RandomCommittee* [100].

Entrada: entrenamiento S, iteraciones M
Salida: un árbol ensamblado $T = t_1, ..., t_M$

if $Classifier\ is\ Randomizable$ **then**
  classifiers = $MakeCopies(learner, M)$
  random_gen = $randomNumberGenerator(S, seed)$
  if $classifier\ can't\ handle\ weights$ **then**
    $S_w = resampleWithWeights(random\_gen)$
  end if
  for $i = 1$ to M **do**
    $classifiers_i.seed = random\_gen.seed()$
    generar un árbol: $T_i = buildClassifier(classfiers_i, S_w)$
  end for
  return T
else
  mensaje : $el\ aprendiz\ debe\ implementar\ el\ método\ aleatorio$
end if

Cuando la variable aleatoria *X* se hace sin reemplazo, tiene una distribución hipergeométrica con parámetros *N (Población)*, *n (número de sorteos)*, y *K (número de éxito en la población N)*.

$$h(x; N, n, K) = \begin{cases} \frac{\binom{K}{x} \times \binom{N-K}{n-x}}{\binom{N}{n}}; & x = 1, 2, ..., min(n, k) \\ 0; & otracosa \end{cases} \tag{2.5}$$

La predicción final es un promedio directo de las predicciones generadas por los clasificadores base individuales en las iteraciones (M).

$$f(x) = P(X = x) = \frac{1}{M} \sum_{1}^{M} h(x; N, n, K) \tag{2.6}$$

El algoritmo **RandomCommittee** tiene la ventaja de ser rápido debido a su naturaleza aleatoria; sin embargo, sólo puede ser usado con algoritmos con mecanismos similares, tales como: *RandomTree (Árboles al azar), REPTree (Aprendizaje rápido del árbol de decisiones), LADTree (árbol que utiliza la estrategia LogitBoost) y ExtraTree (árboles extremadamente aleatorios)*.

2.5.3. MultiboostAB

MultiBoosting es una extensión altamente efectiva que combina *AdaBoost* con *Wagging* para formar comités de decisión, aprovecha tanto el alto sesgo y la reducción de varianza de *AdaBoost* como la reducción de varianza superior de *wagging*. El uso de J48 (J48) como algoritmo base de aprendizaje *MultiBoosting* produce comités de decisión con el error más bajo que *AdaBoost* o *Wagging* de forma significativa, otra ventaja que ofrece es la posibilidad de ejecución paralela [101].

MultiBoost, toma como argumento un único comité n de tamaño $\lfloor\sqrt{T}\rfloor$, desde el cual, por defecto establece el número de subcomités y el tamaño de esos subcomités a $I_i = \lceil i \times T/n \rceil$. Para facilitar la implementación, se establece un índice de miembros del subcomité final objetivo, donde a cada miembro del comité final se le asigna un índice, comenzando desde uno. Esto permite la terminación prematura del crecimiento exponencial de un subcomité debido a un error demasiado grande o demasiado bajo, para llevar a un aumento en el tamaño del siguiente subcomité. Si el último subcomité termina prematuramente, se agrega un subcomité adicional con el objetivo de completar la totalidad de los miembros del comité. Si este subcomité adicional tampoco logra alcanzar este objetivo, este proceso se repite y se agregan otros subcomités hasta que se alcance el tamaño total del comité objetivo. El algoritmo 3 muestra la estructura simplificada de meta-clasificador *MultiBoost* resultante.

El error ponderado en el conjunto de entrenamiento se calcula mediante,

$$\epsilon_t = \frac{\sum x_j \in S\prime : C_t(x_j) \neq y_j weight(x_j)}{m} \tag{2.7}$$

Cuando $\epsilon_t > 0.5$ se restablece $S\prime$ a pesos aleatorios extraídos de la distribución continua de *Poisson* y se estandariza a $S\prime$ para sumar n al incremento k.

Además de las propiedades de reducción de sesgo y varianza que este algoritmo puede heredar de los algoritmos base en los que fue inspirado, *Multiboost* introduce un mecanismo de terminación temprana del aprendizaje a nivel de subcomité para evitar errores demasiados grandes. Los clasificadores aprendidos se comportan de manera independiente como el algoritmo *Wagging*, lo que permite el cómputo paralelo, propiedad que *Multiboost* hereda a nivel de subcomité.

Algoritmo 3 Pseudo-código del meta-clasificador Multiboosting [101].

Entrada: S, una secuencia de muestras etiquetadas como m $(x_1, y_1), ..., (x_m, y_m)$ con etiquetas $y_1 \in Y$ como algoritmo base de **aprendizaje**. Entero T especificando el número de iteraciones. Arreglo de enteros I_i especificando las iteraciones en las que cada subcomité $i > 1$ debería terminar.

Salida: el clasificador final:

$$C^*(x) = \underset{y \in Y}{\operatorname{argmax}} \sum_{t:C_t(x)=y} log\frac{1}{\beta_t}$$

1: $S' = S$ con pesos de instancias asignados a igual a 1.
2: set $k = 1$
3: **for** $t = 1$ to T **do**
4: **if** $I_k = t$ **then**
5: reinicia S' a pesos aleatorios extraídos de la distribución continua de Poisson.
6: Estandarizar S' para sumar a n e incrementar k.
7: **end if**
8: $C_t = BaseLearn(S')$.
9: $\epsilon_t = \frac{\sum_{x_j \in S':C_t(x_j) \neq y_j} weight(x_j)}{m}$ *(el error ponderado en el conjunto de entrenamiento)*
10: **if** $\epsilon_t > 0.5$ **then**
11: reiniciar S' a pesos aleatorios extraídos de la distribución continua de Poisson.
12: estandarizar S' para sumar a n.
13: incrementa k.
14: ir al paso 9.
15: **else if** $\epsilon_t = 0$ **then**
16: establecer β_t a 10^{-10}
17: reiniciar S' a pesos aleatorios extraídos de la distribución continua de Poisson.
18: estandarizar S' para sumar a n.
19: incrementa k.
20: **else**
21: $\beta_t = \frac{\epsilon_t}{(1 - \epsilon_t)}$
22: **for** each $x_j \in S$, **do**
23: divide *weight*(x_j) por $2\epsilon_t$, sí $C_t(x_j) \neq y_j$ y $2(1 - \epsilon_t)$ otra cosa.
24: sí $weight(x_j) < 10^{-8}$, establecer $weight(x_j) = 10^{-8}$.
25: **end for**
26: **end if**
27: **end for**

2.6. Árboles clasificadores

Los algoritmos de clasificación utilizados en este estudio se basan en el método de aprendizaje de árboles de decisión, cuyo objetivo es crear un modelo para predecir el valor de una variable objetivo basado en varias variables de entrada. Un árbol de decisión es un modelo predictivo, donde cada nodo interior del árbol corresponde a una de las variables de entrada; existen valores desde los nodos interiores a nodos exteriores llamados hojas. Cada hoja representa un valor de la variable objetivo, dados los valores de las variables de entrada representadas por la trayectoria desde la raíz hasta la hoja. El aprendizaje en los árboles se basa en las decisiones dentro del árbol para pasar de las observaciones sobre una variable de entrada a las conclusiones sobre el valor objetivo de la variable. Los modelos basados en árboles, donde la variable de destino puede tomar un conjunto discreto de valores, se denominan árboles de clasificación. [102].

2.6.1. Random Forest

Random Forest o bosques de decisión aleatoria son un método de aprendizaje en conjunto para la clasificación, regresión y otras tareas. La idea detrás del algoritmo es construir árboles de decisión pequeños con pocas características y poco costo computacional [103], [104]. Se puede formar un aprendizaje fuerte, combinando muchos árboles de decisión pequeños y débiles promediado utilizando la técnica de mayoría de votos. Esta combinación se lleva a cabo de tal manera que el modelo producido por varios árboles en un conjunto (bosque), funcione mejor que el original.

El algoritmo **Random Forest** utiliza el método de agregación por arranque (Bootstrap) o empaquetamiento (bagging) para construir múltiples árboles de decisión mediante el remuestreo repetido de los datos de entrenamiento con reemplazo y mayoría de votos de los árboles para una predicción en consenso[105]. Cada modelo toma un subconjunto de datos de entrenamiento muestreados al azar, de modo que los participantes produzcan diferentes modelos que puedan promediarse de manera prudente.

Los bosques de decisión aleatoria corrigen el sobreajuste de los árboles de decisión tradicionales en su conjunto de entrenamiento. Esto se debe a que los árboles tradicionales crecen demasiado profundo, por lo que tienden a aprender patrones muy irregulares: se ajustan demasiado a sus conjuntos de entrenamiento, es decir, tienen un sesgo bajo,

pero una varianza muy alta (compensación entre sesgos y varianzas). Los bosques aleatorios superan esto promediando múltiples árboles de decisión profundos, entrenando diferentes partes del mismo conjunto de entrenamiento, con el objetivo de reducir la varianza. Esto se produce a cambio de un pequeño aumento del sesgo y una cierta pérdida de interpretabilidad, pero en general mejora enormemente el rendimiento en el modelo final [106].

Algoritmo 4 Pseudocódigo del árbol clasificador Random Forests [107].

Precondición: Un conjunto de entrenamiento $S := (x_1, y_1), ..., (x_n, y_n))$, y características F, y el número de árboles en el bosque B.

```
1: procedure RANDOMFOREST(S, F)
2:     H = {}
3:     for i ∈, ..., B do
4:         S^i = Una muestra de bootstrap de S
5:         h_i = RANDOMIZED_TREE_LEARN(S^i, F)
6:         H = H ∪ {h_i}
7:     end for
8:     return H
9: end procedure
```

```
1: procedure RANDOMIZED_TREE_LEARN(S, F)
2:     for cada nodo do
3:         f = un subconjunto muy pequeño de F
4:         Dividir la mejor característica en f
5:     end for
6:     return El árbol aprendido
7: end procedure
```

El algoritmo 4 muestra como el *Random Forests* trabajan en dos pasos (funciones): en el primer paso, se selecciona una muestra de bootstrap del conjunto de entrenamiento *S* para cada árbol del bosque, donde $S^{(i)}$ denota cada muestreo del Bootstrap. En el segundo paso se aprende un árbol de decisión, utilizando un algoritmo de aprendizaje de árbol de decisión modificado *RamdomizedTreeLearn*. El algoritmo en lugar de examinar todas las posibles particiones de característica en cada nodo del árbol, selecciona algún subconjunto de las características $f \subseteq F$, donde *F* es el conjunto de características. El nodo entonces se divide en la mejor característica en *f* en lugar de *F*. En la práctica *f* es mucho más pequeño que *F*. Decidir sobre qué característica dividir es a menudo el aspecto más costoso computacionalmente del aprendizaje del árbol de decisión. Al reducir el conjunto de características, el aprendizaje del árbol se acelera considerablemente.

2.6.2. ExtraTrees(Extremely Randomized Trees)

El algoritmo *ExtraTrees (Extremely Randomized Trees)* construye un conjunto de árboles de decisión o regresión no podados según el procedimiento clásico de arriba hacia abajo. Sus dos principales diferencias con otros métodos de ensamblaje basados en árboles son: la división de nodos eligiendo puntos de corte al azar y el uso de todo del conjunto (en lugar de una réplica de auto-aprendizaje) para cultivar los árboles [108].

El procedimiento de partición *ExtraTrees* para atributos numéricos tiene dos parámetros: *K*, el número de atributos seleccionados aleatoriamente en cada nodo y n_{min}.El tamaño mínimo de la muestra para dividir un nodo. Se utiliza varias veces con la muestra de aprendizaje original (completa) para generar un modelo de conjunto (denotado por *M* el número de árboles de este conjunto).

EL algoritmo 5 muestra como el ExtraTrees construye los árboles seleccionando aleatoriamente atributos K del conjunto de entrenamiento sin reemplazo, usando todos los atributos como candidatos $\{a_1, ..., a_k\}$; se generan divisiones aleatorias de K usando el conjunto de entrenamiento de $\{s_1, ..., s_k\}$; de esos conjuntos, seleccione el mejor split s_* con la mejor puntuación; luego dividir el entrenamiento S en dos S_l(izquierda) y S_r(derecha) según la prueba de s_* y reconstruye los árboles T_l y T_r, y a continuación, identificar el mejor nodo s_* y asociar los árboles que se generaron a los dos lados del nodo para formar la estructura de la estructura del nodo, resultando así en un árbol final de t.

La medida de la puntuación en la clasificación es una normalización particular de la ganancia de información; esta medida viene dada por,

$$Score_c(s, S) = \frac{2I_c^s(S)}{H_s(S) + H_c(S)} \tag{2.8}$$

donde $H_c(S)$ es la entropía (log) de la clasificación en S; $H_s(S)$ es la entropía dividida y $2I_c^s(S)$ es la información mutua del resultado dividido y la clasificación. Las predicciones de los árboles se agregan para obtener la predicción final por mayoría de votos en problemas de clasificación y promedio aritmético en problemas de regresión.

La forma de trabajo en *split* o puntos de cortes y los atributos combinados con el promedio del conjunto son capaces de reducir la varianza más fuertemente que los esquemas de aleatorización con árboles débiles usados por otros métodos. El uso de la muestra de aprendizaje original completa en lugar de réplicas de Bootstrap está orientado

Algoritmo 5 Pseudocódigo del árbol clasificador Extremely Randomized Trees [108].

Entrada: el subconjunto de aprendizaje local S correspondiente al nodo que queremos dividir.
Salida: una división $[a < a_c]$ o nada.

procedure $split\_a\_node$(S)
  if $Stop\_split(S)$ es **true then**
    return nada
  else
    Seleccionar K Atributos $\{a_1, ..., a_k\}$ entre todos los no constantes (en S) atributos del candidato;
    Dibuja K en divisiones $\{s_1, ..., s_k\}$ donde $s_i = Pick\_a\_random\_split(S, a_i),\ \forall i = 1, ..., K$;
  end if
  return una partición s_* donde $Score(s_*, S) = max_{i=1,...,K}\ Score(s_i, S)$.
end procedure

Entrada: un subconjunto S y un atributo a.
Salida: una partición.

procedure $Pick\_a\_random\_split(S, a)$
  dejar a^s_{max} y a^s_{min} como el valor máximo y mínimo de a en S;
  Dibuja un punto de corte aleatorio a_c uniformemente en $[a^s_{min}, a^s_{max}]$;
  Devolver la partición $[a < a_c]$
end procedure

Entrada: un subconjunto S.
Salda: un booleano.

procedure $Stop\_split$(S)
  if $|S| < n_{min}$ **then**
    return TRUE;
  end if
  if todos los atributos son constantes en S **then**
    return $TRUE$;
  end if
  if la salida es constante en S **then**
    return TRUE;
  end if
  return FALSE;
end procedure

a minimizar los sesgos. La complejidad del procedimiento de crecimiento de árboles equilibrados es del orden de $N\ log(N)$ con respecto al aprendizaje del tamaño de la muestra, al igual que otros procedimientos de cultivo de árboles. Sin embargo, con la simplicidad del procedimiento de división de nodos, se espera que el factor constante sea mucho menor que en otros métodos basados en ensambles que optimizan localmente los puntos de corte.

2.6.3. J48 (C4.5)

EL algoritmo *J48* genera un árbol de decisión *C4.5* sin podar o podado. *J48* construye árboles de decisión a partir de un conjunto de datos de entrenamiento de la misma manera que lo hace el algoritmo *ID3*, utilizando el concepto de entropía de la información. Los datos de formación son un grupo $D = d_1, d_2, ..., d_n$ de ejemplos ya clasificados. en el algoritmo 6 cada ejemplo $a_i = a_1, a_2, ...a_n$ es un vector que representa los atributos o características. Los datos de entrenamiento se incrementan con un vector $D = d_1, d_2, ..., d_n$ que representa la clase a la que pertenece cada muestra.

Algoritmo 6 Pseudocódigo del árbol clasificador J48 (C4.5) [109].

Entrada: un conjunto de datos con atributos valuados D
Salida: El árbol de decisión

1: $Tree = \{\}$
2: **if** D es $pure$ OR $otro\ criterio\ de\ paro\ encontrado$ **then**
3: $\quad finalizado$
4: **end if**
5: **for all** attributo $a \in D$ **do**
6: $\quad$ *Computar los criterios de la teoría del información si se divide en a*
7: **end for**
8: a_{best} =Mejor atributo según los criterios calculados anteriormente
9: $Tree$ = Crear un nodo de decisión que pruebe a_{best} en la raíz
10: D_v = subconjuntos inducidos de D basado en a_{best}
11: **for all** D_v **do**
12: $\quad Tree_v = C4.5(D_v)$
13: $\quad$ Adjuntar $Tree_v$ a la rama correspondiente del árbol
14: **end for**
15: **return** Árbol

En cada nodo del árbol, J48 elige un atributo de los datos que divide más efectivamente el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el estandarizado para la adquisición de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia

de información normalizada se elige cómo parámetro de decisión. El algoritmo J48 se divide recursivamente en sublistas más pequeñas.

Este algoritmo tiene las siguientes características: todas las muestras de la lista pertenecen a la misma clase, cuando esto sucede, simplemente crea un nodo de hoja para el árbol de decisión diciendo que se elige esa clase; en este caso, ninguna de las características proporciona ninguna ganancia de información, J48 crea un nodo de decisión por encima del árbol usando el valor esperado de la clase; e instancia de la clase no vista anteriormente, una vez más, J48 crea un nodo de decisión más arriba en el árbol con el valor esperado. J48 incorpora un mecanismo de poda de árboles una vez que estos se han inducido basados en una hipótesis de la probabilidad de crecimiento del árbol en determinadas ramas. Con esto se evitan árboles demasiados grandes y complejos que ajusten demasiado los datos provocando un sobre entrenamiento en el modelo aprendido [109], [110].

2.6.4. SimpleCart (Classification and Regression Trees)

El algoritmo *SimpleCart* se basa en árboles de clasificación y regresión de *Breiman et al.* [111]. Un *árbol CART* es un árbol de decisión binario que se construye dividiendo un nodo en dos nodos hijo repetidamente, comenzando con el nodo raíz que contiene toda la muestra de aprendizaje [110]. La idea básica del crecimiento de árboles es elegir una división entre todas las posibles divisiones en cada nodo para que los nodos hijos resultantes sean los *"más puros"*. En este algoritmo, sólo se consideran las *particiones univariadas*, es decir, cada división depende del valor de una sola variable de predicción. Todas las particiones posibles incluyen las posibles particiones de cada predictor.

Si X es una variable categórica nominal de las categorías I, hay $2_{I-1} - 1$ posibles divisiones para este predictor. Si X es una variable categorial ordinal o continua con K valores diferentes, hay $K - 1$ particiones diferentes en X. Un árbol se cultiva a partir del nodo raíz utilizando repetidamente los siguientes pasos en cada nodo.

El algoritmo 7, primero encuentra la mejor división de cada predictor el cual puede ser continuo y ordinal; los valores son ordenados desde el más pequeño hasta el más grande. Para el predictor ordenado, revisa cada valor desde arriba para examinar cada punto de división del candidato (denominado v, si $x \leq v$, el caso va al nodo hijo izquierdo, de lo contrario, va a la derecha.) Cada predictor nominal examina cada subconjunto posible de categorías (llámese A, si $x \in A$, el caso va al nodo hijo de la izquierda, de lo

Algoritmo 7 Pseudocódigo del árbol clasificador CART.

Entrada: conjunto de datos de entrenamiento etiquetados $D = \{(x_i, y_i)\}_{i=1}^{N}$
salida: Árbol de clasificación o regresión.

```
FIT_TREE(0, D, node)
procedure FIT_TREE(depth,R,node)
    if la_tarea_es_clasificación then
        node.prediction := la etiqueta más común en R
    else
        node.prediction := media del vector de salida de los puntos de datos en R
    end if
    (i*, z*, R_L, R_R := SPLIT(R)
    if no se cumplen los criterios de división y paro then
        node.test := x_i^* < z^*
        node.left := FITTREE(depth + 1, R_L, node)
        node.right := FITTREE(depth + 1, R_R, node)
    end if
    return node
end procedure
```

contrario, va a la derecha), para encontrar la mejor división. Posteriormente, busca la mejor partición del nodo entre las mejores particiones encontradas en el inicio del paso, y elige la que maximice el criterio de partición. El nodo se divide utilizando la mejor división que se encuentra en el paso anterior, si no se cumplen las reglas de interrupción o paro.

2.7. Algoritmos de gran asertividad

Existen un grupo de algoritmos bien conocidos y documentados en libro [110], que han demostrado excelente desempeño en tareas de clasificación; sin embargo, no se ha encontrado alguno capaz de dar resultados de discriminación de clase altamente precisos para cualquier conjunto de datos. Los algoritmos paramétricos recientes utilizados para clasificar diferentes grupos basados características extraídas de datos de la marcha incluyen: *los K Vecinos más Cercanos*, *Redes Neuronales Multicapa Perceptrón y Máquinas de Soporte Vectorial*.

2.7.1. Máquinas de Soporte Vectorial (SVM)

Las Máquinas de Soporte Vectorial (SVM) son un conjunto de algoritmos de aprendizaje supervisado en la que una máquina traza cada característica como un punto en el espacio *n-dimensional* (donde n es el número de características) con el valor de cada característica siendo una coordenada particular basada en una opción seleccionada del *kernel*. La clasificación se realiza encontrando el *hiper-espacio* que mejor diferencie las dos clases. Los Vectores de Apoyo son simplemente las coordenadas de la observación individual. Con *SVM* la formación es relativamente fácil, no existe un óptimo local, se adapta bien a los datos de alta dimensión y el compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente [4], [112], [113].

2.7.2. Los K vecinos más cercanos (KNN)

El algoritmo de *los K vecinos más cercanos* (**KNN**, del inglés *K-Nearest Neighbors*) se basa en la similitud de las características, en donde las características que están fuera de la muestra se asemejan mucho al conjunto de entrenamiento y determinan cómo se clasifica un punto de datos dado; en la clasificación, un objeto se clasifica por el voto mayoritario de sus vecinos, y el objeto se asigna a la clase más común entre sus vecinos más cercanos. *KNN* es un algoritmo de aprendizaje no paramétrico y perezoso, lo que significa que no hace ninguna suposición sobre la distribución de datos subyacente, y que la fase de formación es rápida porque sólo se aproxima localmente y todo el cálculo se aplaza hasta la clasificación [114], [115].

El algoritmo **KNN** se basa en tres componentes principales: número de vecinos *K*, el tipo de distancia a implementarse para encontrar la distancia entre elementos (Euclidiana, Manhattan y Minkowski) y un método de búsqueda del vecino más cercano. El algoritmo **KNN** ha sido utilizado en diversos problemas de categorización de enfermedades [5], [113].

2.7.3. Redes neuronales multicapa (MLP)

La red neuronal multicapa de Perceptron (MLP) es un clasificador que utiliza la retropropagación para clasificar las instancias; generalmente tiene una capa de entrada, una capa de salida y una o más capas ocultas; los nodos de esta red son todos sigmoides; y

los números aleatorios se utilizan para establecer los pesos iniciales de las conexiones entre nodos, y también para reordenar los datos de entrenamiento.

Las capas ocultas de la red neural se configuran en $(atributos + clases)/2$ y el número de épocas de entrenamiento se fija en promedio 500. Este algoritmo ha sido implementado en la clasificación de otras enfermedades neurodegenerativas con buen desempeño [6], [116].

2.8. Comentarios finales

La marcha es el mecanismo por el cual un humano es capaz de desplazarse independientemente y autónoma; la pérdida de las habilidades motoras representa un decremento en la calidad de vida de cada individuo; por lo tanto, resulta necesario el desarrollo de investigaciones enfocados en tecnologías que permitan el análisis objetivo de las alteraciones motoras producidas por enfermedades neurodegenerativas que afectan progresivamente a los pacientes.

Las enfermedades de Huntington y Ataxias Hereditarias afectan de manera progresiva la marcha alterando los patrones del caminar conforme la enfermedad avanza y deterioran significativamente la habilidad de los pacientes para desplazarse; ambos grupos patológicos presentan un conjunto de alteraciones similares por lo que algunos especialistas suelen confundir estas patologías en sus comienzos.

La marcha ha sido objeto de estudio desde diferentes perspectivas principalmente desde el área médica. Tecnologías basadas en acelometría, goniometría, videometría, electromiografía, etc., han sido desarrolladas para asistir el análisis del desempeño con diferentes objetivos. Sin embargo, todas estas tecnologías son hechas a la medida y difíciles de implementar, así como carentes de portabilidad. Los sensores de movimiento son tecnologías que han tomado popularidad en el análisis de la marcha de pacientes sanos por su portabilidad, fácil uso y bajo costo; desarrollar métodos con estas tecnologías para cuantificar y analizar la marcha en pacientes con enfermedades neurodegenerativas puede ayudar al seguimiento de la progresión de la enfermedad.

El reconocimiento de patrones permite agrupar información que mantienen una relación entre sus diferentes valores; las técnicas actuales permiten predecir valores futuros (regresión) o indicar la pertenencia a un grupo (clasificación). Los algoritmos de ensambles tales como meta-clasificadores permiten implementar varios algoritmos en

las tareas de clasificación, lo que permite unir las fortalezas de diversos enfoques con la finalidad de mejorar los resultados finales. Encontrar la combinación de algoritmos que permita una alta precisión en la clasificación características de la marcha extraídas de datos obtenidos con sensores de movimiento es una de la motivación de este trabajo.

Capítulo 3

Trabajos relacionados

El diagnóstico de enfermedades neurodegenerativas se realiza mediante la observación médica directa de la forma de caminar de las personas. En los últimos años, la tecnología ha desempeñado un papel importante en el apoyo a los especialistas de la salud para realizar estos diagnósticos. Los trabajos recientes para el análisis de la marcha atáxica en pacientes con la enfermedad de Huntington y Ataxias Hereditarias basado en sensores de movimiento se dividen en dos tipos: aquellos enfocados a cuantificar las variables espacio-temporales del marcha para compararlas con la información de pacientes sanos, otro grupo de trabajo se ha enfocado a encontrar patrones de la marcha que permitan diferenciar entre grupos de pacientes con diversas patologías reconociendo automáticamente las características del caminar más representativas de cada grupo usando técnicas clasificación empleadas en minería de datos y/o aprendizaje automático.

Los sensores de movimiento se han implementado para cuantificar y evaluar las características de la marcha en pacientes con diversas patologías [27], [44], [117]; su uso incluye dispositivos con diferentes capacidades de medición, desde dispositivos con un solo sensor [118] a teléfonos inteligentes con múltiples tipos de sensores [119], [120]. Investigaciones recientes muestran que los sensores de los teléfonos inteligentes iPhone son lo suficientemente fiables y precisos para evaluar e identificar los patrones cinemáticos de la marcha en sujetos sanos [83], [84]. Los estudios relacionados con la evaluación de la marcha y la salud han demostrado la capacidad de los sensores del iPhone para adquirir con precisión parámetros cuantificados de la marcha con un nivel suficiente de consistencia, específicamente en la posición del tobillo, y de una manera cómoda, portátil y transportable [85], [121], [122].

Algunos trabajos recientes se han centrado en encontrar evidencia de alteraciones en

los patrones de marcha, y comparados con escalas médicas para establecer el progreso de las enfermedades neurodegenerativas (ND) como la enfermedad de Parkinson (PD); [23], [123]; enfermedad de Huntington (HD) [22], [124]; y las Ataxias Hereditarias (HA) [25], [35].

Los datos brutos procedentes directamente de los sensores deben prepararse para su procesamiento a fin de mejorar los resultados de la clasificación. Este pre-procesamiento de datos se ha realizado con uno de los siguientes enfoques: extracción de características de la marcha (cinemáticas) relacionadas con un solo paso o una secuencia de ellos, como la longitud, la frecuencia, la velocidad, la cadencia, etc. [123]-[125]; tratar los datos de los sensores como una señal digital (flujo de información de una fuente) y tomar los valores como frecuencia de la señal, frecuencia de muestreo, valores máximos y mínimos, etc. [123], [126]; y tomando directamente los ciclos de zancadas de los datos brutos [3], [38], establecer cuándo un sujeto pertenece a una patología utilizando algoritmos de clasificación.

3.1. Reconocimiento de patrones de la marcha en pacientes con la enfermedad de Huntington

La mayoría de los resultados publicados de trabajos recientes sobre la clasificación de pacientes con enfermedades neurodegenerativas se obtuvieron con datos de sensores de presión del conjunto de datos públicos *"PhysioNet"* [127] o han recolectado y construido su propio conjunto de datos privados.

Los resultados basados en el conjunto de datos públicos "PhysioNet", incluyen los de *Iram S. et al* en [128] que encontraron que el clasificador de *Quadratic Bayesian Normal* alcanzó una tasa de error general más baja del 65 % (23/40) comparado con otros algoritmos y sus mejores resultados en la clasificación de los pacientes con la enfermedad de Huntington fue del 50 % (5/10); *Banie M. et al* en [1] obtuvieron una precisión de 86.957 % de reconocimiento de todas las clases con el clasificador *Quadratic Bayesian Normal* y para los pacientes con la enfermedad de Huntington alcanzaron el 85.714 % con clasificadores de *Árboles de Decisión*. En [2], se obtuvo una precisión del 88.674 % en la clasificación de los pacientes con enfermedad de Huntington combinando diversos algoritmos meta-clasificadores: *RandomSubSpace & Bagging*, *Bagging & PART* y *CVParameterSelection & Bagging*.

Los resultados recientes obtenidos con los datos de cinco sensores de movimiento del conjunto de datos propietarios se pueden encontrar en [38], donde se obtuvo el 81.04 % al clasificar correctamente datos de diversos grupos patológicos. En ese trabajo, se logró reconocer el mayor número de pacientes según su patología, el resultado para los pacientes con HD fue del 78.78 % utilizando los meta-clasificadores *Logitboost & RandomSubspace*.

Mannini A. et al. en [3] utilizaron el clasificador *Modelos Ocultos Markov (HMM)* y *Máquina de Soporte Vectorial (SVM)*, con un conjunto de datos de sensores de movimiento colocados en la cintura y en los vástagos para alcanzar el 90.5 % de precisión en el reconocimiento de la marcha de todos los grupos; sin embargo, para los pacientes con Huntington el puntaje alcanzado fue más bajo (88.2 %).

TABLA 3.1: Resultados de los algoritmos de clasificación de la HD obtenidos en trabajos anteriores.

Fecha	Ref.	Dataset	Sensores	Algoritmos	HD	Exactitud
2011	[1]	Público	Sensores de Presión	Clasificadores	20	85.71 %
2012	[128]	Público	Sensores de Presión	Clasificadores	15	50.00 %
2014	[2]	Público	Sensores de Presión	Meta-Clasificadores	10	88.67 %
2015	[38]	Propietario	Acelerómetros	Meta-Clasificadores	13	78.78 %
2016	[3]	Propietario	Acelerómetro y Giroscopio	Clasificadores	17	88.20 %

Podemos observar en la Tabla 3.1, cómo los resultados obtenidos en estudios anteriores han ido mejorando continuamente dependiendo del conjunto de datos y de los algoritmos de tratamiento. En el último trabajo con sensores de presión (fila 3) se obtuvieron los mejores resultados con algoritmos meta-clasificadores. Los resultados con algoritmos de clasificación (fila 5) fueron mejores que los tratados con meta-clasificadores (fila 4), cuando se utilizó la información de entrada más enriquecida del conjunto de datos de los sensores de movimiento.

3.2. Reconocimiento de patrones de la marcha en las Ataxias Hereditarias

En trabajos recientes centrados en los datos de la marcha de enfermos con HA, el análisis de datos se ha realizado utilizando datos brutos de los sensores (series temporales originales) como entrada a los algoritmos de clasificación. [38], o extrayendo de los datos brutos las características de la marcha relacionadas con los parámetros espacio-temporales [5], estadísticas en el dominio del tiempo [6] y basado en el Modelo Markov Oculto (HMM) con dominio de frecuencia [4].

El reconocimiento de los patrones de marcha basados en sensores en pacientes con ataxias hereditarias ha sido abordado por algunos autores. Estos estudios incluyen pacientes con ataxia cerebelosa (CA), ataxia espinocerebelosa (SCA), ataxia de Friedreich (FA) y ataxia cerebelosa de inicio temprano en niños (EOA). Los algoritmos de aprendizaje automático (clasificadores) utilizados en las tareas de clasificación incluyen principalmente: Redes neuronales artificiales (ANN), máquinas vectoriales de soporte (SVM), los K Vecinos más Cercanos (KNN), Naive-Bayes classifier (NB) y Meta-clasificadores múltiples o de ensamblaje. [4], [6], [38].

Como podemos ver en el resumen de trabajos recientes en la Tabla 3.2 (última columna), la precisión media resultante es bastante similar, pasando de 75.78 % en [38] a 72.9 % sobre la población con ataxia de aparición temprana (EOA) en [4]. Esto es independientemente de que los datos utilizados en esos trabajos procedan de un número diferente de pacientes (columna 5), utilizando sensores diferentes en tipo, número y posición en el cuerpo (columna 3), y de que hayan sido procesados con diferentes algoritmos de clasificación (columna 4). De los resultados de la precisión media se desprende que el 25 % de los pacientes no fueron correctamente reconocidos, lo que indica que existe un amplio margen de mejora.

3.3. Áreas de oportunidades encontradas

Se observa que no existe un método estandarizado para la recolección y análisis de datos de la marcha de pacientes con las enfermedades neurodegenerativa, existe una diversidad de algoritmos implementados y la precisión de la discriminación en diversos grupos patológicos aún no supera el 90 % para cualquiera de las dos enfermedades.

TABLA 3.2: Reconocimiento del patrón de marcha basado en sensores aplicado a pacientes con AH en trabajos previos.

Fecha	Ref.	Sensores	Algoritmos	Pacientes	Precisión
2015	[5]	Sensores de presión	KNN	30 CA	73.3 %
2015	[38]	Cinco acelerómetros.	Meta-clasificadores.	22 SCA	75.8 %
2016	[6]	Un acelerómetro y giroscopio.	ANN Multicapa.	1 FA	74.0 %
2017	[4]	Tres acelerómetros y tres giroscopios.	SVM.	10 EOA	72.9 %

Por lo anterior, el objetivo de este trabajo es doble: la mejora de los resultados recientes en la clasificación de los pacientes con *HD versus HC* y *HA versus HC*; y a la vez se busca reducir el número de dispositivos para adquirir los datos de la marcha, así como identificar las características de la marcha implicadas en el proceso. Para lograr este objetivo se utiliza un conjunto de datos de la marcha obtenidos de pacientes mexicanos y sujetos sanos, utilizando dos sensores de movimiento que se pueden llevar puestos en los tobillos de cada uno de los sujetos.

Este trabajo representa el primer paso hacia la generación de herramientas tecnológicas enfocadas en la monitorización continua en tiempo real de la progresión a largo plazo de la enfermedad en los pacientes con HD y HA.

Capítulo 4

Metodología y herramientas

Esta investigación implica el muestreo, el procesamiento y la evaluación de datos para reconocer a los pacientes con HD mediante clasificadores. Después de recolectar los datos, trabajamos con la información en seis etapas representadas en la Figura. 4.1:

1. La preparación de los datos brutos que deben utilizarse se describe en los datos de la marcha de procesamiento (sec. 4.2);
2. La segmentación y extracción de zancadas establecidas (sec. 4.3);
3. Con los datos procesados, se extraen las características de cada paciente con el procedimiento de extracción de las características de la marcha (sec. 4.4);
4. La configuración del entrenamiento de los meta-clasificadores y la evaluación del rendimiento de los modelos se describen en la estrategia de evaluación del modelo (sec. 4.8);
5. El rendimiento de los clasificadores se compara en el protocolo de análisis de datos de la marcha (sec. 4.8); y
6. La comparación de los errores de clasificación y la selección del clasificador con mejor rendimiento se explican en la sección de Evaluación de los resultados de la clasificación (sec. 4.9).

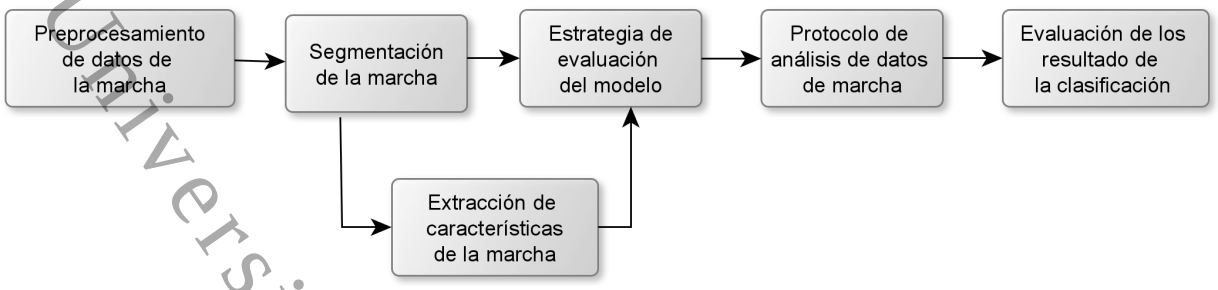


FIGURA 4.1: Esquema de recopilación y clasificación de datos de marcha.

4.1. Recolección de la información

4.1.1. Sujetos

Este estudio se realizó de acuerdo con el Instituto Nacional de Neurología y Neurocirugía "Manuel Velasco Suárez"(INNN-MVS) de la Ciudad de México, quienes estudian y tratan pacientes con enfermedades neuro-degenerativas con trastornos de la marcha como: enfermedad de Parkinson (PD), Esclerosis Múltiple (ELA), enfermedad de Huntington y diversas Ataxias Hereditarias.

El experimento involucró un total de 39 sujetos voluntarios del INNN-MVS: 11 pacientes con enfermedad de Huntington, 14 pacientes con Ataxias hereditarias y 14 sujetos sanos como control. Los pacientes habían sido diagnosticados con la enfermedad por los especialistas, y los controles eran personas sanas sin enfermedades neurodegenerativas existentes. La información fisiológica de los pacientes se detalla en la Tabla 4.1.

TABLA 4.1: Información de la población que participó en el estudio.

Variable	HD(n=11)	AH(n=14)	HC(n=14)
Edad (años, promedio, $\pm$ sd*)	48.8 ± 19.7	43.20 ± 23.06	51.13 ± 3.48
Sexo (Masculino, Femenino)	4, 3	7, 7	7, 7
Peso (kg, promedio$\pm$ sd*)	61.4 ± 9	58.58 ± 9.80	70.58 ± 12.30
Estatura (mts, promedio $\pm$ sd*)	1.62 ± 8.4	1.60 ± 0.11	1.64 ± 0.10

* Desviación estándar

4.1.2. Herramientas

Los datos de la marcha se recolectaron utilizando sensores de movimiento de dos teléfonos inteligentes *iPhone 5S*. El acelerómetro y giroscopio son de tres ejes, de $\pm$ 2g a $\pm$ 16 g de libertad, con un rango de voltaje de 1.6V a 3.6V [129]. Los iPhones fueron fijados en los tobillos de cada sujeto como se observa en la figura 4.2. Los dispositivos fueron configurados para auto calibrarse usando la aplicación *Compas* contenido dentro del sistema operativo del Smartphone (iOS). Se usó la aplicación *VibSensor* [130] con una tasa de muestreo para la recolección de 100Hz (aproximadamente diez muestras por segundo).

El procesamiento de los datos que requerían una mayor demanda de recursos de cómputo fueron realizados con una computadora *Dell* con un procesador Xenon Intel, 12 GB de memoria RAM, ejecutando el sistema operativo Linux (Fedora 25). Para realizar las tareas con menor demanda de recursos de procesamiento se usó una computadora HP con un procesador AMD-A10 (10 *compute cores* 4C+6G) , 12 GB de memoria RAM, ejecutando el sistema operativo Windows 10.

La limpieza, procesamiento, elaboración de gráficas relacionadas a la información se realizó con las siguientes herramientas: *R* (versión 3.1), *Python* (versión 3.5) y *Matlab* (versión R2016b). Las tareas de selección de características e implementación de algoritmos de clasificación fueron asistidas por la herramienta de minería de datos *Weka* (versión 3.8) [100]. La redacción de documentos relacionados a los resultados obtenidos se realizó con *Latex* (versión TexLive 2016).

4.1.3. Procedimiento de adquisición de datos

El INNN-MVS permitió habilitar un laboratorio de marcha en un espacio de 20 m de largo por 3 m de ancho; esto es lo suficientemente grande como para recolectar datos de la marcha y garantizar que los pacientes con trastornos de la marcha se desplacen cómodamente.

Los datos se recolectaron en un período de ocho días cuando los pacientes se sometieron a un chequeo médico. La adquisición de datos se planificó cuidadosamente para tener en cuenta a los pacientes con alteraciones de la marcha y trastornos motores (pérdida de equilibrio, marcha anormal, precisión y velocidad de los movimientos).

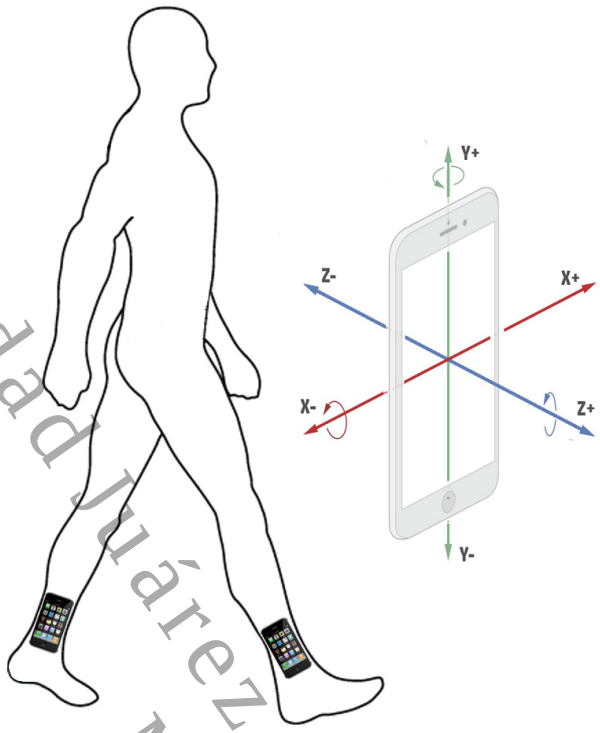


FIGURA 4.2: Colocación de los teléfonos inteligentes en los pacientes.

Los sensores de movimiento colocados en los tobillos de los sujetos en cada caminata fueron adecuados porque reducen las obstrucciones que pueden ocurrir al caminar, debido a la incomodidad de usar dispositivos y el grado de las alteraciones motoras presentadas por los pacientes. El personal médico del INNN apoyó en el momento de recopilación de datos de la marcha con el objetivo generar confianza en los pacientes y prevenir accidentes.

Los datos binarios se almacenaron en la nube mediante el dispositivo y se descargaron en la computadora una vez finalizada la adquisición de datos; se diseñó una carpeta por sujeto con sus archivos de datos para dar forma al Dataset de Gait.

4.2. Procesamiento de los datos

Los datos recolectados por los sensores de teléfonos inteligentes (iPhones) deben estar preparados para identificar adecuadamente la información de paso y la extracción de características; el pre-procesamiento se inspiró en los métodos implementados en

[131]-[133]. El procedimiento incluye: eliminar valores atípicos en la señal, la calibración de los datos, normalización, independencia de la orientación y suavizado de la señal.

4.2.1. Calibración de los datos

Los datos del acelerómetro se recolectan a una velocidad de muestreo variable a pesar de establecer una frecuencia de muestreo fija para la adquisición, esto es porque existe una pequeña diferencia de tiempo en que se realiza la llamada de registro de datos y el tiempo de grabación en el dispositivo; para calibrar los datos a un muestreo a intervalos de tiempo constante, se implementa *la interpolación Lineal*. Esta *interpolación* se obtiene del polinomio interpolador de Newton de primer grado, procedimiento que se utiliza para estimar valores del intervalo de una función en el que se conocen los valores entre los que se encuentra el punto buscado. La ecuación de *la interpolación lineal* es de la forma,

$$a\prime = a_j + \frac{(a_{j+1} - a_j)(t\prime - t_j)}{t_{j+1} - t_j} \tag{4.1}$$

donde a_j es la aceleración muestreada en el tiempo de muestreo t_j y $a\prime$ la nueva muestra calibrada en tiempo constante $t\prime$.

4.2.2. Normalización de los datos

Los acelerómetros están influenciados por la fuerza gravitacional, lo que significa que cuando los dispositivos están inmovilizados, los valores medios de aceleración del acelerómetro deben ser iguales a la constante gravitacional. Sin embargo, cuando el paciente camina con los dispositivos en los tobillos, estos están en constante movimiento en acorde al desplazamiento de los sujetos y la aceleración cambia constantemente registrando los valores de aceleración de los movimientos de las personas y otros valores lineales como la gravedad. Los efectos de constante de señal deben ser eliminados porque no forman parte de la representación de la marcha en la señal; para esto se usa la *ecuación 4.2 de normalización cero* sobre cada uno de los ejes, esta se describe como,

$$A_j(t) = a_j(t) - \mu_j, \quad j \in \{a_x, a_y, a_z\} \tag{4.2}$$

donde a_j es la aceleración recolectada en el tiempo t y μ_j es la aceleración promedio y $A_j(t)$ es la aceleración normalizada.

4.2.3. Independencia de orientación de los ejes

El sistema de coordenadas de los sensores tridimensionales se establecen en relación con los Smartphones. Sin embargo, cuando se recolectan los datos los valores obtenidos en cada eje del sensor son relativos a la posición y como el dispositivo está sujetado al cuerpo del sujeto; la serie de movimientos y la forma de desplazarse de cada individuo puede provocar que variaciones en la dirección de los ejes en la que se registra la información [134]. Para resolver los cambios de la orientación de ejes de los dispositivos se calcula la *magnitud de la aceleración* conocida como (L_2norm), la cual permite unificar en un solo vector las aceleraciones de cada uno de los ejes, se calcula con en la ecuación 4.3,

$$mag_A(i) = \sqrt{A_x(i)^2 + A_y(i)^2 + A_z(i)^2} \tag{4.3}$$

en donde A corresponde a la aceleración en cada eje *x*, *y* y z de cada muestra en el tiempo (i), y $mag_A(i)$ señal invariante obtenida.

4.2.4. Filtrado de los datos (suavizado)

El suavizado de datos se refiere a técnicas para eliminar ruidos o comportamientos no deseados en los datos, mediante la detección de valores atípicos identificado en puntos que son significativamente diferentes del resto de los datos. *El promedio móvil* (moving average) es una técnica común de suavizado de datos que desliza una ventana a lo largo de los datos, calculando la media de los puntos dentro de cada ventana. Este ayuda a reducir variaciones insignificantes de un punto de datos a otro. El suavizado de los datos por *promedio móvil* es un proceso equivalente al filtro de paso bajo con la respuesta del suavizado por su diferencia. Un filtro de promedio móvil suaviza los datos al reemplazar cada punto de datos con el promedio de los puntos de datos vecinos definidos dentro un lapso de la señal; la ecuación 4.4, el *promedio móvil* se define como:

$$y_s(i) = \frac{1}{2N+1}(y(i+N) + y(i+N-1) + ... + y(i-N)) \tag{4.4}$$

donde N es el número de puntos de datos vecinos a cada lado de $y_s(i)$, $2N + 1$ es el lapso y $y_s(i)$ como el valor suavizado para cada dato [135].

4.3. Segmentación de la marcha

Hollman et al. en [136] encontraron que *un mínimo de 10 zancadas contienen la información necesaria como medida suficiente para recolectar las características de la marcha*; así, la información suavizada obtenida la sección 4.2.4 se utiliza para identificar las 10 zancadas correspondientes (ciclos de marcha)

Ruy et al. [137]) identificaron que los cambios de la señal de aceleración de los acelerómetros (positiva a negativa), corresponden al contacto del talón con el suelo; se observó que, cuando la zancada comienza, la aceleración aumenta y cuando la zancada termina, la aceleración disminuye.

Basado en lo anterior se propuso identificar la información de la señal de los acelerómetros correspondientes a 10 zancadas del punto medio de la caminata, con un algoritmo que considere los valores mínimos y máximos de la magnitud de la señal suavizada para determinar el final de cada zancada y poder extraer la información necesaria.

4.4. Extracción de características

A partir de los segmentos de información extraídos en la sección 4.3 se identificaron las características representativas de la marcha con la herramienta de evaluación de la marcha asistida por ordenador (*IGAIT*) [138]. Esta herramienta proporciona una plataforma interactiva y fácil de usar para visualizar los datos de aceleración. *IGAIT* permite extraer veintiocho características: seis espacio-tiempo, quince relacionados con la frecuencia y siete de regularidad y simetría de paso. Las características extraídas se describen a continuación.

4.4.1. Raíz Media Cuadrática (RMS)

La Raíz Media Cuadrática (RMS) es una medida de la magnitud de una cantidad en un conjunto de datos. Para las mediciones de aceleración, indica la intensidad del movimiento. Los valores de la RMS se extraen en las tres direcciones de aceleración X, Y y Z (AP, VT, y ML), se calculan respectivamente utilizando Ecuación. 4.5,

$$RMS_d = \sqrt{\sum_{i=1}^{N}(x_{di} - \bar{x}_d)^2/N} \tag{4.5}$$

en donde x_{di} corresponde a la señal de un eje, $\bar{x}_d$ el promedio de la señal y N el total de datos.

4.4.2. Densidad Espectral de Potencia Integral (IPSD)

El análisis de la señal de frecuencia (análisis espectral) se utiliza para calcular la magnitud de la energía o la potencia de los movimientos correspondientes a la frecuencia con la que se repiten. Este análisis es importante para identificar grupos de personas con movimientos específicos y es independiente de la ubicación de los sensores de movimiento. Esto incluye la estimación de la Densidad Espectral de Potencia Integral (IPSD) usando Ecuación. 4.6 y 4.7;

$$PSD(e^{j\omega}) = \frac{1}{2\pi N}\left|\sum_{i=1}^{N} x_i e^{-j\omega i}\right|^2 \tag{4.6}$$

$$IPSD = \int_0^{\pi} PSD(\omega)d\omega, \qquad donde\ 0 \leq \omega \leq \pi \tag{4.7}$$

ω es la frecuencia angular, x_i es la aceleración en el eje VT, AP o ML, y N es el número total de muestras de aceleración.

4.4.3. Densidad Espectral de Potencia Acumulada (CPSD)

La frecuencia con el valor máximo de PSD es la frecuencia principal. Sin embargo, el análisis de la señal usa cierta cantidad de información pueden revelar valores que no

es posible encontrar cuando se usa el total de la información. La densidad espectral de potencia acumulada (CPSD) puede calcularse usando cierto porcentaje de la información. La CPSD se calcula con la ecuación 4.8 usando diferentes cuartiles (0.5, 0.75, 0.90 y 0.99) de potencia de la señal,

$$CPSD(\omega) = \int_0^{\omega} PSD(\omega) d\omega \tag{4.8}$$

aquí ω es la frecuencia angular de la potencia de la señal.

4.4.4. Coeficientes de Autocorrelación

El *coeficiente de autocorrelación* se refiere a la correlación de una serie temporal con sus propios valores pasados o futuros. Los *coeficientes de autocorrelación* no sesgados de los datos de aceleración son necesarios para escalar la regularidad y la simetría de la caminata. La estimación de los *coeficientes de autocorrelación* se determina automáticamente o se establecen manualmente por los límites máximos de aceleración para cada uno de los ejes de los acelerómetros con la ecuación 4.9),

$$f_c(t) = \frac{1}{N + |t|} \sum_{i=1}^{N-|t|} x_i x_{i+t} \tag{4.9}$$

donde $x_i (i = 1, 2, ..., N)$ es el valor de aceleración, $f_c(t)$ son coeficientes de autocorrelación. t es el tiempo de retardo ($t = -N, -N + 1, ..., 0, 1, 2, ...N$). Cuando el desfase t es igual a la periodicidad de la aceleración x_i, se encontrará un pico en la serie $f_c(t)$.

4.5. Selección de características

Las actividades para obtener un alto resultado en la clasificación se basan en elegir las mejores características que servirán de entrada tareas del reconocimiento de patrones. El objetivo de la selección de atributos es seleccionar el subconjunto más pequeño de atributos de tal manera que el porcentaje de clasificación no se vea afectado significativamente y la distribución de clases resultante sea lo más parecida posible a la original. Un atributo se considera relevante cuando no afecta al concepto de destino de ninguna manera; por lo tanto, es irrelevante o redundante.

4.5.1. Reducción del conjunto de características

Los métodos de selección de atributos reducen la dimensión del conjunto de atributos a un subconjunto de atributos relevante que mejoran la exactitud de la clasificación. Los métodos de selección pueden clasificarse en tres grupos clases debido a la forma en que se evalúan los atributos.

Los *métodos de filtrado* evalúan la relevancia de las características observando sólo las propiedades intrínsecas de los datos. Los métodos con este enfoque son independientes de los algoritmos de clasificación, son fácilmente escalables a conjuntos de datos de muy alta dimensión, son computacionalmente simples y rápidos. Los *métodos wrapper* utiliza un algoritmo de clasificación para medir la relevancia de los atributos; en general, sus resultados son mejores que los métodos de filtrado porque el proceso de selección de atributos está optimizado para el algoritmo de clasificación que se va a utilizar. Los métodos híbridos utilizan una combinación de los dos criterios de evaluación anteriores en diferentes etapas del proceso de búsqueda.

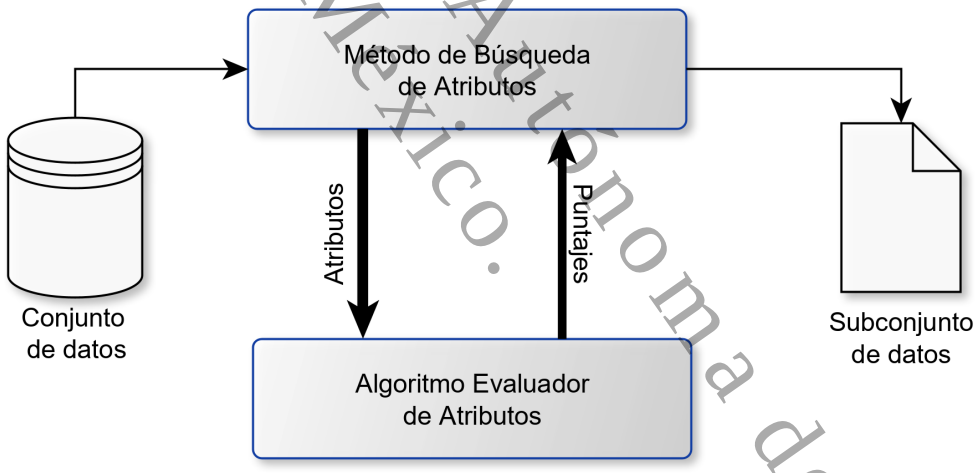


FIGURA 4.3: Procedimiento general de selección de atributos.

La primera parte de los experimentos fueron enfocados a reducir la cantidad de características que en comparaciones binarias: *HD vs HC* y *HA vs HC* por separado. Para reducir el número de atributos, se empleó una revisión exhaustiva de cada uno de los métodos mencionados anteriormente usando los algoritmos propuestos en la herramienta *Weka* [100] como se observa en la figura 4.3. Se introdujo todo el conjunto de características obtenida en el algoritmo usado como método de búsqueda, posteriormente se

le asignó un algoritmo evaluador de atributos, el cual devuelve las puntuaciones asignada a cada atributo, finalmente el método de búsqueda seleccionó las características con las mejores puntaciones para obtener un nuevo subconjunto de datos que servirá de entrada en las tareas de clasificación.

4.5.2. Identificación de características mínimas con la mejor precisión

Con el subconjunto de datos reducido obtenido en la sección 4.5, se procede a encontrar las características mínimas necesarias que permitan mantener una excelente exactitud en la clasificación binaria por grupos.

La idea general de la estrategia que se implementó se basa en el algoritmo *Hill Climbing*, el cual comienza con una solución arbitraria del problema y luego trata de encontrar una mejor solución variando gradualmente un solo elemento de la solución. Si el cambio produce una solución mejor, se realiza otro cambio incremental en la nueva solución, repitiendo este proceso hasta que no se puedan encontrar mejoras. Dado que esta técnica pertenece a la familia de búsqueda local, la idea es reducir el espacio de búsqueda para encontrar una solución óptima [139], [140]. Se usa una estrategia de selección hacia adelante para procesar listas reducidas de atributos para encontrar el número mínimo de atributos que permitan la mejor precisión del algoritmo de la siguiente manera:

1. Comenzamos con un conjunto vacío,
2. Aumentamos la cardinalidad en un atributo y todos los elementos de este subconjunto son evaluados con el algoritmo de aprendizaje correspondiente, y
3. Continuamos con la siguiente cardinalidad hasta que no haya mejora en la calidad del algoritmo de aprendizaje.

Después del último paso, se obtuvo un subconjunto reducido de características para cada algoritmo con las características más relevantes y el mejor rendimiento de precisión.

4.6. La clasificación

El reconocimiento de patrones permite descubrir propiedades del conjunto de datos de los sensores de movimiento que pertenecen a un grupo en particular; algoritmos de clasificación de minería de datos han sido implementados en estas tareas La clasificación puede incluir el uso de uno o más algoritmos a la vez con el uso de meta-clasificadores o mecanismos de ensamble de algoritmos. Nuestro experimento incluye procesos de clasificación de ambos enfoques.

4.6.1. Clasificación para obtener la exactitud promedio con Meta- clasificadores

Los meta-clasificadores usan las predicciones de clasificadores individuales en uno para mejorar la precisión de final. Un meta-clasificador implementa otro clasificador como base de aprendizaje para generar un conjunto de clasificadores; esto añade un paso de procesamiento adicional para evitar errores de los clasificadores anteriores, obteniendo mejores resultados en nuevas predicciones.

Los meta-clasificadores han demostrado un buen rendimiento en el reconocimiento de las características a partir de los datos sobre la marcha obtenidos de pacientes con la enfermedad de Huntington en trabajos anteriores [38]. Los árboles de decisión de aprendizaje utilizan un árbol de decisión para pasar de las observaciones sobre una variable de entrada a las conclusiones sobre el valor objetivo de la variable. Los modelos de árbol, en los que la variable objetivo puede tomar un conjunto discreto de valores, se denominan árboles de clasificación [102]. Los clasificadores de árboles se han implementado en gran cantidad de investigaciones como en [141]-[144] y los meta-clasificadores han mostrado un buen desempeño en el reconocimiento de varias enfermedades [145]-[148]. Por tal motivo proponemos árboles de clasificación como base de aprendizaje en diversos meta-clasificadores.

La estrategia de clasificación se observa en la figura 4.4. Cada meta-clasificador se le asigna un árbol de decisión como clasificador base y un número de clasificadores a implementar (iteraciones). Las iteraciones generan múltiples algoritmos de clasificación a partir del clasificador base; múltiples combinaciones de meta-clasificadores existentes en Weka se toman en cuenta para encontrar las combinaciones con mejores resultados.

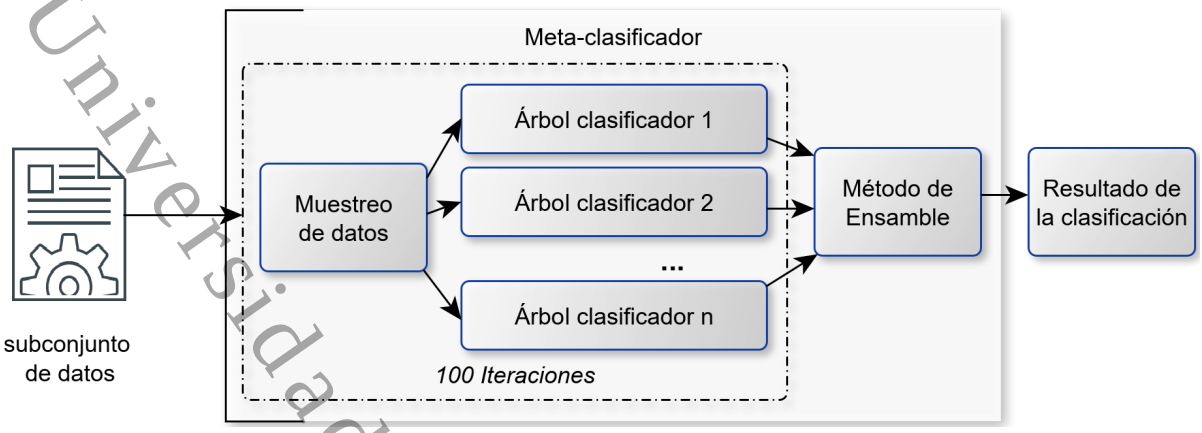


FIGURA 4.4: Proceso de clasificación

La identificación de la combinación de meta-clasificadores y los árboles de clasificación que muestren mejores resultados se realiza tomando en cuenta solamente el 20 % de los datos después del procesamiento sin considerar zancadas; discriminando aquellos que sobrepasen el 88.6 % de exactitud en la clasificación, los pares resultantes son usados en tareas de clasificación usando el conjunto de características, para seleccionar el par que tenga un mejor desempeño.

Posteriormente, las combinaciones de algoritmos con mejor desempeño se usan en tareas de clasificación para encontrar un número mínimo de atributos, los resultados obtenidos se comparan con el desempeño de algoritmos usando en los trabajos revisados en estado del arte.

4.7. Validación de los modelos aprendidos

La terminología utilizada en el aprendizaje automático para hablar de lo bien que un modelo de aprendizaje automático aprende y se generaliza a nuevos datos, es *Underfitting* y *Overfitting*. *Underfitting* se refiere a un modelo que no puede modelar los datos de entrenamiento ni generalizar a nuevos datos. *Underfitting* no es un modelo adecuado porque tendrá un rendimiento bajo en los datos de entrenamiento y validación, es fácilmente detectable y se puede corregir implementado nuevos métodos de tratamiento de la información y extracción de características, así como la implementación de algoritmos alternativos de selección de características y aprendizaje automático. El *Overfitting*

ocurre cuando los datos se ajustan demasiado a los modelos de entrenamiento mostrando un excelente rendimiento; sin embargo, en las pruebas que se realizan el desempeño del modelo es bajo con datos desconocidos por los modelos. La evaluación de la calidad del modelo permite saber hasta qué punto el modelo entrenado está preparado para predecir datos desconocidos; las técnicas de remuestreo permiten averiguar el *Overfitting* cuando se evalúan modelos aprendidos por cierto algoritmo.

La técnica más común de *evaluación es la validación cruzada (CV)*, el cual divide el conjunto de datos: en datos de entrenamiento y validación; con el primero se genera un modelo y con el segundo se realiza una prueba de rendimiento. Nosotros consideramos dos tipos de *validación cruzada*: *validación cruzada de k veces (k-folds CV)*, y *la validación dejando uno fuera (Leave-one-out CV, LOOCV)*, ambas de manera *estratificada* para obtener una distribución uniforme de los datos [96].

La *validación cruzada de K-veces* evalúa el modelo particionando los datos de entrenamiento en subconjuntos desarticulados de tamaño *k* del mismo tamaño $D_1, D_2, ..., D_k$; se utilizan subconjuntos de $1/k$ para las pruebas, mientras que los subconjuntos de $k - 1/k$ se utilizan para el entrenamiento. Este proceso se repite k veces usando cada subconjunto de $1/k$ como el conjunto de prueba; Los k resultados obtenido se promedian para obtener una sola estimación. En este sentido, el procedimiento de validación cruzada $k - veces$ virtualiza de forma efectiva los conjuntos de entrenamiento y validación o de pruebas. Los valores de $k = 5$ y $k = 10$ son especialmente los más utilizados [97]. En aplicaciones reales con acceso a un conjunto finito de muestras pequeñas, el concepto de validación cruzada proporciona una mayor precisión [98].

La *validación dejando uno fuera (LOOCV)* es un tipo especial de validación cruzada donde k es igual a la cantidad total de instancias. Lo anterior permite establecer cada instancia participe al menos una vez como elemento de prueba. En conjuntos de datos muy limitados LOOCV es mucho más precisos que otros métodos [149].

4.8. Evaluación y de comparación de modelos

La precisión de la clasificación se da en términos de varias medidas, las más comúnmente utilizadas para el análisis del rendimiento son tomadas en cuenta para asegurarnos que los resultados de clasificación son los más confiables.

4.8.1. Exactitud de la clasificación

La *exactitud* es una métrica que permite conocer la fracción de predicciones correctas que el modelo obtuvo; se considera una medida de rendimiento y se da en términos de porcentajes de las instancias correctamente clasificadas (ICC). La exactitud considera todas las instancias clasificadas y se calcula con la ecuación, 4.10

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.10}$$

en donde VP (verdaderos positivos) y VN (verdaderos negativos), corresponde a las instancias que fueron clasificadas dentro de la clase a la cual pertenecen; mientras que FP (falsos positivos) y FN (falsos negativos) representan las instancias que fueron asignadas en una clase diferente a la que pertenecen por el clasificador.

4.8.2. Matriz de confusión

La matriz de confusión es una técnica para resumir el rendimiento de un algoritmo de clasificación, esta muestra detalladamente la asignación a una clase particular de cada una de la instancias.

En una matriz de confusión C, tal que $C_{i,j}$ es igual a la cantidad de instancias participantes en la clasificación, i representa el grupo al que pertenece y j el grupo al que fue asignado por el clasificador. La tabla 4.2 muestra la estructura de la estructura de la matriz de confusión, las filas representan la cantidad de instancias por clases y las columnas las clases que asignó el clasificador a cada instancia.

TABLA 4.2: Definición de la matriz de confusión

a (0)	b (1)	<—	Clasificado como
$C_{0,0}$	$C_{0,1}$	a =	Valores reales para la clase a (0)
$C_{1,0}$	$C_{1,1}$	b =	Valores reales para la clase b (1)

La matriz de confusión se lee de la siguiente manera:

- $C_{0,0}$ son Verdaderos Positivos (**VP**), que corresponden a la cantidad de instancias que se predijeron correctamente dentro de la clase a la que pertenecen (instancias de la clase *a* clasificados dentro de *a*).

- $C_{0,1}$ son los Falsos Positivos (**FP**), que representan a la cantidad de instancias que pertenecen a la clase *a*, pero que fueron asignados erróneamente por el clasificador dentro de la clase *b*.
- $C_{1,1}$ son los Verdaderos Negativos (**VN**), que corresponden a la cantidad de instancias de la clase *b* que fueron ubicadas correctamente por el clasificador en la misma clase *b*.
- $C_{1,0}$ son los Falsos Negativos (**FN**), que corresponden a la cantidad de elementos de clase *b* que fueron ubicados erróneamente por clasificador en la clase *a*.

4.8.3. Tasa de Verdaderos Positivos y Negativos

La Tasa de Verdaderos Positivos ***(TVP)***, representa la probabilidad de que una instancia etiquetada como positiva resulte ser positiva en una prueba de clasificación; la **TVP** es referida como una métrica *sensibilidad* y como sinónimo de *exhaustividad*. La **TVP** toma en cuenta los verdaderos positivos (VP) y los falsos negativos (FN) que pertenecen a la misma clase. La ecuación 4.11 define a la **TVP** como,

$$TVP = \frac{VP}{VP + FN} \tag{4.11}$$

La Tasa de Verdaderos Negativos ***(TVN)***, representa la probabilidad de que una instancia etiquetada como negativas, resulte ser negativa en una prueba de clasificación; la **TVN** es referida como una métrica de *especificidad*; la cual, toma en cuenta los verdaderos negativos (VN) y los falsos positivos (FP) que pertenecen a la misma clase. La ecuación 4.12 define a la **TVN** como,

$$TVN = \frac{VN}{VN + FP} \tag{4.12}$$

4.8.4. Tasa de Falsos Positivos y Negativos

*La Tasa de Falsos Positivos (**TFP**)* corresponde a la probabilidad de que se produzca una falsa alerta; debido a que se obtiene un resultado positivo de una instancia que fue etiquetada como negativa. La **TFP** está representada en la ecuación 4.13 como:

$$TFP = \frac{FP}{FP + VN} \tag{4.13}$$

en donde VN son los verdaderos negativos y FP son los falsos positivos.

*La tasa de falsos negativos (**TFN**) o tasa de errores* corresponde a la probabilidad de que una instancia verdadera positiva (VP) no sea determinada como verdadera positiva por el clasificador. la **TFN** se calcula con la ecuación 4.14 como,

$$TFN = \frac{FN}{FN + VP} \tag{4.14}$$

donde *FN* es cantidad de falsos negativos y *VP* la cantidad de verdaderos positivos ($FN + VP$ es el número total de positivos).

La suma de la **TVP** y **TFP** para una clase dada es igual a unidad.

4.8.5. Precisión

La *precisión* es la relación que existe entre el número total de documentos recuperados que son relevantes y cantidad total de elementos recuperados. La *precisión* se expresa como una relación entre los verdaderos positivos (VP) y el número total de positivos que predice un modelo. La *precisión* puede entonces ser representada en términos de verdaderos positivos (VP y FP), como se muestra en la ecuación 4.15,

$$Precision = \frac{VP}{VP + FP} \tag{4.15}$$

en donde VP son los verdaderos positivos y FN son los falsos negativos.

4.8.6. Exhaustividad

*La Exhaustividad o **Recall*** una medida del rendimiento de la recuperación de información, que mantiene una relación entre el número total de elementos recuperados que son relevantes y el número total de elementos relevantes en un conjunto de datos. Esta medida es equivalente a la *TVP* y se define en la ecuación 4.16,

$$Recall = \frac{VP}{VP + FN} \tag{4.16}$$

en donde VP son los verdaderos positivos y FN los falsos negativos.

4.8.7. Medida F

La ***Medida-F*** o ***Measure-F*** es una medida del rendimiento de la recuperación de información; es la combinación equilibrada de la precisión y la exhaustividad usando una media armónica ponderada, como se muestra en la ecuación 4.17,

$$F - measure = 2 \times (\frac{Recall \times Precision}{Recall + Precision}) \tag{4.17}$$

4.8.8. Promedio Ponderado

Promedio ponderado utiliza ponderaciones proporcionales a las frecuencias de clase de los datos en el cálculo del promedio. El promedio ponderado de la *TVP* se ha utilizado para determinar el rendimiento del algoritmo, ya que representa el éxito del algoritmo para clasificar correctamente al miembro de cada clase.

4.8.9. Análisis de la curva ROC

La *curva de Característica Operativa del Receptor* (ROC, del inglés "Receiver Operating Characteristic") o simplemente *curva ROC*, es una representación gráfica de la sensibilidad contra la especificidad a medida que cambia el umbral de discriminación en un clasificador binario(valor a partir del cual se decide que un caso es positivo). La *sensibilidad* o *TVP* mide la proporción de positivos correctamente clasificados; la *especificidad* o *TVN* verdadera mide la proporción de negativos correctamente clasificados. La TVP se gráfica contra la TVN para visualizar el comportamiento de la clasificación del algoritmo.

Cuando un clasificador emite una puntuación proporcional de que una instancia pertenece a la clase positiva, la disminución del umbral aumentará la TVP y TVN. Los cambios de los umbrales de un valor máximo a un valor mínimo producen una curva lineal por partes desde el punto $(0, 0)$ hasta $(1, 1)$, donde el valor mínimo aceptable es una diagonal. Una curva con los valores en el eje y cercanos a la unidad indican un mejor desempeño, mientras que, una curva más cerca a la diagonal indican que el algoritmo esta adivinando al azar. la *curva ROC* permite visualizar un umbral de decisión que minimiza las tasas de clasificaciones erróneas en función de las clases y la distribución del costo; además permite comparar dos curvas de dos algoritmos para identificar

las regiones en que un clasificador es mejor que otro; así como también saber cuál tiene un peor desempeño.

El *área bajo la curva ROC (AUC, del inglés "the area under the curve")* es un valor estadístico de relevancia, debido a que la curva se encuentra un cuadrado (1×1), se tiene que $0 \leq AUC \leq 1$. Si el *AUC* es equivalente a la unidad, el clasificador asigna mayor puntuación a los positivos que a los negativos, en caso de el AUC sea cero ocurre lo contrario. Una *AUC* de 0.5 indica que el clasificador asignó la misma puntuación a los valores positivos y negativos, o bien, que las clases tienen una distribución similar de los datos; lo anterior, significa que clasificador trabajo manera aleatoria.

La interpretación de cada punto del *AUC* representa la probabilidad de que un valor positivo escogido al azar reciba una puntuación más alta que un negativo escogido al azar. El total de la puntuación obtenida por el AUC es equivalente a una versión normalizada de la prueba de la *suma de rangos de Wilcoxon-Mann-Whitney*, que prueba la hipótesis nula de dos muestras de mediciones ordinales que se extraen de una sola distribución.

4.8.10. Estadística Kappa

La *estadística Kappa o simplemente Kappa* se utiliza a menudo como una medida de fiabilidad entre dos clasificadores humanos. El primer "Calificador" refleja los datos verdaderos obtenidos (los valores reales de cada instancia a clasificar), a partir de los datos etiquetados del conjunto de datos, y el otro "Calificador" es el clasificador de aprendizaje automático utilizado para realizar la clasificación.

La estadística Kappa mide la concordancia entre la predicción del clasificador y el valor real de las instancias clasificadas; a menudo se utiliza como una medida de fiabilidad del clasificador. *Kappa* mide el acuerdo de predicción con la clase verdadera; un valor Kappa de 1 significa un acuerdo completo, mientras que el valor Kappa de 0 es comparable a "adivinanzas aleatorias"(similar al valor ROC=0.5). Esta métrica compara la *Exactitud Observada* (P_o) y la *Exactitud Esperada* (P_e), tal como se aprecia en la ecuación 4.18.

$$Kappa = \frac{(P_o - P_e)}{(1 - P_e)} \tag{4.18}$$

La *Exactitud Observada* (P_o) es simplemente el número de instancias que fueron clasificadas correctamente a lo largo de toda la matriz de confusión; es decir, es decir aquellas instancias que fueron etiquetadas como clase **a** y **b** y que fueron clasificadas como clase **a** y **b** respectivamente; La *Exactitud Observada* (P_o) se describe en la ecuación 4.19,

$$P_o = \frac{VP + VN}{N} \tag{4.19}$$

en donde VP son los verdaderos positivos, VN son los verdaderos negativos y N el número total de instancias.

La *Exactitud esperada*(P_e) se define como el valor esperado para cualquier clasificador aleatorio basado en la matriz de confusión. La *Exactitud Esperada* está directamente relacionada con el número de instancias de cada clase en conjunto con el número de instancias que el clasificador estuvo de acuerdo con la etiqueta real asignada. La *Exactitud Esperada* se calcula sumando los *valores marginales* de cada clase entre el número total de instancias. Los *valores marginales* representan el producto de instancias reales por las instancias clasificadas. La *Exactitud Esperada* se calcula con la ecuación 4.20,

$$P_e = \frac{\frac{(VP + FP) \times (VP + FN)}{N} + \frac{(VN + FN) \times (VN + FP)}{N}}{N} \tag{4.20}$$

en donde VP son los verdaderos positivos, VN son los verdaderos negativos, FP son los falsos positivos, FN son los falsos negativos y N el número total de instancias.

El valor de Kappa sugiere que un clasificador está adivinando al azar incluso cuando la *Precisión* y la *Exhaustividad* están cerca de la unidad. por lo tanto, es menos engañosa que una simplemente *Exactitud* (en caso de azar, una precisión observada del 80 % es mucho menos impresionante con una precisión esperada del 75 % frente a una precisión esperada del 50 %).

4.8.11. Coeficiente de correlación de Matthews

El *Coeficiente de correlación de Matthews (**CCM**)* es una medida de la calidad en las clasificaciones binarias incluso cuando los datos están desbalanceados. El *CCM* evalúa la correlación entre las instancias con sus etiquetas reales y el resultado obtenido por el

clasificador. Cuando el *CCM* se acerca a la unidad representa que existe una concordancia entre los valores reales y los obtenidos por el clasificador; un *CCM* igual a 0 representa una predicción al azar y un resultado negativo se relaciona con un desacuerdo entre los valores reales y la salida del clasificador. El *CCM* se conoce como el coeficiente Π, este se calcula usando la matriz de confusión con la ecuación 4.21,

$$CCM = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP) \times (VP + FN) \times (VN + FP) \times (VN + FN)}} \tag{4.21}$$

4.9. Análisis de los errores en la clasificación

Los índices de errores se utilizan para evaluar qué tan bien los resultados de la predicción satisfacen la distribución de los valores reales. El *Error Medio Absoluto* (**MAE**, del inglés *"Mean Absolute Error"*) mide la magnitud media de los errores en un conjunto de predicciones, sin considerar su magnitud; generalmente mide la precisión de las variables continuas. La *Raíz de Error Cuadrado Medio* (**RMSE**, del inglés "Root Mean Square Error") mide la magnitud media del error; ambos MAE y RMSE pueden utilizarse conjuntamente para diagnosticar la variación de los errores en un conjunto de pronósticos. Denotemos el valor real como α y el valor estimado usando algún algoritmo como $\hat{\alpha}$. Todas las estadísticas de error comparan los valores reales con sus estimaciones, pero lo hacen de una manera ligeramente diferente. Esto significa "qué tan lejos" están los valores estimados del valor real de α. Se puede observar que a veces se utilizan raíces cuadradas y a veces valores absolutos, esto se debe a que cuando se utilizan raíces cuadradas los valores extremos tienen más influencia en el resultado. Los indicadores asociados con el error de clasificación se pueden calcular con fórmulas 4.22 y 4.23.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{\alpha}_i - \alpha_i| \tag{4.22}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{\alpha}_i - \alpha_i)^2} \tag{4.23}$$

El *error absoluto relativo* (**RAE**, *del inglés "relative absolute error"*) indica cómo se relacionan los valores residuales del modelo (media) con los valores o la variabilidad de la

función destino (desviación media) en sí misma, dentro de una medida de rendimiento. El RAE debe ser inferior a 1 para cualquier modelo razonable, y preferiblemente cercana a 0. La ecuación 4.24 se define el RAE como,

$$RAE = \frac{\sum_{i=1}^{N} |\hat{\alpha}_i - \alpha_i|}{\sum_{i=1}^{N} |\bar{\alpha}_i - \alpha_i|} \tag{4.24}$$

La *Raíz Cuadrada del Error Relativo* (**RRSE**, *del inglés "Root Relative Square Error"*) es relativo a lo que hubiera sido si se hubiera usado un predictor simple. Por lo tanto, RRSE toma el *error cuadrático total* y lo normaliza dividiendo el error cuadrado total del predictor simple; la RRSE reduce el error a las mismas dimensiones que la cantidad predicha ([97]). La RRSE se define en la ecuación 4.25,

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{\alpha}_i - \alpha_i)^2}{\sum_{i=1}^{N} (\bar{\alpha}_i - \alpha_i)^2}} \tag{4.25}$$

El RMSE siempre será mayor o igual al MAE; cuanto mayor sea la diferencia entre ellos, mayor será la varianza en los errores individuales de la muestra. Si el RMSE=MAE, entonces todos los errores son de la misma magnitud. En RAE y RRSE esas diferencias se dividen por la variación de α por lo que tienen una escala de 0 a 1 y si multiplicamos este valor por 100 obtenemos similitudes en una escala de 0 a 100 (es decir, porcentaje). Los valores de $\sum(\bar{\alpha}_i - \alpha_i)^2$ o $\sum |\bar{\alpha}_i - \alpha_i|$ nos dicen, que tanto α es diferente de su valor medio. Se trata de cuánto α difiere de sí mismo (en comparación con la varianza). Por tal motivo, las medidas se denominan "relativas" porque proporcionan un resultado relacionado con la escala de α.

La correlación es la medida de cuánto α y $\hat{\alpha}$ están relacionados entre -1 y 1, donde 0 no es una relación, 1 es una relación lineal muy fuerte y -1 es una relación lineal inversa (es decir, los valores más grandes de α indican valores más pequeños de $\hat{\alpha}$, o viceversa).

Capítulo 5

Resultados

De acuerdo con el método propuesto dividimos los experimentos en seis momentos:

1. Preprocesamiento de la información con la finalidad de preparar los datos para utilización.
2. Identificación de un algoritmo para la segmentación de la marcha y la extracción de datos equivalente a 10 zancadas.
3. Extracción de características de la marcha basado en los datos obtenidos de los sensores.
4. Validar los datos brutos de los acelerómetros usando los algoritmos de clasificación al mismo tiempo que se seleccionan aquellos que presentaron mejores resultados.
5. Clasificación de las características de los enfermos con Huntington y Sujetos de control (HD vs HC).
6. Clasificación de las características de los enfermos con Ataxias Hereditarias y sujetos de control (HA vs HC) usando un mínimo de atributos.

5.1. Preprocesamiento de la información

Se seleccionaron los archivos binarios de los tobillos derecho e izquierdo de cada sujeto. Los datos sin procesar de cada archivo correspondientes al acelerómetro, giroscopio y el registro de tiempo fueron extraídos usando la herramienta CSVKit (python). Los datos del acelerómetro fueron graficados para identificar los valores atípicos en el inicio y

fin del registro de los datos y que no estaban relacionados a la información de marcha; estos datos fueron eliminados.

Los datos fueron adquiridos a un ritmo de muestreo de 100Hz (100 muestras por segundo); sin embargo, algunos de los registros en el acelerómetro se realizaron milésimas antes de los 10 microsegundos, mientras que los registros de giroscopio mostraban una desfase de tiempo de milésimas después de los 10 microsegundo. Por lo que se realizó la *calibración de los datos* a tiempos constantes de 10 microsegundos con la fórmula 4.1. Posteriormente se procedió a la eliminación de valores lineales como es la gravedad usando la *normalización cero* (ecuación 4.2).

La adquisición de los datos de los acelerómetros se realizó en tres direcciones relativas a la ubicación del dispositivo en tobillo de sujetos. Sin embargo, las alteraciones de marcha y los problemas de desplazamiento que presentaban cada uno de los datos recolectados de cada eje podría variar ligeramente a la dirección en la que fue fijada. Se eliminaron las *invariantes de la señal* con respecto a la ecuación 4.3 para poder fusionar los valores de las magnitudes de las aceleraciones de los ejes en un solo vector y así facilitar el proceso de segmentación de la marcha.

Finalmente, sobre vector obtenido en el paso anterior se aplicó el algoritmo de *promedio móvil* (ecuación 4.4) con la finalidad de suavizar los datos usando el valor del $span = 5$ puntos para definir la ventana de deslizamiento.

5.2. Algoritmo para la segmentación de la marcha

Investigaciones realizadas sobre marcha han identificado con un mínimo de 10 zancadas representan los patrones de la marcha de una persona [136]. Por lo tanto, se propuso identificar la información de los acelerómetros correspondientes a 10 zancadas del punto medio de la caminata.

Los cambios de aceleración de la señal en los acelerómetros (positiva a negativa) estudiados por *Ruy et al.* en [137] fueron la base para el diseño del algoritmo 8, el cual permite identificar el inicio de la zancada basado en los valores mínimos y máximos de la magnitud de la señal suavizada con la ayuda del algoritmo estándar *Peak to Peak* para determinar el inicio de cada zancada y poder extraer la información necesaria.

Los datos suavizados con el *promedio móvil* sirvieron de entrada para el algoritmo diseñado (línea 1). La función $findpeaks$ (línea 2) encuentra todos los picos máximos prominentes que tienen una importancia relativa de al menos la desviación estándar de todos los datos ($std(data)$); esta función devuelve dos arreglos: pks que contiene los picos encontrados y $locs$ que contiene las posiciones de los picos encontrados en el vector de datos mag_A. Los picos mínimos prominentes son buscados en la línea 3, usando el mismo procedimiento de la línea anterior, pero con los datos invertidos. Posteriormente con la función $findmax(pks_{min})$ se localiza el pico mínimo más prominente y se almacena en max_{peak}, este valor sirve inicio de la búsqueda de los picos mínimos prominentes que representa la conclusión de cada zancada. El vector $start$ (línea 6) controlado por la variable j (línea 5) son usados para almacenar los índices de cada pico mínimo prominente identificado como inicio de la zancada en el vector de datos. Las líneas 7-14 realizan la búsqueda de los picos mínimos más prominentes hacia adelante, los cuales representan la desaceleración relacionados directamente con el contacto del pie con el suelo, lo que es un indicador de que una zancada ha terminado; se establece la regla que cada pico mínimo prominente señalado como fin de la zancada debe ser mayor que la desviación estándar los picos contenidos en pks_{min} y que los picos mínimos anterior y posterior; además, el pico máximo prominente contiguo debe ser mayor que el pico máximo anterior, esto , en relación a que existe una mayor aceleración cuando se inicia una nueva zancada, mientras que, la finalización está relacionada con una disminución progresiva de la aceleración; en caso de que la regla se cumpla el índice de la localización del pico en el vector de datos se almacena en el vector $start$; este proceso se repite en sentido contrario. Los índices de los picos detectados son ordenados (línea 15) y se obtiene la cantidad de picos encontrados (línea 16). Los 10 índices de finalización de las zancadas son seleccionados en la línea 17 y se almacenan en el vector $data_{stride}$.

El inicio y fin del segmento de datos que corresponde a las 10 zancadas fueron extraídas y almacenados en diferentes archivos para su posterior implementación en la extracción de características.

5.3. Características extraídas de la marcha

La extracción de características se realizó con la herramienta *iGAIT* usando los datos en la segmentación de la marcha. *iGAIT* fue configurada indicando una frecuencia de muestreo de 10 microsegundos, con una distancia de 10 metros y un umbral de 0.4 para

Algoritmo 8 Procedimiento para encontrar el inicio de cada zancada

```
1: data ← mag_A
2: pks_max, locs_max ← findpeaks(data, std(data))
3: pks_min, locs_min ← findpeaks(−data, std(−data))
4: max_peak ← findmax(pks_min)
5: j ← 0
6: start(j) ← (0)
7: for i ← pos(max_peak) : length(pks_min) do
8:     if pks_min(i) ≥ std(pks_min) and
       (pks_min(i) ≥ pks_min(i + 1) and
       pks_min(i) ≥ pks_min(i − 1)) then
9:         if pks_max(i) ≥ pks_max(i + 1) and
           pks_max(i) ≥ pks_max(i − 1) then
10:            start(j) ← locs_min(i)
11:            j ← j + 1
12:        end if
13:    end if
14: end for
        ...
15: start ← sort(start)
16: length ← size(start)
17: data_stride ← data(start(length/2 − 5), start(length/2 + 5))
```

sujetos sanos y 0.3 para sujetos con una marcha irregular (fig. 5.1). Dado que se está analizando un patrón de marcha irregular, probamos diferentes valores umbral para encontrar el mejor de acuerdo con los resultados de detección del evento de marcha, como se sugiere en [138].

Los datos equivalentes a 10 zancadas extraídas con el algoritmo 8 fueron introducidos en la herramienta por cada sensor y por paciente Los coeficientes de autocorrelación son necesarios para escalar la regularidad y simetría de la marcha, por lo tanto, se determinaron automáticamente y manualmente para los casos en donde la información fuera irregular, señalando los límites máximos de aceleración para cada uno de los ejes de los acelerómetros.

Un total de 56 características fueron obtenida en este paso, los cuales fueron guardado en archivos individuales; por lo que se diseñó un método para obtener la información de manera automática de cada uno de los archivos de cada paciente en un solo conjunto de datos formado por un total de 39 instancias con 56 atributos.

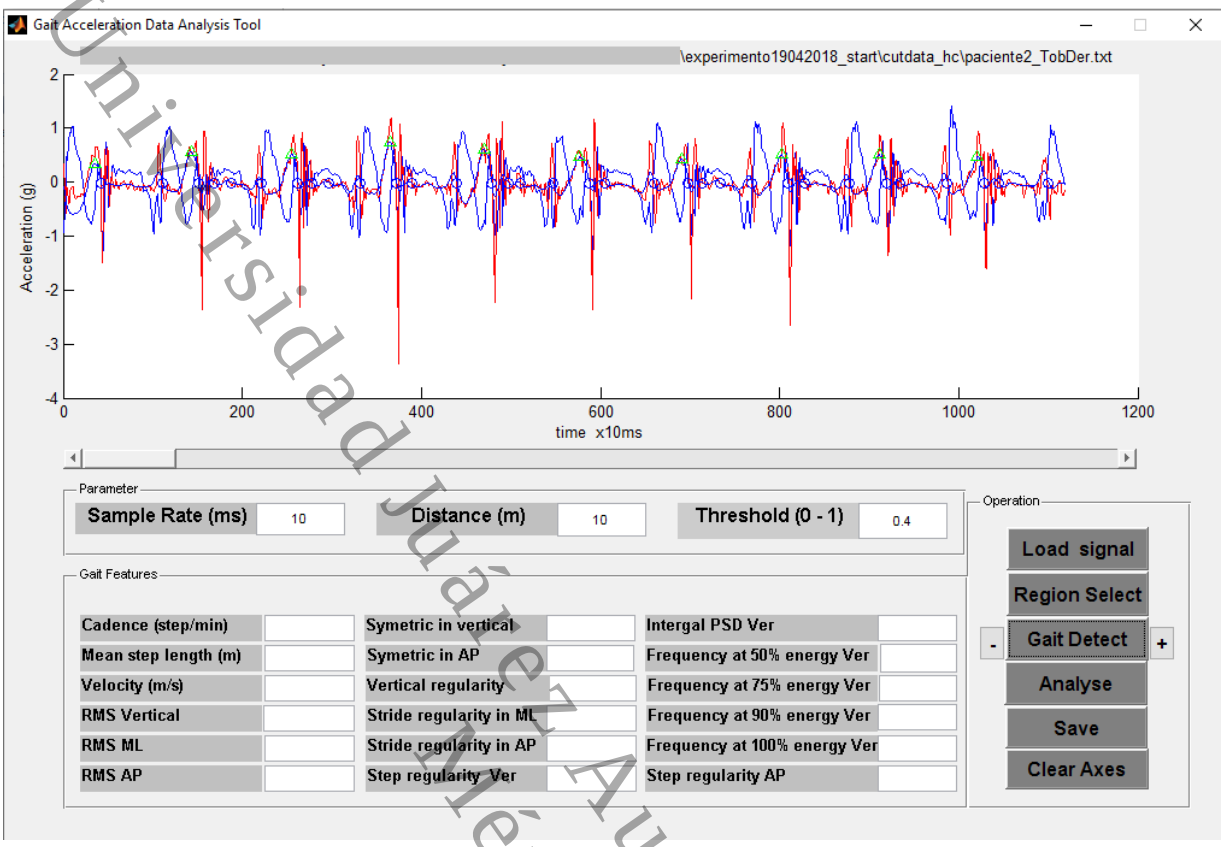


FIGURA 5.1: Configuración de la herramientas iGAit.

5.4. Reconocimientos de patrones de la marcha en enfermos con Huntington

Realizamos dos experimentos: el primero tiene por objeto mejorar los resultados de la clasificación a partir de los datos brutos de los sensores de movimiento de teléfonos inteligentes situados en los tobillos, mientras que el segundo tiene por objeto identificar las características de la marcha que intervienen en el proceso. Se tomaron la información de siete sujetos sanos y siete enfermos con HD, la información fisiológica de los pacientes se observa en la Tabla 5.1.

TABLA 5.1: Características de la población de estudio de HD vs HC.

Variable	**Pacientes(n=7)**	**Controles(n=7)**
Edad (años, promedio $\pm$ sd*)	48.8 ± 19.7	47.8 ± 11.4
Sexo (Masculino:Femenino)	4 : 3	3 : 4
Peso (kg, promedio$\pm$ sd*)	61.4 ± 9	62.4 ± 12.7
Estatura (cm, promedio$\pm$ sd*)	162.4 ± 8.4	162.7 ± 8.0

* Desviación estándar.

5.4.1. Procesamiento y segmentación de la marcha en Pacientes con HD

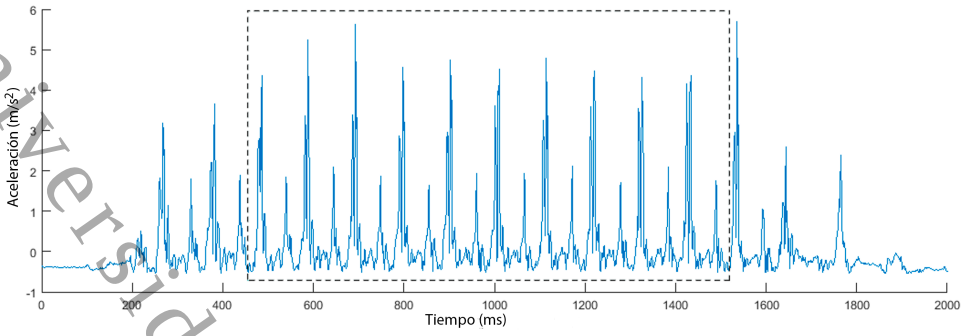
La Figura 5.2a muestra los datos del acelerómetro de un enfermo con Huntington tal como se recolectó con los teléfonos inteligentes, se puede observar una señal con valores insignificantes en los extremos y una señal bastante irregular. La Figura 5.2b muestra los datos del mismo enfermo después de aplicar el preprocesamiento y la segmentación de la marcha de la cual se toma una ventana de 10 pasos a partir de los datos de la marcha, se observa una señal más regular y se puede observar claramente el espacio de las 10 zancadas y la composición de cada una de ellas.

5.4.2. Selección de las características de la marcha a partir del conjunto de datos de la marcha HD

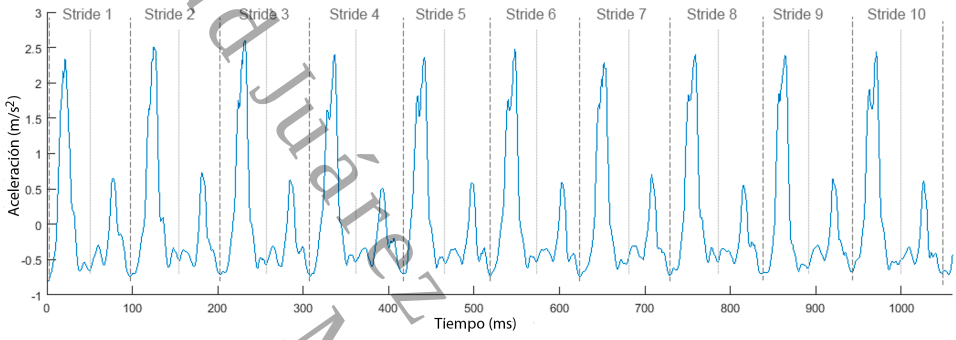
Las 28 características resultantes de *iGait* fueron introducidas en una serie de algoritmos de selección de atributos como se indica en la sección 4.4. La combinación de algoritmos *CfsSubsetEval & BestFirst* y *CfsSubsetEval & GreedyStepwise* seleccionaron las mismas 11 características de la marcha, estas características fueron probadas en los algoritmos con una validación cruzada estratificada de 10 veces. Los valores promedio y de desviación estándar de estas características para ambas clases se muestran en la Tabla 5.2.

Las características seleccionadas de la marcha fueron las siguientes:

- **La Raíz Media Cuadrática** (**RMS**, del inglés *Root Mean Squared*) de los 3 ejes para el sensor derecho, mientras que para el sensor izquierdo solo fueron seleccionados los RMS del eje Anterior-Posterior (AP) y Vertical (VER).



(A) Señal del acelerómetro antes del preprocesamiento.



(B) Ventana de 10 pasos obtenida tras aplicar la segmentación de pasos.

FIGURA 5.2: Resultados del protocolo de segmentación de pasos de un paciente con HD.

- **La frecuencia integral acumulada (IPSD)** en 75 % en el eje VER, en 90 % en el eje AP y 100 % en el eje ML para el sensor derecho; mientras que para cada sensor izquierdo solo seleccionaron los valores en 90 %.
- La **simetría en el eje VER** del sensor Izquierdo, la cual se estima como la relación armónica para cada uno de los tres ejes [150].
- La **Regularidad de la zancada en AP**, la cual se define como la auto covarianza normalizada para un desfase de tiempo en una zancada estimada [151]

5.4.3. Selección de los algoritmos y validación de los datos

Los meta-clasificadores han demostrado un buen rendimiento en el reconocimiento de las características de la marcha de las personas con enfermedades neurodegenerativas trabajos previos (Tabla 3.1).

TABLA 5.2: Características seleccionadas incluyen datos de los sensores izquierdo (L) y derecho (R).

Característica de la marcha	Control		Huntington	
	Promedio	Dev. Std.	Promedio	Dev. Std.
RMS in AP (R)	0.44	0.14	0.32	0.16
RMS in AP (L)	0.44	0.15	0.32	0.78
RMS in ML (R)	0.70	0.20	0.42	0.97
RMS in VER (R)	0.56	0.11	0.43	0.97
RMS in VER (L)	0.62	0.11	0.45	0.11
IPSD at 75 % in VER (R)	6.56	2.75	4.44	0.96
IPSD at 90 % in AP (R)	21.43	2.78	23.88	7.35
IPSD at 90 % in AP (L)	20.42	4.02	20.84	2.61
IPSD at 100 % in ML (R)	25.98	4.59	28.10	4.13
Symmetry in VER (L)	0.64	0.22	0.34	0.23
Stride regularity in AP (L)	-0.36	0.06	0.21	0.19

Todas las combinaciones posibles de los meta-clasificadores con los clasificadores de árboles listados dentro la Tabla 5.3, fueron probados con la herramienta "Weka" y los datos brutos de los acelerómetros con la finalidad de validar los datos y obtener las combinaciones con mejores resultados.

El proceso para encontrar las combinaciones de algoritmos que demuestren un mejor rendimiento se observa en la Figura 5.3, el proceso inicia con la configuración de una combinación de *meta-clasificador & clasificador*; posteriormente se realizó el proceso de entrenamiento y se compara la cantidad de *Instancias Correctamente Clasificadas (ICC)* con el umbral establecido (88.6 % según los trabajos relacionados), en caso de que la *ICC* no supere el umbral se cambia el árbol clasificador y se repite el proceso, si el umbral es superado la combinación es seleccionada y se repite el proceso en caso de que existan más árboles clasificadores por probar; en caso contrario, se cambia el meta-clasificador, en caso de que todos los meta-clasificadores propuestos hayan participado al menos una vez se termina el proceso y se listan las combinaciones seleccionadas.

Un total de 225 combinaciones de *Meta-clasificador & clasificador* con una validación cruzada de 10 veces fueron ejecutadas. El gráfico 5.4 muestra que los meta-clasificadores

TABLA 5.3: Lista de meta-clasificadores y algoritmos de clasificación de árboles.

Meta-clasificadores	**Árboles Clasificadores**
Decorate (Decrt)	ADTree
Bagging (Bagg)	DesicionStump (DcStp)
LogitBoost(LogBst)	ExtraTree(XtrTr)
RandomSubSpace (RSS)	FT
Multiboost (MltBst)	HoeffdingTree (HfdTr)
Ordinalclassclassifier (RdlClsf)	J48
RotationForest (RotFst)	J48Consolidate (J48Cnst)
Dagging (Dagg)	J48Graft
Grading (Gradg)	LADTree
RealAdaBoost (RlaBst)	LMT
Multiclassifier (MulClr)	NBTree
RandomCommittee (RanCom)	RandomForest (RndFrt)
CVParameterSelection (CVPS)	RandomTree (RndTr)
MultiBoostAB (MBstAB	REPTree (RpTr)
RecedIncrementalLogitBoost (RiLogbst)	SimpleCart (SmpCrt)

que mostraron un mejor desempeño fueron: *Logitboost*, *Randomcommittee* y *MultiboostAB*; los árboles clasificadores mejor clasificados fueron: R*andomForest*, *ExtraTree*, *SimpleCart* y *LMT*.

Cada combinación de *meta-clasificador & clasificador* fueron ejecutados indicando un número de modelos a construir (iteraciones) con la finalidad de obtener la mejor precisión. Las combinaciones de algoritmos excelente desempeño fueron: *LogitBoost & RandomForest* (45 iteraciones), *RandomCommittee & ExtraTree* , *MultiBoostAB & SimpleCart* y *MultiBoostAB & J48* (100 iteraciones). La Figura 5.5 muestra el comportamiento del rendimiento basados en la exactitud en los algoritmos al usar diferentes iteraciones.

La exactitud promedio de los resultados de la clasificación usando datos brutos se reporta en Tabla 5.4. Se observa que los resultados superan el 92 %, lo que equivale que de 14 sujetos participantes en el experimento sólo las instancias correspondientes a un sujetos fueron clasificados en una clase diferente a la que pertenecen. La combinación *LogitBoost & RandomForest* alcanza el porcentaje de instancias correctamente clasificadas (94.1102 %) y *MultiBoostAB & J48* el más bajo (93.7479 %); sin embargo, la diferencia es

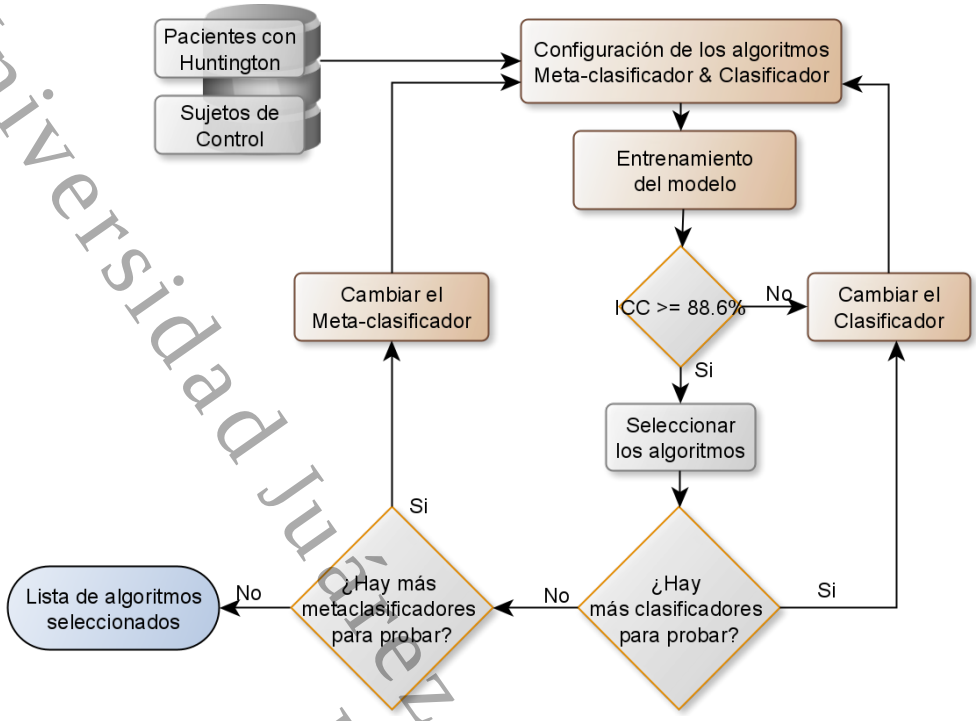


FIGURA 5.3: Proceso de selección de los algoritmos clasificadores.

de 0.6923 %, por lo que los resultados son muy competitivos.

TABLA 5.4: Resultados de la clasificación binaria entre sujetos enfermos y sanos utilizando datos brutos.

Meta-clasificador & Clasificador	**Exactitud**
LogitBoost & RandomForest	94.4402 %
RandomCommittee & ExtraTree	93.8688 %
MultiBoostAB & SimpleCart	93.8633 %
MultiBoostAB & J48	93.7479 %

5.4.4. Clasificación de enfermos con HD usando características

Las 11 características obtenidas en la sección 5.4.2 sirvieron como datos de entrada de los algoritmos que mostraron mejor desempeño en la sección 5.4.3, con la finalidad de averiguar que tan bueno es el rendimiento de clasificación usando características en comparación con los datos brutos.

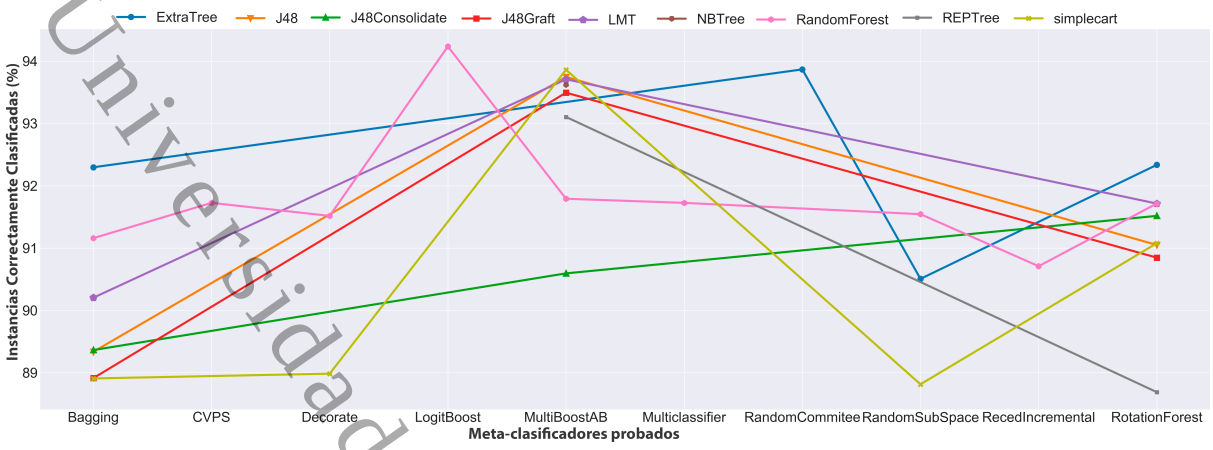


FIGURA 5.4: Comparación del rendimiento de los *meta-clasificador & clasificador* .

La validación del modelo para estos resultados se confirmó ejecutando el método *LOOCV*, sobre todo el conjunto de datos de la marcha; ejecutamos cada par *meta-clasificador & clasificador* 14 veces de forma estratificada; en cada fase 13 sujetos fueron utilizados para el entrenamiento y un sujeto que se sometió a prueba, de modo que cada sujeto fue utilizado como prueba sólo una vez.

5.4.4.1. Exactitud de la clasificación de pacientes con HD vs Sujetos de control

La exactitud promedio de la clasificación se reporta en la Tabla 5.5. Los clasificadores *LogitBoost & RandomForest* y *MultiBoostAB & J48* obtuvieron un resultado similar (92.8571 %), mientras que *RandomCommittee & ExtraTree* obtuvo un 85.7143 % y *MultiBoostAB & SimpleCart* exhibe el resultado con menor precisión (78.5714 %). Estos resultados demuestran la capacidad de los algoritmos para discriminar correctamente a los miembros de cada clase. Estos resultados son un criterio suficiente para la selección de algoritmos, pero no demuestra que se tenga un mejor rendimiento.

Los resultados reportados en las Tablas 5.4 y 5.5 de ambos experimentos fueron muy cercanos, con una diferencia máxima de aproximadamente el 1 % entre los algoritmos *LogitBoost & RandomForest* y *MultiBoostAB & J48* , excepto para los otros dos pares de algoritmos que su diferencia fue mayor para *RandomCommittee & ExtraTree* (8 % aprox.) y la diferencia más grande para *MultiBoostAB & SimpleCart* de hasta 15 %. Sin embargo el 1 % de diferencia entre los algoritmos con mejor resultado no es significativa en este caso, puesto que cada paciente representa 7.143 %. En cambio, *RandomCommittee*

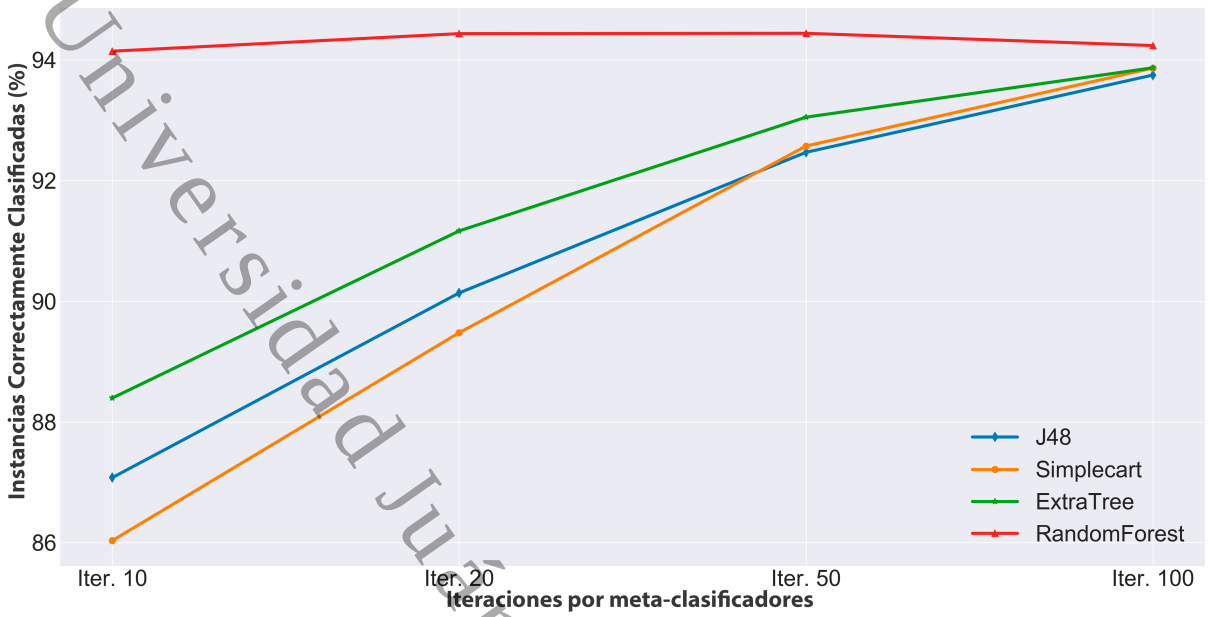


FIGURA 5.5: Comportamiento del desempeños de los algoritmos con las iteraciones

TABLA 5.5: Resultados de la clasificación binaria entre sujetos con HD y sanos utilizando 11 características de la marcha.

Meta-clasificador & Clasificador	**Exactitud**
Logitboost & RandomForest	92.8571 %
RandomCommittee & ExtraTree	85.7143 %
Multiboost-AB & SimpleCart	78.5714 %
Multiboost-AB & J48	92.8571 %

& ExtraTree obtuvo una diferencia de una instancia correctamente clasificada y para *MultiBoostAB & SimpleCart* fue de dos instancias correctamente clasificadas.

5.4.4.2. Matriz de confusión de HD vs HC

La Matriz de Confusión, las filas indican la clase real (datos reales), las columnas indican la salida del clasificador. En la matriz de confusión de la Tabla 5.6 las etiquetas representadas son: HC (controles sanos) y HD (Enfermos con Huntington). Por ejemplo, la Matriz de Confusión para los algoritmos *LogitBoost & RandomForest* se interpreta como: 8042 casos de Control se clasificaron correctamente como Control, mientras que

326 casos de Huntington se clasificaron erróneamente como Control; las instancias correctamente clasificadas se encuentran en la diagonal de arriba a la izquierda hacia abajo a la derecha, mientras que los errores del clasificador se observan en la diagonal de abajo a la derecha hacia arriba a la izquierda.

El número total de instancias en la Tabla 5.7 es de caso 8,200. Cada modelo clasifica un número ligeramente superior de casos de HD debido a la alteración de la marcha de los pacientes. Los valores correctamente clasificados son mucho más altos que los erróneos. El valor medio de la muestra por cada 10 pasos es: para HC es de 1,247; mientras que para HD es de 1,353 instancias. La clase HD obtiene el mayor número de instancias correctamente clasificadas para todos los algoritmos. *LogitBoost & RandomForest* obtuvieron los porcentajes más altos, ya que el 94.44 % de los casos se clasificaron correctamente.

TABLA 5.6: Matriz de confusión de meta-clasificadores con datos brutos.

Meta-clasificador & clasificador		**HC**	**HD**
Logitboost & RandomForest	HC	8042	686
	HD	326	9148
RandomCommittee & ExtraTree	HC	8068	660
	HD	456	9018
MultiBoost-AB & SimpleCart	HC	8102	626
	HD	491	8983
MultiBoost-AB & J48	HC	8090	638
	HD	500	8974

La Tabla 5.7 muestra la matriz de confusión de la clasificación usando 11 características de la marcha. El número total de instancias es de 14, (siete por cada clase). Los resultados de la clasificación son los mismos para los algoritmos *LogitBoost & RandomForest* y *MultiBoostAB & J48* con un caso de HC mal clasificado como HD. *RandomCommittee & ExtraTree* tiene dos casos clasificados erróneamente, uno de cada clase, y *MultiBoostAB & SimpleCart* clasificó mal tres casos de HD como HC.

5.4.4.3. Tasa de VP, tasa de FP, Precisión, Recall y la Medida-F

Los resultados de rendimiento de los algoritmos seleccionados se muestran en Tablas 5.8 y 5.9. Se puede notar que la precisión se da en términos de varias medidas, y tomamos aquí las más comúnmente utilizadas para el análisis de rendimiento. Podemos

TABLA 5.7: Matriz de confusión de meta-clasificadores para los datos de las características de la marcha.

Meta-clasificador & clasificador		**HC**	**HD**
Logitboost & RandomForest	HC	6	1
	HD	0	7
RandomCommittee & ExtraTree	HC	6	1
	HD	1	6
MultiBoost-AB & SimpleCart	HC	7	0
	HD	3	4
MultiBoost-AB & J48	HC	6	1
	HD	0	7

observar que los resultados obtenidos con las características de la marcha no son tan suaves como los obtenidos con los datos brutos, esto se debe a que los primeros están escalonados por el número de sujetos mal clasificados. En el caso del algoritmo *LogitBoost & RandomForest* , por ejemplo, el valor 1.0 en VP indica que todos los elementos de la clase HD fueron clasificados correctamente, mientras que para la clase HC el valor 0.85 indica que un sujeto fue clasificado erróneamente (no hay valores intermedios).

La *tasa de VP*, también conocida como **sensibilidad** muestran la frecuencia de instancias correctamente clasificadas. los resultados en general fueron superiores a 0.921 para los datos brutos y 0.571 para las características de la marcha; mientras que la tasa de falsos positivos, los resultados fueron inferiores a 0.079 para los datos brutos e inferiores a 0.429 para las características de la marcha. Es importante notar que en la Tabla 5.9 para los algoritmos *LogitBoost & RandomForest* , la clase HD tiene un resultado de la unidad *tasa VP*, lo que significa que todos los miembros de esa clase fueron clasificados correctamente, mientras que el resultado de la *tasa FP* fue 0.143 ya que un elemento de la clase HC fue erróneamente clasificado como HD.

La *precisión* mostró un valor superior a 0.93 para los datos brutos y superior a 0.70 para las características de la marcha; se observa una precisión mayor al trabajar con datos basados en características. *LogitBoost & RandomForest* muestran ligeras diferencias en ambas Tablas; mientras que *MultiBoostAB & J48* preserva su puntaje sobre los algoritmos *RandomCommittee & ExtraTree* y *MultiBoostAB & SimpleCart* . La *exhaustividad* mostró valores más altos al trabajar con datos brutos (tabla 5.8) que los valores de las tablas 5.9; se observa una similitud entre estos valores la *tasa VP*.

El promedio ponderado de la tasa de VP se ha utilizado para determinar el rendimiento

TABLA 5.8: Precisión detallada por clase con datos brutos.

Tasa TP	Tasa FP	Precisión	Recall	Medida-F	Clase
Logitboost & RandomForest					
0.921	0.034	0.961	0.921	0.941	HC
0.966	0.079	0.930	0.966	0.948	HD
0.944	0.057	0.945	0.944	0.944	Avg.
RandomCommittee & ExtraTree					
0.924	0.048	0.947	0.924	0.935	HC
0.952	0.076	0.932	0.952	0.942	HD
0.939	0.062	0.939	0.939	0.939	Avg.
MultiBoost-AB & SimpleCart					
0.928	0.052	0.943	0.928	0.936	HC
0.948	0.072	0.935	0.948	0.941	HD
0.939	0.062	0.939	0.939	0.939	Avg.
MultiBoost-AB & J48					
0.927	0.053	0.942	0.927	0.934	HC
0.947	0.073	0.934	0.947	0.94	HD
0.937	0.063	0.938	0.937	0.937	Avg.

del algoritmo, ya que representa el éxito del algoritmo para clasificar correctamente al miembro de cada clase. Todos los resultados de los algoritmos tienen valores muy cercanos para todas las medidas. Por ejemplo, el promedio ponderado de la tasa de VP, Precisión, Recall y la Medida-F tienen una distancia máxima de 0.07 para todos los algoritmos con datos brutos, lo que se traduce en lo bien adaptados que están los *meta-clasificador & clasificador* para la clasificación de los conjuntos de datos de la marcha en HD. Sin embargo, los resultados de la media ponderada de las características de la marcha están más separados entre sí, incluso si se trata de un solo sujeto de diferencia mal clasificado.

Los valores de precisión detallados se ordenan con valores más altos en la parte superior de la Tabla 5.8. Basado en el promedio ponderado, los valores de precisión más altos son para los algoritmos *LogitBoost & RandomForest*. Resultados similares se presentan en la tabla 5.5 para los algoritmos *LogitBoost & RandomForest*, sin embargo, *Multiboost-AB & J48* obtuvieron un resultado similar.

TABLA 5.9: Precisión detallada por clase con 11 características de marcha.

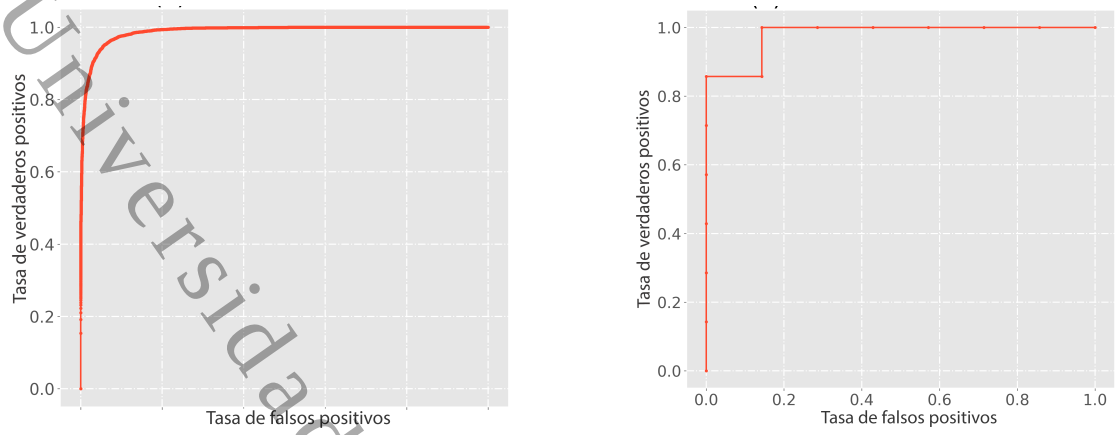
Tasa VP	Tasa FP	Precisión	Recall	Medida-F	Clase
Logitboost & RandomForest					
0.857	0.000	1.000	0.857	0.923	HC
1.000	0.143	0.875	1.000	0.933	HD
0.929	0.071	0.938	0.929	0.928	Avg.
RandomCommittee & ExtraTree					
0.857	0.143	0.857	0.857	0.857	HC
0.857	0.143	0.857	0.857	0.857	HD
0.857	0.143	0.857	0.857	0.857	Avg.
MultiBoost-AB & SimpleCart					
1.000	0.429	0.700	1.000	0.824	HC
0.571	0.000	1.000	0.571	0.727	HD
0.786	0.214	0.850	0.786	0.775	Avg.
MultiBoost-AB & J48					
0.857	0.000	1.000	0.857	0.923	HC
1.000	0.143	0.875	1.000	0.933	HD
0.929	0.071	0.938	0.929	0.928	Avg.

5.4.4.4. Análisis de la curva ROC

La *curva ROC* es una métrica de desempeño de un clasificador binario; en una clasificación binaria muestra el comportamiento del clasificador al cambiar los umbrales de decisión. Normalmente el umbral para dos clases es de 0.5. Un algoritmo que funciona mejor que el azar estará por encima de la diagonal (la línea de azar TVP=TFP).

EL comportamiento de la curva ROC para el algoritmo *LogitBoost & RandomForest* se muestra en la Figura 5.6; cada uno de los experimentos tiene una excelente predicción y no hay diferencias relevantes entre ellos. En ambos experimentos los algoritmos clasifican correctamente las instancias al cambio de umbrales. Lo anterior se puede identificar porque la curva se encuentra por encima de la diagonal, la curva esta más cerca de la unidad en el eje Y.

Los algoritmos presentan un mejor desempeño cuando el área de debajo de la curva se acerca a la unidad. La Tabla 5.10 muestra que el área ROC para el experimentos con datos brutos es mejor que usando las características en cada caso. Las diferencias entre



(A) Gráfico ROC usando Datos brutos. (B) Gráfica ROC usando características.

FIGURA 5.6: Gráfica ROC del proceso de clasificación de HD y HC.

la área para datos brutos es de tan solo 0.015 mientras para que las características es 0.1. En el experimento de clasificación usando características los algoritmos *LogitBoost & RandomForest* y *MultiBoostAB & J48* obtienen los mismos resultados en la exactitud, precisión, recall y la Medida-F, sin embargo, el área de la curva ROC es superior para *LogitBoost & RandomForest* , lo que demuestra un mejor rendimiento basados en esta métrica.

TABLA 5.10: Áreas ROC al clasificar HD vs HC.

Meta-clasificador & Clasificador	**Datos brutos**	**Características**
Logitboost & RandomForest	0.988	0.939
RandomCommittee & ExtraTree	0.985	0.898
Multiboost-AB & SimpleCart	0.979	0.816
Multiboost-AB & J48	0.974	0.929

5.4.4.5. Estadística Kappa

La *estadística Kappa* es calculada usando la Exactitud Observada (Po) y la Exactitud Esperada (Pe). La Exactitud Observada es simplemente el número de instancias que fueron clasificadas correctamente a lo largo de toda la matriz de confusión (Tablas 5.6 y 5.7) divido entre la cantidad de instancias participantes en la clasificación.

La Exactitud Esperada está directamente relacionada con el número de instancias de cada clase (Control y Huntington), junto con el número de instancias que el clasificador de aprendizaje automático acordó con el conjunto de datos. La Exactitud Esperada se calcula multiplicando la frecuencia marginal de ambas clases entre la cantidad de instancias totales.

En la matriz de confusión (Tabla 5.6), El clasificador *Logitboost & RandomForest* reconoció 8,042 instancias como Huntington de 9,148 instancias etiquetadas, por lo tanto la Exactitud Esperada es $(8,042 + 9,148)/18,202 = 0.9444$. Los valores marginales para la clase de control son de $4,012.521$ y para Huntington es de $5,118.523$, por lo tanto, la exactitud esperada es de $(4,012.521 + 5,118.523)/18,202 = 0.5016$; este resultado es debido a que los resultados de ambos clasificadores son muy cercanos. La estadística kappa se calcula con la ecuación 4.18 utilizando la Exactitud Observada y la Exactitud Esperada, lo que da un resultado de $(0.8870 - 0.5017)/(1 - 0.5017) = 0.8884$.

Los resultados de la estadística Kappa de todos los meta-clasificadores se muestran en la Tabla 5.11. Se observa que el experimento con datos brutos kappa es similar para todos algoritmos, siendo el mayor para *LogitBoost & RandomForest* . En el experimento con características, *LogitBoost & RandomForest* y *MultiBoostAB & J48* obtiene una puntuación más alta (0.8571) y *MultiBoostAB & SimpleCart* el más bajo (0.5714) muy cercano a 0.5 por lo que se asume los resultados son débiles.

TABLA 5.11: Estadística Kappa en los procesos de clasificación de HD vs HC.

Meta-clasificador & Clasificador	**Datos Brutos**	**Características**
LogitBoost & RandomForest	0.8884	0.8571
RandomCommittee & ExtraTree	0.8771	0.7143
Multiboost-AB & SimpleCart	0.8770	0.5714
Multiboost-AB & J48	0.8747	0.8571

5.4.4.6. El coeficiente de correlación de Matthews

El coeficiente de correlación de Matthews (CCM) indica que los valores cercanos a la unidad tienen una fuerte relación entre los valores reales en el conjunto de datos y los pronosticados. Los índices de correlación entre los valores de experimentos con datos brutos son muy similares (0.011 puntos de diferencia entre los valores extremos),

mientras que los valores de experimento con características tiene una diferencia relativamente mayor (0.234). Lo anterior, lo que significa una fuerte relación entre los valores reales en el conjunto de datos y los predichos por el algoritmo. Se observan valores muy parecidos, lo que significa una fuerte relación entre los valores reales en el conjunto de datos y la predicción del algoritmo.

Se observa una puntuación mayor para *MultiBoostAB & J48* a pesar de compartir resultados similares en otras medidas con los algoritmos *LogitBoost & RandomForest*. El pésimo resultado para esta medida lo obtuvieron los algoritmos *Multiboost-AB & SimpleCart*, una puntuación muy cercana a 0.5.

TABLA 5.12: Coeficiente de correlación de Matthews en los procesos de clasificación de HD vs HC.

Meta-clasificador & Clasificador	**Datos Brutos**	**Características**
LogitBoost & RandomForest	0.886	0.866
RandomCommittee & ExtraTree	0.877	0.714
Multiboost-AB & SimpleCart	0.877	0.632
Multiboost-AB & J48	0.875	0.929

5.4.5. Medidas de evaluación (MAE, RMSE, RAE, RRSE)

Las medidas de error son los indicadores de cuán bien los resultados de la predicción se ajustan a la distribución de los valores reales, cuanto mayor es la diferencia entre ellos, mayor es la varianza en los errores individuales. Valores resultantes obtenidos con fórmulas 4.22 y 4.23 muestran que los valores RMSE son siempre mayores que los valores MAE. La mayor diferencia entre RMSE y MAE es para el algoritmo *MultiBoostAB & J48* (0.184) con datos brutos y para el algoritmo *MultiBoostAB & SimpleCart* (0.234) con características de marcha. Las diferencias más pequeñas entre ellos fueron para el algoritmo *RandomCommittee & ExtraTree* en ambos casos, 0.39 con datos brutos y 0.49 con características de marcha.

La Tabla 5.13 muestra que los valores para los algoritmos *LogitBoost & RandomForest* son: MAE=0.05, que mide la precisión de las variables continuas y RMSE=0.22 la magnitud media del error. La diferencia entre ellos muestra una variación en la predicción de 0.16, lo que representa el valor más bajo de todos los clasificadores.

TABLA 5.13: Puntuación en MAE, RMSE, RAE y RRSE[1] para los datos brutos.

Meta-clasificador & Clasificador	MAE	RMSE	RAE %	RRSE %
LogitBoost & RandomForest	0.056	0.223	11.37	44.63
RandomCommi-ttee & ExtraTree	0.252	0.291	50.60	58.33
Multiboost-AB & SimpleCart	0.061	0.244	12.31	48.83
Multiboost-AB & J48	0.063	0.247	12.57	49.57

[1] MAE (Error absoluto medio), RMSE (Error de la raíz cuadrática), RAE (Error absoluto relativo), RRSE (Raíz cuadrática del error relativo).

Basándonos en estos indicadores podemos decir que el algoritmo de mejor rendimiento es *LogitBoost & RandomForest* que tuvo la mayor correlación y estimaciones de error más pequeñas. Podemos ver en las Tablas 5.13 y 5.14 que los resultados obtenidos para la exactitud, colocan este algoritmo como el mejor para predecir la enfermedad de Huntington en las personas que usan dispositivos sensores de movimiento en los tobillos.

TABLA 5.14: Puntuación en MAE, RMSE, RAE y RRSE para las características de la marcha.

Meta-clasificador & Clasificador	MAE	RMSE	RAE %	RRSE %
LogitBoost & RandomForest	0.074	0.267	13.85	50.14
RandomCommi-ttee & ExtraTree	0.329	0.378	61.74	70.86
Multiboost-AB & SimpleCart	0.233	0.467	43.60	87.15
Multiboost-AB & J48	0.071	0.267	13.39	50.11

5.4.6. Hallazgos y observaciones

Este primer experimento validó los métodos de pre-procesamiento de la información y se diseñó un algoritmo para la segmentación de la marcha y selección de datos del acelerómetro equivalente a diez zancadas. Los datos obtenidos sirvieron de base en la búsqueda de la combinación de algoritmos *meta-clasificador & clasificador* que permitieron un rendimiento superior a los obtenidos en los trabajos relacionados en donde solamente se usaron meta-clasificadores. También se validaron los pares de algoritmos *meta-clasificador & clasificador* usando datos brutos de los acelerómetros. Posteriormente

se evaluó el rendimiento de los algoritmos usando un conjunto de características. Para finalizar se compararon los resultados clasificadores de ambos experimentos.

El método de pre-procesamiento identificado fue excelente debido a que se obtuvieron resultados superiores a los publicados en los trabajos relacionados [131]-[133]. La técnica de suavizado de los datos fue ideal para el algoritmo de segmentación de la marcha, ya que la cantidad de pasos obtenidos coincidía con otros experimentos realizados y las características extraídas se adaptaron a los algoritmos implementados.

El principal hallazgo de este experimento fue la propuesta del uso combinado de algoritmos *meta-clasificador & clasificador* , los cuales demostraron una exactitud global de clasificación de 94.5 % con datos brutos y del 92.8 % con características de marcha, superior a los alcanzados con los meta-clasificadores *Bagging, RandomSubspace, and LogitBoost* en [2] con sensores de presión y en [38] con sensores de movimiento. La adición de árboles clasificadores usados en otros experimentos que incluyen enfermedades neurodegenerativas como en [152], [153], nos permitió obtener una mayor exactitud incluso superior la obtenida en [3], que hasta donde se ha reportado es el mejor rendimiento obtenido en la clasificación de enfermos con HD vs Sujetos sanos. Esto permite confirmar una de las suposiciones iniciales acerca de la combinación de diversos algoritmos produce mejores resultados que usar algoritmos individuales.

El segundo hallazgo fue que sólo dos sensores de movimiento (acelerómetro y giroscopio), uno por tobillo, fueron suficientes para obtener mejores resultados que los trabajos anteriores; a diferencia de [38] y [3] que utilizaron cinco y tres dispositivos respectivamente. El menor número de dispositivos permite que nuestra propuesta consuma menos recursos y reduce los efectos negativos al recolectar los datos en pacientes con problemas motores. El uso de varios tipos de sensores de movimiento (en un mismo dispositivo), arrojaron datos más significativos durante la marcha de los sujetos que cuando sólo se usaba un tipo de sensor, lo que llevó a una mejor caracterización del patrón de marcha.

El tercer hallazgo se relaciona con la fortaleza en el rendimiento de los algoritmos; estós producen una exactitud similar al trabajar con datos brutos y características de la marcha. La implementación de tareas de clasificación usando once características son suficientes para alcanzar una exactitud de 92 % (13 sujetos de 14) para reconocer pacientes con HD. La combinación de los algoritmos *LogitBoost & RandomForest* y *MultiBoostAB & J48* permiten reconocer todos los pacientes enfermos. Este hecho es importante porque en aplicaciones de monitorización continua es necesario minimizar la cantidad de

información para un buen desempeño.

Suponemos que los resultados de la clasificación de los sujetos con HD fueron superiores a la clasificación de los sanos HC; lo anterior, se debió a que estos pacientes tienen movimientos exagerados e incontrolados, que se reflejaron en los datos recolectados; como resultado los clasificadores fueron más eficientes en la discriminación de información entre ambos grupos.

5.5. Reconocimientos de patrones de la marcha en enfermos con Ataxias Hereditarias

En este apartado se usaron los datos de un grupo distinto de pacientes; sin embargo, el procesamiento de los datos brutos la segmentación de la marcha y el procedimiento de selección de características y métricas de evaluación del desempeño de los algoritmos se implementaron los discutidos en la sección de 4.1.2.

5.5.1. Sujetos participantes

Las muestras de datos fueron recolectados de pacientes del Instituto Nacional de Neurología y Neurocirugía "Manuel Velasco Suárez"(INNN-MVS) de la ciudad de México usando el protocolo establecido en la sección 4.1. Veintiocho sujetos de INNN-MVS participaron en el estudio, catorce pacientes con ataxias hereditarias (HA) y catorce sujetos sanos como Control (HC); los pacientes habían sido diagnosticados con la enfermedad de HA por los especialistas, y los controles eran personas sanas sin enfermedades neurodegenerativas existentes. La información fisiológica del paciente (edad, sexo, estado de salud, etc.) que aparece en la Tabla 5.15. El protocolo de adquisición de datos se planificó cuidadosamente para tener en cuenta el deterioro de la marcha del paciente y las alteraciones motoras (pérdida del equilibrio, marcha anormal, precisión y velocidad de los movimientos).

TABLA 5.15: Características de la población de estudio HA y HC.

Variable	Pacientes (n=14)	Controles (n=14)
Edad (años, promedio, ± sd*)	43.20 ± 23.06	51.13 ± 3.48
Sexo (Hombre:Mujer)	7:7	7:7
Peso (kg, promedio ± sd*)	58.58 ± 9.80	70.58 ± 12.30
Altura (cm, promedio ± sd*)	1.60 ± 0.11	1.64 ± 0.10

* Desviación estándar

5.5.2. Selección de características de Pacientes con HA

Los datos obtenidos con la segmentación de la marcha fueron introducidos en la herramienta *iGAIT* para extraer un total de 56 características de la marcha usando los parámetros establecidos. La reducción del número de atributos se basó en el algoritmo *Ranker* como método de búsqueda y probamos diversos algoritmos para obtener las características que produjeran la más exactitud en la clasificación.

Los algoritmos que produjeron un mejor resultado fueron los basados en correlaciones y clasificadores (J48) tomados del paquete de selección de atributos de *Weka* [100]. El algoritmo *CorrelationAttributeEval*, utiliza la correlación de *Pearson* entre atributo y clase, obteniendo una correlación global para un atributo nominal utilizando una media ponderada. El algoritmo *CorrelationAttributeEval* se ejecutó con la opción *-E weka.attributeSelection.CorrelationAttributeEval*. El algoritmo *ClassifierAttributeEval* evalúa cada atributo basándose en la precisión estimada por el clasificador seleccionado. Se ejecutó con la opción *-L* para evaluar cada atributo midiendo el impacto de dejarlo fuera del conjunto completo en lugar de considerar su valor de forma aislada, la opción *weka.classifiers.trees.J48* fue usada para seleccionar el algoritmo *J48* como aprendiz base y *-F 5 el número de pliegues de validación* cruzada a utilizar para estimar la precisión.

Los atributos con *rankeo* menores de cero fueron descartados, lo que permitió reducir de 56 a 21 atributos mejores puntuados. Ambos algoritmos proporcionaron una lista con atributos similares. Los 21 atributos fueron graficados usando un diagrama de bloques como se muestra en la Figura 5.7, los atributos que traslaparon sus valores entre las clases (izquierda) fueron pruebados para descartar ya que no aportaban al proceso de clasificación; caso contrario aquellos atributos que presentaban diferencias significativas fueron considerados como los que más aportaban a la discriminación de ambos

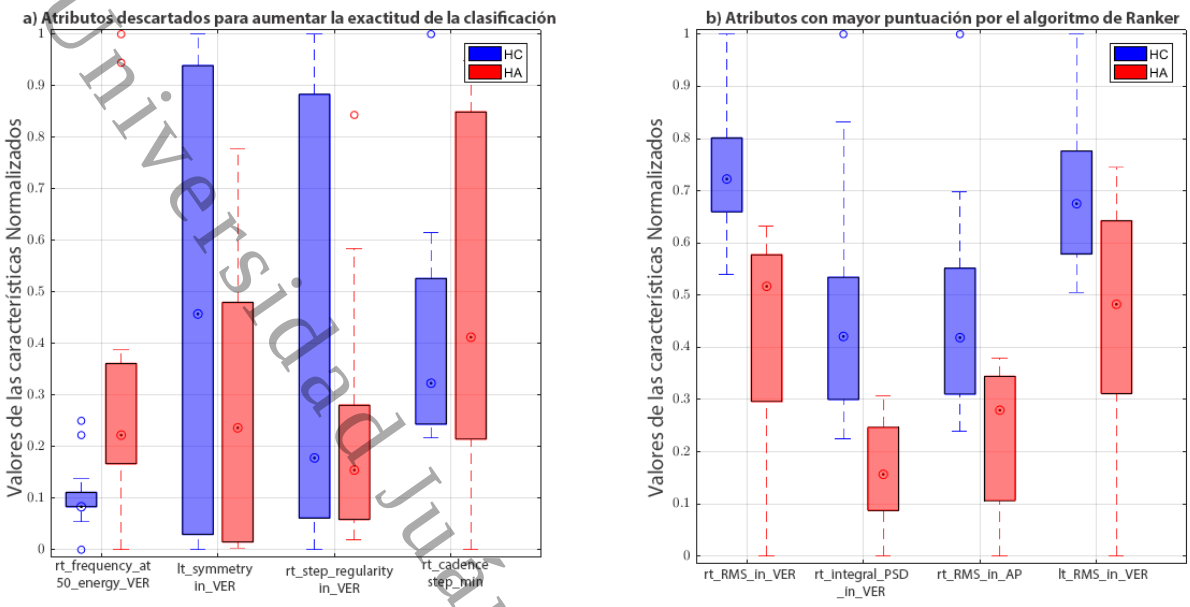


FIGURA 5.7: Comparación de las características obtenidas con los selectores de atributos.

grupos.

El procedimiento aplicado nos permitió observar que un total seis atributos presentaban un traslape entre los valores, y por lo tanto no aportaban al conjunto de datos, reduciendo la lista de características de la marcha a 15 como se muestra en la Tabla 5.16. La mayoría de las características seleccionadas se relacionaban con la raíz cuadrada media y la integral de la densidad espectral de ambos sensores.

5.5.3. Resultados de la clasificación de sujetos con Ataxias Hereditarias y Sanos

Las combinaciones de *meta-clasificador & clasificador* identificados en la sección 5.4.3 fueron implementadas con las características seleccionadas. Los meta clasificadores basados en regresión logística aditiva (LogiBoost), comité aleatorio (RandomCommittee) y multiboosting (MultiBoostAB) combinados con los árboles clasificadores: Bosques al azar (RandonForest), Árboles extremadamente Aleatorios (ExtraTree) y J48: fueron los que mejores resultados demostraron después de aplicar una CV 10 veces para averiguar el *overfitting* en las predicciones. Finalmente, se realizó una LOOCV estratificada de 28 veces, por cada pliegue se utilizaron 27 instancias para el entrenamiento y una

TABLA 5.16: Características seleccionadas para la tarea de clasificación.

No	Característica	Descripción
1	rt_velocity	Velocidad de marcha desde el tobillo derecho.
2	rt_RMS_VER	Raíz cuadrada media en el eje Z del tobillo derecho.
3	rt_RMS_ML	Raíz cuadrada media en el eje Y del tobillo derecho.
4	rt_RMS_AP	Raíz cuadrada media en el eje X del tobillo derecho.
5	rt_integral_PSD_VER	Espectral de potencia integral en el eje Z del tobillo derecho.
6	rt_integral_PSD_ML	Espectral de potencia integral en el eje Y del tobillo derecho.
7	rt_Integral_PSD_AP	Espectral de potencia integral en el eje X del tobillo derecho.
8	rt_step_regularity_AP	Paso regularidad estimada en el eje X del tobillo izquierdo.
9	lt_velocity	Velocidad de marcha desde el tobillo izquierdo.
10	lt_RMS_VER	Raíz cuadrada media en el eje Z del tobillo izquierdo.
11	lt_RMS_ML	Raíz cuadrada media en el eje Y del tobillo izquierdo.
12	lt_RMS_AP	Raíz cuadrada media en el eje X del tobillo izquierdo.
13	lt_integral_ PSD_VER	Espectral de potencia integral en el eje Z del tobillo izquierdo.
14	lt_integral_ PSD_ML	Espectral de potencia integral en el eje Y del tobillo izquierdo
15	lt_Integral_ PSD_AP	Espectral de potencia integral en el eje X del tobillo izquierdo.

instancia se dejó fuera para prueba, de modo que cada sujeto se utilizó como prueba sólo una vez.

5.5.3.1. Exactitud de la clasificación

Los resultados de la exactitud en la clasificación se reportan en la Tabla 5.17. Los resultados después de aplicar la validación cruzada muestra una diferencia de 3.47 % entre la *CV* de 10 veces y *LOOCV* lo equivalente a un sujeto; *LogitBoost & RandomForest* y *RandomCommittee & ExtraTree* incrementaron las precisiones cuando se aplicó el *LOOCV*; mientras que *MultiboostAB & J48* mantuvo su nivel de precisión.

Se observa que los algoritmos *RandomCommittee & ExtraTree* presentaron la exactitud más alta (96.86 %), mientras que *LogitBoost & RandomForest* una precisión baja (82.14 %), comparados con los resultados de la tabla 5.9 se puede apreciar los resultados son mejor para *LogitBoost & RandomForest* que *RandomCommittee & ExtraTree* , lo que permite afirmar que el primer par de algoritmos es mejor para clasificar pacientes con HD y el segundo trabaja mejor las características de pacientes con HA.

TABLA 5.17: Exactitud de los algoritmos en la clasificación binaria mediante el uso de meta & clasificadores al clasificar HA.

Algoritmos	10-fold CV	LOOCV
Logitboost & RandomForest	78.57 %	82.14 %
RandomCommittee & ExtraTree	92.86 %	96.86 %
Multiboost-AB & J48	89.29 %	89.29 %

5.5.3.2. Matriz de confusión

En la matriz de confusión en la Tabla 5.18, las filas indican la clase verdadera (datos reales) y las columnas indican la salida del clasificador, aquí Sujetos de control tenían la etiqueta HC y enfermos con Ataxias Hereditarias tenían HA. La *Matriz de confusión* muestra que el algoritmo *LogitBoost & RandomForest* obtuvo 12 controles saludables clasificados correctamente y dos mal clasificados; 11 pacientes con Ataxias Hereditarias fueron clasificados correctamente y tres mal clasificados como controles saludables; este algoritmo clasificó 23 instancias correctamente (82.14 %) y cinco pacientes fueron clasificados en una clase diferente a la que pertenecían (17.86 %). *RandomCommittee &*

ExtraTree clasificó correctamente todos los pacientes sanos y solamente un sujeto enfermo con HA fue clasificado como Control (HC) lo que indica un margen de 3.47 % de error en la clasificación. *MultiBoostAB & J48* clasifica de misma manera que *LogitBoost & RandomForest* a los sujetos sanos y los sujetos enfermos son clasificados al igual que *RandomCommittee & ExtraTree* , lo que equivale a una tasas de error del 10.37 %.

Los valores de las instancias correctamente clasificadas son mucho más altos que los de las erróneas en todos los algoritmos; *RandomCommittee & ExtraTree* tienen más instancias correctamente clasificadas y *Logitboost & RandomForest* demuestra menos rendimiento. Los algoritmos fueron más efectivos al reconocer las características de las personas sanas que las de los enfermos; con una mínima diferencia de una instancia.

TABLA 5.18: Matriz de confusión de meta-clasificadores para los datos de las características de la marcha de pacientes Atáxicos

Meta-Clasificadores		**HC**	**HA**
Logitboost & RandomForest	HC	12	2
	HA	3	11
RandomCommittee & ExtraTree	HC	14	0
	HA	1	13
MultiBoost-AB & J48	HC	12	2
	HA	1	13

5.5.3.3. Tasa de VP, Tasa de FP, Precisión, Recall y la Medida-F

La *tasa de VP* de cada clase indica las veces que el clasificador estuvo de acuerdo con las etiquetas de cada instancia, mientras que la *tasa de FP* se refiere a las veces que fue incapaz de reconocer correctamente las instancias. La *tasa de VP* de una clase se complementa con la *tasa de FP* de la clase contraria, la suma de ambas debe ser la unidad, lo cual indica que todos los elementos han sido correctamente clasificados. En la Tabla 5.19 *RandomCommittee & ExtraTree* tienen el valor más alto para la tasa de VP en ambas clases (1.0 y 0.929). *LogitBoost & RandomForest* y *MultiBoostAB & J48* fueron capaces de reconocer pacientes sanos en la misma medida (0.857); sin embargo, *MultiBoostAB & J48* tuvo una mejor TVP el reconocimiento de HA (0.929) que *LogitBoost & RandomForest* (0.786). El promedio ponderado de la *tasa de VP* demuestra un rendimiento superior para *RandomCommittee & ExtraTree* (0,964) superior a los demás algoritmos y muy cercano a la exactitud (tabla 5.17).

La *Precisión* y el *Recall* están relacionados con la probabilidad de que el algoritmo pueda reconocer correctamente una instancia respectivamente; *RandomCommittee & ExtraTree* y *MultiBoostAB & J48* tienen la mejor precisión para la etiqueta *HC*; sin embargo, el promedio de la *Medida-F* es superior a *RandomCommittee & ExtraTree* (0.964), esta medición debe estar de acuerdo con la exactitud (Tabla 5.17). Basado en el promedio ponderado, *RandomCommittee & ExtraTree* presenta un desempeño superior al 0.9 , mientras que los demás algoritmos presentan un promedio menor.

TABLA 5.19: Precisión detallada por clase en la clasificación de HC y HA.

TVP	**TFP**	**Precisión**	**Recall**	**Medida-F**	**Clase**
Logitboost & RandomForest					
0.857	0.214	0.800	0.857	0.828	HC
0.786	0.143	0.846	0.786	0.815	HA
0.821	0.179	0.823	0.821	0.821	Avg.
RandomCommittee & ExtraTree					
1.000	0.071	0.933	1.000	0.966	HC
0.929	0.000	1.000	0.929	0.963	HA
0.964	0.036	0.967	0.964	0.964	Avg.
MultiBoost-AB & J48					
0.857	0.071	0.923	0.857	0.889	HC
0.929	0.143	0.867	0.929	0.897	HA
0.893	0.107	0.895	0.893	0.893	Avg.

5.5.3.4. Medición de área en la curva ROC

El *gráfico ROC* muestra el desempeño de los algoritmos a medida que cambian los umbrales de discriminación; con un umbral mayor existe la probabilidad de tener una mayor tasa de FP, un umbral menor aumentará la tasa de VP. El gráfico 5.8, muestra la relación entre la *tasa de VP* (sensibilidad) y la *tasa de FP*(especificidad). Se observa que los algoritmos *RandomCommittee & ExtraTree* y *MultiBoostAB & J48* tienen un desempeño igual cuando se tiene un umbral bajo, mientras que *LogitBoost & RandomForest* inicia con un desempeño deficiente; sin embargo, cuando la *tasa de VP* supera el 0.9 los algoritmos tienen un desempeño similar; posteriormente *RandomCommittee &*

ExtraTree tiene un rendimiento superior; con la *tasa de FP* cerca del 0.35 *LogitBoost & RandomForest* supera en desempeño a *MultiBoostAB & J48* .

El área bajo la curva (AUC) indica el desempeño de una clasificación binaria a medida que cambia el umbral de probabilidad en una curva ROC, los valores cercanos a la unidad indican un excelente desempeño; mientras que valor por debajo del 0.5 indica un desempeño deficiente. El *área ROC* es mejor para *RandomCommittee & ExtraTree* (0.986) y *LogitBoost & RandomForest* tiene el más bajo desempeño (0.936).

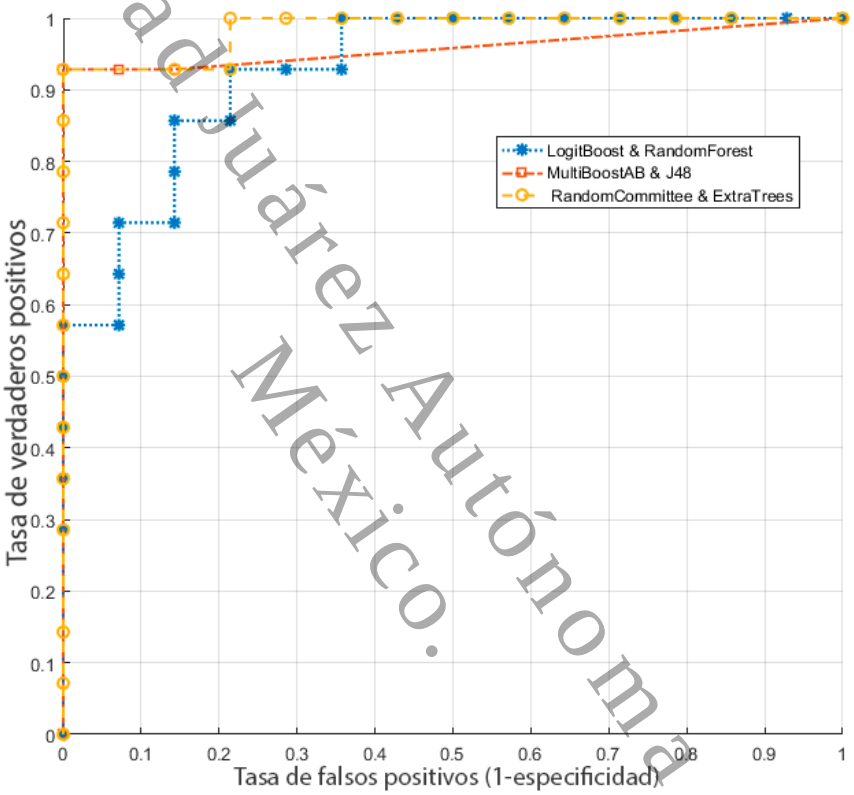


FIGURA 5.8: Gráfico ROC para la clasificación de resultados HA vs HC.

5.5.3.5. La estadística Kappa (Kappa) y el coeficiente de correlación de Matthews (CCM)

La estadística Kappa se calcula utilizando la exactitud observada y la exactitud esperada. La precisión observada es igual a la exactitud (Tabla 5.17. La Precisión Esperada está directamente relacionada con las instancias de cada etiqueta (HC y HA) junto con las instancias que el clasificador acordó con el conjunto de datos.

El clasificador *LogitBoost & RandomForest* tiene 14 instancias etiquetadas como HC y 15 clasificadas como HC; 14 instancias etiquetadas como HA y 13 clasificadas como HA. Los valores marginales son $7.5(14\times15/28)$ para los controles (HC) y $6.5(14\times13/28)$ para el grupo de enfermos (HA). Por lo tanto, la Precisión Esperada es de $0.50((7.5+6.5)/28)$. Finalmente, la estadística kappa es calculada aplicando $(0.8214-0.5)/(1-0.5)=0.643$. La Exactitud esperada resultó ser del 50 %, como siempre será el caso cuando cualquiera de los dos "calificadores" clasifique cada clase con la misma frecuencia en una clasificación binaria (número muy reducido de instancias).

La *estadística Kappa* de los meta-clasificadores se muestran en la Tabla. 5.20; se observa que el mayor acuerdo de predicción con exactitud es para *RandomCommittee & ExtraTree* (0.928); seguido por *MultiBoostAB & J48* (0.78) y valor más bajo para *LogitBoost & RandomForest* (0.643). Todos los valores *Kappa* fueron superior a 0.5.

TABLA 5.20: Estadística kappa y CCM en precisión para las características de la marcha.

Meta-clasificador & Clasificador	**Kappa**	**CCM**
LogitBoost & RandomForest	0.643	0.645
RandomCommittee & ExtraTree	0.928	0.931
Multiboost-AB & J48	0.786	0.788

Los valores de los coeficientes de correlación de Matthews en la última columna de la Tabla 5.20, se calcularon con datos de la matriz de confusión (Tabla 5.18). Podemos afirmar que *RandomCommittee & ExtraTree* tiene una predicción casi perfecta ($+1$) y *LogitBoost & RandomForest* un valor más bajo (0.645). Tanto los resultados de *kappa* como los de *CCM* tienen valores similares para todos los algoritmos, ya que ambos miden la precisión de las puntuaciones de clasificación. Estos resultados son menores que exactitud (Tabla 5.17); sin embargo, no estas diferencias no son significativas.

5.5.3.6. Análisis de errores (MAE, RMSE, RAE, RRSE)

Las medidas de error son los indicadores de cuán bien los resultados de la predicción se ajustan a la distribución de los valores reales, cuanto mayor es la diferencia entre ellos, mayor es la varianza en los errores individuales. Los resultados en la Tabla 5.21 muestran que las estimaciones en el *RMSE* son superiores que los valores de *MAE*

para todos los algoritmos; *MultiBoostAB & J48* tienen la magnitud de error más baja (0.105) y *RandomCommittee & ExtraTree* tiene los valores superiores (0.234); sin embargo, comparando el valor *RMSE* de *RandomCommittee & ExtraTree* tienen la magnitud de error más baja (0.289) y *LogitBoost & RandomForest* más alta (0.408). Lo anterior indica que *RandomCommittee & ExtraTree* tiene una variación menor en la magnitud de los errores (0.055) que los otros algoritmos (0.22 promedio); por lo tanto, los errores son casi de igual magnitud.

TABLA 5.21: Puntuación en MAE, RMSE, RAE y RRSE para la clasificación de HA y HC.

Meta-clasificador & Clasificador	**MAE**	**RMSE**	**RAE %**	**RRSE %**
LogitBoost & RandomForest	0.172	0.4083	33.32	78.95
RandomCommittee & ExtraTree	0.234	0.289	45.30	55.95
Multiboost-AB & J48	0.105	0.320	20.25	61.91

El valor *RAE* en la predicción es mayor para *RandomCommittee & ExtraTree* (45.30), mientras que *MultiBoostAB & J48* permanece bajo (20.25); sin embargo, *RRSE* es relativamente bajo para *RandomCommittee & ExtraTree* (55.95) y mayor para *LogitBoost & RandomForest* (78.95); *RandomCommittee & ExtraTree* mantiene una diferencia cercana entre los dos resultados (10.65), mientras que *MultiBoostAB & J48* tiene una diferencia mucho mayor (41.66); aunque se esperaría que *RandomCommittee & ExtraTree* mantuviera un *RAE* bajo los errores permanece constante; sin embargo, los demás algoritmos muestran una variabilidad cuatro veces mayor. Basándonos en indicadores analizados podemos decir que el algoritmo de mejor rendimiento es *RandomCommittee & ExtraTree* para predecir la enfermedad de Ataxias Hereditarias en personas que utilizan dispositivos sensores de movimiento en los tobillos; sin embargo, debido a la aleatoriedad del meta-clasificador, se debe tener cuidado al incluir nuevas instancias desconocidas para el proceso de entrenamiento.

5.5.4. Reconocimiento de patrones usando los atributos mínimos

Después de encontrar las características básicas que permiten tener una exactitud de clasificación aceptable con las instancias de los pacientes versus enfermos con Ataxias Hereditarias; se inició la búsqueda de la cantidad mínima de atributos para mantener

o mejorar el porcentaje de exactitud obtenida en la sección anterior. El objetivo de esta selección de atributos fue encontrar el subconjunto de características más reducido, pero sin comprometer el porcentaje de clasificación y que la distribución de clases resultante sea lo más parecida posible a la original. Un atributo se considera esencial si no es irrelevante (no afecta al concepto de objetivo de ninguna manera) o redundante (no añade nada nuevo al objetivo).

Usamos algoritmos de tipo *wrapper* para obtener un conjunto reducido de atributos como espacios de búsqueda, el cual involucró el algoritmo *Ranker* como método de búsqueda y los algoritmos evaluador de atributos *AttrEvalSelection* y *ClassiffierAttributeEval* para obtener puntuaciones individuales por atributo. Con el nivel de importancia obtenido con el algoritmo *Ranker* para cada evaluadores de atributos se estableció un umbral del importancia mayor a cero para descartar atributos con puntuaciones bajas, posteriormente se tomaron los atributos que coincidían en ambas lista y sus puntuaciones fueron ordenadas de mayor a menor [154].

Posteriormente implementamos la estrategia del algoritmo *Hill Climbing* para encontrar la cantidad mínima de atributos. Se inició el proceso de búsqueda con un solo atributo, tomando primeramente aquellos que tuvieran en promedio una mejor puntuación dada por el *Ranker*. El procedimiento fue el siguiente: (a) empezamos con un conjunto de atributos vacío, (b) aumentamos la cardinalidad en uno atributo gradualmente con la finalidad de encontrar un mejor rendimiento de clasificación; y (c) continuamos con la siguiente cardinalidad si el cambio producía una solución mejor; este proceso fue repetitivo e incremental hasta no encontrar mejoras en los resultados del algoritmo de aprendizaje.

En este experimento tomamos la mejor combinación de algoritmos *RandomCommittee & ExtraTree* (RCMT) e incluimos otros algoritmos que se usaron en el estado del arte como es *los K Vecinos más cercanos (KNN)*, *Máquina de soporte vectorial (SVM)* y *Redes neuronales Perceptron Multicapa (MLP)*. Consideramos la exactitud de los algoritmos como un parámetro de mejora. Los criterios de no mejora se cumplen con diferentes subconjuntos de características como se muestra en la Tabla 5.22. Los algoritmos *SVM* y *MLP* con seis características dejaron de mostrar una mejora; mientras que *RCMT* con cinco características y KNN con dos características; sin embargo, sólo *RCMT* y *MLP* alcanzaron 100 % de exactitud. También podemos observar que el número de características de ambos tobillos parece más bien equilibrado, excepto en el caso del *SVM*, que muestra un predominio de la tobillo derecho.

TABLA 5.22: Precisión de los algoritmos para las características más destacadas.

(A) SVM

Características	**%**
R_Integral_PSD_VER	78.57
R_Integral_PSD_AP	82.14
L_Integral_PSD_AP	89.29
R_Frequency_at_50_energy _VER	85.71
R_Frequency_at_75_energy _AP	89.29
R_Step_regularity_in_AP	92.86
R_Mean_steplength	89.29
R_RMS_VER	89.29
L_RMS_VER	89.29
R_Integral_PSD_ML	92.86

(B) KNN

Características	**%**
R_RMS_VER	75.00
L_Integral_PSD_AP	96.43
R_Frequency_at_75_energy _ML	96.43
R_Integral_PSD_VER	92.86
R_Integral_PSD_AP	89.29
L_RMS_AP	89.29
L_Step_regularity_in_VER	92.86
R_RMS_AP	92.86
L_Stride_regularity_in_VER	92.86
R_RMS_ML	92.86

(C) RCMT

Características	**%**
R_RMS_VER	75.00
R_Integral_PSD_AP	89.29
L_Integral_PSD_AP	96.43
L_Stride_regularity_in_VER	100.00
L_RMS_AP	96.43
L_Step_regularity_in_VER	100.00
R_RMS_AP	100.00
L_Cadence_step_min	100.00
R_Integral_PSD_VER	96.43
R_RMS_ML	92.86

(D) MLP

Características	**%**
R_RMS_VER	89.29
R_Integral_PSD_AP	96.43
R_Integral_PSD_VER	89.29
R_Frequency_at_50_energy _VER	89.29
R_RMS_AP	85.71
L_Integral_PSD_VER	100.00
L_Integral_PSD_AP	100.00
R_Velocity	100.00
L_RMS_VER	100.00
L_RMS_ML	100.00

Las características que coincidían en la clasificación más alta de todos los algoritmos fueron: la *intensidad de aceleración (RMS)*, y *la potencia de señal (Integral_PSD)* en ambos tobillos con predominio en los ejes *VER* y *AP*. No se encontró ningún subconjunto de características mejor cuando se aplicó la estrategia iterativa de selección progresiva propuesta sobre los algoritmos *KNN* y *RCMT*. Sin embargo, encontramos un subconjunto reducido de cuatro características para el algoritmo *SVM* que alcanza el *92.86 %* de exactitud en el rendimiento y tres subconjuntos de características que permiten una exactitud del *100 %* para el algoritmo *MLP*, como se observa en la Tabla 5.23.

TABLA 5.23: Subconjuntos adecuados de características de marcha para cada algoritmo.

(A) SVM

Características	**%**	**ICC**	**HC**	**HA**
R_integral_PSD_VER	78.57	22	10	12
R_Integral_PSD_AP	82.14	23	10	13
R_frequency_at_50_energy_VER	78.57	22	12	10
R_frequency_at_75_energy_AP	92.86	26	13	13

(B) KNN

Características	**%**	**ICC**	**HC**	**HA**
R_RMS_VER	75.00	21	12	9
L_Integral_PSD_AP	96.43	27	14	13

(C) RCMT

Características	**%**	**ICC**	**HC**	**HA**
R_RMS_VER	75.00	21	12	9
R_Integral_PSD_AP	89.29	25	13	12
L_Integral_PSD_AP	96.43	27	14	13
L_Stride_regularity_in_VER	100.00	28	14	14

(D) MLP

Características	**%**	**ICC**	**HC**	**HA**
R_RMS_VER	89.29	25	12	13
R_Integral_PSD_AP	96.43	27	13	14
L_Integral_PSD_AP	100.00	28	14	14

TABLA 5.24: El subconjunto de características de la marcha que permite la correcta clasificación de todos los participantes.

No	Características	Descripción
1	R_RMS_VER	Raíz cuadrada media en el eje Y del tobillo derecho.
2	R_Integral_ PSD_AP	Densidad espectral de potencia integral en el eje X del tobillo derecho.
3	L_Integral_ PSD_AP	Densidad espectral de potencia integral en el eje X del tobillo izquierdo.
4	R_Stride_regularity _in_VER	Regularidad estimada de la zancada en el eje X del tobillo derecho.

La combinación propuesta de criterios mínimos prominentes y una estrategia de selección de características permitió los siguiente hallazgos: La característica más relevante es *R_RMS_VER* (la medida estadística de la magnitud de la aceleración del sensor derecho) del tobillo en el eje vertical (arriba y abajo), lo que permite a *MLP* alcanzar una precisión de casi el *90 %* en la discriminación correcta de los *sujetos de control HC* y de los *pacientes de HA*. Ninguna otra característica por sí sola logra una precisión tan alta; sin embargo, la potencia total de la señal en el *dominio de la frecuencia (Integral_PSD)* contribuye a su mejora en todos los algoritmos. Las características del *eje ML (Media-lateral)* no fue relevantes para ninguno de los algoritmos.

Con sólo dos características, *KNN* y *MLP* alcanzaron el *96 %* en exactitud, mientras que *RMCT* alcanzó el *89,29 %*. *KNN* clasificó erróneamente a un paciente *HA*; mientras que, *MLP* clasificó erróneamente a un sujeto de control, *RMCT* obtuvo un *96 %* con tres características, equivocándose al clasificar un paciente HA. Para MLP, se tuvo que añadir R_Integral_PSD_AP (la potencia total de la señal para el movimiento del tobillo derecho hacia atrás y hacia delante del eje X) para permitir que todos los pacientes con AH se clasificaran correctamente, pero fue necesario utilizar la misma característica para el tobillo izquierdo (L_Integral_PSD_AP) para separar correctamente ambas clases.

Se observa en la Tabla 5.24 un subconjunto de características que permitieron la clasificación correcta de todos los participantes. Además de las tres características de la marcha requeridas por *MLP*; el algoritmo *RCMT* necesitaba la *regularidad de la zancada* en el *eje VER* para el tobillo izquierdo *(L_Stride_regularity _in_VER)* para alcanzar el *100 %* de exactitud. El mismo número de características de ambos tobillos fueron seleccionadas en *RCMT*, pero hay un predominio del tobillo derecho para *MLP*.

TABLA 5.25: Precisión de clasificación binaria y tiempo de entrenamiento + validación.

Algoritmo	**Características**	**LOOCV**	**Tiempo (s)**
SVM	4	92.86 %	0.88 + 0.21
KNN	2	96.43 %	0.44 + 0.07
RCET	4	100.00 %	1.24 + 0.71
MLP	3	100.00 %	0.56 + 2.63

Se ejecutaron todos los algoritmos de clasificación sobre el correspondiente subconjunto de características de la marcha. Los resultados de la clasificación muestran que la partición aleatoria en un CV de 10 veces no tiene ningún impacto que pudiera ser revelado por *LOOCV* en la evaluación de la precisión, dando los mismos resultados. Podemos observar en la Tabla 5.25 que todos los algoritmos lograron una precisión superior al *90 %* con el subconjunto de atributos correspondiente, pero *KNN* consumió sólo 51 segundos para lograr su mejor resultado *(96,43 %)* con dos características de marcha, mientras que para obtener una precisión del *100 % RCET* se necesitaron cuatro características y casi dos minutos y *MLP* con tres características tomó más de tres minutos.

En aplicaciones como la monitorización continua, en las que es necesario detectar cambios en la forma de andar, es importante que los dispositivos sean utilizados constantemente mucho tiempo, ya que tendrá que repetirse muchas veces, especialmente en la monitorización en tiempo real; por lo que se requieren soluciones tecnológicas eficientes.

Los coeficientes de correlación de *Matthews* entre la exactitud observada y la predicha (5.26), muestran que *MLP* y *RCET* tuvieron una correlación de uno (+1) lo que significa alta precisión en las clases positiva y negativa. Las mediciones de error son indicadores de cuán bien los resultados de la predicción concuerdan con la distribución del valor real, cuanto mayor es la diferencia entre ellos, mayor es la varianza en los errores individuales. Podemos ver en la Tabla 5.26 que *SVM* y *KNN* tienen la mayor diferencia en los valores de *MAE* y *RMSE*, lo que significa una mayor incidencia de errores individuales. El algoritmo *MLP* tiene los valores más bajos en ambas medidas, lo que es consistente con la precisión obtenida.

TABLA 5.26: Evaluación de errores en la clasificación de 28 participantes.

Algoritmos	**ICC**	**CCM**	**MAE**	**RMSE**
SVM	26	0.86	0.0714	0.2673
KNN	27	0.93	0.0380	0.1888
RCET	28	1.00	0.1375	0.1945
MLP	28	1.00	0.0290	0.0710

5.5.5. Hallazgos y observaciones

De una ventana de 10 pasos definida alrededor del centro del conjunto de datos de la marcha, se extrajeron 56 características de la marcha espacio-temporal. Para reducir el número de estas características se revisaron los algoritmos evaluadores de atributos para la clasificación supervisada con la ayuda del algoritmo selector de atributos *Ranker*; lo anterior, para seleccionar aquellos que resulten en el mayor número de atributos comunes en los primeros lugares. Basándose en la frecuencia de las características de la marcha que aparecen en la parte superior de las listas sugeridas por el selector de atributos, se seleccionaron 21 de ellas que tuvieran una puntación superior a cero.

Las 21 características fueron gráficas usando un *Boxplot*, para realizar un análisis de la distribución de valores con la finalidad encontrar características con datos irregulares que afecten el resultado obtenido por los algoritmos clasificadores. El *Boxplot* permitió identificar que las características con cuartiles amplios y valores atípicos más extendidos, tenían una desviación estándar más amplia, por lo que producían resultados engañosos en los procesos de clasificación; sin embargo, las características con cuartiles más cerrados tenían datos más compactos y menos variabilidad y traslapamiento con los datos de la clase contraria, en consecuencia, abonaban a mejorar el desempeño de los clasificadores. Lo anterior nos llevó a eliminar seis características que tenían cuartiles más amplios y mayor traslapamiento en los datos con respecto a la clase contraria; por lo tanto, se redujeron de 21 a 15 características de la marcha.

Descubrimos que 15 características de la marcha eran las más representativas de los patrones de marcha de los sujetos e identificamos tres combinaciones de *meta-clasificador & clasificador* con buen desempeño en las tareas de clasificación. Los resultados de los algoritmos con alta precisión en la clasificación binaria fueron: *LogitBoost & RandomForest* (82.14 %), *MultiBoostAB & J48* y *RandomCommittee & ExtraTree* (96.43 %). La combinación de *MultiBoostAB & J48* y *RandomCommittee & ExtraTree* reconoció a *13* de *14*

pacientes con *AH*, lo que corresponde a un 92.6 %; sin embargo, *RandomCommittee & ExtraTree* presentaron mejores medidas de rendimiento, tiempo de ejecución y precisión. En este trabajo, se alcanzó una alta precisión en el reconocimiento de los patrones de marcha de la ataxia hereditaria (92.6 %), incluso por encima de lo que se ha logrado en otros trabajos reportados (86,7 %). Con esta investigación se confirma que la combinación de *meta-clasificador & clasificador* tienen un mejor rendimiento que los clasificadores individuales, incluso cuando estos últimos han sido modificados para adaptarse a una enfermedad específica.

En la búsqueda de la cantidad mínima de atributos para discriminar correctamente a dos clases de individuos la estrategia de basada en el algoritmo *Hill climb* permitió reducir 15 características de la marcha a un número mínimo según el algoritmo de clasificación implementado. Los algoritmos *KNN* y *MLP* obtuvieron un 96 % con dos características; sin embargo, *MLP* sólo requirió una característica adicional del sensor de tobillo derecho para alcanzar un 100 % de precisión; es importante notar que *RCMT* alcanzó una precisión exacta con cuatro características. Por lo tanto, cuando el tiempo de ejecución es más importante se recomienda usar *KNN* ya que los resultados se obtienen en menor tiempo (0.44 segundos); sin embargo, cuando la precisión sea de mayor importancia se sugiere usar *MLP*, incluso cuando se disponen de más recursos para computar se podría usar *RCMT* ya que ambos demostraron una precisión exacta.

Capítulo 6

Conclusiones y trabajo a futuro

El término sensores de movimiento (MS) o inerciales hace referencia al conjunto de dispositivos que permiten cuantificar aceleración lineal y angular de un cuerpo basados en su movimiento y el cuerpo al que están sujetos conocido como principio de inercia. Los dispositivos inteligentes de uso cotidiano como son los teléfonos móviles (smartPhones) cuentan con este tipo de sensores para realizar algunas de sus funciones. Los MS de los *SmartPhones empleados en esta investigación* han sido comparados con dispositivos de medición de la marcha considerados como estándares y los resultados han demostrado un desempeño similar exponiendo así su capacidad para tareas de mediciones de los datos de la marcha de las personas. El reconocimiento de patrones de la marcha usando datos de los MS han sido estudiado en diversas investigaciones considerando diversos grupos de pacientes con diversas patologías con alteraciones en los patrones de la marcha. Los estudios reportados empleando aprendizaje automático para diferencias diversas patologías basados en los patrones de la marcha han reportado una precisión de 90 % de reconocimiento de pacientes en general y un 88.2 % al reconocer a los pacientes con Huntington; mientras que para pacientes con algún tipo de Ataxias Hereditaria se ha obtenido un 78.78 % de precisión al diferencias entre enfermos y sanos.

El correcto tratamiento de los datos brutos de los sensores y el uso de la cantidad mínima de sensores fueron esenciales para el éxito de esta investigación. Se realizaron incontables prueba de la información de sensores ubicados en: el brazo, cintura, piernas y tobillos para determinar aquellos que permiten recolectar mejor la señales que representan las alteraciones de la marcha en los grupos de Ataxias Hereditarias y enfermos con Huntington; se encontró que los sensores de los tobillos presentaban un mejor rendimiento y eran ideales para recolectar los datos eficientemente en los pacientes con

trastornos de la marcha.

El procedimiento implementado para el pre-procesamiento de los datos brutos de los sensores incluyó: (a) la eliminación de valores atípicos en el inicio y fin de los datos recolectados relacionados a la falta de movimiento antes de iniciar la caminata y la finalización de la misma; (b) la calibración de datos a intervalo constante de tiempo debido a las diferencias que existían entre la adquisición y la grabación de la información dentro del dispositivo; (c) la aplicación de una normalización cero con la finalidad de eliminar valores constantes en la aceleración que no eran parte de la marcha, tales como la gravedad; (d) el cálculo de la señal invariante para obtener un solo vector de datos de los tres ejes para minimizar los cambios de direcciones; y (e) el suavizado de la señal usando el filtro pasa bajo *promedio móvil* para eliminar ligeramente ruidos en la señal.

Se diseñó un algoritmo para identificar el inicio de cada una de las zancadas realizadas por los sujetos con la finalidad de seleccionar aproximadamente 10 zancadas consideradas como la información representativa de la marcha de un sujeto. El algoritmo estándar *Peak to Peak* fue implementado para identificar los picos mínimos prominentes que representaban el inicio y fin de una zancada; posteriormente se tomaron los datos que se encontraban alrededor del centro del conjunto de datos de cada sensor de cada paciente.

La extracción de características de los datos obtenidos con el algoritmo diseñado, se realizó con el herramienta asistida por computadora *iGAIT*. Esta herramienta permitió obtener 28 características por sensor, un total de 56 características por paciente fueron usadas en el proceso. La selección de atributos y las tareas de clasificación fueron llevadas a cabo con la Herramienta *Weka*. En el proceso de clasificación de los sujetos de control (HC) y sujetos enfermos con Huntington (HD), las combinaciones de algoritmos selectores de atributos *CfsSubsetEval & BestFirst* y *CfsSubsetEval & GreedyStepwise* seleccionaron un subconjunto de 11 características en la misma manera. Las características seleccionadas estaban relacionadas con: el *RMS* en los ejes *AP* y *VER* de ambos sensores, el *RMS* en el eje *ML* del sensor derecho, la *IPSD* al *90 %* en el eje *AP* del ambos sensores, el *IPSD* al *75 %* en el eje *VER* del sensor derecho, el *IPSD* al *100 %* del eje *ML* del sensor derecho, la *simetría* en el eje *VER* en el sensor izquierdo y la *regularidad del paso* del eje *AP* en el sensor derecho. Se realizaron pruebas de clasificación con un total de 225 combinaciones de *Meta-clasificador & árbol clasificador* con una validación cruzada de 10 veces para encontrar las combinaciones con mejores resultados. Los meta-clasificadores que mostraron un mejor desempeño fueron: *Logitboost, Randomcommittee*

y *MultiboostAB*; los árboles clasificadores con excelente desempeño fueron: *RandomForest, ExtraTree, SimpleCart* y *LMT*. Los pares de combinaciones con excelentes resultados después de aplicar una validación cruzada dejando uno fuera (LOOCV) fueron: *LogitBoost & RandomForest* (92.8571 %), *MultiBoostAB & J48* (92.8571 %) *RandomCommittee & ExtraTree* (85.7143 %). *LogitBoost & RandomForest* y *MultiBoostAB & J48* recocieron el 100 % de los sujetos enfermos y 85.7143 % de los sujetos de control; mientras que *RandomCommittee & ExtraTree* reconoció el 85.7143 % de cada grupo respectivamente; En este trabajo se supera la exactitud obtenida por *Mannini et al.* en [3] de 90.5 % hasta 92.9 %, una diferencia de 2.4 % en la clasificación general; sin embargo nosotros obtuvimos 100 % de reconocimientos de pacientes enfermos mientras que *Mannini et al.* alcanzaron un 88.2 % lo que representa una diferencia del 11.8 %

La clasificación binaria de pacientes con Ataxias Hereditarias y Sujetos de control (HC) se realizó siguiendo el proceso de pre-procesamiento, segmentación de la marcha y extracción de características establecidos. La reducción del número de atributos se basó en el algoritmo *Ranker* como método de búsqueda y probamos diversos algoritmos para obtener las características que produjeran la más exactitud en la clasificación; los combinaciones de selectores de atributos con similares características seleccionadas fueron: *Ranker & CorrelationAttributeEval* y *Ranker & ClassifierAttributeEval & J48 Classifier* un total de 21 atributos; los cuales fueron graficados y usando un diagrama de cajas por grupos para descartar aquellos que presentaran mayor traslapamiento y que mejoran el rendimiento de los clasificadores; un subconjunto de 15 características fueron tomadas en cuenta para las tareas de clasificación. Las características seleccionadas estaban relacionadas con la información de los tres ejes de ambos sensores: *el RMS, la integral IPSD y la velocidad de la marcha*. Para encontrar los mejores pares de algoritmos clasificadores *meta-clasificador & clasificador* primero aplicamos una validación cruzada de 10 veces para obtener las tres combinaciones con los mejores resultados, posteriormente aplicamos la validación dejando uno fuera para asegurar el desempeño del algoritmo. las combinaciones de *LogitBoost & RandomForest* obtuvo el 82.14 %, *RandomCommittee & ExtraTree* 96.86 % y *MultiBoostAB & J48* 89.29 %; *RandomCommittee & ExtraTree* reconocio el 100 % del sujetos de control, mientras que *LogitBoost & RandomForest* y *MultiBoostAB & J48* solo reconocieron el 85.71 % (12 de 14); *RandomCommittee & ExtraTree* y *MultiBoostAB & J48* reconocieron el 92.86 % de sujetos enfermos (13 de 14), mientras que *LogitBoost & RandomForest* reconoció solamente el 78.57 % (11 de 14). El reconocimiento de sujetos es superior al reportado por *Pradhan et al.* en [5] quien reconoció al 90.90 % de todos los sujetos, nosotros mejoramos ese resultado en 5.96 % y solo alcanzó

a reconocer el 86.70 % de los pacientes enfermos mientras que nosotros alcanzamos a reconocer el 92.86 % lo que presenta una mejora del 6.16 %.

Los resultados obtenidos en esta investigación son superiores a los alcanzados por *Sanchez de la Cruz et al.* en [38] ya que obtuvieron un 92.04 % al clasificar los HC, mientras que nosotros obtuvimos un 92.86 % lo que representa una mejora de 0.82 %; ellos obtuvieron un 75.78 % al clasificar HA, mientras que nosotros 89.29 % lo que representa una mejora del 13.51 %; y finalmente, ellos obtuvieron un 78.78 % al clasificar HD, mientras que nosotros obtuvimos un 81.82 % lo que representa una mejora del 3.04 %; la mejora en la clasificación global es de 6.93 % con respecto al trabajo anterior sobre la clasificación en general. Cabe mencionar que nuestros resultados fueron usando datos de dos sensores mientras que el trabajo anterior se llevó a cabo con cinco sensores, por lo que nuestra propuesta consume menos recursos para recolectar información y procesamiento de la información.

La búsqueda de soluciones que puedan funcionar en dispositivos con pocos recursos nos llevó a realizar experimentos enfocados a usar una cantidad de mínima de características, en donde se identificó que los algoritmos *KNN* alcanzaron una precisión del 96.43 % usando solo dos características, mientras que *MLP* alcanza un exactitud del 100 % usando solo tres características; el primero se ejecuta en un tiempo mínimo de 0.44 segundos y el segundo en 0.56 segundos; lo anterior es vital cuando se tiene limitaciones capacidad de procesamiento, se tiene que optimizar la cantidad de información de la marcha y la velocidad de transmisión de información debe ser moderada. Por lo tanto, este último está orientado al desarrollo de herramientas que puedan asistir al seguimiento y monitoreo de la progresión a largo plazo desde casa usando dispositivos de la vida diaria como son los teléfonos inteligentes.

6.1. Cumplimiento de los objetivos

El conjunto de datos recolectado de información de los pacientes y personas sanas estuvo conformado por seis sensores: dos en los tobillos, dos encima de rodilla, uno en la cintura y uno en el brazo derecho; se realizaron pruebas de los datos de cada sensor; posteriormente con combinaciones de pares de sensores, siendo los datos de sensores colocados en los tobillos los que mostraron un mejor resultado en proceso.

La determinación de las técnicas de procesamiento de datos fué bastante efectiva según los resultados obtenidos; se previó que el método incluyera: limpieza de los datos, calibración, normalización y aplicación de un filtro de suavizado mínimo. Posteriormente se desarrolló un algoritmo que permitiera segmentar la marcha para identificar el inicio de cada zancada, con la finalidad de identificar un segmento de datos equivalente a 10 zancadas que se encontraban alrededor de punto medio de los datos recolectados para evitar datos insignificante al inicio y final de los datos colectados. Las características obtenidas basadas en análisis del tiempo y la frecuencia resultaron ser efectivas y significativas.

El exhaustivo proceso de selección de características resultó ser excelente, al seleccionar una cantidad reducida de características. Se usó una combinación de evaluadores de atributos con selectores de atributos en todas sus combinaciones posibles para cubrir el mayor número de posibilidades, esto permitió obtener una efectividad de alrededor del 90 % al clasificar sanos y enfermos. Para identificar la cantidad mínima de características necesarias para reconocer eficientemente a los enfermos de los sanos se implementó una estrategia similar al *ascenso de colinas* para encontrar los atributos mínimos que preservarán la efectividad de los algoritmos de aprendizaje automático implementados; lo anterior sobre el número atributos obtenidos con los algoritmos selectores de atributos.

La identificación de los algoritmos basados en una combinación de meta-clasificadores y árboles clasificadores con excelente desempeño en las tareas de clasificación, se realizó usando el 20 % de la información total de las muestras; lo que permitió discriminar los pares de combinaciones con excelente desempeño. Lo anterior permitió obtener un total de tres meta-clasificadores que combinados con cuatro diferentes árboles presentaban excelentes resultados ($> 93\,\%$) al clasificar enfermos con Huntingon versus personas sanas. Lo anterior fue aplicado sobre el conjunto de datos de enfermos con Ataxias Hereditarias versus Sujetos de Control obteniendo subconjunto de características similares; este subconjunto fue utilizado para realizar la búsqueda de características mínimas que permitieran una excelente precisión; para esto se usó el par de algoritmos de clasificación que obtuvo un mejor resultado en el proceso anterior (RandomCommittee & ExtraTrees), además de incluir algoritmos tradicionales como: SVM, KNN y MLP. KNN alcanzó un 96 % con solamente dos características; mientras que MLP necesito tres características para alcanzar 100 % de exactitud; los demás algoritmos necesitaron cuatro características para alcanzar altos puntajes, SVM obtuvo un 92 % de precisión mientras que RandomCommittee & ExtraTrees obtuvo 100 % de efectividad.

Las preguntas de investigación planteadas fueron respondidas y los objetivos establecidos fueron alcanzados, debido a que se comprobo que los dos sensores ubicados en los tobillos proporcionan información con mayor relevancia que en otros lugares del cuerpo humano; se estableció un procedimiento para procesamiento de la información de los sensores y fue posible segmentar la marcha de manera eficiente; las 56 características extraídas basadas en el análisis de la frecuencia y el tiempo fueron las correctas para obtener excelentes desempeño en la clasificación; por otro lado, la búsqueda exhaustiva de las características basados en algoritmos evaluadores y selectores de atributos permitió obtener las características más significativas para la investigación. Finalmente Todo lo anterior permitió que los algoritmos fueran eficientes al discriminar entre enfermos y sanos obteniendo resultados superiores a los encontrados en la literatura; esto fue corroborado usando diversos algoritmos, los cuales, obtuvieron excelentes resultados en la precisión en la predicción.

6.2. Investigaciones futuras

La diferenciación entre los patrones de la marcha usando algoritmos de clasificación puede ayudar a identificar las alteraciones características de la marcha en cada grupo patológico que no son perceptibles al ojo humano, esta información puede ser de utilidad para el especialista en la identificación de la aparición de alteraciones motoras relacionadas con alguna de estas enfermedades mejorando el diagnóstico de la enfermedad. Las enfermedades degenerativas HD y HA son progresivas, por lo que los síntomas en los pacientes afectarán cada vez más sistema motor, en consecuencia, las alteraciones se volverán incapacitantes; por lo que es importante proporcionar un seguimiento continuo de la progresión de las alteraciones de la marcha. La implementación de una infraestructura tecnológica que permita el monitoreo continuo resulta de importancia para mejorar la atención del paciente y cuantificar el progreso de estas enfermedades.

Los resultados obtenidos en esta investigación contribuyen a los trabajos futuros enfocados al seguimiento continuo de la progresión de las alteraciones en los patrones de la marcha; por ejemplo, monitorizar la progresión de la enfermedad detectando una empeoramiento que se refleja en los cambios en la forma de caminar de los sujetos de estudio; esta línea de estudio necesita un sistema de monitorización con transferencia

de datos en tiempo real para su diseño. Otra línea de estudio que promueve los resultados es la detección temprana de la enfermedad; usando técnicas de aprendizaje automático es posible detectar cambios sutiles tempranos en los patrones de la marcha que no son visibles con la observación clínica por los especialistas. Finalmente, los resultados de este estudio también podrían ser aprovechados para el seguimiento de pacientes sometidos a algún tipo de rehabilitación u otra intervención terapéutica que implique cambios temporales en los patrones de la marcha.

Bibliografía

[1] M. Banaie, M. Pooyan y M. Mikaili, «Introduction and application of an automatic gait recognition method to diagnose movement disorders that arose of similar causes», *Expert Systems with Applications*, vol. 38, n.º 6, págs. 7359-7363, 2011, ISSN: 09574174. DOI: 10.1016/j.eswa.2010.12.091.

[2] M. A. W. Eddy Sanchez-Delacruz, Francisco Acosta-Escalante y M.-C.J. J. Hernández-Nolasco José Adán, Pancardo Pablo, «Gait Recognition in the Classification of Neurodegenerative Diseases», en *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services*, B. J. Hervás Ramon, Lee Sungyoung, ed., Springer International Publishing, 2014, cap. Ubiquitous, págs. 128-135, ISBN: 978-3-319-13101-6. DOI: 10.1007/978-3-319-13102-3_23.

[3] A. Mannini, D. Trojaniello, A. Cereatti y A. M. Sabatini, «A Machine Learning Framework for Gait Classification Using Inertial Sensors: Application to Elderly, Post-Stroke and Huntington's Disease Patients.», *Sensors (Basel, Switzerland)*, vol. 16, n.º 1, pág. 134, ene. de 2016, ISSN: 1424-8220. DOI: 10.3390/s16010134.

[4] A. Mannini, O. Martinez-Manzanera, T. F. Lawerman, D. Trojaniello, U. D. Croce, D. A. Sival, N. M. Maurits y A. M. Sabatini, «Automatic classification of gait in children with early-onset ataxia or developmental coordination disorder and controls using inertial sensors», *Gait & Posture*, vol. 52, págs. 287-292, feb. de 2017, ISSN: 0966-6362. DOI: 10.1016/J.GAITPOST.2016.12.002.

[5] C. Pradhan, M. Wuehr, F. Akrami, M. Neuhaeusser, S. Huth, T. Brandt, K. Jahn y R. Schniepp, «Automated classification of neurological disorders of gait using spatio-temporal gait parameters», *Journal of Electromyography and Kinesiology*, vol. 25, n.º 2, págs. 413-422, 2015, ISSN: 18735711. DOI: 10.1016/j.jelekin.2015.01.004.

[6] R. LeMoyne, F. Heerinckx, T. Aranca, R. De Jager, T. Zesiewicz y H. J. Saal, «Wearable body and wireless inertial sensors for machine learning classification of gait for people with Friedreich's ataxia», *BSN 2016 - 13th Annual Body Sensor Networks Conference*, págs. 147-151, 2016. DOI: 10.1109/BSN.2016.7516249.

[7] Samsung Electronics Co. Ltd., *Samsung Health*, 2017. dirección: https://play.google.com/store/apps/details?id=com.sec.android.app.shealth.

[8] Apple Inc, *ResearchKit and CareKit*, 2017. dirección: https://www.apple.com/ca/researchkit/https://www.apple.com/lae/researchkit/{\%}0Ahttps://www.apple.com/ca/researchkit/https://www.apple.com/ca/researchkit/{\%}0Ahttps://www.apple.com/lae/researchkit/{\%}0Ahttps://www.apple.com/ca/researchkit/.

[9] P. Menaspà, *Effortless activity tracking with Google Fit*, dic. de 2015. DOI: 10.1136/bjsports-2015-094925.

[10] M. Singh, M. Singh e I. Engineering, «Neuro-Degenerative Disease Diagnosis using Human Gait : A Review», *Computer Science & Electronics Journals*, vol. 7, n.º 1, págs. 16-20, 2013.

[11] Y. M. Bordelon, «Huntington Disease», *Neurologic Clinics*, vol. 31, n.º 4, págs. 1085-1094, nov. de 2013, ISSN: 07338619. DOI: 10.1016/j.ncl.2013.05.004.

[12] A Delval y P Krystkowiak, *Locomotion et maladie de Huntington*, 2010. DOI: 10.1016/j.neurol.2009.05.013.

[13] K. Duff, L. J. Beglinger y J. S. Paulsen, *"PRE-SYMPTOMATIC"HUNTINGTON'S DISEASE*, 2008. DOI: 10.1016/S0072-9752(07)01255-9.

[14] S. Jayadev y T. D. Bird, «Hereditary ataxias: overview», vol. 15, n.º 9, págs. 673-683, sep. de 2013, ISSN: 1098-3600. DOI: 10.1038/gim.2013.28.

[15] A. J. Mcgarry, K. Biglan y F. Marshall, «Chapter 75 - Huntington Disease», 2015. DOI: 10.1016/10.1016/B978-0-12-410529-4.00075-9.

[16] Jane S. Paulsen, «Early Detection of Huntington's Disease», *Future Neurology*, págs. 85-104. 2010. dirección: http://www.medscape.com/viewarticle/715374{_}4.

[17] S. Kwak, «Huntington disease», *Jankovic*, vol. 60 Suppl 4, n.º 5, págs. 417-421, 2011, ISSN: 00471852. DOI: 10.1038/nrdp.2015.52.

[18] M. Gudesblatt y D. Tarsy, «Huntington ' s Disease : A Clinical Review», *NEUROLOG REVIEWS*, vol. 19, n.º May, s1-s8, 2011.

[19] M. Danoudis y R. Iansek, «Gait in Huntington¿s disease and the stride length-cadence relationship.», *BMC neurology*, vol. 14, n.º 1, pág. 161, 2014, ISSN: 1471-2377. DOI: 10.1186/s12883-014-0161-8.

[20] J Schwabova, T Maly, J Laczo, A Zumrova, V Komarek, Z Musova y F Zahalka, «Application of a Scale for the Assessment and Rating of Ataxia (SARA) in

Friedreich's ataxia patients according to posturography is limited.», *Journal of the neurological sciences*, vol. 341, n.º 1-2, págs. 64-7, jun. de 2014, ISSN: 1878-5883. DOI: 10.1016/j.jns.2014.04.001. dirección: http://www.sciencedirect.com/science/article/pii/S0022510X14002196.

[21] B. Woehrlen, M. Catherine, R. Ibarra, S. Nadja, A. Ochoa, L. Martínez Ruano y U. Rodríguez Ortiz, «Taxonomy of ataxias. algorithm of the lack of rhythm», *Archivos de Neurociencias*, vol. 21, n.º 3, págs. 6-13, 2017.

[22] A. Dalton, H. Khalil, M. Busse, A. Rosser, R. van Deursen y G. ÓLaighin, «Analysis of gait and balance through a single triaxial accelerometer in presymptomatic and symptomatic Huntington's disease», *Gait and Posture*, vol. 37, n.º 1, págs. 49-54, 2013, ISSN: 09666362. DOI: 10.1016/j.gaitpost.2012.05.028.

[23] D. Trojaniello, A. Ravaschio, J. M. Hausdorff y A. Cereatti, «Comparative assessment of different methods for the estimation of gait temporal parameters using a single inertial sensor: application to elderly, post-stroke, Parkinson's disease and Huntington's disease subjects.», *Gait & posture*, vol. 42, n.º 3, págs. 310-6, 2015, ISSN: 1879-2219. DOI: <.

[24] D. Trojaniello, A. Cereatti, E. Pelosin, L. Avanzino, A. Mirelman, J. M. Hausdorff y U. Della Croce, «Estimation of step-by-step spatio-temporal parameters of normal and impaired gait using shank-mounted magneto-inertial sensors: application to elderly, hemiparetic, parkinsonian and choreic gait.», *Journal of neuroengineering and rehabilitation*, vol. 11, n.º 1, pág. 152, 2014, ISSN: 1743-0003. DOI: 10.1186/1743-0003-11-152. dirección: http://www.jneuroengrehab.com/content/11/1/152.

[25] A. Matsushima, K. Yoshida, H. Genno, A. Murata, S. Matsuzawa, K. Nakamura, A. Nakamura y S.-i. Ikeda, «Clinical assessment of standing and gait in ataxic patients using a triaxial accelerometer», *Cerebellum & Ataxias*, n.º August, págs. 1-7, 2015, ISSN: 2053-8871. DOI: 10.1186/s40673-015-0028-9. dirección: http://dx.doi.org/10.1186/s40673-015-0028-9.

[26] P. Kutílek, V. Socha, O. Čakrt y Z. Svoboda, «Assessment of postural stability in patients with cerebellar disease using gyroscope data.», *Journal of bodywork and movement therapies*, vol. 19, n.º 3, págs. 421-8, jul. de 2015, ISSN: 1532-9283. DOI: 10.1016/j.jbmt.2014.09.005.

[27] M. Iosa, P. Picerno, S. Paolucci y G. Morone, «Wearable inertial sensors for human movement analysis.», *Expert review of medical devices*, vol. 4440, n.º June,

págs. 1-19, 2016, ISSN: 1745-2422. DOI: 10.1080/17434440.2016.1198694. dirección: http://www.ncbi.nlm.nih.gov/pubmed/27309490.

[28] *Cuentame INEGI*, Accesado el 08 de Junio de 2019, Ciudad de México, México, jun. de 2015. dirección: https://www.inegi.org.mx/temas/estructura/.

[29] Naldy Rodríguez, *Ataxia. La rara enfermedad que diezma a Tlaltetela*, Veracruz, México, nov. de 2016. dirección: http://www.eluniversal.com.mx/articulo/estados/2016/11/2/ataxia-la-rara-enfermedad-que-diezma-tlaltetela.

[30] *Tlaltetela, el pueblo donde la gente se paraliza | México | Edición América | Agencia EFE*, Veracruz, México, jun. de 2017. dirección: https://www.efe.com/efe/america/mexico/tlaltetela-el-pueblo-donde-la-gente-se-paraliza/50000545-3301219.

[31] M. P. Parsons y L. A. Raymond, «Huntington Disease», en *Neurobiology of Brain Disorders*, Elsevier, 2015, págs. 303-320, ISBN: 9780123982704. DOI: 10.1016/B978-0-12-398270-4.00020-3.

[32] G. Oz, C. D. Nelson, D. M. Koski, P.-G. Henry, M. Marjanska, D. K. Deelchand, R. Shanley, L. E. Eberly, H. T. Orr y H. B. Clark, «Noninvasive Detection of Presymptomatic and Progressive Neurodegeneration in a Mouse Model of Spinocerebellar Ataxia Type 1», *Journal of Neuroscience*, vol. 30, n.º 10, págs. 3831-3838, mar. de 2010, ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.5612-09.2010. dirección: http://www.jneurosci.org/content/jneuro/30/10/3831.full.pdfhttp://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5612-09.2010.

[33] M. Naqvi y a Kumar, «Step Counting Using Smartphone-Based Accelerometer», *International Journal*, vol. 4, n.º 05, págs. 675-681, 2012. dirección: http://www.enggjournals.com/ijcse/doc/IJCSE12-04-05-266.pdf.

[34] M. Bennasar, Y. Hicks, S. Clinch, P. Jones, A. Rosser, M. Busse y C. Holt, «Huntington's Disease Assessment Using Tri Axis Accelerometers», en *Procedia Computer Science*, vol. 96, 2016, págs. 1193-1201. DOI: 10.1016/j.procs.2016.08.163. dirección: www.sciencedirect.comhttp://linkinghub.elsevier.com/retrieve/pii/S1877050916319731.

[35] S. Shirai, I. Yabe, M. Matsushima, Y. M. Ito, M. Yoneyama y H. Sasaki, «Quantitative evaluation of gait ataxia by accelerometers.», *Journal of the neurological sciences*, vol. 358, n.º 1-2, págs. 253-8, nov. de 2015, ISSN: 1878-5883. DOI: 10.1016/j.jns.2015.09.004.

[36] C. Gavriel, A. A. C. Thomik, P. R. Lourenço, S. Nageshwaran, S. Athanasopoulos, A. Sylaidi, R. Festenstein y A. A. Faisal, «Towards neurobehavioral biomarkers for longitudinal monitoring of neurodegeneration with wearable body sensor networks», en *International IEEE/EMBS Conference on Neural Engineering, NER*, vol. 2015-July, 2015, págs. 348-351, ISBN: 9781467363891. DOI: 10.1109/NER.2015.7146631.

[37] K. Terayama, R. Sakakibara y A. Ogawa, «Wearable gait sensors to measure ataxia due to spinocerebellar degeneration», *Neurology and Clinical Neuroscience*, vol. 6, n.º 1, págs. 9-12, ene. de 2018, ISSN: 20494173. DOI: 10.1111/ncn3.12174. dirección: http://doi.wiley.com/10.1111/ncn3.12174.

[38] E. Sánchez-delacruz, F. Acosta-escalante, C. Boll-woehrlen, A. Hernández-nolasco y M. A. Wister, «Categorización de enfermedades neurodegenerativas a partir de biomarcadores de la marcha», *Komputer Sapiens*, vol. II, págs. 17-20, 2015.

[39] G. E. Barchini, «Métodos i+ d de la informática», *Revista de Informática Educativa y Medios Audiovisuales*, vol. 2, n.º 5, págs. 16-24, 2005.

[40] R Hernández Sampieri, «Metodología de la investigación, sexta edición méxico», *DF, Editores, SA de CV*, 2014.

[41] J. D. Lesmes, *Evaluación clínico-funcional del movimiento corporal humano*. Editorial M{é}dica Panamericana, 2007, ISBN: 9789589181614.

[42] W. Tao, T. Liu, R. Zheng y H. Feng, «Gait Analysis Using Wearable Sensors», *Sensors*, vol. 12, n.º 12, págs. 2255-2283, feb. de 2012, ISSN: 1424-8220. DOI: 10.3390/s120202255.

[43] A. I. Agudelo, T. J. Briñez, V. Guarín y J. P. Ruiz, «Marcha: descripción, métodos, herramientas de evaluación y parámetros de normalidad reportados en la literatura», *CES Movimiento y Salud*, vol. 1, n.º 1, págs. 29-43, 2013.

[44] J. Rueterbories, E. G. Spaich, B. Larsen y O. K. Andersen, *Methods for gait event detection and analysis in ambulatory systems*, 2010. DOI: 10.1016/j.medengphy.2010.03.007.

[45] C. Cifuentes, F. Martínez y E. Romero, «ANÁLISIS TEÓRICO Y COMPUTACIONAL DE LA MARCHA NORMAL Y PATOLÓGICA : UNA REVISIÓN», *Revista med*, vol. 18, n.º 2, págs. 182-196, 2010, ISSN: 0121-5256.

[46] J. Perry, J. R. Davids y col., «Gait analysis: Normal and pathological function», *Journal of Pediatric Orthopaedics*, vol. 12, n.º 6, pág. 815, 1992.

[47] D. Galimberti y E. Scarpini, *Neurodegenerative diseases*. Springer, 2018.

[48] S. I. Ahmad, *Neurodegenerative diseases*. Springer Science & Business Media, 2012, vol. 724.

[49] C. Dominguez, *Neurodegenerative diseases*. Springer Science & Business Media, 2010, vol. 6.

[50] J. G. N. Adolfo M. Bronstein, Thomas Brandt, Marjorie H. Woollacott, *Clinical Disorders of Balance, Posture and Gait*, 2d edition, J. G. N. Bronstein, Adolfo M, Thomas Brandt, Marjorie H. Woollacott, ed. Arnold, nov. de 2004, págs. 81-87.

[51] E. Buckley, C. Mazzà y A. McNeill, «A systematic review of the gait characteristics associated with Cerebellar Ataxia», *Gait & Posture*, vol. 60, págs. 154-163, feb. de 2018, ISSN: 0966-6362. DOI: 10.1016/J.GAITPOST.2017.11.024.

[52] F. Cardoso, *Huntington Disease and Other Choreas*, 2009. DOI: 10.1016/j.ncl.2009.04.001.

[53] L. Jones y A. Hughes, «Pathogenic mechanisms in huntington's disease», *International Review of Neurobiology*, 2011, ISSN: 00747742. DOI: 10.1016/B978-0-12-381328-2.00015-8.

[54] F. Aydin y Z. Aslan, «Classification of neuro-degenerative diseases using machine learning methods», *Intelligent Systems and Applications in Engineering*, vol. 5, n.º 1, págs. 1-9, 2017. dirección: www.atscience.org/IJISAE.

[55] J. M. Rodríguez Pupo, Y. Díaz Rojas, Y. Rojas Rodríguez, Y. Rodríguez Batista y E. Núñez Arias, «Actualización en enfermedad de Huntington», *Correo Científico Médico*, vol. 17, págs. 546-557, 2013, ISSN: 1560-4381.

[56] P Soliveri, C Mariotti, D Paridi, D Monza, C Tomasello, M Panzeri, F Taroni, A Albanese y F Girotti, «G07 Differences between Huntington's disease and spinocerebellar ataxia types 1 and 2 on cognitive and behavioural profile», *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 81, n.º Suppl 1, A33.1-A33, sep. de 2010, ISSN: 0022-3050. DOI: 10.1136/jnnp.2010.222646.7.

[57] S. Bech, T. Petersen, A. Nørremølle, A. Gjedde, L. Ehlers, H. Eiberg, L. E. Hjermind, L. Hasholt, E. Lundorf y J. E. Nielsen, «Huntington's disease-like and ataxia syndromes: identification of a family with a de novo SCA17/TBP mutation.», *Parkinsonism & related disorders*, vol. 16, n.º 1, págs. 12-5, ene. de 2010, ISSN: 1873-5126. DOI: 10.1016/j.parkreldis.2009.06.006.

[58] S. A. Schneider y T. Bird, *Huntington's Disease, Huntington's Disease Look-Alikes, and Benign Hereditary Chorea: What's New?*, jul. de 2016. DOI: 10.1002/mdc3.12312.

[59] M. Singh, M. Singh e I. Engineering, «Neuro-Degenerative Disease Diagnosis using Human Gait : A Review», *Computer Science & Electronics Journals*, vol. 7, n.º 1, págs. 16-20, 2013.

[60] R. Du, «Impact of Caryolanemagnolol on Gait and Functional Mobility on Individuals with Huntington's Disease», *Tropical Journal of Pharmaceutical Research*, vol. 14, n.º 9, págs. 1713-1717, 2015, ISSN: 1596-9827. DOI: 10.4314/tjpr.v14i9..

[61] T. Cruickshank, A. Reyes, L. Peñailillo, J. Thompson y M. Ziman, «Factors that contribute to balance and mobility impairments in individuals with Huntington's disease», *Basal Ganglia*, vol. 4, n.º 2, págs. 67-70, 2014, ISSN: 22105336. DOI: 10.1016/j.baga.2014.04.002.

[62] Huntington Study Group, «Unified Huntington's Disease Rating Scale: Reliability - and-Consistenc», *Movement Disorders*, vol. 11, n.º 2, págs. 136-142, 1996, ISSN: 0885-3185. DOI: 10.1002/mds.870110204.

[63] A Barbeau, R. Duvoisin, F Gerstenbrand, J. Lakke, C. Marsden y G Stern, «Classification of extrapyramidal disorders. proposal for an international classification and glossary of terms.», *Journal of the neurological sciences*, vol. 51, n.º 2, pág. 311, 1981.

[64] N. C. Reynolds, J. B. Myklebust, T. E. Prieto y B. M. Myklebust, «Analysis of gait abnormalities in Huntington disease», *Archives of Physical Medicine and Rehabilitation*, vol. 80, n.º 1, págs. 59-65, 1999, ISSN: 00039993. DOI: 10.1016/S0003-9993(99)90308-8.

[65] L. Seeberger, «Huntington's Disease», en *Encyclopedia of Movement Disorders*, SpringerReference_98361, 2003, págs. 1-14, ISBN: 9780123741059. DOI: 10.1016/B978-0-12-374105-9.00407-X.

[66] M. O. Miller-Keane, *Encyclopedia & dictionary of medicine, nursing & allied health.* 7.ª ed. Saunders, 2003, pág. 2344, ISBN: 9781455726240.

[67] J. Stephenson, T. Zesiewicz, C. Gooch, L. Wecker, K. Sullivan, I. Jahan y S. H. Kim, «Gait and balance in adults with Friedreich's ataxia», *Gait and Posture*, vol. 41, n.º 2, págs. 603-607, 2015, ISSN: 18792219. DOI: 10.1016/j.gaitpost.2015.01.002.

[68] L. Schöls, S. Peters, S. Szymanski, R. Krüger, S. Lange, C. Hardt, O. Riess y H. Przuntek, «Extrapyramidal motor signs in degenerative ataxias.», *Archives of neurology*, vol. 57, n.º 10, págs. 1495-500, oct. de 2000, ISSN: 0003-9942. DOI: 10.1001/archneur.57.10.1495.

[69] P. J. Garcia Ruiz, D. Mayo, J. Hernandez, S. Cantarero y C. Ayuso, «Movement disorders in hereditary ataxias», *Journal of the Neurological Sciences*, vol. 202, n.º 1-2, págs. 59-64, oct. de 2002, ISSN: 0022510X. DOI: 10.1016/S0022-510X(02)00211-3.

[70] T. Klockgether y M. Abele, «Hereditary ataxias», en *Handbook of Clinical Neurophysiology*, C, vol. 4, Elsevier B.V., 2004, págs. 655-673, ISBN: 9780444513595. DOI: 10.1016/S1567-4231(04)04038-9.

[71] A. M. Dueñas, R. Goold y P. Giunti, «Molecular pathogenesis of spinocerebellar ataxias», *Brain*, vol. 129, n.º 6, págs. 1357-1370, 2006, ISSN: 14602156. DOI: 10.1093/brain/awl081.

[72] T. D. Bird, «Hereditary Ataxia Overview - Review», en *Genetic Counseling*, et al. Adam MP, Ardinger HH, Pagon RA, ed., Seattle (WA): University of Washington, Seattle, nov. de 2008, págs. 1-39, ISBN: 2372-0697. DOI: NBK1138[bookaccession].

[73] M. E. Busse, C. M. Wiles y A. E. Rosser, «Mobility and falls in people with Huntington's disease.», *Journal of neurology, neurosurgery, and psychiatry*, vol. 80, n.º 1, págs. 88-90, ene. de 2009, ISSN: 1468-330X. DOI: 10.1136/jnnp.2008.147793.

[74] A. Fasano, J. Herzog, J. Raethjen, F. E. M. Rose, M. Muthuraman, J. Volkmann, D. Falk, R. Elble y G. Deuschl, «Gait ataxia in essential tremor is differentially modulated by thalamic stimulation», *Brain*, vol. 133, n.º 12, págs. 3635-3648, dic. de 2010, ISSN: 00068950. DOI: 10.1093/brain/awq267.

[75] K. Y. Chen y J. DAVID R BASSETT, «The technology of accelerometry-based activity monitors: Current and future», *Medicine & Science in Sports & Exercise*, vol. 37, n.º 11, S490-S500, 2005.

[76] P. A. Oberg, T. Togawa y F. A. Spelman, *Sensors in medicine and health care*. John Wiley & Sons, 2006, vol. 3, págs. 245-252.

[77] A. Godfrey, R. Conway, D. Meagher y G. ÓLaighin, «Direct measurement of human movement by accelerometry», *Medical engineering & physics*, vol. 30, n.º 10, págs. 1364-1386, 2008.

[78] M. J. Mathie, A. C. Coster, N. H. Lovell y B. G. Celler, «Accelerometry: Providing an integrated, practical method for long-term, ambulatory monitoring of human movement», *Physiological measurement*, vol. 25, n.º 2, R1, 2004.

[79] «A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring», *Sensors*, vol. 10, n.º 8, págs. 7772-7788, ago. de 2010, ISSN: 1424-8220. DOI: 10.3390/s100807772.

[80] O. J. Woodman, «An introduction to inertial navigation», University of Cambridge, Computer Laboratory, inf. téc., 2007.

[81] A. Allan, *Basic sensors in ios: Programming the accelerometer, gyroscope, and more*. .ºReilly Media, Inc.", 2011.

[82] Apple. Inc, *Framework Core Motion*, 2019. dirección: `https://developer.apple.com/documentation/coremotion` (visitado 28-06-2019).

[83] A. Galán-Mercant, F. J. Barón-López, M. T. Labajos-Manzanares y A. I. Cuesta-Vargas, «Reliability and criterion-related validity with a smartphone used in timed-up-and-go test», *BioMedical Engineering OnLine*, vol. 13, n.º 1, pág. 156, ago. de 2014, ISSN: 1475-925X. DOI: `10.1186/1475-925X-13-156`. dirección: `https://doi.org/10.1186/1475-925X-13-156`.

[84] B. Sun, Y. Wang y J. Banda, «Gait characteristic analysis and identification based on the iphone's accelerometer and gyrometer», *Sensors*, vol. 14, n.º 9, págs. 17 037-17 054, 2014, ISSN: 1424-8220. DOI: `10.3390/s140917037`. dirección: `http://www.mdpi.com/1424-8220/14/9/17037`.

[85] Y. Fujiki, P. Tsiamyrtzis e I. Pavlidis, «Making sense of accelerometer measurements in pervasive physical activity applications», en *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, ép. CHI EA '09, New York, NY, USA: ACM, 2009, págs. 3425-3430, ISBN: 978-1-60558-247-4. DOI: `10.1145/1520340.1520497`. dirección: `http://doi.acm.org/10.1145/1520340.1520497`.

[86] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian y W. Grundfest, «Implementation of an iPhone as a wireless accelerometer for quantifying gait characteristics», en *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, vol. 2010, IEEE, ago. de 2010, págs. 3847-3851, ISBN: 978-1-4244-4123-5. DOI: `10.1109/IEMBS.2010.5627699`.

[87] R. LeMoyne, T. Mastroianni, A. Hessel y K. Nishikawa, «Ankle Rehabilitation System with Feedback from a Smartphone Wireless Gyroscope Platform and Machine Learning Classification», en *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, IEEE, dic. de 2015, págs. 406-409, ISBN: 978-1-5090-0287-0. DOI: `10.1109/ICMLA.2015.213`.

[88] J. M. De Sa, *Pattern recognition: Concepts, methods and applications*. Springer Science & Business Media, 2012.

[89] M. N. Murty y V. S. Devi, *Pattern recognition: An algorithmic approach*. Springer Science & Business Media, 2011.

[90] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[91] G. Dougherty, *Pattern recognition and classification: An introduction*. Springer Science & Business Media, 2012.

[92] I. Guyon, S. Gunn, M. Nikravesh y L. A. Zadeh, *Feature extraction: Foundations and applications*. Springer, 2008, vol. 207.

[93] M. Nixon y A. S. Aguado, *Feature extraction and image processing for computer vision*. Academic Press, 2012.

[94] M. Mohri, A. Rostamizadeh y A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[95] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[96] H Jiawei, M. Kamber, J. Han, M. Kamber y J. Pei, *Data Mining: Concepts and Techniques*. 2012, pág. 745, ISBN: 978-0-12-381479-1. DOI: 10.1016/B978-0-12-381479-1.00001-0. arXiv: arXiv:1011.1669v3.

[97] P. Cichosz, «Classifcation model evaluation», en *DATA MINING ALGORITHMS EXPLAINED USING R*, Chichester, UK: John Wiley & Sons, Ltd, 2015, págs. 189-233, ISBN: 9781118950951. DOI: 10.1002/9781118950951.

[98] C. C. Aggarwal, *Data classification : algorithms and applications*. 2015, pág. 704, ISBN: 9781466586758.

[99] J. Friedman, T. Hastie y R. Tibshirani, «Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)», *Ann. Statist.*, vol. 28, n.º 2, págs. 337-407, abr. de 2000. DOI: 10.1214/aos/1016218223. dirección: https://doi.org/10.1214/aos/1016218223.

[100] Machine Learning Group at the University of Waikato, *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*, 2014. dirección: https://www.cs.waikato.ac.nz/ml/weka/http://www.cs.waikato.ac.nz/ml/weka/ (visitado 23-01-2019).

[101] G. I. Webb, «MultiBoosting: a technique for combining boosting and wagging», *Machine Learning*, vol. 40, n.º 2, págs. 159-196, 2000, ISSN: 08856125. DOI: 10.1023/A:1007659514849.

[102] L. Rokach y O. Maimon, *Data mining with decision trees: Theory and applications*. World scientific, 2014.

[103] T. K. Ho, «Random decision forests», en *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, IEEE, vol. 1, 1995, págs. 278-282.

[104] T. Ho, «The random subspace method for constructing decision forests», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, n.º 8, págs. 832-844, 1998, ISSN: 01628828. DOI: 10.1109/34.709601. eprint: 34.709601.

[105] L. Breiman, «Bagging predictors», *Machine learning*, vol. 24, n.º 2, págs. 123-140, 1996.

[106] J. Friedman, T. Hastie y R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[107] L. Breiman, «Random forests», *Machine Learning*, vol. 45, n.º 1, págs. 5-32, 2001, ISSN: 08856125. DOI: 10.1023/A:1010933404324. arXiv: /dx.doi.org/10.1023{\%}2FA{\%}3A1010933404324 [http:].

[108] P. Geurts, D. Ernst y L. Wehenkel, «Extremely randomized trees», *Machine learning*, vol. 63, n.º 1, págs. 3-42, 2006.

[109] J. R. Quinlan, *C4. 5: Programs for machine learning*. Elsevier, 2014.

[110] X. Wu y V. Kumar, *The top ten algorithms in data mining*. CRC press, 2009.

[111] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[112] C.-C. Chang y C.-J. Lin, «LIBSVM», *ACM Transactions on Intelligent Systems and Technology*, vol. 2, n.º 3, págs. 1-27, abr. de 2011, ISSN: 21576904. DOI: 10.1145/1961189.1961199.

[113] S. Raghu y N. Sriraam, «Classification of focal and non-focal eeg signals using neighborhood component analysis and machine learning algorithms», *Expert Systems with Applications*, vol. 113, págs. 18-32, 2018.

[114] S. B. Imandoust y M. Bolandraftar, «Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background», *International Journal of Engineering Research and Applications*, vol. 3, n.º 5, págs. 605-610, 2013.

[115] D. W. Aha, D. Kibler y M. K. Albert, «Instance-based learning algorithms», *Machine learning*, vol. 6, n.º 1, págs. 37-66, 1991.

[116] R. Das, «A comparison of multiple classification methods for diagnosis of Parkinson disease», *Expert Systems with Applications*, vol. 37, n.º 2, págs. 1568-1572, mar. de 2010, ISSN: 09574174.

[117] R. Martinez-Mendez, M. Sekine y T. Tamura, «Detection of anticipatory postural adjustments prior to gait initiation using inertial wearable sensors», *Journal of NeuroEngineering and Rehabilitation*, vol. 8, n.º 1, pág. 17, 2011, ISSN: 1743-0003. DOI: 10.1186/1743-0003-8-17.

[118] T. Shirakawa, M. Kamiura, A. Takagi y H. Sato, «Analysis for the gait patterns of healthy subjects during March», en *Procedia Computer Science*, vol. 24, Elsevier Masson SAS, 2013, págs. 167-174. DOI: 10.1016/j.procs.2013.10.040.

[119] Y. Watanabe, «Toward Application of Immunity-based Model to Gait Recognition Using Smart Phone Sensors: A Study of Various Walking States», *Procedia Computer Science*, vol. 60, págs. 1856-1864, 2015, ISSN: 18770509. DOI: 10.1016/j.procs.2015.08.296.

[120] W. Tao, T. Liu, R. Zheng, H. Feng, B. Sun, Y. Wang, J. Banda, S. Sprager y M. B. Juric, «Gait characteristic analysis and identification based on the iPhone's accelerometer and gyrometer», *Sensors (Switzerland)*, vol. 12, n.º 9, págs. 17 037-17 054, 2014, ISSN: 14248220. DOI: 10.3390/s140917037.

[121] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian y W. Grundfest, «Implementation of an iphone as a wireless accelerometer for quantifying gait characteristics», en *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, ago. de 2010, págs. 3847-3851. DOI: 10.1109/IEMBS.2010.5627699.

[122] R. LeMoyne, T. Mastroianni, A. Hessel y K. Nishikawa, «Ankle rehabilitation system with feedback from a smartphone wireless gyroscope platform and machine learning classification», en *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, dic. de 2015, págs. 406-409. DOI: 10.1109/ICMLA.2015.213.

[123] J. Klucken, J. Barth, P. Kugler, J. Schlachetzki, T. Henze, F. Marxreiter, Z. Kohl, R. Steidl, J. Hornegger, B. Eskofier y J. Winkler, «Unbiased and Mobile Gait Analysis Detects Motor Impairment in Parkinson's Disease», en, *PLoS ONE*, vol. 8, n.º 2, M. Toft, ed., e56956, feb. de 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0056956. dirección: http://dx.plos.org/10.1371/journal.pone.0056956 (visitado 04-12-2017).

[124] D. Trojaniello, A. Cereatti, A. Ravaschio, M. Bandettini y U. Della Croce, «Assessment of gait direction changes during straight-ahead walking in healthy elderly and Huntington disease patients using a shank worn MIMU.», *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2014, págs. 2508-11, ago. de 2014, ISSN: 1557-170X. DOI: 10.1109/EMBC.2014.6944132.

[125] A. Mannini, D. Trojaniello, U. Della Croce y A. M. Sabatini, «Hidden Markov model-based strategy for gait segmentation using inertial sensors: Application to elderly, hemiparetic patients and Huntington's disease patients», en *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, ago. de 2015, págs. 5179-5182, ISBN: 978-1-4244-9271-8. DOI: 10.1109/EMBC.2015.7319558. dirección: http://ieeexplore.ieee.org/document/7319558/.

[126] D. Joshi, A. Khajuria y P. Joshi, «An automatic non-invasive method for Parkinson's disease classification», *Computer Methods and Programs in Biomedicine*, vol. 145, págs. 135-145, jul. de 2017, ISSN: 18727565. DOI: 10.1016/j.cmpb.2017.04.007.

[127] N. I.o.B. I. National Institute of General Medical Sciences (NIGMS) y B. (NIBIB), *PhysioNet, the research resource for complex physiologic signals*. dirección: https://physionet.org/ (visitado 18-05-2017).

[128] S. Iram, D. Al-jumeily, P. Fergus, M. Randles y A. Hussain, «Computational Data Analysis for Movement Signals Based on Statistical Pattern Recognition Techniques for Neurodegenerative Diseases», 2012.

[129] Apple, *iPhone 5s - Technical Specifications*, 2014. dirección: https://support.apple.com/kb/SP685?locale=en{_}GBhttps://www.apple.com/uk/iphone-5s/specs/ (visitado 18-10-2017).

[130] Now Instruments + Software, *VibSensor User Guide - Now Instruments + Software*. dirección: http://www.now-instruments.com/get-help/5-vibsensor-user-guide (visitado 05-03-2019).

[131] M. Muaaz y R. Mayrhofer, «Orientation Independent Cell Phone Based Gait Authentication», en *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia - MoMM '14*, 2014, págs. 161-164, ISBN: 9781450330084. DOI: 10.1145/2684103.2684152.

[132] V. Chandel, A. D. Choudhury, A. Ghose y C. Bhaumik, «AcTrak - Unobtrusive activity detection and step counting using smartphones», en *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 131, 2014, págs. 447-459, ISBN: 9783319115689. DOI: 10.1007/978-3-319-11569-6_35. arXiv: 1412.6016.

[133] M. Muaaz y R. Mayrhofer, «Cross pocket gait authentication using mobile phone based accelerometer sensor», en *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*,

vol. 9520, Springer, Cham, feb. de 2015, págs. 731-738, ISBN: 9783319273396. DOI: 10.1007/978-3-319-27340-2_90.

[134] R. Subramanian y S. Sarkar, «Evaluation of Algorithms for Orientation Invariant Inertial Gait Matching», *IEEE Transactions on Information Forensics and Security*, vol. 14, n.º 2, págs. 304-318, feb. de 2019, ISSN: 1556-6013. DOI: 10.1109/TIFS.2018.2850032.

[135] MathWorks, *Filtering and Smoothing Data - MATLAB & Simulink - MathWorks America Latina*. dirección: https://la.mathworks.com/help/curvefit/smoothing-data.html{\#}bq{_}6ys3-1 (visitado 21-03-2018).

[136] J. H. Hollman, K. B. Childs, M. L. McNeil, A. C. Mueller, C. M. Quilter y J. W. Youdas, «Number of strides required for reliable measurements of pace, rhythm and variability parameters of gait during normal and dual task walking in older individuals», *Gait & posture*, vol. 32, n.º 1, págs. 23-28, 2010.

[137] «Adaptive step detection algorithm for wireless smart step counter», en *2013 International Conference on Information Science and Applications, ICISA 2013*, 2013, ISBN: 9781479906031. DOI: 10.1109/ICISA.2013.6579332.

[138] M. Yang, H. Zheng, H. Wang, S. McClean y D. Newell, «Igait: An interactive accelerometer based gait analysis system», *Computer Methods and Programs in Biomedicine*, vol. 108, n.º 2, págs. 715 -723, 2012, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2012.04.004. dirección: http://www.sciencedirect.com/science/article/pii/S016926071200106X.

[139] Y. Saeys, I. Inza y P. Larrañaga, «A review of feature selection techniques in bioinformatics», *Bioinformatics*, vol. 23, n.º 19, págs. 2507-2517, 2007.

[140] A. G. Karegowda, A. Manjunath y M. Jayaram, «Comparative study of attribute selection using gain ratio and correlation based feature selection», *International Journal of Information Technology and Knowledge Management*, vol. 2, n.º 2, págs. 271-277, 2010.

[141] A. Pinto, S. Pereira, H. Correia, J. Oliveira, D. M.L. D. Rasteiro y C. A. Silva, «Brain Tumour Segmentation based on Extremely Randomized Forest with high-level features», en *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, ago. de 2015, págs. 3037-3040, ISBN: 978-1-4244-9271-8. DOI: 10.1109/EMBC.2015.7319032.

[142] V. John, Z. Liu, C. Guo, S. Mita y K. Kidono, «Real-Time Lane Estimation Using Deep Features and Extra Trees Regression», en *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), T. Bräunl, B. McCane, M. Rivera e Y. Xinguo, eds., Switzerland: Springer, Cham, 2016, págs. 721-733, ISBN: 978-3-319-29450-6. DOI: 10.1007/978-3-319-29451-3_57.

[143] M. Soltaninejad, G. Yang, T. Lambrou, N. Allinson, T. L. Jones, T. R. Barrick, F. A. Howe y X. Ye, «Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI», *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, n.º 2, págs. 183-203, feb. de 2017, ISSN: 1861-6410. DOI: 10.1007/s11548-016-1483-3.

[144] A. Pinto, S. Pereira, D. Rasteiro y C. A. Silva, «Hierarchical brain tumour segmentation using extremely randomized trees», *Pattern Recognition*, vol. 82, págs. 105-117, oct. de 2018, ISSN: 00313203. DOI: 10.1016/j.patcog.2018.05.006.

[145] R. Ali, M. H. Siddiqi, M. Idris, B. H. Kang y S. Lee, «Prediction of Diabetes Mellitus Based on Boosting Ensemble Modeling», en *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services*, R. Hervás, S. Lee, C. Nugent y J. Bravo, eds., Belfast, UK: Springer, Cham, 2014, págs. 25-28, ISBN: 9783319131016. DOI: 10.1007/978-3-319-13102-3_6.

[146] A. Salih y A. Abraham, «Novel ensemble decision support and health care monitoring system», *Journal of Network and Innovative*, vol. 2(2014), págs. 041-051, 2014.

[147] J. Pereira, H. Peixoto, J. Machado y A. Abelha, «A Data Mining Approach for Cardiovascular Diagnosis», *Open Computer Science*, vol. 7, n.º 1, págs. 36-40, dic. de 2017, ISSN: 2299-1093. DOI: 10.1515/comp-2017-0007.

[148] M. Mirmozaffari1, A. Alinezhad y A. Gilanpour, «Data Mining Classification Algorithms for Heart Disease Prediction», *International Journal of Computing, Communication and Instrumentation Engineering*, vol. 4, n.º 1, págs. 14-15, ene. de 2017, ISSN: 23491477. DOI: 10.15242/IJCCIE.DIR1116008.

[149] A. M. Molinaro, R. Simon y R. M. Pfeiffer, «Prediction error estimation: A comparison of resampling methods», *Bioinformatics*, vol. 21, n.º 15, págs. 3301-3307, 2005, ISSN: 13674803. DOI: 10.1093/bioinformatics/bti499.

[150] J. Bellanca, K. Lowry, J. VanSwearingen, J. Brach y M. Redfern, «Harmonic ratios: A quantification of step to step symmetry», *Journal of Biomechanics*, vol. 46, n.º 4, págs. 828 -831, 2013, ISSN: 0021-9290.

[151] R. Moe-Nilssen y J. L. Helbostad, «Estimation of gait cycle characteristics by trunk accelerometry», *Journal of Biomechanics*, vol. 37, n.º 1, págs. 121 -126, 2004,

ISSN: 0021-9290. DOI: https://doi.org/10.1016/S0021-9290(03)00233-1. dirección: http://www.sciencedirect.com/science/article/pii/S0021929003002331.

[152] A. K. Tiwari, «Machine Learning Based Approaches for Prediction of Parkinson's Disease», *Machine Learning and Applications: An International Journal*, vol. 3, n.º 2, págs. 33-39, jun. de 2016, ISSN: 23940840. DOI: 10.5121/mlaij.2016.3203.

[153] D. G. Ramani Sivagami Shomona Gracia Jacob, «Feature Relevance Analysis and Classification of Parkinson Disease Tele-Monitoring Data Through Data Mining Techniques», *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, n.º 3, 2012.

[154] J. Novaković, «Toward optimal feature selection using ranking methods and classification algorithms», *Yugoslav Journal of Operations Research*, vol. 21, n.º 1, 2016.