Identificación de los intervalos de confianza, anchos de banda y evolución temporal de las principales señales sísmicas del volcán Llaima

Arequipa Moreta, Xavier Alexander y Villacrés Figueroa, Esteban Eduardo

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Trabajo de Integración Curricular, previo a la obtención del título de Ingeniero en Telecomunicaciones

Ing. Lara Cueva, Román Alcides PhD.

25 de agosto del 2023

Copyleaks

Plagiarism report

MIC_Arequipa_Villacrés_Copyleaks_C...

Scan details

Scan time: August 26th, 2023 at 0:59 UTC

Total Pages: 99

Total Words: 24552

Plagiarism Detection

Types of plagiarism		Words
Identical	0.7%	182
Minor Changes	0.1%	13
Paraphrased	4.6%	1135
Omitted Words	0%	0

AI Content Detection

Text coverage
- AI text
- Human text

Plagiarism Results: (30)

Tesis comportamiento eruptivo del volcan Llaima.pdf 1.8%

http://repositorio.udec.cl/jspui/bitstream/11594/1139/1/tesis...

Luis Franco

Universidad de Concepción Dirección de Postgrado Facultad de Ciencias Químicas Programa Doctorado en Ciencias Geológicas Comportamiento ...

ehernandez.pdf?sequence=1 0.7%

http://mriuc.bc.uc.edu.ve/bitstream/handle/123456789/7946...

Elimar

UNIVERSIDAD DE CARABOBO FACULTAD DE INGENIERIA AREA DE ESTUDIOS DE POSTGRADO MAESTRIA EN INGENIERIA ELECTRICA DESARROLLO DE UN...

Fase Exploratoria SURA LEGAL 0.6%

https://repositorio.konradlorenz.edu.co/bitstream/handle/00...

Alejandro Fandiño Benavides

FUNDACIÓN UNIVERSITARIA KONRAD LORENZ Tesis maestría MODELO DE RECOMENDACION DE ZONAS LABORALES A LOS BENEFICIARIOS DE...

Firmado electrónicamente por: ROMAN ALCIDES LARA CUEVA

.....

Ing. Lara Cueva, Román Alcides PhD.

Director

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Certificación

Certifico que el trabajo de integración curricular: "**Identificación de los intervalos de confianza, anchos de banda y evolución temporal de las principales señales sísmicas del volcán Llaima**" fue realizado por los señores **Arequipa Moreta, Xavier Alexander y Villacrés Figueroa, Esteban Eduardo** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizada en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 25 de agosto de 2023

Firmado electrónicamente por:
ROMAN ALCIDES LARA CUEVA

.....

Ing. Lara Cueva, Román Alcides PhD

C. C.: 1713988218

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

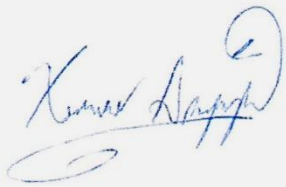
Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Responsabilidad de Autoría

Nosotros, **Arequipa Moreta, Xavier Alexander y Villacrés Figueroa, Esteban Eduardo**, con cédulas de ciudadanía n° 1725175069 y n° 1722168059, declaramos que el contenido, ideas y criterios del trabajo de integración curricular: **Identificación de los intervalos de confianza, anchos de banda y evolución temporal de las principales señales sísmicas del volcán Llaima** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 25 de agosto de 2023



.....

Arequipa Moreta, Xavier Alexander

C.C.: 1725175069

. .....

Villacrés Figueroa, Esteban Eduardo

C.C.: 1722168059

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Telecomunicaciones

Autorización de Publicación

Nosotros **Arequipa Moreta, Xavier Alexander y Villacrés Figueroa, Esteban Eduardo**, con cédulas de ciudadanía n° 1725175069 y n° 1722168059, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular: **Identificación de los intervalos de confianza, anchos de banda y evolución temporal de las principales señales sísmicas del volcán Llaima** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 25 de agosto de 2023

.....

Arequipa Moreta, Xavier Alexander

C.C.: 1725175069

.....

Villacrés Figueroa, Esteban Eduardo

C.C.: 1722168059

Dedicatoria

A mis padres, Fernando y Blanca que han sido determinantes para llegar hasta este punto de mi vida y que han dado todo lo mejor de sí para ser los mejores padres. A mi familia, que me han apoyado sin importar las circunstancias y a mis abuelos, con mención especial a Luis y Margarita, que siempre soñaron con verme triunfar. A mis amigos, por el apoyo emocional brindado y por su ayuda en momentos críticos. El aporte brindado a mi vida por parte de cada uno de ustedes ha sido importante para llegar hasta este punto de mi vida. Para ustedes.

Arequipa Moreta, Xavier Alexander

Este trabajo de titulación está dedicado a mi querida madre Ruth, cuyo amor y guía han sido mi brújula. A mi padre Hugo, cuya sabiduría y determinación ha sido y será mi ejemplo. A mi hermana Raquel, cuyo apoyo inquebrantable ha sido mi fuerza. En este emocionante punto de mi vida, quiero expresar mi profundo agradecimiento a cada uno de ustedes. Y a toda mi familia, su constante aliento ha sido mi impulso. Cada paso en este viaje ha sido moldeado por su amor y apoyo.

Villacrés Figueroa, Esteban Eduardo

Agradecimientos

Mi gratitud se extiende al gran creador de todo por permitirme ser bendecido con tantas cosas buenas en mi vida y permitirme llegar hasta este momento. Agradezco profundamente a mis padres por dar su mejor esfuerzo y brindarme su apoyo constante para que triunfe. No puedo dejar de mencionar a mi familia por su constante apoyo y por estar siempre al pendiente de mí, gracias por su cariño. Quiero expresar mi reconocimiento a mi Tutor Ing. Román Lara, cuya guía, conocimiento, paciencia y apoyo han permitido dar forma a este trabajo. A todos, gracias totales.

Arequipa Moreta, Xavier Alexander

En este trabajo de titulación quiero extender mi más sincero agradecimiento a todos quienes han desempeñado un papel fundamental en mi travesía universitaria. Desde mis compañeros y profesores, hasta mi amada familia, su apoyo ha sido mi ancla en momentos de desafío. A mi madre Ruth, padre Hugo y hermana Raquel, su amor inquebrantable ha sido mi refugio constante. Quiero hacer una mención especial a mi Tutor Ing. Lara Román, cuya orientación, consejos y respaldo han sido pilares cruciales en la creación de este trabajo. Cada paso que he dado lleva la influencia de su experiencia y sabiduría.

Villacrés Figueroa, Esteban Eduardo

Índice de Contenidos

Índice de tablas

Índice de figuras

Resumen

El conocimiento de los volcanes brinda información acerca de su historia eruptiva, como los estilos de erupción y eventos sísmicos presentes, esta información permite evaluar el peligro asociado a la actividad volcánica. La vigilancia y monitorización de un volcán comúnmente se lo realiza de forma visual por medio de personal capacitado en el área y observatorios vulcanológicos; por esta razón es necesario estudiar métodos que permitan identificar de una manera automática posibles erupciones volcánicas mediante esta monitorización y reconocimiento de microsismos, con el fin de anticipar eventos eruptivos devastadores para salvaguardar la mayor cantidad de vidas. En este trabajo de investigación se plantea identificar características para distinguir entre las distintas señales de microsismos entre las que constan: volcano-tectónico, largo período, tremor y tectónico. La caracterización al emplear técnicas de aprendizaje de máquina tradicional (ML, del inglés Machine Learning), selecciona las características principales de los datos recopilados. Para conseguir este objetivo se emplean dos clasificadores de aprendizaje supervisado: Árbol de Decisión (DT, del inglés Decision Tree) y Máquina de Vector Soporte. Además, se emplean los valores de energía de los coeficientes resultantes de la Transformada Wavelet (WT, del inglés Wavelet Transform) multinivel como características de entrada a los modelos de ML, donde, cada característica obedece a un rango en el dominio de la frecuencia. Como resultado se determina que la caracterización de las señales de microsismos está definida por los coeficientes wavelet de detalle cD3, cD5 y cD7, que corresponden al rango de frecuencias de 6.25 a 12.5 Hz, 1.56 a 3.12 Hz y 0.39 a 0.78 Hz respectivamente. Estas características surgen tras determinar los mejores resultados al utilizar una wavelet de la familia Daubechies, 10 niveles de descomposición en la WT y un clasificador DT. Para corroborar los resultados, se reentrena el modelo DT con las tres características principales del dominio de la escala y se logra obtener un 76.2% de exactitud de clasificación.

Palabras clave: Aprendizaje supervisado, microsismos, caracterización, wavelet.

Abstract

Detailed knowledge of volcanoes provides relevant information about their eruptive history, eruption styles, and seismic events, which are necessary to estimate the danger associated with volcanic activity. Traditionally, volcano surveillance and monitoring are conducted visually through volcanic observatories and trained personnel in the field. For this reason, there is a need to study methods that allow for the automatic identification of potential volcanic eruptions through the monitoring and recognition of microseisms, to anticipate potentially devastating eruptive events and thereby safeguard as many lives as possible. This research project aims to identify features for distinguishing among different types of microseismic signals, including volcano-tectonic, long period, tremor, and tectonic signals. The characterization employs feature selection techniques based on traditional machine learning theory. To achieve this goal, two supervised learning classifiers are used: Decision Tree (DT) and Support Vector Machine (SVM). Additionally, the coefficients resulting from the multi-level Wavelet Transform (WT) are used as input features for ML models, where each feature corresponds to a frequency domain range.

As a result, it is determined that the characterization of microseismic signals is defined by the wavelet coefficients cD3, cD5, and cD7, which correspond to frequency ranges of 12.5 to 25 Hz, 3.12 to 6.25 Hz, and 0.78 to 1.56 Hz respectively. These features emerge after obtaining the best results using a Daubechies wavelet, 10 levels of decomposition in the WT, and a DT classifier. To validate the results, the DT model is retrained with the three main scale domain features, achieving a 76.2% accuracy metric in classification.

Keywords: Supervised learning, microseisms, characterization, wavelet transform.

Capítulo I

Introducción

Los volcanes suscitan un interés constante debido a su capacidad para ocasionar efectos significativos y daños colaterales en las áreas circundantes. El volcán Llaima, situado en el Parque Nacional Conguillío, constituye un caso emblemático de un volcán activo que ha atraído la atención de la comunidad científica y de las poblaciones locales debido a su historial de erupciones y a su potencial riesgo. Con una extensión de más de 29 km² de glaciares en sus flancos occidental, suroccidental y oriental, el volcán Llaima muestra una manifestación impresionante de la interacción entre procesos geológicos y climáticos. (Naranjo & Moreno, 2011)

Una mirada al pasado eruptivo del volcán Llaima revela una serie de eventos que han dejado huella en su historia geológica. Desde un voluminoso depósito de flujo piroclástico datado hace 13200 años, el volcán ha experimentado episodios alternantes de erupciones efusivas y explosivas. Este registro eruptivo histórico, que abarca desde 1640 hasta 2009, está marcado por 48 eventos documentados, como en el caso de la erupción de 1640, la cual se destaca como una de las erupciones más intensa registrada en la historia.

La presencia de riesgos y peligros asociados a futuros eventos eruptivos del volcán Llaima agrega un nivel de urgencia a su estudio. La amenaza potencial de caídas de piroclastos, flujos de lava y lahares, a menudo desencadenados por la fusión de la cubierta glaciar y nival, resalta la importancia de comprender y predecir los patrones de actividad de este volcán (Naranjo & Moreno, 2011). Dentro del contexto de este estudio, se identifican señales de microsismos de gran relevancia, entre las que se incluyen las categorías de Volcano-Tectónico (VT, del inglés Volcano-Tectonic), Largo Período (LP, del inglés Long-Period), Tremor (TR) y Tectónico (TC, del inglés Tectonic).

Antecedentes

Varios métodos para la caracterización de microsismos han sido utilizados como el análisis de patrones temporales y magnitudes, el análisis de espectros de frecuencia a detalle y características de onda. Además, se han empleado técnicas avanzadas de procesamiento de señales y ML para identificar y distinguir distintos tipos de microsismos en función de sus propiedades únicas. Para este fin, también se han utilizado estimadores de entropía, análisis de densidad espectral de potencia (PSD, del inglés Power Spectral Density), transformadas de Fourier y Wavelet (Lara-Cueva, Bernal, Saltos, Benitez, & Rojo-Álvarez, 2014), así como técnicas de detección de actividad de voz (VAD, del inglés Voice Activity Detection) (Lara-Cueva, Benítez, Carrera, Ruiz, & Rojo-Álvarez, 2016). Seguido a la caracterización, se han clasificado los eventos mediante algoritmos de ML como DT y SVM, junto con técnicas de aprendizaje profundo con redes neuronales convolucionales (CNN, del inglés Convolutional Neural Networks) (Lara, Lara R, Larco, Carrera, & León, 2021).

Además, en estudios previos, se desea encontrar patrones o características que permitan diferenciar una señal de acuerdo con el tipo de evento. Conforme a una medición de tipo estocástica que abarca un gran número de características que varían con el tiempo, se encontraron diferencias entre cada evento sísmico en un estudio de señales obtenidas del Nevado del Ruiz en Colombia, tal como se presenta en la referencia (Cardenas, Orozco-Alzate, & Castellanos-Dominguez, 2013). En (Ibáñez & Carmona, 2000), se desarrolla otro ejemplo al estudiar los volcanes Stromboli y Etna en términos de ruido de fondo y LP en el caso del primer volcán, y VT y TR en el caso del segundo. Los autores emplearon el modelo oculto de Márkov (HMM, del inglés Hidden Markov Models) para encontrar y analizar un total de 39 parámetros relacionados con características temporales y espectrales, en donde se incluyen coeficientes que reflejan la evolución de la señal en el tiempo y la energía en diferentes bandas de frecuencia. Los resultados fueron de 84% y 86% respectivamente en el porcentaje de

clasificación de cada uno, cantidades que señalan éxito para ambos casos. En contraste, en el estudio realizado por (Alvarez, Henao, & Duque, 2007), se adopta un enfoque diferente al utilizar ventanas de 4s hasta 8s con un filtro pasa banda de 1 a 25 Hz para obtener características temporales y espectrales de los datos recopilados del volcán Colima situado en México. Como resultado, se generaron dos conjuntos de atributos: uno con 84 características y otro con 39 características, que se utilizaron al considerar la presencia o ausencia de armónicos, así como la envolvente espectral.

Justificación e importancia

A lo largo de la historia varias erupciones volcánicas han provocado grandes desastres naturales que han costado vidas humanas y daños materiales. En la actualidad, este peligro sigue latente en el día a día de pobladores que viven junto a varios volcanes activos. Según (Tilling & Beate, 1993), se ha estimado que cerca de 360 millones (aproximadamente el 10% de la población del planeta) de personas viven sobre o cerca de volcanes potencialmente peligrosos.

En la literatura científica, el volcán Vesubio fue el primer volcán con el que se relacionaron los microsismos. En 1847 se estableció un observatorio para estudiar dicho volcán y fue el primero en ser monitorizado con un equipo sísmico (Zobin, 2011). Al transcurrir el tiempo, se han registrado señales sísmicas relacionadas con la actividad volcánica en diferentes lugares del mundo gracias a que el progreso de vigilancia volcánica ha involucrado el uso de instrumentos electrónicos (Tilling & Beate, 1993).

Una erupción volcánica representa un suceso devastador capaz de desencadenar temblores, desprendimientos de suelo, fuegos e incluso tsunamis en formaciones geológicas adyacentes al océano. Domicilios, centros médicos, vecindarios e incluso poblaciones completas corren el riesgo de ser completamente arrasados en caso de que se materialice este evento de considerables proporciones. Esta eventualidad conlleva numerosas repercusiones y

perjuicios irremediables, al incluir la pérdida de vidas humanas a causa de heridas o quemaduras ocasionadas por los restos de edificios colapsados debido a las ondas sísmicas del volcán. Adicionalmente, las muertes derivadas de la lava emitida o de la inhalación de gases venenosos también forman parte de las consecuencias lamentables (Peng, 2008).

Para reducir lo máximo posible los daños y pérdidas causadas tras una erupción volcánica es preciso crear estrategias y métodos de alerta temprana para la población vulnerable debido a una erupción inminente. Un método es el desarrollo de sistemas de monitorización, los cuales usan instrumentos de alta precisión, que pueden registrar las señales sísmicas que viajan por la superficie o el interior de la tierra. Un sistema de reconocimiento automático de microsismos es un componente necesario para países propensos a erupciones volcánicas ya que al identificar si la actividad de los volcanes supera ciertos niveles establecidos, permite posteriormente recomendar el cambio de alerta en un tiempo oportuno y con esto definir el procedimiento de emergencia a seguir para las zonas más vulnerables, sea la evacuación o el refugio en áreas seguras. Además, este sistema estaría en las condiciones de emitir una alerta temprana en un evento inminente con el fin de salvaguardar vidas humanas.

Las señales de los microsismos son detectadas, almacenadas, caracterizadas y clasificadas por diferentes entidades en todo el mundo. El volcán Llaima, ubicado en la Región de La Araucanía de Chile, es monitoreado por el Observatorio Volcanológico de los Andes del Sur, mientras que en el Ecuador, el Instituto Geofísico de la Escuela Politécnica Nacional (IGEPN), capta las señales microsísmicas de los volcanes activos en el país mediante un detector STA/LTA, el cual posee un software especializado para la detección de diferentes sucesos en señales de tiempo, al identificar un breve aumento de la amplitud de la señal de entrada, comúnmente usado para microsismos, explosiones de diferente naturaleza, eventos

previstos en la realización de actividades de seguridad, entre otros (Armijos Sarango, Palacios Serrano, & González Martínez, 2021).

Ante el desafío de analizar la gran cantidad de datos generados por estos sistemas de monitorización y reconocimiento, se ha recurrido a la aplicación de técnicas de Aprendizaje de Máquina (ML, del inglés Machine Learning), como Árboles de Decisión (DT, del inglés Decision Tree) y máquinas de vector soporte (SVM, del inglés Support-Vector Machines). Sin embargo, la mayoría de las investigaciones existentes no han abordado el problema en tiempo real. Las señales más comunes en los volcanes son las LP y VT, mientras que TR y TC son menos comunes. La mayoría de los centros de monitorización volcánica procesa manualmente los datos de las señales, lo que puede resultar complicado debido a que la interpretación y el análisis de dichas señales pueden variar con el tiempo (Wener-Allen, Johnson, Ruiz, & Lees, 2005).

En este contexto, el objetivo general del trabajo de investigación es desarrollar y evaluar técnicas relacionadas con el reconocimiento automático de microsismos del volcán Llaima, basadas en la teoría de ML tradicional. Específicamente, se busca identificar los intervalos de confianza, anchos de banda y evolución temporal de las principales señales sísmicas del volcán. Para alcanzar este objetivo, se procede a emplear características representativas de las señales mediante la extracción de coeficientes con la Trasformada Wavelet. Asimismo, se aplica la Transformada Discreta de Wavelet (DWT, del inglés Discrete Wavelet Transform) multinivel para abordar el problema de resolución temporal-frecuencial que surge al emplear la Transformada de Fourier. El concepto fundamental de la Transformada Wavelet radica en obtener una señal filtrada en la escala al utilizar filtros pasa-bajo y pasa-alto, los cuales eliminan componentes de frecuencia alta o baja. El análisis Wavelet es una técnica que se basa en ventanas de tamaño variable, lo que lo hace adecuado para aplicaciones en las que se requiere extraer información de baja frecuencia con mayor detalle a lo largo de períodos de

tiempo más largos (Castro, 2002). Este concepto permite caracterizar eventos provenientes de señales de microsismos.

La importancia de clasificar señales volcánicas se enfoca en identificar eventos potenciales que resultarán en un posible desastre. En la actualidad, las señales captadas por el detector son clasificadas de manera visual por los analistas encargados de este proceso, mediante la interpretación subjetiva y específica. Estos analistas utilizan las características de las señales en el dominio del tiempo o de la frecuencia para etiquetar el tipo de evento que se presenta. Los observatorios de vulcanología, al obtener una gran cantidad de eventos que generan una enorme carga de trabajo y requieren una gran cantidad de tiempo para la clasificación de los microsismos, pueden presentar errores de clasificación debido al componente humano o una carga masiva de datos represada, problemas solucionados con sistemas modernos, como el sistema de reconocimiento automático de microsismos, mismo que reduce la carga para los trabajadores, aumenta la eficiencia y la fidelidad, evita errores y reduce el tiempo de procesamiento (Lara-Cueva, Bernal, Saltos, Benitez, & Rojo-Álvarez, 2014).

La clasificación por medio de DT, utiliza técnicas de podamiento y validación cruzada para identificar las características más importantes y evitar el sobreajuste en el clasificador. Además, se utiliza un algoritmo de ajuste de los eventos predichos por el clasificador en un proceso de posprocesamiento. El trabajo de investigación se enfoca en la importancia de automatizar la detección, caracterización y clasificación de microsismos para tomar medidas preventivas ante un posible proceso eruptivo y así salvaguardar la integridad de las comunidades directamente afectadas por la actividad volcánica.

Este trabajo de investigación presenta una revisión de trabajos relacionados en detección y clasificación de eventos vulcanológicos, además de describir la base de datos del volcán Llaima utilizada en las investigaciones y exponer detalladamente la metodología

empleada para la caracterización, extracción, selección y reconocimiento de microsismos. Posteriormente, se muestran los resultados experimentales obtenidos mediante diversas técnicas de clasificación y detección. Finalmente, se presentan los resultados y conclusiones obtenidas a lo largo del trabajo de investigación, junto con las posibles líneas de trabajo futuro. Con este estudio, se espera contribuir al avance en el monitoreo y prevención de peligros en el volcán Llaima y, potencialmente, en otros contextos volcánicos mediante el análisis de microsismos.

Considerados todos los elementos expuestos anteriormente, este trabajo de investigación adquiere una relevancia considerable tanto en un contexto académico y científico como en términos prácticos y sociales. El enfoque de este fenómeno trasciende hacia la práctica al abordar un asunto que tienen repercusiones inmediatas en la seguridad y el bienestar de las comunidades próximas al volcán Llaima.

Objetivo

Identificar de los intervalos de confianza, anchos de banda y evolución temporal de las principales señales sísmicas del volcán Llaima.

Actividades

1. Identificación de las propiedades y variables del problema a tener en cuenta.
2. Reunión de la base de datos representativa, con datos provistos por el observatorio Observatorio Volcanológico de los Andes del Sur.
3. Creación de una estructura de datos (temporales, espaciales, otros.) con un soporte común.
4. Pruebas con estimadores espectrales de series temporales y técnicas de re-muestreo.
5. Pruebas y evaluación del desempeño.

Capítulo II

Marco teórico

Volcán Llaima

El volcán Llaima es una de las montañas más importantes y extensas ubicada al sur de la cordillera de Los Andes, posee una latitud de -38.69 y longitud de -71.73. Se ubica en la Región de la Araucanía, Chile, con una altitud de 3,125 metros sobre el nivel del mar, un área basal de $500\ \mathrm{km}^2$ y un volumen de $400\ \mathrm{km}^3$. Este volcán ha mantenido actividad desde el Pleistoceno y ha sido objeto de estudio gracias a sus 49 erupciones documentadas desde 1640 hasta 2003, lo que lo convierte en una valiosa fuente de datos para fenómenos vulcanológicos (SERNAGEOMIN, 2023).

Figura 1

Volcán Llaima

Nota. El gráfico muestra una fotografía del Volcán Llaima. Tomado de (SERNAGEOMIN, 2023), por Felipe Flores.

Actividad sísmica del volcán Llaima

En el lapso del 2007 al 2009, se registraron las actividades más recientes del volcán Llaima, denotadas por un ciclo eruptivo formado por varias fases. Durante la primera fase, que transcurrió desde el 26 de mayo hasta el 31 de diciembre de 2007, se observaron explosiones menores acompañadas de la expulsión de cenizas. Las señales sísmicas captadas durante estas explosiones fueron identificadas como microsismos LP y, en ocasiones, como TR.

La siguiente fase marcó el comienzo de una erupción más significativa dentro del ciclo, con explosiones de mayor intensidad y una liberación notable de energía sísmica. Se calcula que alrededor de 5000 unidades promedio de la amplitud absoluta por minuto (RSAM, por sus siglas en inglés de Real-time Seismic Amplitude Measurement-system) fueron liberadas y atribuidas a un TR de alta energía. En el período que abarcó desde la tercera a la sexta fase, comprendido entre el 2 de enero y el 1 de julio de 2008, la actividad volcánica se caracterizó por episodios intermitentes de emisiones de ceniza y gases. Además, se observaron explosiones laterales aisladas y explosiones menores con expulsión de material fundido en forma de salpicaduras, denominadas comúnmente "spatters". Durante esta etapa, se observó una reducción en la actividad sísmica en la fase 6 (Franco Marín, 2019).

En la fase 7, se registró una reactivación del volcán en términos sísmicos, al producir 5 incidentes eruptivos. Esta fase marcó la última reactivación del volcán antes de que se experimente un súbito descenso tanto en la actividad eruptiva como en la actividad sísmica, lo cual tuvo lugar durante la fase 10 en el mes de abril de 2009.

Monitorización del volcán Llaima

La monitorización instrumental del volcán Llaima, empezó a la par con al análisis de otros volcanes chilenos, como el Villarrica, Calbuco, Osorno y Grupo Carrán-Los Venados, antes del año 2000. En aquel entonces, cada volcán tenía solo una estación que detectaba microsismos LP y los datos se almacenaban localmente sin someterse a un análisis detallado.

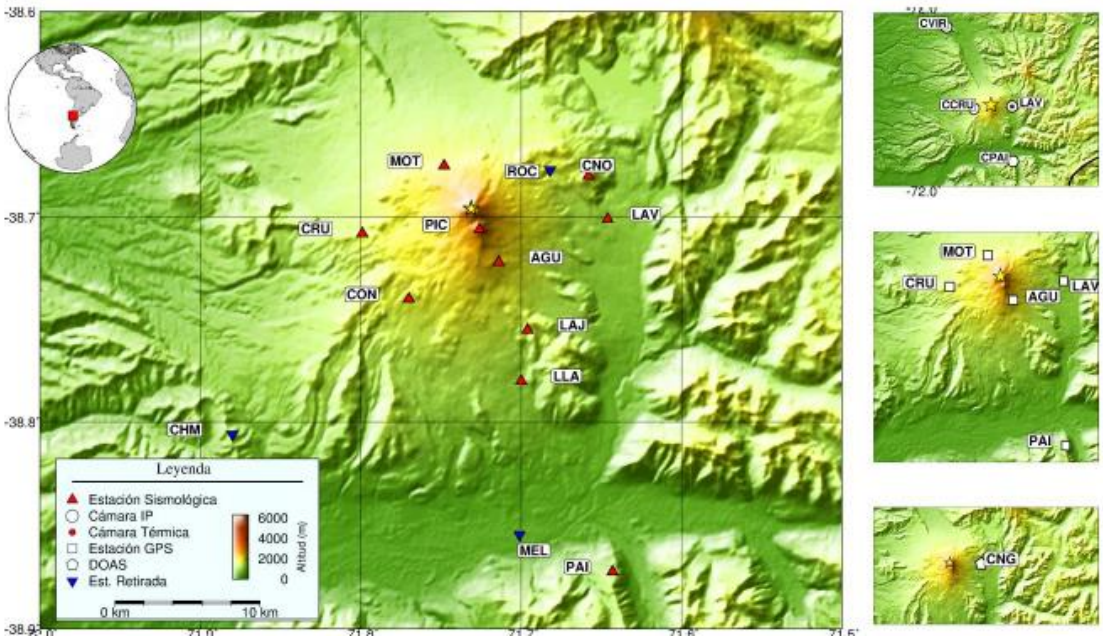
Antes de la erupción de 2008, el volcán Llaima contaba con solo dos estaciones operativas, ubicadas a 9,2 km y 17 km del cráter, respectivamente. Estas estaciones utilizaban un sensor L4C de una componente y un período natural de 1 Hz. Los datos se digitalizaban con una tasa de 50 muestras por segundo y se registraban en formato binario de 16 bits. Después de la erupción de enero de 2008, se instalaron dos estaciones sismológicas adicionales a 7,3 km y 21 km del volcán. En abril de 2008, las cuatro estaciones funcionaban de manera estable, transmitían telemetría corta y almacenaban datos en un nodo informático en la dependencia de Carabineros Chile en Melipeuco.

Luego de las erupciones de los volcanes Llaima y Chaitén en enero y mayo de 2008, se puso en marcha un proyecto para instalar instrumentación en los volcanes más activos de Chile. El Servicio Nacional de Geología y Minería (Sernageomin) estableció el departamento de la Red Nacional de Vigilancia Volcánica (RNVV) para ampliar y modernizar el Observatorio Volcanológico de los Andes del Sur (OVDAS). Gracias a este proyecto, se mejoró la red instrumental de monitorización de los 45 volcanes más peligrosos del país, incluido el volcán Llaima. Se añadieron más estaciones sismológicas, estaciones de GPS, inclinómetros electrónicos, cámaras de observación fija y sensores de gases por absorción espectroscópica (DOAS). Estas nuevas estaciones se ubicaron a distancias que variaban entre 1 km y 21 km con respecto al cráter activo (SERNAGEOMIN, 2023).

Posteriormente, se realizó un procesamiento manual de clasificación y cuantificación de los datos sísmicos, mediante métodos similares a los empleados por otros observatorios volcanológicos a nivel mundial. Los microsismos fueron discriminados, clasificados y localizados según criterios mínimos, la detección en al menos cuatro estaciones y la identificación de al menos cinco fases sísmicas. Los eventos tectónicos cercanos al volcán Llaima o locales se localizaron mediante el programa Hypo71pc, mientras que los LP se ubicaron de acuerdo con la metodología de Battaglia y Aki.

Figura 2

Red instrumental de monitorización del volcán Llaima



Nota. Se representan mediante diversos símbolos, la disposición actual del equipo utilizado para la monitorización de la actividad volcánica desde 2010 (Franco Marín, 2019).

Con el propósito de obtener información sobre las estructuras del sistema magmático del volcán Llaima, se desplegó una red temporal de estaciones de monitorización de microsismos de banda ancha y tres componentes durante el verano de 2015, en colaboración con las universidades Boise State y North Carolina at Chapel Hill de Estados Unidos, junto al OVDAS-RNVV-Sernageomin. Esta red temporal comprendía 26 estaciones distribuidas desde el cráter hasta distancias máximas de 35 km, con los datos almacenados en ordenadores locales. Este experimento permitió obtener imágenes de zonas con contrastes de velocidad mediante el análisis de ruido sísmico ambiental y función receptora.

Para analizar los datos sísmicos digitales correspondientes a los años 2007 a 2009, se convirtieron al formato ASCII. Se realizaron análisis para cuantificar la variación de la amplitud

de las señales, clasificar los microsismos y posteriormente ubicarlos. Los resultados obtenidos concuerdan con análisis basados en inclusiones fundidas en olivinos y proporciona indicios sobre la evolución del magma previa a la erupción de enero de 2008.

Microsismos

Los volcanes no solo emiten gases y materiales volcánicos, sino que también generan sucesos sísmicos volcánicos que son indicadores del comportamiento y cambios que experimentan. Estos eventos pueden ser causados por diversos factores, como el desplazamiento del magma a través de fracturas nuevas o preexistentes, la desgasificación del magma, la fracturación de la roca subterránea debido a cambios abruptos de temperatura, entre otros. Algunos de los eventos estrechamente relacionados con las señales sísmicas volcánicas son LP, TR, VT y HYB (de los Ángeles Linares, Ortiz, & Marreno, s.f.).

LP – Largo periodo

LP, también conocidos como microsismos de baja frecuencia, son fenómenos sísmicos relevantes para el estudio y monitorización de la actividad volcánica. La generación de estos microsismos ocurre debido a la vibración de grietas o canales llenos de fluidos en el interior de los volcanes. Además, se caracterizan por tener una duración más prolongada que los microsismos tradicionales, extendiéndose desde minutos hasta días y están vinculados a procesos magmáticos en el interior del volcán (IGEPN, 2014). Estos eventos se originan cerca de la superficie, a profundidades inferiores a 1 km y presentan magnitudes muy pequeñas con un contenido espectral limitado, junto con frecuencias generalmente en el rango de 1 a 5 Hz, aunque algunos eventos pueden concentrar la mayor parte de su energía entre los 5 y 10 Hz (CENAPRED, 2018).

TR – Tipo tremor

TR en volcanes son fenómenos sísmicos persistentes y no destructivos que generan señales sísmicas con amplitud constante durante un largo período de tiempo, que puede variar

desde minutos hasta horas. Estos eventos están asociados con el movimiento de fluidos en el interior del volcán, especialmente con el ascenso y desplazamiento del magma a través de fracturas y conductos dentro del sistema volcánico. En la clasificación, se identifican diferentes tipos, como los armónicos, espasmódicos, episódicos y los de banda ancha (Mora & Alvarado, 2001).

Las señales sísmicas de baja frecuencia y amplitud de estos eventos pueden persistir durante horas, días e incluso semanas y se considera un indicador importante de una actividad eruptiva inminente, ya que su persistencia y amplitud pueden aumentar significativamente antes de una erupción. No obstante, no todos los TR conllevan a una erupción inminente, ya que pueden estar vinculados a procesos de magma bajo la superficie sin presentarse una erupción.

VT – Volcano tectónicos

VT, son conocidos como de alta frecuencia, y se definen como fenómenos sísmicos relacionados con la actividad volcánica, derivados de la interacción de las fuerzas tectónicas en la corteza terrestre y procesos volcánicos (Yugsi, 2022).Una característica de estos eventos es que tienen una actividad frecuencial característica en el intervalo de 5 a 10 Hz, aunque ocasionalmente pueden ser más altas, además de ser provocados por el movimiento de placas tectónicas y la liberación de energía acumulada en zonas de fallas cercanas a volcanes activos, estos eventos ocurren debido a la fracturación de rocas cercanas a conductos volcánicos, lo que resulta en el rompimiento de rocas y la apertura de grietas en la estructura volcánica. En estos microsismos predominan frecuencias, por lo general ubicadas entre 5 a 15 Hz, aunque ocasionalmente pueden ser más altas, además de ser provocados por el movimiento de placas tectónicas y la liberación de energía acumulada en zonas de fallas cercanas a volcanes activos, estos eventos ocurren debido a la fracturación de rocas cercanas a conductos volcánicos, lo que resulta en el rompimiento de rocas y la apertura de grietas en la estructura volcánica.

TC – Tectónicos

TC en un volcán se refieren a los movimientos sísmicos causados principalmente por fuerzas tectónicas, es decir, las interacciones y movimientos de las placas tectónicas en la corteza terrestre. A diferencia de los eventos sísmicos relacionados con la actividad volcánica interna, como los microsismos y los tremores, los eventos tectónicos en un volcán están relacionados con la actividad tectónica regional o global y no necesariamente con el sistema magmático del propio volcán.

La liberación abrupta de energía acumulada es otra causa de estos eventos tectónicos al originar terremotos con epicentro en las fallas geológicas, fracturas en la corteza terrestre o zonas de choque de placas tectónicas. En el contexto volcánico, los eventos tectónicos pueden ocurrir en las cercanías del volcán debido a la interacción de las placas tectónicas en una región geológicamente activa (Minango, 2022).

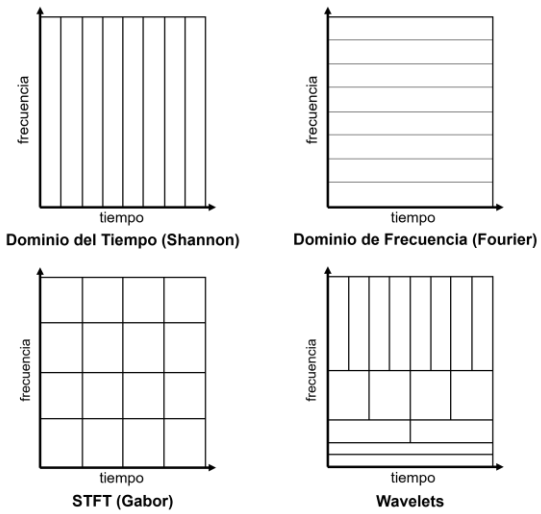
Análisis Wavelet

Gracias al avance de la ciencia y al progreso tecnológico, actualmente se dispone de diversas herramientas para analizar señales. Una de ellas es la transformación de una función o señal temporal, f(t), hacia una representación en un dominio diferente como, por ejemplo: la Transformada de Fourier (TF, del inglés Fourier Transform). El análisis de Fourier ha sido uno de los métodos más importantes para el tratamiento de señales eléctricas; sin embargo, esta representación presenta una notoria desventaja, dado que, al transformar una señal al dominio de la frecuencia, se pierde información temporal, lo que imposibilita la determinación del momento exacto en que ocurrió un evento específico. Esta peculiaridad resulta de gran relevancia en el caso de señales no estacionarias, en las cuales sus características estadísticas son variables a lo largo del tiempo como, por ejemplo, los microsismos volcánicos (Gómez, Silva, & Aponte, 2013).

Con el objetivo de superar esta limitación, Dennis Gabor llevó a cabo una adaptación del análisis de Fourier que se conoce como Transformada de Fourier de tiempo corto (STFT, del inglés Short Time Fourier Transform). Esta modificación considera una ventana de duración finita de la señal original, lo que permite realizar un análisis más localizado en el tiempo. La precisión en tiempo y frecuencia de la STFT está directamente influenciada por la duración y el ancho de banda de la función ventana utilizada. Si se emplea una ventana con mayor duración, se logra una mayor resolución en frecuencia, pero se sacrifica la resolución en tiempo. Además, la STFT carece de flexibilidad ya que mantiene una resolución fija en tiempo y frecuencia una vez se ha determinado la ventana (Kouro & Musalem, 2002) .

Figura 3

Esquema de resolución temporal y frecuencial de las distintas transformaciones.



Nota. El tamaño y la orientación del bloque indican el tamaño de la resolución.

Surgen entonces, las Wavelets. Se trata de una técnica similar a la utilizada con ventanas, pero con la particularidad de contar con regiones de tamaño variable, tal y como se muestra en la Figura 3. Estas funciones se usan en el análisis de señales, para obtener características relacionadas con el espacio, tamaño y dirección, con el objetivo de facilitar el

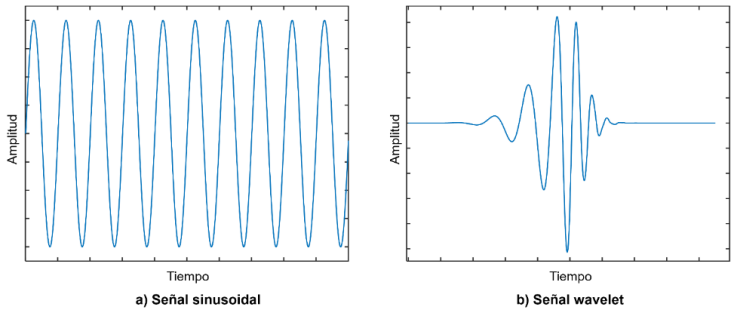
análisis de fenómenos transitorios, no estacionarios o variables en el tiempo (Burrus, Gopinath, & Guo, 1998). El término "Wavelet" se menciona por primera vez por Alfred Haar en su tesis de doctorado presentada en 1909 (Merry , 2005).

Wavelets

Las Wavelets son conjuntos de funciones espaciales que se caracterizan por ser pequeñas ondas con energía concentrada alrededor de un punto en el tiempo, tienen una duración limitada y con un valor medio de cero, lo que la diferencia del análisis de Fourier, en donde, las ondas sinusoidales no poseen una duración finita y se extienden desde menos hacia más infinito. En el análisis de Fourier las ondas sinusoidales son suaves y predecibles, mientras que las Wavelets pueden ser irregulares y asimétricas (GISI - UNAM, 2018). En Fourier, seno y coseno son las funciones base para este análisis. En el proceso de análisis Wavelet, se utilizan como base funciones focalizadas tanto en el dominio de la frecuencia (mediante la dilatación) como en el dominio temporal (a través de la traslación). (Nieto & Orozco , 2008). Esta comparación de características se puede apreciar claramente en la Figura 4, en donde se observa que la energía de la señal Wavelet se concentra alrededor de un punto.

Figura 4

Comparación de señales en el tiempo



Nota. a) Señal sinusoidal y b) señal Wavelet (Daubechies 10). Aunque existen varias familias de wavelet, se compara a la señal seno con una señal de la familia Daubechies 10 con fines ilustrativos.

Una característica distintiva de las Wavelets es su capacidad para representar señales de manera localizada, lo que significa que pueden detectar cambios o eventos específicos en una señal con alta precisión en el tiempo. Esto es especialmente útil en situaciones donde los eventos de interés son transitorios, no estacionarios o fenómenos variables en el tiempo, lo que dificultaría su detección mediante técnicas como el análisis de Fourier. El análisis Wavelet tiene la capacidad de revelar aspectos de las señales que otras técnicas de análisis no pueden percibir y que a menudo son pasados por alto como: la identificación de tendencias, picos de corta duración, puntos de ruptura y discontinuidades en las derivadas de orden superior (Cortés & Cano, 2007).

Wavelets comúnmente usadas

Existen diversas familias de wavelets, cada una con sus propias características y aplicaciones específicas. No existe una wavelet que sea adecuada para resolver todos los problemas, por lo que surgen estas familias de wavelets. Ciertas clases de wavelets pueden ofrecer mejores resultados en ciertas aplicaciones, mientras que en otras pueden no ser tan efectivas. La elección adecuada del tipo de wavelet depende en gran medida del principio de similitud, es decir, se debe seleccionar una wavelet que se asemeje lo más posible a la forma de la señal que se desea analizar.

Las familias de wavelets se pueden clasificar de acuerdo con su nivel de simetría:

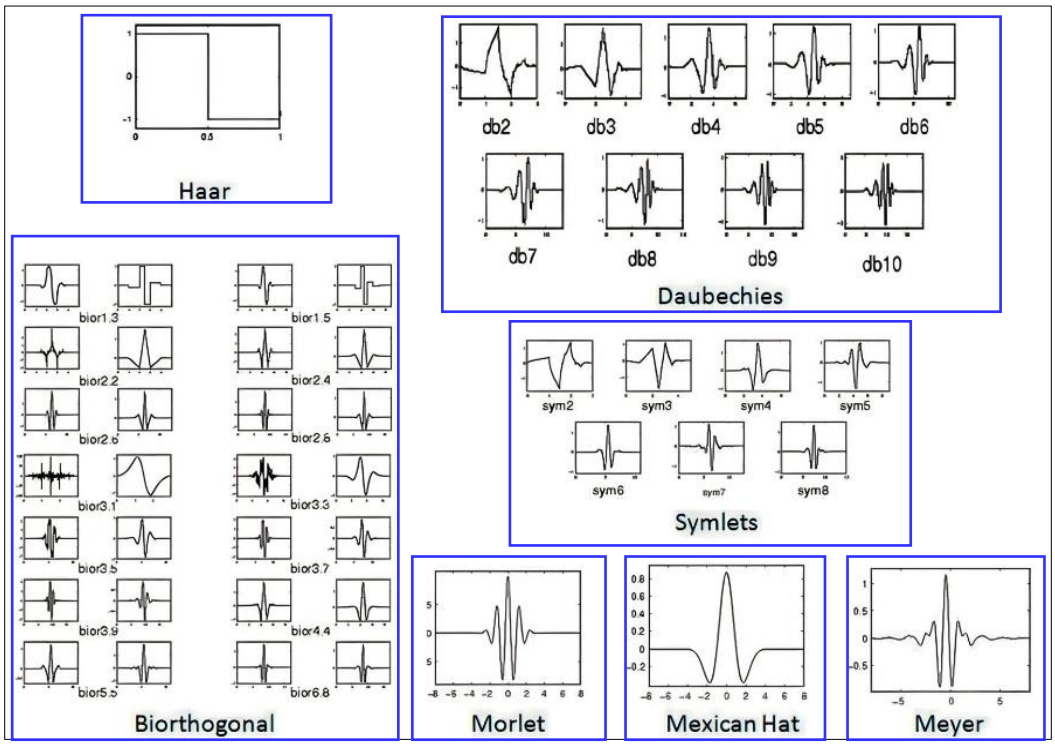
- No simétricas: Estas wavelets tienen forma antisimétrica en su forma básica. Su valor es cero en el punto central y tienen alternancia de signos en ambos lados del origen. Por ejemplo: wavelet Daubechies y wavelet Biortogonal.
- Simétricas: Estas wavelets tienen simetría par o impar en su forma básica. En el caso de las wavelets simétricas par, su forma es simétrica con respecto al eje vertical, mientras que las wavelets simétricas impar tienen simetría con respecto al origen. Por ejemplo: wavelet Haar y wavelet Mortet.

- Quasi-simétricas o de simetría mixta: Estas wavelets tienen características tanto simétricas como antisimétricas en su forma básica. Combinan elementos de simetría y antisimetría para adaptarse a diferentes aplicaciones. Por ejemplo: wavelet Coiflets y wavelet Symlets.

Cada una de estas familias de wavelets tiene propiedades y características específicas que los hacen adecuados para diferentes aplicaciones y tipos de señales. La elección de una familia de wavelets depende del problema específico que se desee resolver y las características de la señal a analizar (Pérez O. , 2004) (MathWorks, n.d.) (Fernández, 2007) . A continuación, en la Figura 5, se pueden observar distintas familias de wavelets en el tiempo.

Figura 5

Familias de wavelets de distintos órdenes disponibles en Matlab®.



Nota. Fuente (Barlowe, 2011)

Transformada Wavelet

La Transformada de Wavelet (WT, del inglés Wavelet Transform) implica la separación de una señal en distintos elementos de frecuencia mediante la aplicación de una función conocida como wavelet o wavelet madre (MW, del inglés Mother Wavelet). Esta función madre $\psi(t)$ genera una familia de funciones que son el resultado de traslaciones y dilataciones en la escala de la función original (Gómez, Silva, & Aponte, 2013). Este proceso de traslación y dilatación se define en la ecuación (1):

$$\psi_{\tau,a} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right), \tag{1}$$

donde τ realiza la traslación y a provee la dilatación (o escalamiento).

La transformada Wavelet permite el estudio de la señal en diversos niveles de detalle, desentrañando información acerca de las distintas frecuencias existentes en la señal mediante las wavelets generadas. Para lograr este objetivo, se requiere que la media del valor ponderado sea nula, que tenga un valor de energía limitada y que cumpla con un criterio ideal. Estos estándares se encuentran reflejados en las ecuaciones (2), (3) y (4):

$$\int_{-\infty}^{+\infty}\psi(t)dt = 0, \tag{2}$$

$$\int_{-\infty}^{+\infty}|\psi(t)|^2 d < \infty, \tag{3}$$

$$C_\psi = \frac{1}{\sqrt{2\pi}}\int_a^b \frac{\left|\widehat{\psi}(\omega)\right|^2}{|\omega|}d\omega < \infty, \tag{4}$$

donde $\widehat{\psi}(\omega)$ es la FT de $\psi(t)$.

La Transformada Wavelet Continua (CWT, del inglés Continuous Wavelet Transform) se define como aquella técnica que realiza una transformación continua de una señal en el

dominio del tiempo y la frecuencia. La CWT en función de la señal temporal y la MW se define por:

$$C(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\, \psi^* \left(\frac{t - \tau}{a} \right) dt, \tag{5}$$

donde $\overline{\psi(t)}$ es el complejo conjugado de $\psi(t)$.

No obstante, cuando se realiza la evaluación computacional de la CWT, se obtienen resultados con alta redundancia para la reconstrucción de la señal debido a la gran cantidad de información generada. Esto provoca un elevado tiempo de cálculo y un considerable costo computacional. Por este motivo, se utiliza la Transformada Wavelet Discreta (DWT, del inglés Discrete Wavelet Transform) la cual es una versión muestreada de la CWT y proporciona información suficiente para el análisis y reconstrucción de la señal, con una notable reducción en el tiempo de procesamiento. Bajo el enfoque DWT los parámetros a y t varían de acuerdo con las ecuaciones (6) y (7), respectivamente:

$$a = a_0^n \text{, donde } n = \mathbb{Z}^+, \tag{6}$$

$$\tau = k\tau_0 a_0^n \text{, donde } n = \mathbb{Z}^+. \tag{7}$$

Con el fin de obtener de manera rápida y optimizar recursos computacionales en los cálculos de los coeficientes Wavelet, se define un algoritmo DWT basado en análisis multiresolución (MRA, del inglés Analysis Multiresolution), creado con el fin de descomponer señales en tiempo discreto. Bajo este enfoque se establecen los valores de $a_0 = 2$ y $\tau_0 = 1$ y se define una descomposición mediante filtros pasa alto y pasabajo. Esto implica que la señal se somete a filtros de paso alto para las componentes de alta frecuencia y filtros de paso bajo para las componentes de baja frecuencia. Esto resulta en una alteración de la resolución de la señal a medida que la escala varía mediante operaciones de interpolación y submuestreo.

Descomposición Wavelet

Para realizar la descomposición Wavelet de una señal se usa el enfoque de Mallat, el cual propone la utilización de la DWT y una estructura de árbol binario basado en filtros conocido como algoritmo de Mallat. Dicho algoritmo propone emplear un banco de filtros que permite obtener la WT de manera breve y con el nivel de información adecuado, al crear una representación adecuada en tiempo y escala.

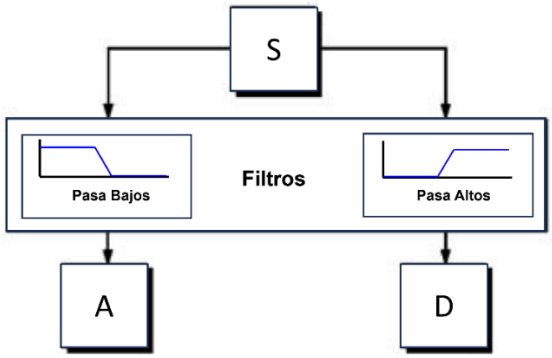
En la mayoría de las señales, la información principal, es decir la identidad de la señal, surgen a partir de sus componentes de baja frecuencia. Por otro lado, las componentes de alta frecuencia añaden características más específicas. Por esta razón, las componentes de una señal se dividen en dos categorías (Kouro & Musalem, 2002):

- Aproximaciones (componentes de baja frecuencia)
- Detalles (componentes de alta frecuencia)

Para lograr esta separación de componentes, se utiliza un proceso de filtrado, donde se aplican filtros específicos para extraer las aproximaciones y los detalles por separado, tal y como se ilustra en la Figura 6. En este enfoque de análisis, como se aprecia en la Figura 6, S es la señal para analizar, la misma que atraviesa dos filtros: el filtro pasabajos, que proporciona la salida A (aproximaciones) y el filtro pasa altos, que ofrece la salida D (detalles). Estos filtros están diseñados de manera complementaria para garantizar que la suma de A y D sea igual a la señal original S. Sin embargo, este método presenta una desventaja, pues duplica el número de datos originales, al generar un par de muestras (A, D) por cada muestra de S, lo que aumenta significativamente el costo matemático y computacional.

Figura 6

Diagrama de descomposición de señales con filtros de un nivel.

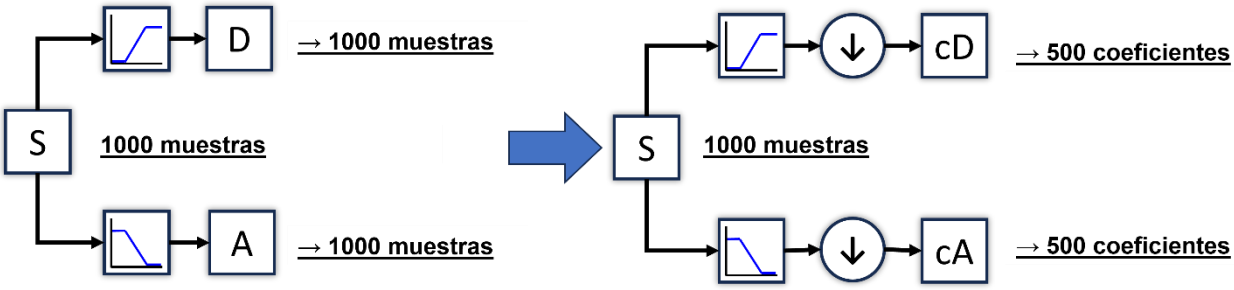


Nota. Fuente (Kouro & Musalem, 2002)

Para abordar esta limitación, se propone un método llamado submuestreo, que permite guardar únicamente la mitad de los puntos (A, D) sin sacrificar información relevante de la señal S. Esta estrategia reduce la cantidad de datos necesarios para el análisis, como se muestra en la Figura 7.

Figura 7

Esquema: descomposición de señales con filtros y método de submuestreo.



Nota. Fuente (Kouro & Musalem, 2002)

En la Figura 7, ↓ se representa la operación de submuestreo. Por otro lado, los coeficientes de detalle (cD, del inglés detail coefficients) y los coeficientes de aproximación (cA,

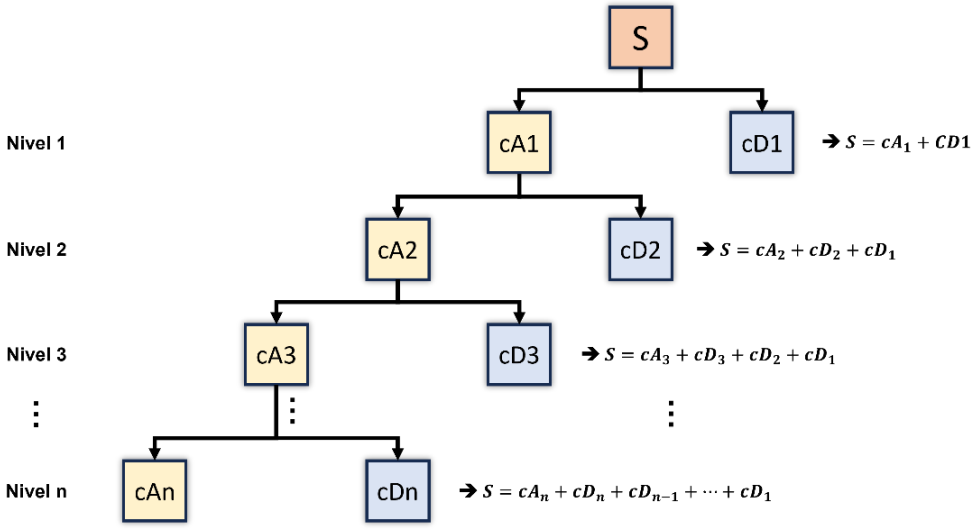
del inglés approximate coefficients) son los nuevos coeficientes obtenidos de la etapa de filtrado y que en conjunto forman la señal original S. Con esto se mantiene la misma cantidad de información o muestras. A modo de ilustración en la Figura 7, la señal original S contiene 1000 muestras, lo que resulta en dos series de aproximadamente 500 datos cada una. La noción de "aproximado" se debe a que, después del proceso de filtrado mediante la convolución de la señal de entrada con la función de transferencia del filtro (discreta), es posible que se agregue una o más muestras adicionales en la salida (Kouro & Musalem, 2002).

Descomposición Wavelet Multinivel

La mayoría de las señales no estacionarias suponen un caso más complejo, puesto que, descomponer simplemente con dos bandas de frecuencias (alta y baja) no es suficiente. Es necesario realizar una descomposición en múltiples niveles en cascada para separar las distintas características y poder analizarlas de manera independiente. De esta forma, surge la idea de utilizar filtros multiniveles para llevar a cabo este proceso de descomposición.

Figura 8

Diagrama de árbol de descomposición wavelet.



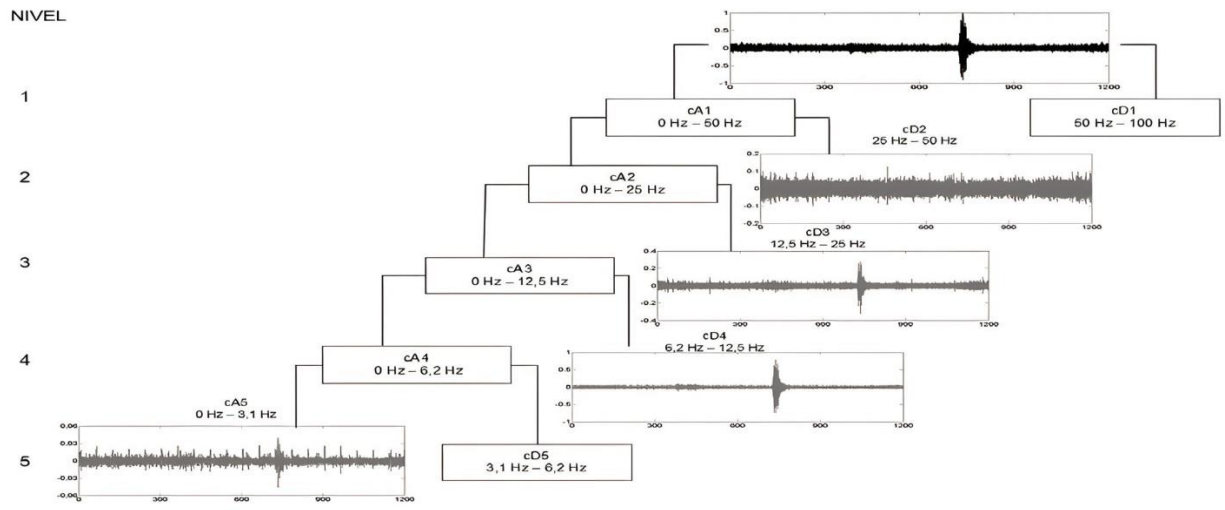
Nota. Elaboración propia.

Es posible lograr una descomposición de una señal en diversos niveles al pasar los coeficientes de escala obtenidos del filtrado previo a través de un conjunto de filtros idénticos, lo que resulta en la obtención de los coeficientes del siguiente nivel. Este proceso se repite consecutivamente hasta alcanzar el nivel de precisión deseado. Esta descomposición multinivel se le conoce como Árbol de descomposición Wavelet, tal cual se ilustra en la Figura 8.

En resumen, bajo este enfoque el funcionamiento de la WT implica filtrar una señal en el dominio del tiempo con el uso de filtros pasa bajo y filtros pasa alto, los cuales descartan características de alta o baja frecuencia de la señal original. Este proceso es repetitivo para las señales resultantes y genera una cascada de filtrado en etapas sucesivas de diferente orden, dependiendo de las repeticiones del proceso.

Figura 9

Diagrama de WT multinivel de un microsismo volcánico LP.



Nota. Fuente (Lara Cueva, Paillacho Salazar, & Villalba Chaluisa, 2017).

Durante la fase inicial del proceso de filtración, se procede a fragmentar la señal original en dos componentes luego de aplicarla a un filtro de paso inferior y otro de paso superior. Esta acción conlleva a la generación de dos variantes de la señal: una que encapsula rangos de

frecuencia desde 0 Hz hasta la mitad del límite máximo de frecuencia, y otra que engloba el intervalo desde la mitad hasta dicho límite superior. A continuación, se puede elegir una de estas dos versiones, generalmente la versión del filtro pasa bajo o inclusive ambas y emplear nuevamente el mismo proceso de división en etapas sucesiva (Martínez & Castro , 2002)s. Se ilustra un ejemplo de este proceso de descomposición en la Figura 9.

Machine Learning **y Algoritmos de Clasificación**

ML pertenece a la inteligencia artificial, la cual desea otorgar a las máquinas la capacidad de emular el comportamiento humano. Los sistemas de inteligencia artificial son usados para ejecutar tareas complejas y resolver problemas de una forma manera similar a las personas. El objetivo principal de la inteligencia artificial es desarrollar modelos computacionales que exhiban comportamientos inteligentes, tales como el reconocimiento visual de escenas, la comprensión de texto en lenguaje natural o la ejecución de acciones en el mundo físico (Marrón, 2021).

En su versión más fundamental, el ML emplea algoritmos que han sido programados para examinar datos de entrada, con el propósito de anticipar los valores de salida en un intervalo adecuado. Conforme se inyectan datos frescos en estos algoritmos, asimilan y perfeccionan sus funciones para elevar su desempeño y por lo tanto, adquieren progresivamente cierta forma de "inteligencia". En ML, se distinguen tres categorías de algoritmos: supervisados, no supervisados y por refuerzo.

- Aprendizaje Supervisado: Implica enseñar a la máquina a utilizar ejemplos. El operador proporciona un conjunto de datos conocidos que contienen entradas y salidas deseadas. El algoritmo es capaz de señalar patrones, es decir utiliza las observaciones para aprender y realizar predicciones supervisadas por el operador.

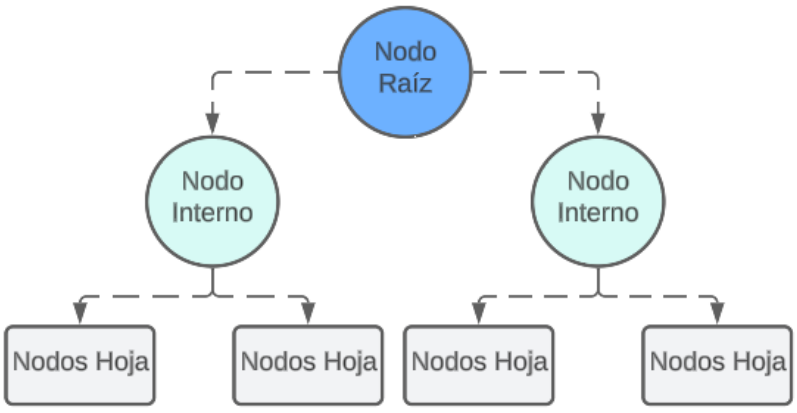
- Aprendizaje No Supervisado: El algoritmo estudia y examina los datos para detectar patrones, dejando de lado el recurso humano. Organiza y estructura los datos a medida que los analiza, lo que conduce a una comprensión más refinada con el tiempo.
- Aprendizaje por Refuerzo: Los algoritmos utilizan el método de prueba y error para aprender a partir de estas interacciones. Se les proporciona un conjunto de acciones y valores finales y a través de la exploración y evaluación, adaptan su enfoque para lograr los mejores resultados posibles en un proceso de aprendizaje guiado por la retroalimentación.

Árbol de Decisión (DT)

Los DT son algoritmos de aprendizaje supervisado sin parámetros que se emplean en labores de clasificación y regresión. Poseen una organización jerárquica con un nodo principal, ramificaciones, nodos intermedios y nodos finales.

Figura 10

Estructura de un árbol de decisión



Nota. Elaboración propia.

Tal como se ilustra en la Figura 10, un DT se inicia con un nodo principal sin ramificaciones entrantes. Las ramas que salen del nodo raíz conducen a los nodos intermedios, también llamados nodos de decisión. Tanto los nodos intermedios como los nodos de decisión llevan a cabo evaluaciones en función de las características disponibles para formar grupos homogéneos. Estos grupos son representados por nodos finales o nodos terminales. Los nodos finales engloban todas las posibles conclusiones dentro del conjunto de datos (IBM, 2022). Este procedimiento se consigue mediante la utilización de métricas de impureza, como la entropía o el índice de Gini, que se define según la ecuación (8). Estas métricas evalúan la uniformidad de las clases en cada subdivisión y tienden a acercarse a valores cercanos a cero cuando la proporción de una de las clases (ρ), es sumamente pequeña (InteractiveChaos, 2023)

$$G = 1 - \sum_{k} p_k^2. \tag{8}$$

La particularidad de este algoritmo radica en su capacidad para segmentar conjuntos de datos de entrenamiento en grupos coherentes, fundamentados en los valores más relevantes de las variables de entrada. La ventaja de este algoritmo es que permite manejar datos numéricos, datos grandes y complejos. Vale la pena mencionar que los árboles de decisión son algoritmos que muestran resistencia ante el ruido y valores atípicos, aunque tienden al sobre ajuste, en especial cuando son profundos y se adaptan en exceso a los datos de entrenamiento (Altamirano, 2021).

Específicamente, en un clasificador DT, cuando se construye un árbol de decisión, el algoritmo evalúa automáticamente qué características son las más relevantes para realizar particiones efectivas en los datos. Las características que se encuentran más arriba en el árbol,

es decir las más cercanas a la raíz, son las que tienen un mayor impacto en la división de los datos en diferentes ramas y éstas son las que más relevancia tienen dentro del modelo DT.

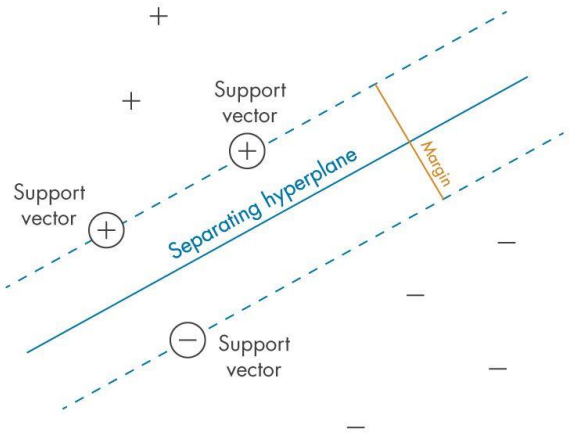
Máquinas de vector de soporte (SVM)

El algoritmo SVM encuentra aplicación en diversos escenarios de clasificación y regresión, esto abarca desde la interpretación médica de señales hasta el procesamiento de lenguaje natural, al pasar por el reconocimiento de imágenes y voz.

El propósito de SVM es identificar un plano en el espacio que divida dos categorías de puntos de datos los más diferentes posibles de una manera eficiente y efectiva. La noción de "de la manera más efectiva posible" implica hallar el plano con el margen más amplio entre ambas categorías, ilustrado mediante los signos "+" y "-" en la siguiente figura. En situaciones donde la separación lineal es factible, el algoritmo encuentra este plano óptimo. No obstante, en la mayoría de las situaciones prácticas, la SVM maximiza el margen de forma flexible, lo que permite un limitado número de clasificaciones incorrectas (MathWorks, 2023). En la figura 11, se muestra un ejemplo de la definición del margen entre clases.

Figura 11

Definición del "margen" entre clases.



Nota. (MathWorks, 2023).

El subconjunto de las observaciones para el entrenamiento se las conoce como vectores de soporte, las cuales identifican la posición del plano de separación. El algoritmo SVM es utilizado en problemas de clasificación binaria, resolviendo incluso problemas multiclase con la misma idea, al convertirlo en una serie de problemas binarios. SVM pertenece a una categoría de algoritmos de ML llamados métodos Kernel, donde una función de Kernel puede transformar las características. Estas funciones asignan los datos a un espacio dimensional diferente, generalmente de mayor dimensión, para facilitar la separación de las clases después de esta transformación. Esto simplifica los límites de decisión no lineales, volviéndolos lineales en el espacio dimensional transformado.

Métricas de desempeño

Una vez realizado el entrenamiento de los diversos algoritmos de clasificación, se evalúa el desempeño a través de métricas específicas. Estas métricas incluyen exactitud, precisión, sensibilidad, especificidad y la tasa de error balanceado, que permiten cuantificar y comparar su rendimiento. La obtención de estas métricas se basa en el análisis de la matriz de confusión generada para cada modelo, donde se detallan los conteos de las predicciones realizadas y los valores reales. La Tabla 1 expresa un modelo de la matriz de confusión, y de ella es posible derivar indicadores clave que brindan una perspectiva integral acerca del rendimiento y la calidad de la clasificación llevada a cabo por cada uno de los algoritmos.

Tabla 1

Matriz de confusión

		Predicción	
		Positivo	Negativo
Observación	Positivo	Verdadero Positivo – VP	Falso Negativo – FN
	Negativo	Falso Positivo – FP	Verdadero Negativo – VN

Nota. En las columnas se reflejan las predicciones del modelo, en contraste, las filas manifiestan las observaciones correspondientes a los datos reales.

- VP – Verdadero Positivo: Se refiere al caso en el que el modelo acertadamente categoriza una instancia como positiva.
- VN – Verdadero Negativo: Indica el acierto del modelo al clasificar una instancia como negativa.
- FP – Falso Positivo: Se produce cuando el modelo clasifica erróneamente una instancia como positiva, a pesar de ser negativa en realidad.
- FN – Falso Negativo: Se manifiesta cuando el modelo erróneamente clasifica una instancia como negativa, a pesar de ser positiva en la realidad.

A continuación, se define las métricas de desempeño:

Exactitud

La métrica de exactitud (A, del inglés Accuracy) brinda un enfoque general del desempeño del modelo y evalúa su capacidad para categorizar de manera correcta los casos positivos y los negativos dentro del conjunto total. Esta métrica brinda una visión global de la precisión del modelo al clasificar los datos.

Precisión

La precisión (P, del inglés Precision) es una medida que evalúa la capacidad del modelo para clasificar los casos positivos. Esta métrica ofrece información sobre la proporción de predicciones positivas que son efectivamente correctas, lo cual resulta valioso para evaluar la capacidad del modelo en la identificación de casos positivos.

Sensibilidad

La sensibilidad (R, del inglés Recall) refleja que tan capaz es el modelo para distinguir de manera precisa, dentro del conjunto de datos, los casos positivos. En otras palabras, la

sensibilidad mide la proporción de casos positivos reales que el modelo logra detectar correctamente.

Especificidad

La especificidad (S, del inglés Specificity) mide la habilidad del modelo para identificar correctamente los casos negativos. En otras palabras, la especificidad evalúa la capacidad del modelo para evitar clasificar incorrectamente casos negativos como positivos.

Tasa de error balanceado

La tasa de error balanceado (BER, del inglés Balance Error Rate) es una métrica que proporciona una visión global del rendimiento del modelo al considerar tanto los falsos positivos como los falsos negativos. La BER es la media aritmética de las tasas de error de las dos clases (positiva y negativa). Representa un equilibrio entre la sensibilidad y la especificidad y ayuda a comprender cómo el modelo se desempeña en general. Una BER más baja indica un mejor rendimiento.

Cálculo de cada métrica

Tabla 2

Fórmula de métricas de rendimiento en modelo ML.

Métrica de Desempeño	Fórmula
Exactitud (%)	$\mathrm{A} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100$
Precisión (%)	$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100$
Sensibilidad (%)	$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100$
Especificidad (%)	$\mathrm{S} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100$
BER	$\mathrm{BER} = 1 - \frac{\mathrm{R} + \mathrm{S}}{200}$

En la Tabla 2, se presentan las fórmulas para calcular cada una de las métricas que se derivan de la matriz de confusión generado por un modelo de clasificación con ML.

Densidad espectral de potencia

La Densidad Espectral de Potencia (PSD, del inglés Power Spectral Density), es una función matemática que proporciona información sobre cómo se distribuye la potencia de una señal en diferentes frecuencias en las que está compuesta. Esto permite identificar el rango de frecuencias donde se concentran los cambios de potencia. Analizar el comportamiento de señales en el dominio de la frecuencia resulta beneficioso, ya que facilita la detección de variaciones en comparación con el dominio del tiempo. Esto posibilita la comparación entre dos grupos y la detección de cambios en el parámetro estudiado (Luengas & Toloza, 2020).

Método Welch

El periodograma no constituye un estimador fiable de la verdadera PSD de un proceso estacionario en sentido amplio. El método de Welch, dirigido a reducir la variabilidad del periodograma, segmenta la serie temporal en fragmentos, que habitualmente presentan superposiciones.

El enfoque de Welch calcula una versión adaptada del periodograma para cada fragmento y luego promedia estas evaluaciones para formar la estimación de la PSD. Ya que el proceso ostenta estacionariedad en sentido amplio y el procedimiento de Welch se vale de estimaciones de PSD de diversos fragmentos de la serie temporal, los periodogramas modificados se asemejan a aproximaciones no correlacionadas de la PSD real y el proceso de promediado reduce la variabilidad.

Usualmente, los fragmentos se ajustan con una función de ventana, como la ventana Hamming, de manera que el método de Welch equivalga a la agrupación de periodogramas alterados. Dado que los fragmentos tienden a superponerse, los valores de datos al inicio y al

final del fragmento, ajustados por la ventana en un fragmento, se encuentran a cierta distancia de los finales de los fragmentos adyacentes. Esto actúa como una salvaguarda contra la pérdida de información debida a la disposición en ventanas (MathWorks, s.f.).

Intervalos de confianza

El intervalo de confianza destinado a la media es un rango que establece un conjunto de valores aceptables para la media de una población. En otras palabras, este intervalo provee un límite superior y uno inferior dentro de los cuales la media de una población puede situarse con un grado de error dado. El intervalo de confianza aplicado a la media se utiliza para estimar dos valores que encierran la media de una población, donde este intervalo es de gran utilidad para aproximar el valor promedio de una población cuando no se dispone de información acerca de todos sus valores (Probabilidad y Estadística.net, 2023).

Matemáticamente, un intervalo de confianza para una media poblacional, en la que la desviación estándar poblacional es conocida, se fundamenta en un concepto derivado del teorema del límite central, en el que se establece que la distribución de las medias obtenidas de diferentes muestras tiende a aproximarse a una distribución normal.

La fórmula del intervalo de confianza para la media nace a partir de un proceso llamado estandarización o normalización de una variable, el cual consiste en aplicar una transformación lineal a una distribución con el fin de que su media sea igual a cero y su desviación estándar sea igual a uno. En otras palabras, la estandarización consiste en restar la media y dividir entre la desviación típica. En este caso puntual, la estandarización de la variable aleatoria es:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), \tag{9}$$

donde:

- $\overline{X}$ es la media muestral.

- μ es la media poblacional.
- $\sigma/\sqrt{n}$ es el error estándar de la media.

Si se resuelve (9) en términos de Z se obtiene:

$$\mu = \overline{X} \pm Z\frac{\sigma}{\sqrt{n}}. \tag{10}$$

Bajo este contexto, para encontrar los intervalos de confianza de una población que sigue una distribución $Z \sim N(0,1)$ se debe encontrar el punto crítico $Z_{\alpha/2}$ para obtener un intervalo que contenga la media poblacional. Al tomar en cuenta que la distribución muestral para las medias es de tipo normal y a su vez, que la distribución normal es simétrica, se puede expresar (10) de la siguiente manera:

$$\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \ \mu \leq \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{11}$$

Por lo tanto, los límites superior e inferior se expresan como:

$$\left(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ , \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{12}$$

Es así como, en (12) se logran obtener los intervalos de confianza para la media μ con una desviación estándar σ conocida para una media μ. Está definido el valor crítico de $Z_{\alpha/2} = 1.96$ para muestras de tamaños grandes con nivel de confianza del 95% y el valor crítico de $Z_{\alpha/2} = 2.576$ para un nivel de confianza del 99%.

Esta fórmula es usada cuando se conoce el valor la desviación estándar de la población. No obstante, este parámetro es desconocido, que es el caso más frecuente. Si se desconoce la desviación estándar poblacional, se emplea la desviación estándar de la muestra.

En este caso, para la media μ con desviación estándar desconocida, el intervalo de confianza es calculado mediante la fórmula (13):

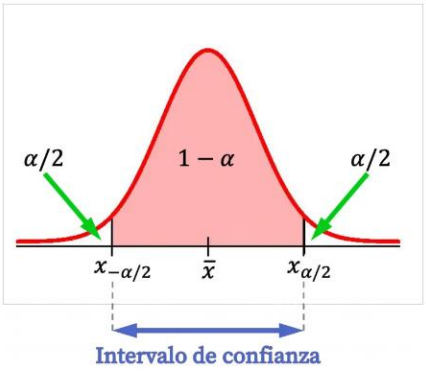
$$\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right), \tag{13}$$

donde:

- $\bar{x}$ es la media de la muestra.
- $t_{\frac{\alpha}{2}}$ es el valor de la distribución t de $\mathrm{Student}$ de $n-1$ grados de libertad con una probabilidad de $\frac{\alpha}{2}$.
- s es la desviación típica de la muestra.
- n es el tamaño de la muestra.

Figura 12

Intervalo de confianza



Nota. Nivel de confianza = $1-\alpha$ (Probabilidad y Estadística.net, 2023).

Selección de Características

La extracción de características resulta ser un recurso valioso en la capacitación de modelos de aprendizaje supervisado, sobre todo cuando se trata del algoritmo SVM. Esta técnica posibilita la identificación de las características más significativas dentro de un conjunto de datos extenso, con la consecuente reducción del tamaño de la base de datos y la

prevención de que características adicionales introduzcan perturbaciones en el proceso de entrenamiento del modelo. El propósito es mejorar el desempeño del modelo entrenado. Algunas ventajas de aplicar este enfoque incluyen:

- Obtención de modelos de entrenamiento más sencillos
- Disminución del costo computacional
- Reducción del sobreajuste

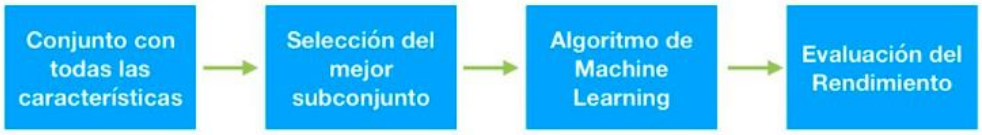
Existen diversos enfoques para extraer características, entre ellas se tienen el método de filtrado (del inglés Filters), el método de incrustación (del inglés Embedded) y el método de envoltura (del inglés Wrapper). Cada enfoque para extraer características es explicado, a continuación:

Método de Filtrado

El método de filtros es comúnmente utilizado como parte del preprocesamiento de datos y selección de características (aprendeIA, 2019). En la Figura 13 se muestra un esquema de ilustra su funcionamiento. Primero, se parte de un conjunto de características y se clasifican según su puntaje estadístico, que indica la correlación con la variable objetivo. Luego, se reduce este conjunto a un subconjunto más relevante. A continuación, se entrena el algoritmo de ML que utiliza estas características seleccionadas y, finalmente, se evalúa el rendimiento del modelo resultante.

Figura 13

Proceso usado por método de filtros



Nota. Fuente (aprendeIA, 2019)

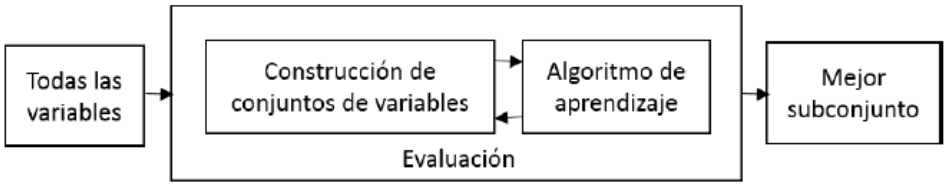
Método Wrapper

La característica central de la selección de atributos a través de enfoques wrappers radica en su incorporación del MAS en la evaluación de cada conjunto de variables. Debido a este aspecto, el proceso se vuelve computacionalmente intensivo. La búsqueda del conjunto óptimo de variables representa un problema NP que se vuelve prohibitivamente complejo a medida que crece el número de variables de entrada. Esto impulsa la necesidad de una estrategia de búsqueda que resalte las soluciones más prometedoras para su evaluación (Dorado Betancourt, 2019).

Adicionalmente, se requiere un criterio de evaluación que determine el rendimiento del MAS logrado con cada conjunto de variables. La estructura general de un enfoque wrapper se presenta en la Figura 14. En ella, se observa que el proceso de selección de atributos involucra al MAS como una entidad opaca, lo que permite la identificación del conjunto de variables más adecuado.

Figura 14

Enfoque de wrappers para la selección de atributos.



Nota. La selección de atributos participa como una caja negra que permite identificar el mejor subconjunto de variables.

Diversos enfoques han sido adaptados para la implementación de enfoques wrappers y la característica primordial que los distingue es el criterio de búsqueda empleado para definir los conjuntos de variables a evaluar. Los primeros enfoques se reconocen como métodos de complejidad exponencial y requieren la realización de un gran número de pruebas. Por ejemplo,

se encuentra la búsqueda exhaustiva, que, a pesar de asegurar una alta probabilidad de hallar la solución óptima, suele ser impracticable debido a los elevados costos computacionales asociados. Como resultado, estos métodos son difíciles de ejecutar, incluso en conjuntos de datos con un número reducido de variables de entrada.

Otra estrategia de búsqueda clásica es el enfoque secuencial, una variante de los algoritmos codiciosos que se apoyan en iteraciones sucesivas. Dentro de esta categoría se incluye el método de selección hacia adelante (forward), que comienza con un conjunto de variables vacío y, en cada paso, agrega una variable que mejore el rendimiento del MAS evaluado. Este proceso continúa hasta que no se observen mejoras al añadir más variables. Aunque el método de selección hacia adelante produce resultados en poco tiempo, su rendimiento a veces es limitado debido a su incapacidad para excluir variables una vez que han sido incorporadas en iteraciones anteriores (Dorado Betancourt, 2019). Por contraste, el método de eliminación hacia atrás (backward) comienza con todas las variables en el conjunto y, a lo largo del proceso, elimina variables de manera que el rendimiento del MAS mejore. La estrategia culmina cuando la exclusión adicional de una variable no genera mejoras significativas. Sin embargo, el método de eliminación hacia atrás tiende a volverse computacionalmente costoso al tratar con MAS que involucran un gran número de variables.

Método Embedded

Los enfoques embebidos surgen como una alternativa computacionalmente más ágil en comparación con los wrappers. Se destacan por llevar a cabo la selección de atributos en paralelo al proceso de entrenamiento y pueden ser incorporados directamente dentro del MAS o funcionar como una extensión del mismo. Estos métodos se pueden clasificar en tres categorías:

En primer lugar, se encuentran los métodos de poda. Estos implican entrenar el MAS con todas las variables y, posteriormente, descartar aquellas que, al ser eliminadas, mantengan

o mejoren el rendimiento del MAS. Este proceso puede ser ejecutado mediante enfoques secuenciales, búsquedas codiciosas o mediante medidas de contribución como vectores de pesos, información mutua, mínima redundancia-máxima relevancia, entre otros. Un ejemplo es la eliminación recursiva con máquinas de soporte vectorial (Dorado Betancourt, 2019).

En segundo lugar, se hallan los MAS con mecanismos de selección de atributos incorporados. En estos casos, durante el proceso de aprendizaje, las variables más informativas son filtradas o priorizadas internamente. Ejemplos de esto son los algoritmos ID3, CART y C4.5.

El tercer tipo son los MAS que emplean modelos de regularización. Estos contienen capacidades que disminuyen los coeficientes de algunas variables específicas, llegando incluso a valores que tienden a cero, simultáneamente se sigue la minimización del error de ajuste en el denominando MAS. Por tanto, dentro de este contexto, el coeficiente valor refleja la influencia de cada variable MAS. Por ejemplo, incluyen la Regresión Lasso o la Red Elástica (RE, del inglés Elastic Net).

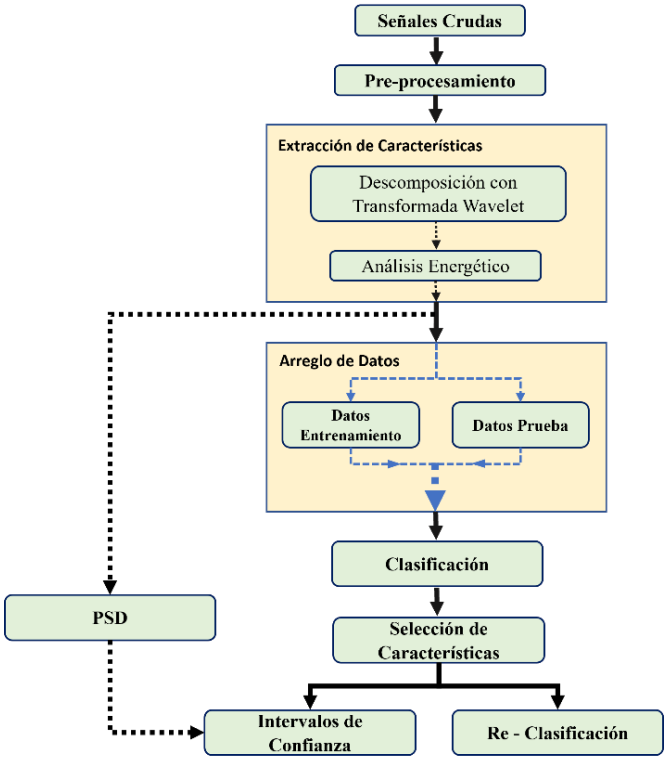
Capítulo III

Materiales y Métodos

Una vez se abordan todos los recursos teóricos necesarios para alcanzar los objetivos establecidos para el presente trabajo de investigación, se procede a detallar la metodología y cuáles son los materiales que se emplean durante su desarrollo. Para alcanzar estos objetivos se emplean simulaciones y programas desarrollados en la plataforma de programación y cálculo numérico llamada Matlab® en su versión R2022a®. Durante el desarrollo de los experimentos, el procedimiento se ejecutó sobre una PC Ryzen 5 con 2.10 GHz y 16 GB de RAM. A continuación, en la Figura 15, se resume el proceso a seguir para cumplir con el objetivo que se plantea para el presente trabajo de investigación.

Figura 15

Resumen de algoritmo de selección de características e intervalos de confianza.



Nota. Proceso establecido para la caracterización de eventos micro sísmicos.

Descripción general del programa

De manera general, el algoritmo desarrollado se expresa de manera escalonada en la Figura 15, y tiene como fin el identificar los intervalos de confianza en frecuencia que permiten caracterizar cada evento en función de la energía de los coeficientes de aproximación (cA) y coeficientes de detalle (cD) obtenidos a partir de la transformada wavelet de todas las señales micro sísmicas. Para ello, se usan técnicas de selección de características basadas en ML que permiten identificar cuáles son las bandas de frecuencia más importantes que permiten distinguir un suceso de otro. Con esto se logra encontrar las bandas de frecuencia únicas de cada evento. Más adelante se explica de manera más detallada cada bloque de la Figura 15.

Base de datos (Señales crudas)

El conjunto de datos utilizado en este estudio proviene del volcán Llaima y consiste en señales registradas en la estación LAV, una de las siete estaciones sísmicas de Llaima. Este conjunto de datos tiene la característica de que cada señal tiene un total de 60 [s] de duración y las mismas se adquirieron a una frecuencia de muestreo de 100Hz, con un total de 6000 muestras cada una. Estas señales han sido previamente sometidas a un proceso de filtrado mediante un filtro pasabanda Butterworth de décimo orden, en un rango de frecuencia [1, 10] Hz, con el propósito de preservar el ancho de banda que abarca el rango de interés.

El conjunto de datos crudos consta de un total de 3592 señales representadas por M y contenidas en la matriz:

$$\mathrm{S} = \left\{\mathrm{s}_1^{\mathrm{T}}, \mathrm{s}_2^{\mathrm{T}}, \dots, \mathrm{s}_{\mathrm{M}}^{\mathrm{T}}\right\}^{\mathrm{T}}, \tag{14}$$

donde s_{M} consta de 6000 muestras cada una. Dichas señales en la matriz S se encuentran organizadas por categoría y filtradas para seleccionar únicamente el segmento que mejor representa el microsismo en cuestión.

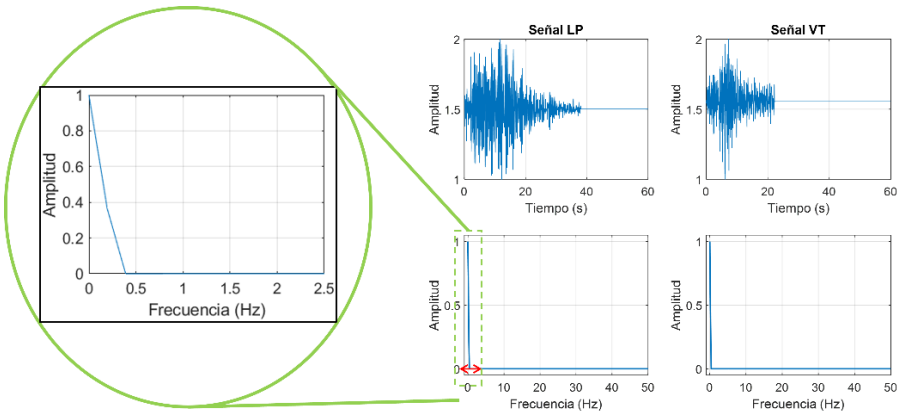
Preprocesamiento

Para el preprocesamiento de señales se establecen dos procesos que se deben implementar al conjunto de datos: normalización y filtrado. El hecho de normalizar las señales de microsismos tiene como fin ajustar y estandarizar las amplitudes de las señales para facilitar su análisis y comparación. Por otro lado, en el contexto de un análisis en frecuencia, las señales de microsismos poseen varios componentes espectrales ruidosos no relacionados que provienen de distintas fuentes. Es por ello que, es importante eliminar o atenuar aquellos componentes no deseados o ruido que puede estar presente en las señales crudas.

En la Figura 16 se representa la necesidad de realizar un proceso de filtrado y normalización de cada señal. En la Figura 16 se observa que las señales en el dominio del tiempo no se encuentran normalizadas por lo que, es importante realizarlo con el fin de que cada registro de evento tenga una media igual a cero y varianza igual a uno $(\mu = 0\,,\ \sigma^2 = 1)$. Por otro lado, se aprecia en la misma Figura 16 que en su espectro de frecuencias se encuentra un pico muy prominente en 0,2 Hz producto de microsismos de baja intensidad generados por el choque de las olas con la masa continental.

Figura 16

Análisis de dos señales crudas en tiempo y frecuencia.

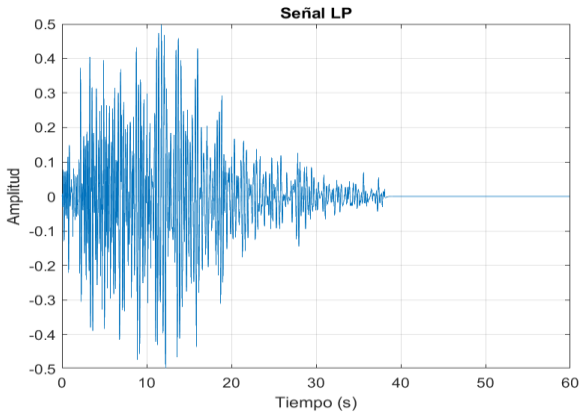


Nota. Cada señal en el dominio de la frecuencia presenta un contenido espectral en 0,2 Hz que pertenece a un ruido no relacionado al evento.

Tal y como se indicó anteriormente, para el proceso de normalización se establece un algoritmo que permita que cada microsismo tenga una media igual a cero y varianza igual a uno ($\mu = 0\,,\ \sigma^2 = 1$), con un rango de amplitud de cero a uno ([0 1]) desplazado hacia su media, tal y como se puede observar en la Figura 17.

Figura 17

Señal LP aleatoria normalizada

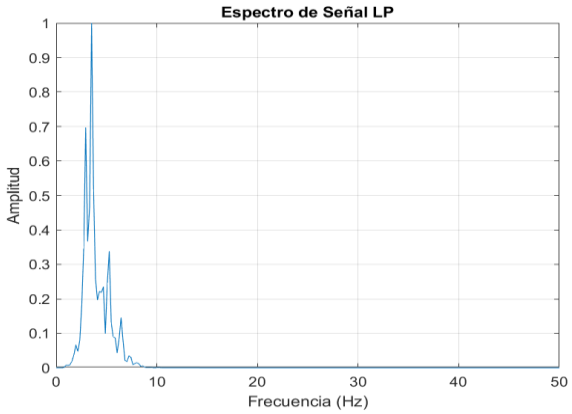


Nota. La señal normalizada tiene un rango de amplitud de cero a uno ([0 1]), sin embargo; ésta se encuentra desplazada hacia su media ($\mu = 0$).

Para el proceso de filtrado se aplica un filtro digital con respuesta finita al impulso (FIR, del inglés Finite Impulse Response), este elimina los componentes en frecuencia no deseados. En este caso, el filtro es de orden de 256. La frecuencia de corte mínima es de 0.7 Hz, mientras que 49.5 Hz es la frecuencia de corte superior. Un ejemplo del nuevo espectro que se obtiene a partir del proceso de filtrado y normalización se puede apreciar en la Figura 18, en donde el espectro de la señal representada en la Figura 16 presenta un notable cambio con respecto a su original no filtrada, para finalmente observar componentes espectrales en frecuencias lejanas a 0 Hz.

Figura 18

Ejemplo de señal LP filtrada.

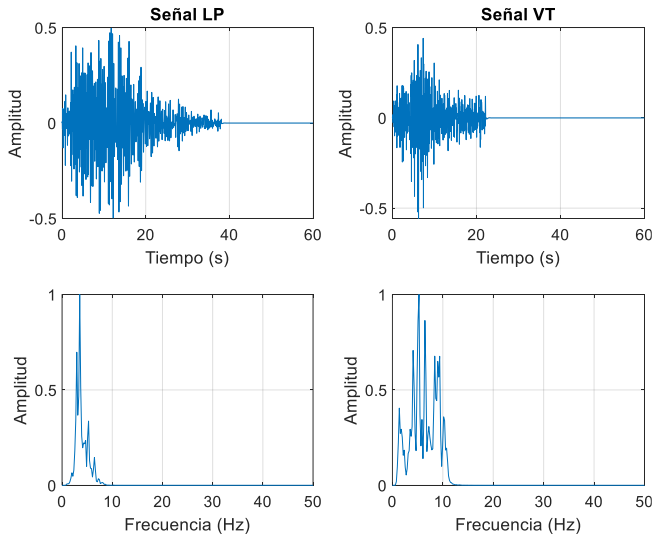


Nota. Al eliminar la prominencia en 0.2 Hz se observan nuevas componentes espectrales.

El resultado comparativo se puede observar en la Figura 19 en donde se aprecia el cambio de las amplitudes en el dominio del tiempo, producto del proceso de normalización y el cambio del espectro de potencias donde se aprecian componentes de frecuencia distintas a las que se observan en la Figura 16.

Figura 19

Señales LP y VT en tiempo y frecuencia después de preprocesamiento.



Nota. Es importante destacar el cambio que se produce en los espectros de potencia.

Expresado de manera matemática, dadas las $M = 3592$ señales de microsismos contenidas en la matriz S; el filtro FIR y el proceso de normalización se aplican en cada señal s_M. Con ello, se obtienen los vectores h_M como resultado de aplicar el operador $h_M\{s_M\}$. Dicho operador representa de forma matemática la aplicación del filtro y la normalización a cada señal cruda. Con esto, se genera la matriz H que contiene dichas señales:

$$H = \{h_1^T, h_2^T, h_3^T, \dots, h_M^T\}^T. \quad (15)$$

Descomposición con Transformada Wavelet

A través de la WT multinivel, se procede a obtener los coeficientes de aproximación (cA) y de detalle (cD) de cada una de las señales filtradas. Para el presente trabajo de investigación, se establece un análisis inicial enfocado en 6, 8 y 10 niveles de descomposición.

Dada la matriz H, se aplica la DWT a la misma. Esto genera los vectores $w_M = w_M\{h_M\}$, donde $w_M\{h_M\}$ representa el operador de la WT multinivel conformado por los coeficientes cD y cA. Este conjunto de vectores se encuentra en la matriz de coeficientes Wavelet W:

$$W = \{w_1^T, w_2^T, w_3^T, \dots, w_M^T\}^T. \quad (16)$$

La Tabla 3 detalla la información relevante de cada nivel de descomposición wavelet para el proceso de WT. En este trabajo, todas las señales presentan su componente de mayor frecuencia en 50 Hz, por lo que, el rango de frecuencia base a analizar es de 0 a 50 Hz. En el primer nivel de descomposición, al realizar la WT, se obtienen los coeficientes iniciales de aproximación y detalle. En este nivel, los coeficientes de aproximación (cA1) resultan del filtrado de la señal dentro del rango de frecuencias de 0 a 25 Hz con un filtro pasa bajo. Por otro lado, los coeficientes de detalle (cD1) surgen del filtrado en el rango de frecuencias de 25 a 50 Hz mediante un filtro pasa alto. Al avanzar al nivel 2 de descomposición, los coeficientes de

aproximación (cA2) se generan mediante el filtrado de las señales en el intervalo de 0 a 12.5 Hz, mientras que los coeficientes de detalle (cD2) se originan del filtrado de las mismas señales en el rango de 12.5 a 25 Hz. Este proceso se repite de manera consecutiva hasta alcanzar el nivel de descomposición deseado.

Tabla 3

División en niveles de frecuencia de la descomposición wavelet.

Nivel de Descomp.	Coeficiente de Descomp.	Rangos de Frecuencia [Hz]	Ancho de Banda [Hz]
1	**cD1**	25 - 50	25
2	**cD2**	12.5 - 25	12.5
3	**cD3**	6.25 – 12.5	6.25
4	**cD4**	3.12 – 6.25	3.12
5	**cD5**	1.56 – 3.12	1.56
6	**cD6**	0.78 – 1.56	0.78
7	**cD7**	0.39 – 0.78	0.39
8	**cD8**	0.195 – 0.39	0.195
9	**cD9**	0.097 – 0.195	0.097
10	**cD10**	0.0485 – 0.097	0.0485
10	**cA10**	0 – 0.485	0.0485

Nota. En la descomposición wavelet el ancho de banda y los rangos de frecuencia se reducen a la mitad en cada aumento del nivel de descomposición.

De acuerdo con el principio del diagrama de árbol de descomposición wavelet, que permite evitar la redundancia de información y disminuir el costo matemático y computacional, se procede a ignorar los coeficientes de aproximación, tal y como se indica en la Tabla 3, puesto que los únicos coeficientes de aproximación que aportan la información necesaria son aquellos que se encuentran en el último nivel de descomposición. Para este caso, se toman en cuenta los coeficientes de aproximación del nivel 6, 8 y 10 en sus respectivos casos de análisis.

El detalle de los coeficientes que se toman en cuenta en cada caso de análisis se detalla en la Tabla 4.

Tabla 4

Coeficientes de WT que se toman en cuenta en los distintos casos de análisis.

Nivel de Descomp.	Coeficientes de Descomposición	Dimensión
6	cA6, cD5, cD4, cD3, cD2	3592 x 5
8	cA8, cD7, cD6, cD5, cD4, cD3, cD2	3592 x 7
10	cA10, cD9, cD8, cD7, cD6, cA5, cD4, cD3, cD2	3592 x 9

Nota. Los coeficientes de aproximación se descartan debido a la redundancia de información.

Por último, se procede a escoger las MW. Por principio de similaridad, se escoge como MW a las familias de wavelets Daubechies y Symlets debido a su gran parecido con las señales de análisis. Por lo tanto, se realizan descomposiciones wavelet de 6, 8 y 10 niveles con dos familias de MW: Daubechies y Symlets de orden 10.

Análisis energético

La presente etapa consiste en obtener los valores de energía a partir de los coeficientes wavelet. Para proceder con el cálculo de la energía de la matriz W, a través de la ecuación:

$$\mathrm{E} = \sum_{\mathrm{i}=1}^{\mathrm{M}} |\mathrm{w}[\mathrm{i}]|^2, \tag{17}$$

donde w representa cada coeficiente wavelet de cada señal. De esta forma se genera la matriz E que contiene las energías:

$$\mathrm{E} = \left[\mathrm{e}_1^{\mathrm{T}}, \mathrm{e}_2^{\mathrm{T}}, \dots, \mathrm{e}_{\mathrm{M}}^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{17}$$

donde cada e_{M} representa $\mathrm{e}_{\mathrm{M}} = \left[\mathrm{e}_{\mathrm{M},1}, \mathrm{e}_{\mathrm{M},2}, \dots, \mathrm{e}_{\mathrm{M,N}}\right]$ y para $\mathrm{N} = 1, 2, \dots, 9$ características.

Tabla 5

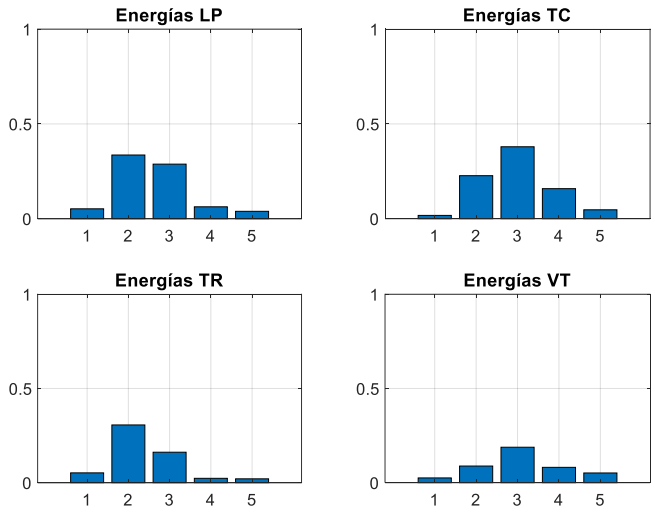
Esquema de valores a obtener tras el análisis energético.

Nivel de Descomp.	Evento	Valores de Energía	Dimensión $[m \times n]$
6	$\mathbf{LP_{m\times n}}$	$e_{cA6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{1310 \text{ x } 5}$
	$\mathbf{TC_{m\times n}}$	$e_{cA6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{1488 \text{ x } 5}$
	$\mathbf{TR_{m\times n}}$	$e_{cA6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{490 \text{ x } 5}$
	$\mathbf{VT_{m\times n}}$	$e_{cA6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{304 \text{ x } 5}$
8	$\mathbf{LP_{m\times n}}$	$e_{cA8}, e_{cD7}, e_{cD6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{1310 \text{ x } 5}$
	$\mathbf{TC_{m\times n}}$	$e_{cA8}, e_{cD7}, e_{cD6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{1488 \text{ x } 5}$
	$\mathbf{TR_{m\times n}}$	$e_{cA8}, e_{cD7}, e_{cD6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{490 \text{ x } 5}$
	$\mathbf{VT_{m\times n}}$	$e_{cA8}, e_{cD7}, e_{cD6}, e_{cD5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{304 \text{ x } 5}$
10	$\mathbf{LP_{m\times n}}$	$e_{cA10}, e_{cD9}, e_{cD8}, e_{cD7}, e_{cD6}, e_{cA5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{1310 \text{ x } 5}$
	$\mathbf{TC_{m\times n}}$	$e_{cA10}, e_{cD9}, e_{cD8}, e_{cD7}, e_{cD6}, e_{cA5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{1488 \text{ x } 5}$
	$\mathbf{TR_{m\times n}}$	$e_{cA10}, e_{cD9}, e_{cD8}, e_{cD7}, e_{cD6}, e_{cA5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{490 \text{ x } 5}$
	$\mathbf{VT_{m\times n}}$	$e_{cA10}, e_{cD9}, e_{cD8}, e_{cD7}, e_{cD6}, e_{cA5}, e_{cD4}, e_{cD3}, e_{cD2}$	$E_{304 \text{ x } 5}$

Por lo tanto, con los coeficientes identificados en la Tabla 6, se procede a obtener la energía para cada caso de análisis y posteriormente se comparan estos niveles de energía obtenidos para cada tipo de evento. Una vez se obtienen los niveles de energía para cada caso de análisis, se procede a graficar y comparar las energías de cada microsismo, tal y como se ilustra en la Figura 20, en donde se observan los niveles de energía de los microsismos LP, TC, TR y VT al realizar la transformada wavelet de quinto nivel.

Figura 20

Valores medios de energía de coeficientes wavelet con MW Daubechies10.



Nota. Los valores medios de energía de cada coeficiente se encuentran normalizados.

Además, se obtiene los diagramas de caja con los que se obtiene las siguientes informaciones:

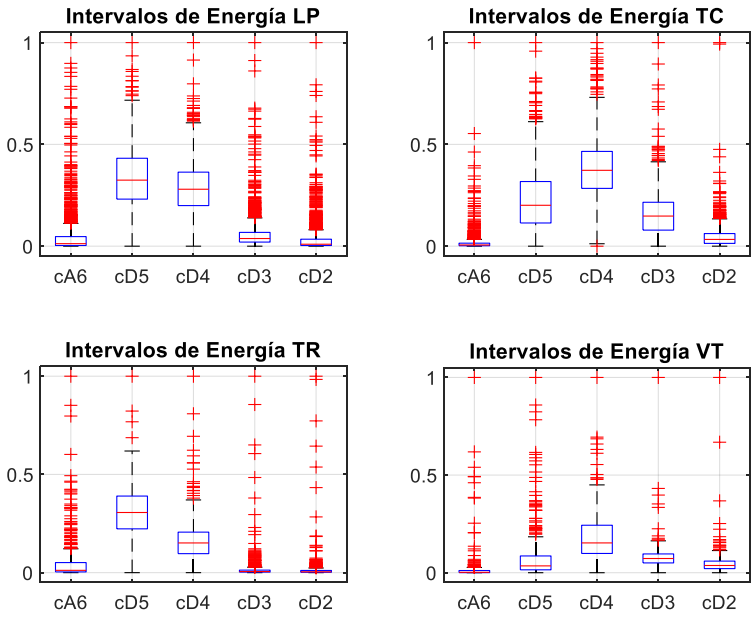
- Mediana (Q2): La línea en el centro de la caja representa la mediana, que es el valor que separa los datos en dos mitades iguales.
- Cuartiles (Q1 y Q3): El rango intercuartil (IQR), se extiende desde el cuartil (Q1) hasta el cuartil (Q3). Estos cuartiles indican los valores que separan el 25% inferior y superior de los datos, respectivamente.
- Rango intercuartil (IQR): La diferencia entre Q3 y Q1. Se utiliza para identificar valores atípicos.
- Bigotes (Whiskers): Los bigotes se extienden desde los bordes de la caja hasta los valores mínimo y máximo dentro del rango de 1.5 veces el IQR. Son considerados valores atípicos los que se encuentran fuera del rango y se representan en forma de puntos individuales.

- Valores atípicos: Puntos individuales más allá de los bigotes que podrían ser valores extremadamente altos o bajos en comparación con el resto de los datos, es decir, los puntos rojos que se observan en la Figura 21.
- Densidad de datos: La concentración de datos dentro de la caja y la presencia de valores atípicos proporcionan información sobre la densidad de los datos en diferentes partes de la distribución.

Como ejemplo ilustrativo, se observa en la Figura 21 el diagrama de caja para las energías de los cuatro microsismos.

Figura 21

Diagrama de caja a partir de las energías de WT con MW Daubechies10.



Nota. Los valores obtenidos en cada coeficiente se encuentran normalizados.

Selección de niveles de descomposición

Tal y como se menciona previamente, se establece un análisis con 6, 8 y 10 niveles de descomposición. En este punto del desarrollo del trabajo de investigación, se establece un

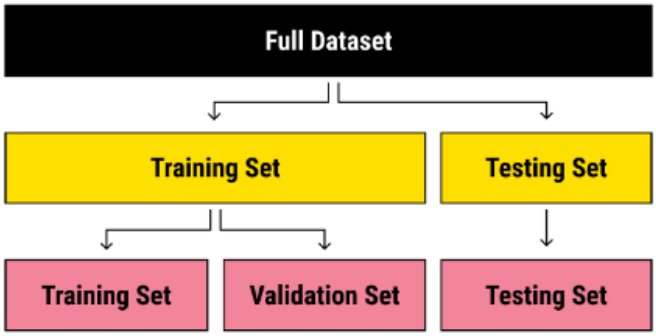
proceso de análisis manual con el fin de determinar cuál es el nivel de descomposición de WT adecuado que ayuda a diferenciar un microsismo del resto. Preliminarmente, en la Figura 20, se observa que los distintos valores medios de energía que caracterizan a cada evento son similares para LP y TR. Además, puesto que la finalidad del trabajo de investigación es encontrar características únicas de cada evento mediante la WT, se evidencia que es apropiado aumentar el número de características iniciales, es decir, generar más coeficientes de aproximación y detalle mediante la WT en niveles superiores. Debido a que ciertos eventos son similares en energía de wavelet y a que existen pocas características iniciales, se procede a descomponer en niveles superiores: 8 y 10 niveles de descomposición. Una vez se determine cuál es el nivel de descomposición adecuado, el resto se descarta.

Arreglo y partición de características

Una vez que se selecciona el nivel adecuado de descomposición en la WT, se procede a preparar los coeficientes wavelet (características) de manera que sean adecuados para ingresar a un algoritmo de ML, es decir, los coeficientes se etiquetan de acuerdo con el evento al que pertenecen y se particionan los datos de acuerdo con el esquema de la Figura 22, en donde se observa la manera correcta de particionar los datos.

Figura 22

Esquema de particionamiento de datos de entrada para un modelo ML.



Nota. Fuente (Sydorenko , 2021).

En este caso, los coeficientes wavelet obtenidos anteriormente son los datos de entrada al modelo ML. Además, los porcentajes de partición de los datos que se establece para el presente trabajo es de 80/20 tal y como se expresa en la Tabla 6, en donde se ilustra la manera y la proporción correcta de dividir los datos.

Tabla 6

Energías WT que se toman en cuenta en los distintos casos de análisis.

Nivel de Descomp.	Porcentaje	Coeficientes de Descomposición	Dimensión
Dataset	**100 %**	$\mathrm{D} = \left[\mathrm{E_{LP}}^{\mathrm{T}}, \mathrm{E_{TC}}^{\mathrm{T}}, \mathrm{E_{TR}}^{\mathrm{T}}, \mathrm{E_{VT}}^{\mathrm{T}} \right]^{\mathrm{T}}$	3592 x 9
Train	**80 %**	$D_{Train} = [D_{0.8M\times 9}]$	2874 x 9
Test	**20 %**	$D_{Test} = [D_{0.2M\times 9}]$	718 x 9

Nota. Recordar que M es 3592.

Clasificación

Para distinguir entre diferentes eventos y seleccionar las características más importantes, se utilizan métodos que se basan en el ML. Por medio de estas técnicas, se realiza una clasificación entre los cuatro microsismos mencionados. Al final de este proceso, se obtienen varios parámetros que ayudan a identificar un conjunto principal de características. Estas características fundamentales son utilizadas por el modelo de aprendizaje automático para llevar a cabo la clasificación de los microsismos. Para el desarrollo de este trabajo se consideran dos algoritmos de ML para realizar la clasificación: DT y SVM.

Clasificación DT

En términos de selección de características, los DT son un método de selección de características que pertenece al grupo de métodos embedded. Por lo tanto, para encontrar las características principales se emplea clasificación DT y se procede a identificar los nodos

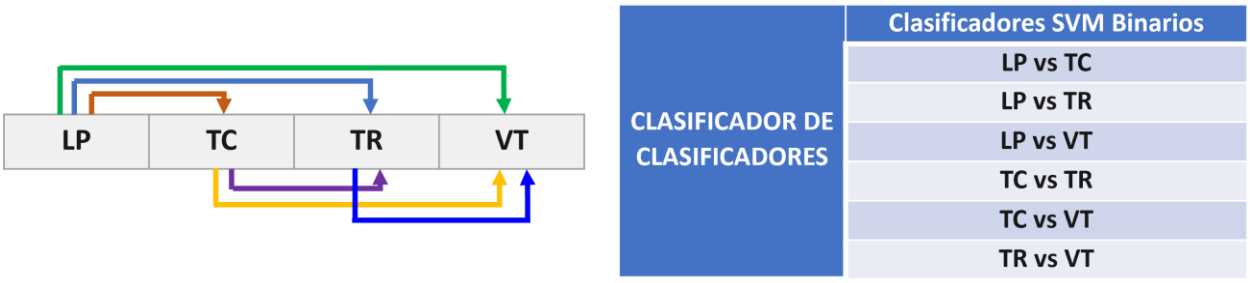
raíces del árbol que permiten la clasificación multi clase. Con esta identificación, se concluye la selección de características para este método.

Clasificación SVM

Un modelo de selección de características basado en SVM pertenece al grupo de métodos wrapper. En el caso de SVM, para seleccionar las características más importantes se procede a analizar el rendimiento del clasificador SVM como criterio para determinar qué características son las más influyentes en la clasificación. Para esto, se debe entender que si se considera que los métodos wrapper se ejecutan en bucles iterativos, donde se genera una estructura de clasificación denominada clasificador de clasificadores en donde se combinan sub clasificadores binarios bajo un enfoque de uno contra uno. Para ilustrar este concepto, se puede apreciar en la Figura 23 que se crea un clasificador binario para cada par único de clases.

Figura 23

Estructura del clasificador SVM basado en el enfoque uno contra uno.



Nota. Cada clasificador binario tiene un vector de pesos para las características.

Como se aprecia en la Figura 23, se tienen 4 clases de salida, por lo que se generan todos los posibles pares: LP vs. TC, LP vs. TR, LP vs. VT, TC vs. TR, TC vs. VT y TR vs. VT. Esto da como resultado un total de 6 clasificadores binarios, cada uno con su propio vector de pesos para las características.

Dentro de este marco, para encontrar las características principales se procede a determinar la relevancia de cada característica en cada clasificador binario y determinar cuál es el más influyente al momento de realizar una clasificación con el modelo SVM completo. Bajo esta metodología, se propone encontrar las características más importantes con SVM.

Selección de características

Una vez identificada la forma en que se van a extraer las características más importantes de los modelos DT y SVM, se procede a realizar la selección de características como tal. Para ello, se establece seleccionar únicamente las tres características que más influyen en la decisión de clasificación de cada modelo. Es importante recordar que, cada característica se traduce en bandas de frecuencia debido a que cada característica representa un coeficiente de descomposición wavelet producto de la WT, y a la vez, cada coeficiente tiene ligado un rango de frecuencias dentro de su PSD. Estos rangos de frecuencia se proceden a representar de manera gráfica con la finalidad de observar su relación con la PSD media de cada evento y con los intervalos de confianza.

Densidad espectral de potencia

Paralelamente al desarrollo del resto del algoritmo establecido en la Figura 15, se procede a obtener la PSD para cada evento a partir de las energías calculadas anteriormente. Para ello se emplea el método de Welch con un número de puntos NFFT igual a 512, un valor de solapamiento de 0.5 y con una ventana de tipo Hamming. Finalmente, cada PSD correspondiente a un evento en específico se normaliza con el fin de comparar apropiadamente los resultados obtenidos.

Es así como el operador $\mathrm{p_M\{e_M\}}$ expresa de manera matemática la aplicación de la PSD Welch a las energías de la WT. El resultado se guarda en la matriz P:

$$P = \{p_1^T, p_2^T, p_3^T, \dots, p_M^T\}^T \quad (16)$$

donde p_M representa la PSD Welch de las energías de cada coeficiente cD y cA, que a su vez fueron extraídas de cada señal normalizada y filtrada h_M.

Intervalos de confianza

Para cada evento y a partir de las PSD, se procede a calcular los intervalos de confianza para la media. El intervalo de confianza se lo calculo para un rango de valores aceptables que abarcan la media de un conjunto de los datos obtenidos de los microsismos. Este intervalo suministra un límite superior y un límite inferior que delimitan la posible ubicación de la media de las energías wavelet, con un cierto grado de margen de error. Para determinar los límites se emplea el método de la distribución t-student. Los límites inferior y superior del intervalo se calculan en función de la media, la desviación estándar y el nivel de confianza especificado. El grado de confianza se establece al 99% en este caso.

Evaluación de características seleccionadas

En etapas anteriores, se reduce el conjunto inicial de características a tan solo 3. En esta etapa de evaluación, se lleva a cabo una nueva clasificación al emplear los mismos algoritmos de ML. Sin embargo, la diferencia radica en que solo emplean las 3 características seleccionadas por el modelo DT y las 3 características seleccionadas por el modelo SVM.

Estas características, que conforman un grupo más reducido, se utilizan como entrada en los modelos DT y SVM, respectivamente. Luego, se documentan las métricas de rendimiento obtenidas en este proceso de evaluación. El objetivo fundamental de esta sección es evaluar el rendimiento de clasificación cuando se usan únicamente estas 3 características.

Capítulo IV

Pruebas y resultados

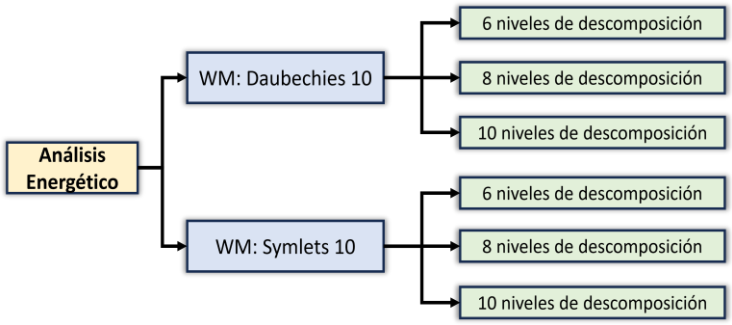
Análisis energético

Para alcanzar los resultados energéticos, es importante recordar que las señales atraviesan una serie de etapas previas. Estas etapas incluyen la normalización, el filtrado y la obtención de coeficientes de detalle y aproximación mediante la WT. Estos procesos se omiten en la sección de Resultados debido a que no aportan información adicional más allá de lo detallado en el Capítulo III.

Después de aclarar este punto, el proceso que lleva a la obtención de datos para un análisis más profundo comienza con el análisis energético. En este proceso, se calcula la energía a partir de los coeficientes de detalle y aproximación obtenidos mediante la descomposición wavelet multinivel. En este caso, se consideran 6, 8 y 10 niveles.

Figura 24

Proceso de análisis energético.



Nota. Se establecen dos líneas de análisis energético con el fin de comparar y determinar la MW que genere mejores resultados.

Además, como se detalla en el Capítulo 3, este estudio incluye la aplicación de la WT con MW de Daubechies y Symlets. Se eligen estas wavelets basándose en el principio de similaridad, como se detalla previamente. El proceso del análisis energético se presenta de

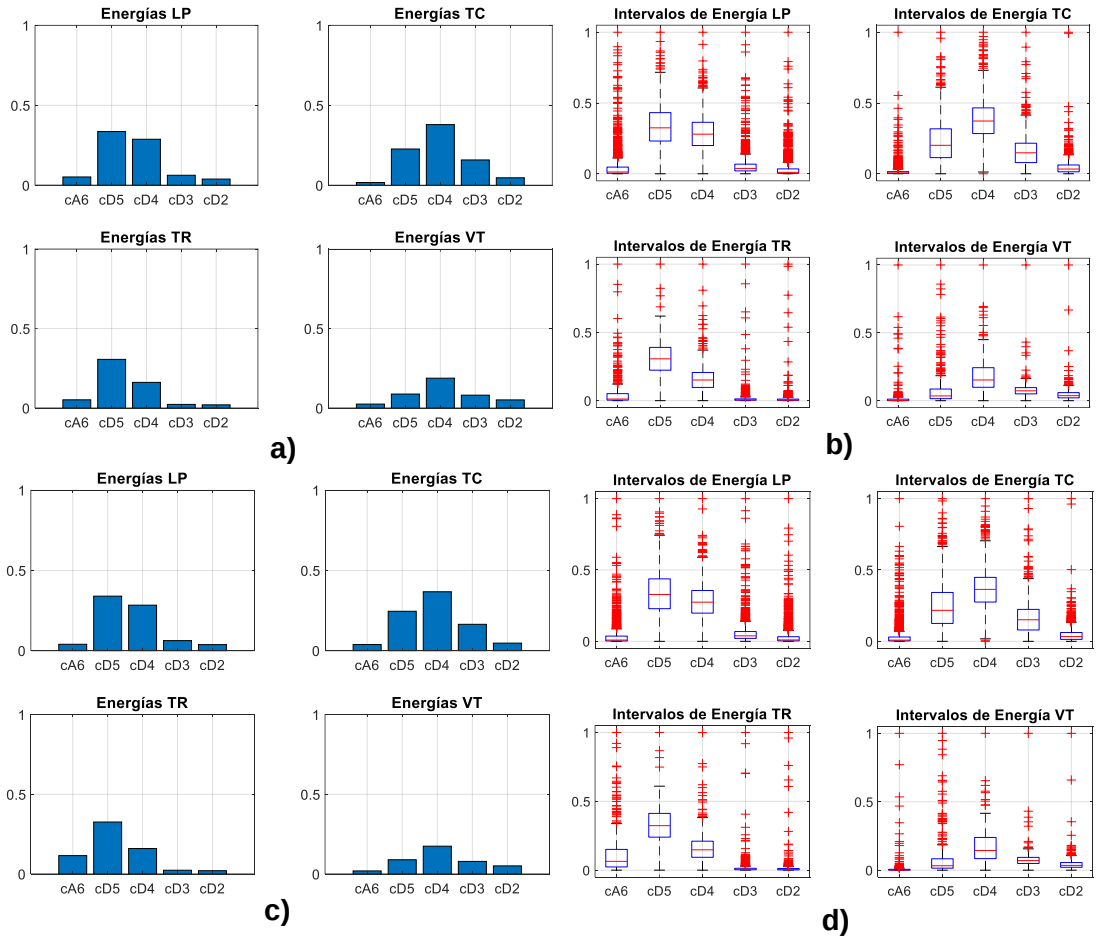
manera resumida en la Figura 24, en donde se observa que para cada tipo de MW se desarrolla la descomposición multinivel con WT con 6, 8 y 10 niveles.

Nivel de descomposición 6

Para un nivel de descomposición wavelet de seis niveles, se obtienen cinco bandas de frecuencia, cuyos valores medios normalizados pueden observarse en la Figura 25.

Figura 25

Análisis de energías de coeficientes obtenidos mediante WT de 6 niveles.



Nota. Fila superior: MW Daubechies y fila inferior: MW Symlets. Donde: a) y c) Valores medios de energía de coeficientes wavelet. Y b) y d) diagrama de caja y bigotes para distribución de datos. Para una correcta comparación, todos los valores se encuentran normalizados.

En la Figura 25, se puede observar que en a) y c) destacan las medias de cada nivel energético, con lo cual se establece una caracterización previa. Visualmente, por medio de la comparación de las bandas cD5 (1.56 a 3.12 Hz) y cD4 (3.12 a 6.25 Hz) se puede distinguir un evento de otro. Sin embargo, existe una alta posibilidad de confundir un LP con un TR debido a que la similitud de las amplitudes medias de energías. Además, como se puede observar en la Figura 25, en b) y d) específicamente, se tienen representados los diagramas de cajas, de donde se extrae los datos estadísticos representados en la Tabla 7 y se realiza una comparación entre los análisis con MW Daubechies y Symlets para un LP.

Tabla 7

Datos estadísticos extraídos del diagrama de caja para evento LP.

	Media		Q1		Q2		Q3		IQR		Outliers	
MW / Coef.	dB	Sym	dB	Sym	dB	Sym	dB	Sym	dB	Sym	dB	Sym
cA6	0,052	0,040	0,003	0,003	0,012	0,009	0,003	0,003	0,044	0,034	271	291
cD5	0,336	0,340	0,231	0,229	0,324	0,328	0,231	0,229	0,201	0,209	9	9
cD4	0,288	0,283	0,200	0,198	0,280	0,274	0,200	0,198	0,164	0,159	16	16
cD3	0,063	0,062	0,020	0,021	0,038	0,038	0,020	0,021	0,048	0,048	126	120
cD2	0,039	0,038	0,003	0,003	0,010	0,010	0,003	0,003	0,031	0,029	269	262

Nota. Datos extraídos para $\mathrm{LP}_{1310\times5}$. Q1, Q2, Q3: cuartiles. IQR: rango intercuartil. Outliers: Número de valores atípicos. MW dB: Daubechies, MW Sym: Symlets.

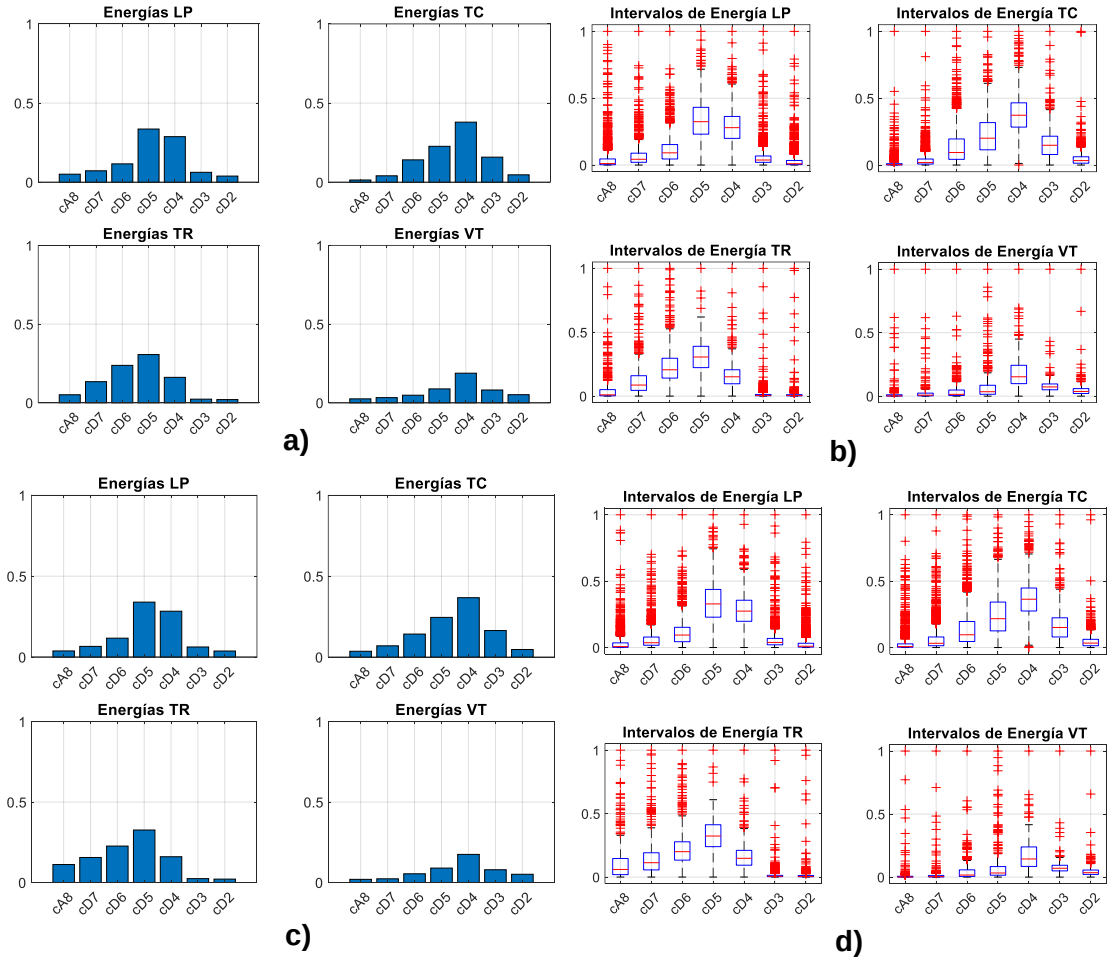
Al examinar la Tabla 7 se observa que entre el análisis con MW Daubechies y Symlets no existe una gran diferencia de valores; sin embargo, se debe destacar el número de valores atípicos existentes en las bandas cA6, cD3 y cD2 (0 a 0.78 Hz, 6.25 a 12.5 Hz, 12.5 a 25 Hz respectivamente) para ambos métodos de análisis. Dentro de estos rangos de frecuencias existen más valores atípicos ya que, por ejemplo, para el caso del análisis con MW Symlets en la banda cA6 existen 291 de 1310 datos que representan valores atípicos. Esto es importante ya que puede afectar las métricas de rendimiento en la clasificación con ML.

Nivel de descomposición 8

Para un nivel de descomposición wavelet de ocho niveles, se obtienen 7 bandas de frecuencia, cuyos valores medios normalizados pueden observarse en la Figura 26.

Figura 26

Análisis de energías de coeficientes obtenidos mediante WT de 8 niveles.



Nota. Fila superior: MW Daubechies y fila inferior: MW Symlets. Donde: a) y c) Valores medios de energía de coeficientes wavelet. Y b) y d) diagrama de caja y bigotes para distribución de datos. Para una correcta comparación, todos los valores se encuentran normalizados.

En la Figura 26, se puede observar que en las representaciones a) y c), resaltan las energías promedio de cada coeficiente wavelet, lo que establece una caracterización

preliminar. Inicialmente, al comparar las bandas cD6 (0.78 a 1.56 Hz), cD5 (1.56 a 3.12 Hz) y cD4 (3.12 – 6.25 Hz), es posible distinguir entre diferentes eventos. Además, tanto en el análisis con Daubechies como en Symlets, se obtienen valores medios muy similares, hasta el punto en que los gráficos en a) y c) son casi idénticos. Además, como se puede observar en la Figura 26, en c) y d) específicamente, se tienen representados los diagramas de cajas, de donde se extrae los datos estadísticos representados en la Tabla 8 en donde se realiza una comparación entre los análisis con MW Daubechies y Symlets para un LP.

Tabla 8

Datos estadísticos extraídos del diagrama de caja para evento TC.

	Media		Q1		Q2		Q3		IQR		Outliers	
MW / Coef.	dB	Sym	dB	Sym	dB	Sym	dB	Sym	dB	Sym	dB	Sym
cA6	0,014	0,036	0,001	0,002	0,003	0,007	0,001	0,002	0,009	0,026	300	325
cD5	0,041	0,070	0,007	0,014	0,019	0,032	0,007	0,014	0,038	0,066	205	201
cD4	0,141	0,142	0,043	0,046	0,094	0,096	0,043	0,046	0,153	0,151	110	102
cD3	0,227	0,245	0,114	0,125	0,201	0,217	0,114	0,125	0,204	0,217	18	21
cD2	0,379	0,367	0,284	0,275	0,373	0,364	0,284	0,275	0,182	0,173	17	19

Nota. Datos extraídos para $\mathrm{TC}_{1488\times5}$. Q1, Q2, Q3: cuartiles. IQR: rango intercuartil. Outliers: Número de valores atípicos. MW dB: Daubechies, MW Sym: Symlets.

Al analizar la Tabla 8 se observa que entre el análisis con MW Daubechies y Symlets no existe una gran diferencia de valores en las bandas de alta frecuencia como son cD4, cD3 y cD2 (3.12 a 6.25 Hz, 6.25 a 12.5 Hz, 12.5 a 25 Hz respectivamente), pero no sucede lo mismo en los coeficientes de baja frecuencia como en cA8 y cD7 (0 a 0.195 Hz, 0.39 a 0.79 Hz respectivamente). Esto se comprueba previamente al observar la Figura 26 y comparar a) con c). Además, se observa que existen un gran número de valores atípicos existentes en las bandas de frecuencia baja cA8, cD7 y cD6 para ambos métodos de análisis. Dentro de estos rangos de frecuencias existen más valores atípicos ya que, por ejemplo, para el caso del análisis con MW Symlets en la banda cA10 existen 201 de 1488 datos que son valores atípicos.

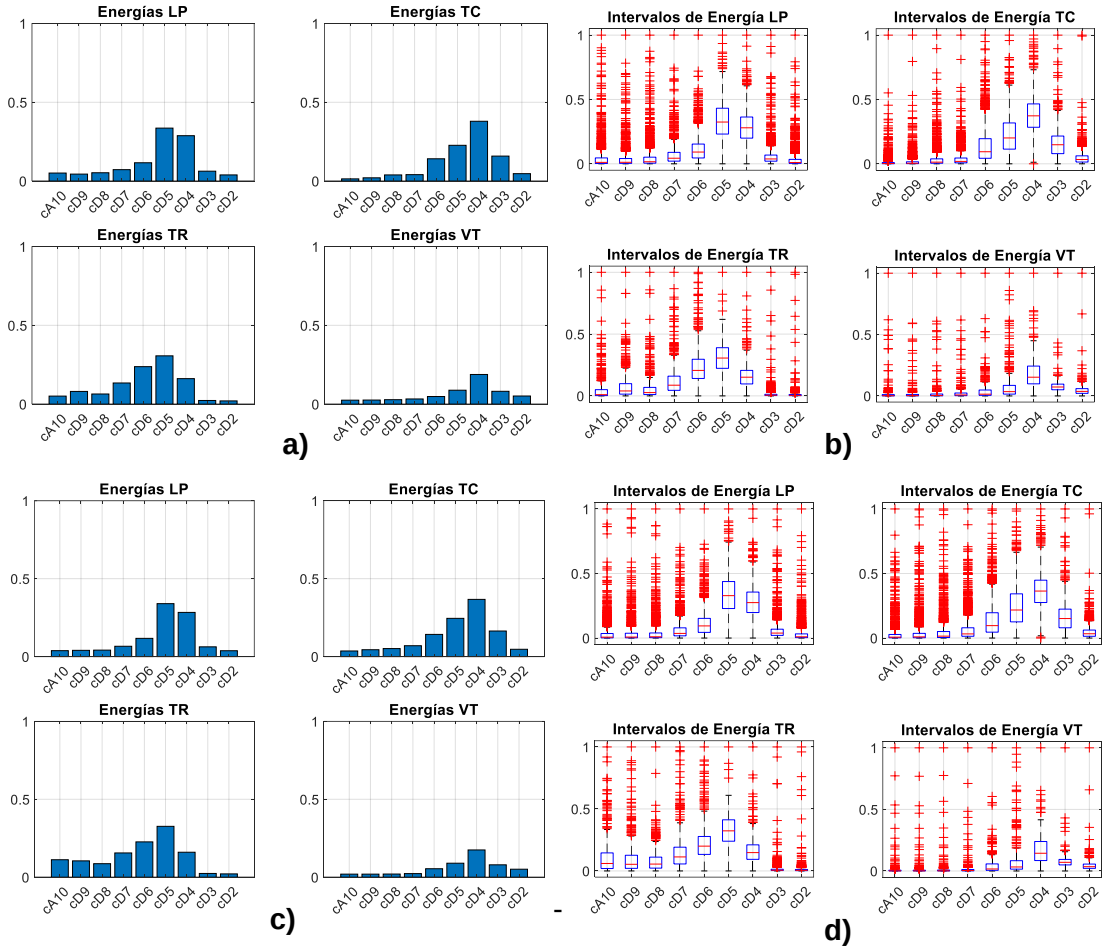
Para este caso no es muy relevante este hecho, ya que los microsismos tienen actividad espectral en frecuencias mucho mayores. Por último, se observa en la Tabla 8 que el número de valores atípicos se reduce en las bandas de interés como lo son cD5, cD4 y cD3.

Nivel de descomposición 10

Para un nivel de descomposición wavelet de 10 niveles, se obtienen 9 bandas de frecuencia, cuyos valores medios normalizados pueden observarse en la Figura 27.

Figura 27

Análisis de energías de coeficientes obtenidos mediante WT de 10 niveles



Nota. Fila superior: MW Daubechies y fila inferior: MW Symlets. Donde: a) y c) Valores medios de energía de coeficientes wavelet. Y b) y d) diagrama de caja y bigotes para distribución de datos. Para una correcta comparación, todos los valores se encuentran normalizados.

En la Figura 27, se puede observar que en las representaciones a) y c), resaltan las energías promedio de cada coeficiente wavelet, lo que establece una caracterización preliminar. De igual manera que en el caso anterior, al comparar las bandas cD6 (0.78 a 1.56 Hz), cD5 (1.56 a 3.12 Hz) y cD4 (3.12 a 6.25 Hz), es posible distinguir entre diferentes eventos. Además, tanto en el análisis con Daubechies como en Symlets, se obtienen valores medios muy similares, lo que hace que los gráficos en a) y c) sean muy similares.

Tabla 9

Datos estadísticos extraídos del diagrama de caja para evento TR.

	Media		**Q1**		**Q2**		**Q3**		**IQR**		**Outliers**	
MW / Coef.	**dB**	**Sym**	**dB**	**Sym**	**dB**	**Sym**	**dB**	**Sym**	**dB**	**Sym**	**dB**	**Sym**
cA10	0,051	0,112	0,002	0,019	0,010	0,112	0,002	0,019	0,048	0,126	122	43
cD9	0,081	0,105	0,014	0,021	0,040	0,105	0,014	0,021	0,085	0,106	60	59
cD8	0,064	0,087	0,011	0,023	0,027	0,087	0,011	0,023	0,058	0,089	70	41
cD7	0,134	0,156	0,045	0,056	0,086	0,156	0,045	0,056	0,115	0,135	54	37
cD6	0,238	0,227	0,141	0,133	0,206	0,227	0,141	0,133	0,155	0,145	26	26
cD5	0,306	0,326	0,223	0,240	0,306	0,326	0,223	0,240	0,167	0,173	4	4
cD4	0,162	0,160	0,097	0,093	0,151	0,160	0,097	0,093	0,110	0,117	16	13
cD3	0,023	0,024	0,003	0,003	0,006	0,024	0,003	0,003	0,010	0,011	86	92
cD2	0,020	0,021	0,002	0,002	0,004	0,021	0,002	0,002	0,009	0,010	83	80

Nota. Datos extraídos para $\mathrm{TR}_{490\times 9}$. Q1, Q2, Q3: cuartiles. IQR: rango intercuartil. Outliers: Número de valores atípicos. MW dB: Daubechies, MW Sym: Symlets.

Al examinar la Tabla 9, se puede notar que no existe una notable diferencia en los valores entre el análisis realizado con las wavelets Daubechies y Symlets en las bandas de alta frecuencia, tales como cD4, cD3 y cD2 (3.12 a 6.25 Hz, 6.25 a 12.5 Hz, 12.5 a 25 Hz respectivamente). Sin embargo, esto no ocurre en los coeficientes de baja frecuencia como cA10 y cD9 (0 - 0.0485 Hz, 0.097 a 0.195 Hz respectivamente). Esta observación se confirma previamente al comparar los TR en las representaciones a) y c) de la Figura 27.

Por otro lado, en la Tabla 9 se evidencia que el número de valores atípicos es mayor en las bandas de frecuencia mínima y máxima. Por ejemplo, en el caso del método de análisis con Daubechies, la banda de frecuencia más baja (cA10) cuenta con 122 valores atípicos de un total de 490 datos, mientras que la banda de frecuencia más alta (cD2) tiene 83 valores atípicos de 490 datos. En las bandas de interés, como cD5, se encuentran apenas 4 valores atípicos en ambos casos de análisis.

Clasificación

Previo al análisis, para determinar la proporción de datos a usar en los clasificadores, se procede a realizar un barrido de entrenamientos con distintas divisiones de datos para entrenamiento y prueba. Un primer entrenamiento, se establece en la Tabla 10, en donde se puede destacar que la mejor relación de datos es de 90/10 %; sin embargo, para evitar problemas de sobre entrenamiento se ignora esta división de datos.

Tabla 10

Métricas con variación de partición de datos y algoritmo DT ajuste automático.

Partición de conjunto de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	78,5	77,9	91,9	73,8	0,2
Train: 60% - Test: 40%	80,6	81,2	92,6	76,1	0,2
Train: 70% - Test: 30%	79,5	79,3	92,3	76,3	0,2
Train: 80% - Test: 20%	83,3	80,1	94.1	80,8	0.13
Train: 90% - Test: 10%	83,3	82,5	93,7	79,8	0,15

Al analizar la Tabla 10 y la Tabla 11, se aprecia que la relación de datos que ofrece las mejores métricas de evaluación es de 80% para entrenamiento y 20% para prueba. Es por este motivo que de uso de la proporción de datos 80/20 % se sustenta en el hecho de que se obtienen los mejores resultados.

Tabla 11

Métricas con variación de partición de datos y algoritmo DT con podamiento.

División de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	80,5	82,2	92,4	74,6	0,2
Train: 60% - Test: 40%	82,7	83,7	93,4	78,1	0,1
Train: 70% - Test: 30%	84,2	83,1	94,2	80	0,1
Train: 80% - Test: 20%	84,5	83,9	94,3	80,9	0.12
Train: 90% - Test: 10%	84,1	82,4	94,2	80,2	0,1

Además, al evaluar las métricas que se obtienen del modelo SVM en distintos escenarios de división de datos, en la Tabla 12, se observa que la proporción con la que se obtiene la mayor precisión, y en general las mejores métricas, es el modelo que propone dividir los datos en 80/20.

Tabla 12

Métricas con variación de partición de datos y algoritmo SVM.

Partición de conjunto de datos	A (%)	P (%)	S (%)	R (%)	BER
Train: 50% - Test: 50%	85,2	84,8	94,4	80,4	0,1
Train: 60% - Test: 40%	86,1	86,4	94,7	80,7	0,1
Train: 70% - Test: 30%	86,3	87,3	94,7	80,9	0,1
Train: 80% - Test: 20%	87,7	88,0	95,4	82,8	0,11
Train: 90% - Test: 10%	84,7	81,9	94,4	81,2	0,1

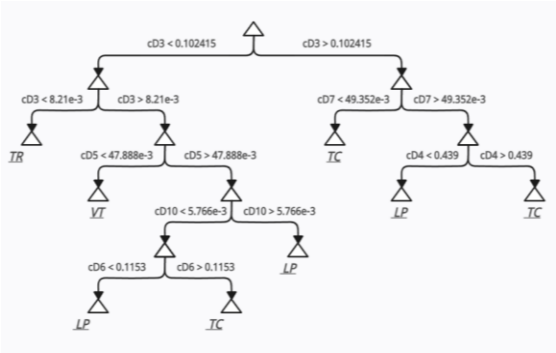
Para el proceso de extracción de características se emplean dos enfoques: basado en Wrapper y otro denominado embedded. Para aplicar el primero se utiliza un clasificador multiclase basado en DT, mientras que, para aplicar el método embedded se emplea un clasificador basado en algoritmos SVM.

Árbol de desición (DT)

El primer clasificador es el modelo DT, cuyas métricas se analizan con el fin de determinar con qué MW se obtiene el mejor resultado.

Figura 28

Nodos raíz en modelo DT con validación cruzada.



Nota. Las tres características principales identificadas son cD3, cD5 y cD7.

En la Figura 28, se puede observar un ejemplo con los nodos principales del árbol que se obtiene al emplear el modelo DT. A continuación, se exponen las métricas de desempeño que se obtienen del entrenamiento de un modelo de clasificación multiclase en tres modalidades de configuración distintas: ajuste automático, validación cruzada y podamiento.

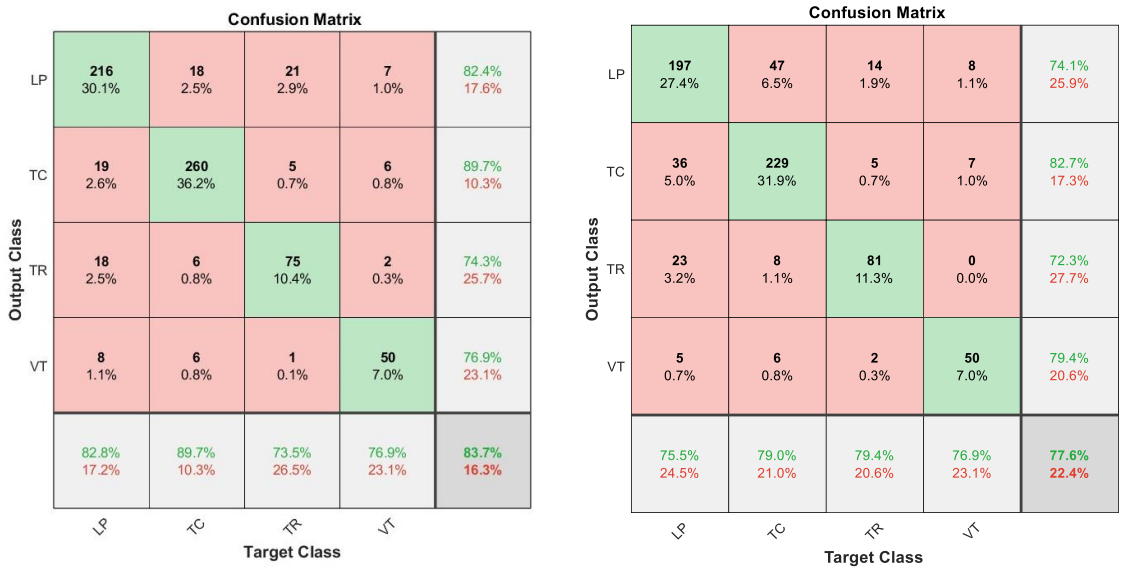
A partir de este punto, el análisis se centra en los resultados obtenidos mediante las nueve características generadas por la Transformada WT de diez niveles con las funciones MW Daubechies y Symlets, dado que con estos parámetros se obtienen los mejores resultados.

Configuración con ajuste automático. Para este modelo de DT se usa una configuración predeterminada en el cual únicamente se colocan como entrada las características y las etiquetas de los datos de entrenamiento. Adicionalmente, se le indica al modelo que use el comando ‘OptimizeHyperparameters','auto' para que el algoritmo ajuste automáticamente los mejores valores de los hiperparámetros, como, por ejemplo: profundidad máxima del árbol, mínimo de ejemplos por hoja, mínimo de ejemplos por división, criterios de

división, número máximo de hojas, etc. Posterior a crear el modelo, se calcula y se crea un gráfico de matriz de confusión que permite comparar las predicciones que realizó el modelo con las etiquetas reales de los datos de prueba (test). En la Figura 29, se observan las matrices de confusión obtenidas para los dos métodos de análisis: Daubechies 10 y Symlets 10.

Figura 29

Matrices de Confusión con características de entrada



Nota. Resultados generados con: a) MW Daubechies10 y b) MW Symlets10. El modelo DT que genera estos resultados tiene una configuración por defecto que optimiza y mejora las métricas automáticamente.

En la Figura 29 se presenta los dos resultados diferentes en cuanto a la exactitud de clasificación. Se obtiene un 83,7% si los coeficientes de la WT con Daubechies 10 son las características. Para el modelo generado con Symlets 10 se obtiene un 77.6% de exactitud. Además, se observa que ambos modelos tienen la mayor precisión de clasificación para el TC, ya que se obtiene un porcentaje de precisión del 89.7% con Daubechies 10 y el 82.7% con Symlets 10. Adicionalmente, en la Tabla 13 se puede observar que los eventos que presentan

el menor valor de BER son los TC bajo el método de análisis de Daubechies 10 con un valor de 0.09.

Tabla 13

Métricas obtenidas para el modelo DT con configuración de ajuste automático.

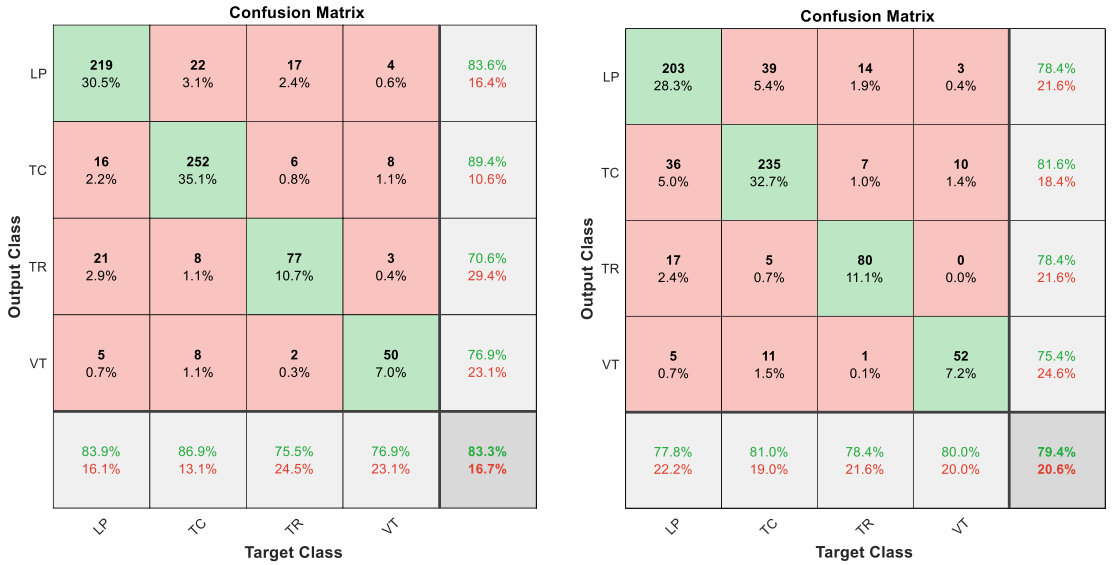
Métrica	A (%)		P (%)		S (%)		R (%)		BER	
MW / Evento	Db	Sym	Db	Sym	Db	Sym	Db	Sym	Db	Sym
LP	87	82	82	74	90	85	83	76	0,14	0,2
TC	92	85	90	83	93	89	90	79	0,09	0,16
TR	93	93	74	72	96	95	74	79	0,15	0,13
VT	96	96	77	79	98	98	77	77	0,13	0,13
General	84	78	81	77	94	92	81	78	0,13	0,15

Nota. 'Db' hace referencia a modelo con MW Daubechies y 'Sym' al modelo con MW Symlets.

Configuración con validación cruzada. En la Figura 30, se observan las matrices de confusión obtenidas para los dos métodos de análisis: Daubechies 10 y Symlets 10.

Figura 30

Matrices de Confusión con DT Validación cruzada.



a) Daubechies **b) Symlets**

En este modelo de DT se crea un clasificador multiclase que utiliza un enfoque de validación cruzada con 20 particiones (kFold). Además, se configura al modelo mediante el comando 'AlgorithmForCategorical','Exact' para que el modelo use el algoritmo exacto para manejar características categóricas. Esto ayuda a encontrar las divisiones óptimas en cada nodo y mejorar las decisiones tomadas en el árbol. Posterior a crear el DT, se calcula y se crea un gráfico de matriz de confusión.

Tabla 14

Métricas obtenidas para el modelo DT con validación cruzada.

Métrica	**A (%)**		**P (%)**		**S (%)**		**R (%)**		**BER**	
MW / Evento	**Db**	**Sym**	**Db**	**Sym**	**Db**	**Sym**	**Db**	**Sym**	**Db**	**Sym**
LP	88	84	84	78	91	88	84	78	0,13	0,17
TC	91	85	89	82	93	88	87	81	0,1	0,16
TR	92	94	71	78	95	96	76	78	0,15	0,13
VT	96	96	77	75	98	97	77	80	0,13	0,11
General	83	79	80	78	94	92	81	79	0,13	0,14

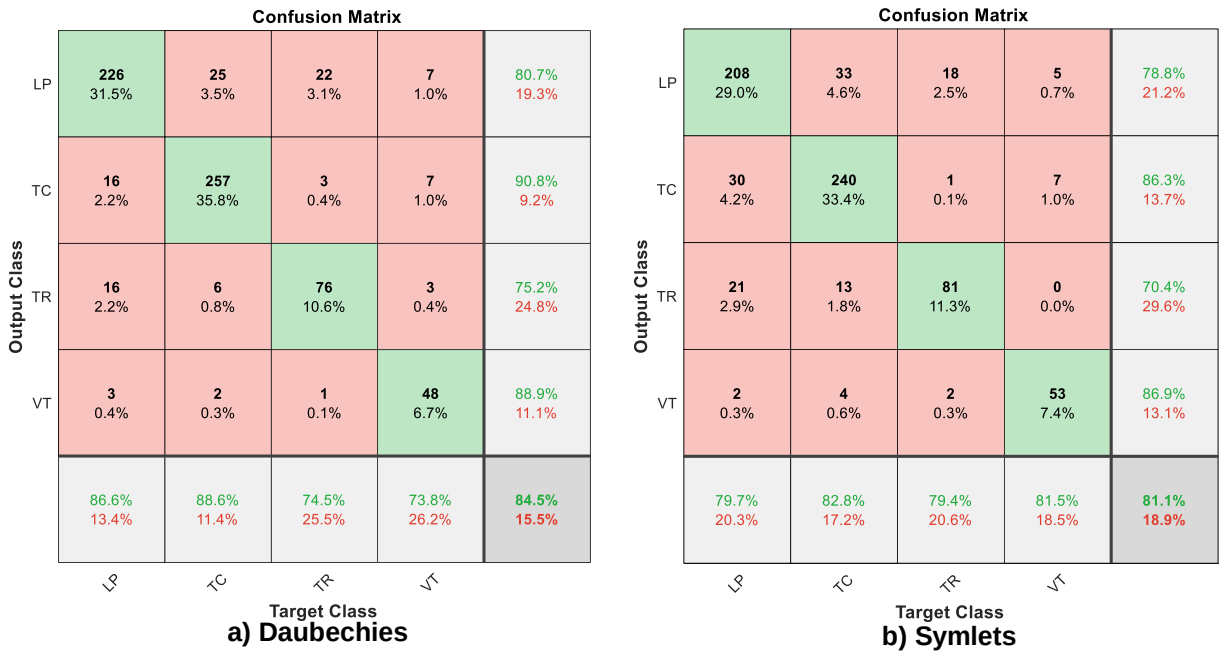
Nota. 'Db' hace referencia a modelo con MW Daubechies y 'Sym' al modelo con MW Symlets.

Como se puede apreciar en las matrices de confusión de la Figura 30, se obtienen resultados distintos en lo que respecta a la exactitud de clasificación puesto que con Daubechies 10, se logra un índice de clasificación del 83,3%. Por otro lado, para el modelo basado en Symlets 10, se alcanza una exactitud del 79,4%. Además, se debe destacar que ambos modelos presentan su mayor nivel de precisión en la clasificación del TC, en donde se obtiene un porcentaje de precisión del 89,4% por medio de una MW Daubechies 10 y un 81,6% mediante Symlets 10. También es importante mencionar que en la Tabla 14 se visualiza cómo los eventos con el menor valor de BER corresponden a los TC bajo el enfoque de análisis con Daubechies 10, ya que se tiene un valor de BER igual a 0,1.

Configuración con podamiento. Para este modelo de DT se emplea una configuración con podamiento y se modifican varios parámetros con el fin de encontrar los mejores resultados. Para ello se establece un proceso de poda que emplea iteraciones de validación cruzada con el fin de probar varias posibles combinaciones de subárboles mediante el comando 'SubTrees','All', con el comando 'TreeSize','min' se especifica que se debe seleccionar el nivel de poda que brinde el árbol más pequeño (menos complejo) mientras mantiene un rendimiento adecuado y de entre todas estas validaciones se selecciona el que tiene mejores métricas de rendimiento mediante el comando ‘bestlevel’.

Figura 31

Matrices de Confusión con DT Podamiento.



En la Figura 31 se presentan las matrices de confusión resultantes de los dos enfoques de análisis: por medio de Daubechies 10 y Symlets 10, con lo cual, se facilita la comparación entre las predicciones efectuadas por cada modelo.

Tabla 15

Métricas obtenidas para el modelo DT con podamiento.

Métrica	A (%)		P (%)		S (%)		R (%)		BER	
MW / Evento	Db	Sym	Db	Sym	Db	Sym	Db	Sym	Db	Sym
LP	88	85	81	79	88	88	87	80	0,13	0,16
TC	92	88	91	86	94	91	89	83	0,09	0,13
TR	93	92	75	70	96	95	75	79	0,15	0,13
VT	97	97	89	87	99	99	74	82	0,14	0,1
General	85	81	84	81	94	93	81	81	0,12	0,13

Nota. 'Db' hace referencia a modelo con WM Daubechies y 'Sym' al modelo con WM Symlets.

Tal como se puede observar en las matrices de confusión presentadas en la Figura 31, se obtienen resultados diversos en cuanto a la precisión de la clasificación. Por un lado, al utilizar el método de análisis con Daubechies 10, se logra un porcentaje de clasificación del 84,5%. Por otro lado, al considerar el enfoque de análisis con Symlets 10, se alcanza una precisión del 81.1%.

Al observar la Tabla 15, se debe destacar los eventos que tienen mayor precisión. Mediante Daubechies 10 se logra un porcentaje de precisión del 90,8% en el TC, mientras que con Symlets 10 se obtiene un 86,6% en el VT.

Además, es importante notar que en la Tabla 15 se refleja cómo los eventos con el valor de BER más bajo se encuentra en los TC bajo el análisis con Daubechies 10, ya que se obtiene un valor de BER igual a 0.1.

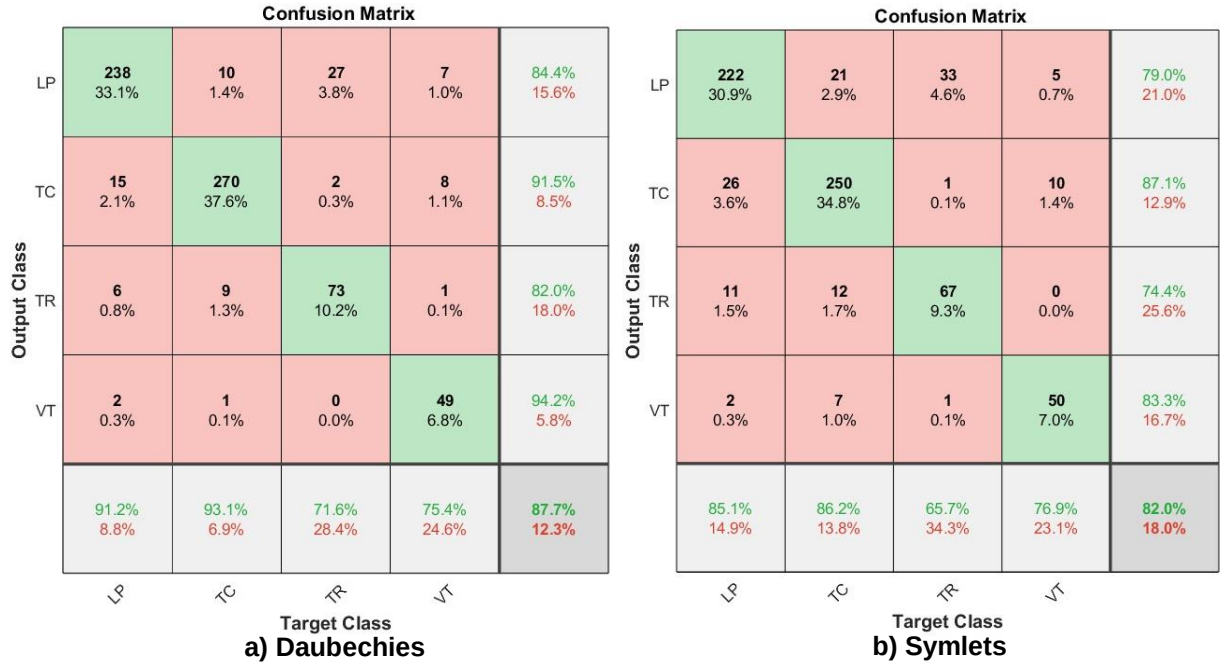
Máquinas de vector soporte (SVM)

El último método de extracción de características que se emplea en el presente trabajo de investigación se basa en entrenar un modelo de clasificación multiclase mediante máquinas de soporte vectorial basado en una estructura de clasificación denominada clasificador de clasificadores en donde se combinan sub clasificadores binarios bajo un enfoque de uno contra

uno. La configuración de entrenamiento del modelo que se usa es la misma que se encuentra por defecto; pero añadido un parámetro de optimización. Se usa el comando 'OptimizeHyperparameters','auto' para que el algoritmo ajuste automáticamente los mejores valores de los hiperparámetros, como, por ejemplo: parámetros del Kernel, parámetros de regularización, parámetros de optimización del hiperplano, etc.

Figura 32

Matrices de Confusión con SVM.



a) Daubechies **b) Symlets**

En la Figura 32, se puede observar una representación gráfica de las métricas de interés. Además, se puede extraer y comparar los valores de exactitud bajo los dos métodos de análisis. Al utilizar el método de análisis con Daubechies 10, se alcanza un porcentaje de clasificación del 88 %, el más alto obtenido durante el desarrollo del trabajo. Por otro lado, al considerar el enfoque de análisis con Symlets 10, se alcanza una precisión del 82 %.

Tabla 16

Métricas obtenidas para el modelo SVM.

Métrica	A (%)		P (%)		S (%)		R (%)		BER	
MW / Evento	Db	Sym	Db	Sym	Db	Sym	Db	Sym	Db	Sym
LP	91	86	84	79	90	87	91	85	0,09	0,14
TC	94	89	92	87	94	91	93	86	0,06	0,11
TR	94	92	82	74	97	96	72	66	0,16	0,20
VT	97	97	94	83	100	99	75	77	0,13	0,12
General	88	82	88	81	95	93	83	79	0.11	0,14

Nota. ‘Db’ hace referencia a modelo con MW Daubechies y ‘Sym’ al modelo con MW Symlets.

Se debe añadir que, al analizar la Tabla 16 se observa que ambos modelos tienen una mayor métrica de precisión de clasificación para el TC, ya que se obtiene un porcentaje de precisión del 91.5% con Daubechies 10 y el 87.2% con Symlets 10. Además, en la Tabla 16 se puede observar que los eventos que presentan el menor valor de BER son los TC bajo el método de análisis de Daubechies 10 con un valor de 0.06.

Selección de características

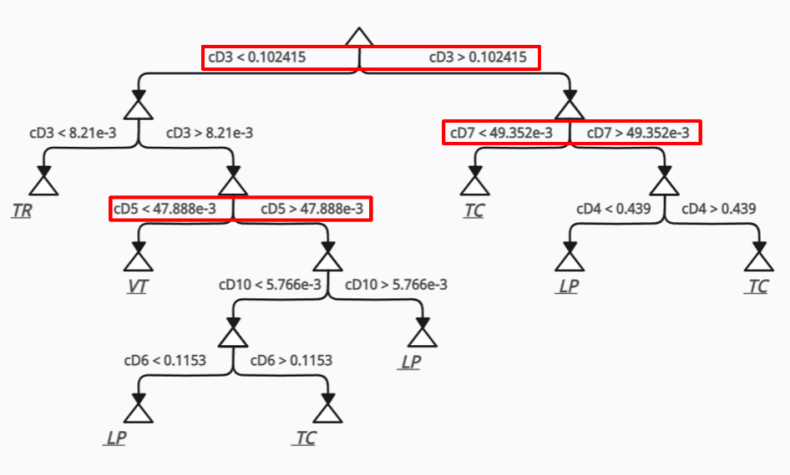
Una vez que los modelos de aprendizaje automático se han entrenado por medio de los nueve coeficientes wavelet como características de entrada, se procede a determinar cuáles de estos coeficientes son los más influyentes al momento de clasificar un evento. En el caso del DT, se seleccionan los tres coeficientes más significativos en los nodos raíz, los cuales permiten identificar un evento en cada una de las tres variantes de modelo DT: configuración con ajuste automático, validación cruzada y podamiento. En cuanto al modelo SVM, para identificar los principales coeficientes que influyen en la clasificación se implementa un enfoque de "clasificador de clasificadores" mediante la unión de clasificadores binarios uno contra uno, como se ilustra en la Figura 23.

A partir de este punto, se ignora cualquier análisis relacionado con los coeficientes wavelet generados mediante MW Symlets. En su lugar, se centra la atención en los resultados obtenidos por medio de MW Daubechies, ya que este método demostró ofrecer las mejores métricas de rendimiento al clasificar los microsismos.

DT con ajuste automático. Los resultados de los nodos raíz y sus pesos se pueden observar en la Figura 33. De esta ilustración se extraen los nodos principales que dan origen al árbol de decisiones, es así como se identifican a los coeficientes cD3, cD5 y cD7 como los coeficientes fundamentales.

Figura 33

Nodos raíz en modelo DT con análisis Daubechies.



Nota. Las tres características principales identificadas son cD3, cD5 y cD7.

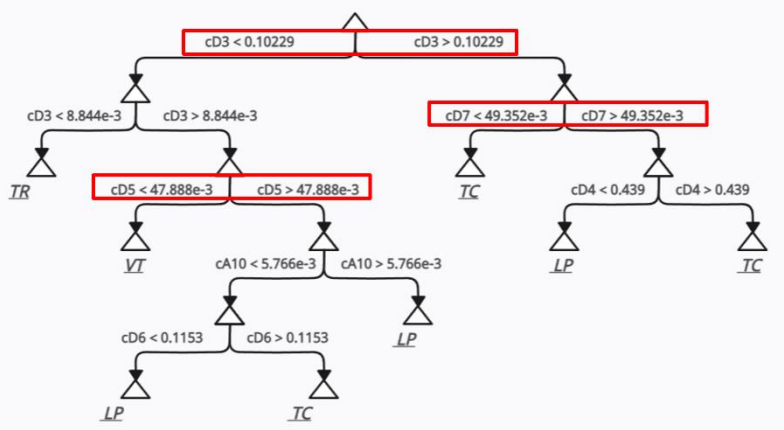
El nodo raíz más importante se compone del coeficiente cD3 que comprende el rango de frecuencias entre 6.25 a 12.5 Hz. Este nodo establece una regla fundamental de clasificación y es que si $\mathrm{cD3} < 0.102415$ y a su vez $\mathrm{cD} < 8.21 \cdot 10^{-3}$, entonces se llega a la conclusión de que el evento analizado es un TR.

Los siguientes coeficientes más importantes son cD3 y cD7 los cuales se pueden apreciar en el segundo nivel del árbol en la Figura 33. Sin embargo, debido a que cD3 se repite y el objetivo es encontrar tres coeficientes distintos, cD3 se descarta y se procede a tomar el coeficiente del siguiente nivel. Es así como, se establecen como siguientes características principales a los coeficientes cD5 y cD7, que comprenden los rangos de frecuencias de 1.56 a 3.12 Hz y 0.39 a 0.78 Hz respectivamente.

DT con validación cruzada. Los resultados de los nodos raíz junto con sus ponderaciones se pueden apreciar en la Figura 34. A partir de esta representación gráfica, se derivan los nodos fundamentales que forman la base del árbol de decisión. Es así como se realiza la identificación de los coeficientes cD3, cD5 y cD7 como los elementos clave.

Figura 34

Nodos raíz en modelo DT con validación cruzada.



Nota. Las tres características principales identificadas son cD3, cD5 y cD7.

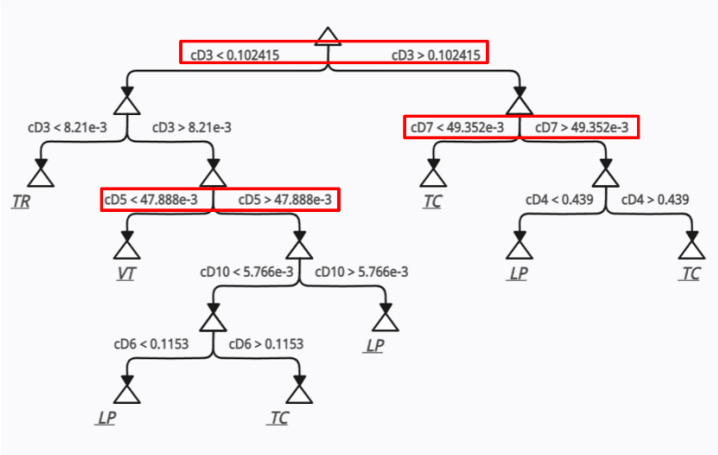
El nodo central de mayor relevancia está formado por el coeficiente cD3, que abarca el intervalo de frecuencias de 6.25 a 12.5 Hz. Este nodo establece una regla primordial de clasificación: si $\mathrm{cD3} < 0.10229$ y simultáneamente $\mathrm{cD3} < 8.844 \cdot 10^{-3}$, se concluye que el evento bajo análisis es un TR. Al igual que en el caso anterior, los siguientes coeficientes más

significativos son cD3 y cD7 los cuales se pueden apreciar en el segundo nivel del árbol en la Figura 34. Sin embargo, debido a que cD3 se repite y el objetivo es encontrar tres coeficientes distintos, cD3 se descarta y se procede a tomar el coeficiente del siguiente nivel. En consecuencia, los coeficientes cD5 y cD7, que abarcan los intervalos de frecuencias de 1.56 a 3.12 Hz y 0.39 a 0.78 Hz, respectivamente, son considerados como las características primordiales.

DT con podamiento. Los resultados de los nodos raíz y sus pesos se pueden observar en la Figura 35. En esta representación gráfica, se identifican los nodos fundamentales que forman la raíz del árbol de decisión. Es así como se realiza la identificación de los coeficientes cD3, cD5 y cD7 como las características principales para este caso.

Figura 35

Nodos raíz en modelo DT con podamiento.



Nota. Las tres características principales identificadas son cD3, cD5 y cD7.

El nodo central de mayor importancia se forma a partir del coeficiente cD3, que comprende el rango de frecuencias entre 6.25 y 12.5 Hz. Este nodo establece una regla fundamental para la clasificación: si $\mathrm{cD3} < 0.10229$ y al mismo tiempo $\mathrm{cD3} < 8.844 \cdot 10^{-3}$, se determina que el evento a evaluar es un TR. Similar al caso anterior, los siguientes coeficientes

más significativos son cD3 y cD7, los cuales se observan en el segundo nivel del árbol en la Figura 35. Sin embargo, dado que cD3 se repite y el objetivo consiste en encontrar tres coeficientes distintos, cD3 se excluye y se procede a seleccionar el coeficiente del nivel siguiente nivel. Como resultado, los coeficientes cD5 y cD7, que comprende los intervalos de frecuencias de 1.56 a 3.12 Hz y 0.39 a 0.78 Hz respectivamente, se consideran como las siguientes características principales.

SVM. En cuanto al modelo SVM, para identificar los principales coeficientes que influyen en la clasificación se implementa un enfoque de "clasificador de clasificadores" conformado por la unión de clasificadores binarios uno contra uno.

Tabla 17

Importancia de características del modelo SVM.

Clasificador de clasificadores	**Importancia de Característica**		
SVM Binarios	**Primero**	**Segundo**	**Tercero**
LP vs TC	cD4	cD2	cD5
LP vs TR	cD3	cD2	cD9
LP vs VT	cD2	cD7	cD3
TC vs TR	cD4	cD2	cD3
TC vs VT	cD4	cD2	cD3
TR vs VT	cD3	cD2	cD4
FRECUENCIA	**cD4 3.12 a 6.25 Hz**	**cD2 12.5 a 25 Hz**	**cD3 6.25 a 12.5 Hz**

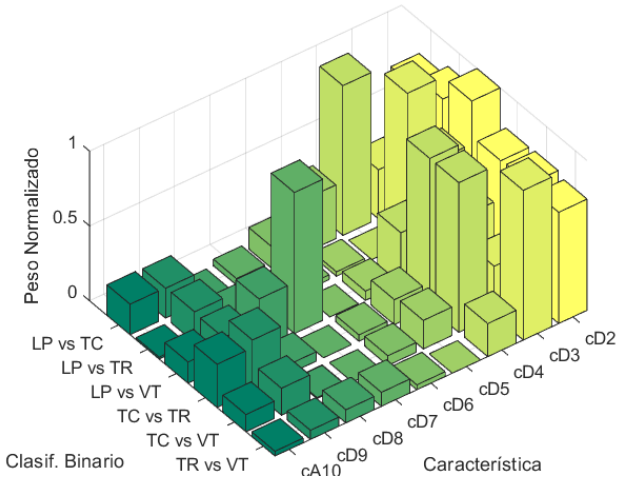
Nota. La jerarquía de importancia se basa en las características más influyentes.

Como se puede observar en la Tabla 17, en cada clasificador binario se analizan los pesos asociados a cada característica, se ordenan de mayor a menor y se extrae los tres resultados mayores de cada clasificador binario. En la Tabla 17 se expresan las tres características más importantes. Esto significa que, para clasificar un evento mediante un modelo SVM multiclase, la característica que más influye en la toma de decisiones es el

coeficiente cD4, el cual comprende el rango de frecuencias desde 3.12 a 6.25 Hz. Los coeficientes que le siguen en cuanto a importancia son las características cD2 y cD3, que corresponden a los intervalos de frecuencias de 12.5 a 25 Hz y 6.25 a 12.5 Hz respectivamente.

Figura 36

Pesos de características con respecto a clasificadores binarios SVM.

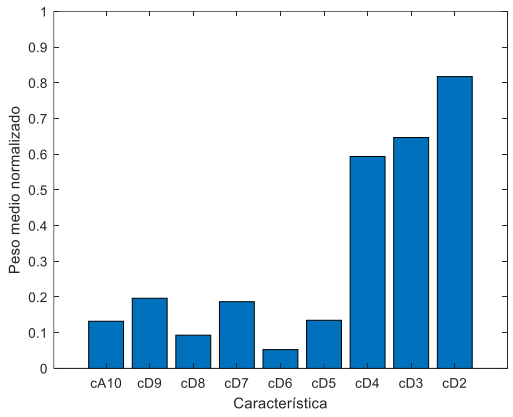


Nota. Los pesos se encuentran normalizados

Las afirmaciones que se plantean a partir de la Tabla 17, se sustentan en la Figura 36, en donde se observa que los pesos con mayor relevancia (mayor peso normalizado), se concentran en las bandas cD4, cD2 y cD3. De igual manera, en la Figura 37, se aprecian los pesos medios que tiene cada característica en el clasificador SVM.

Figura 37

Peso medio de las características que se obtienen con el clasificador SVM



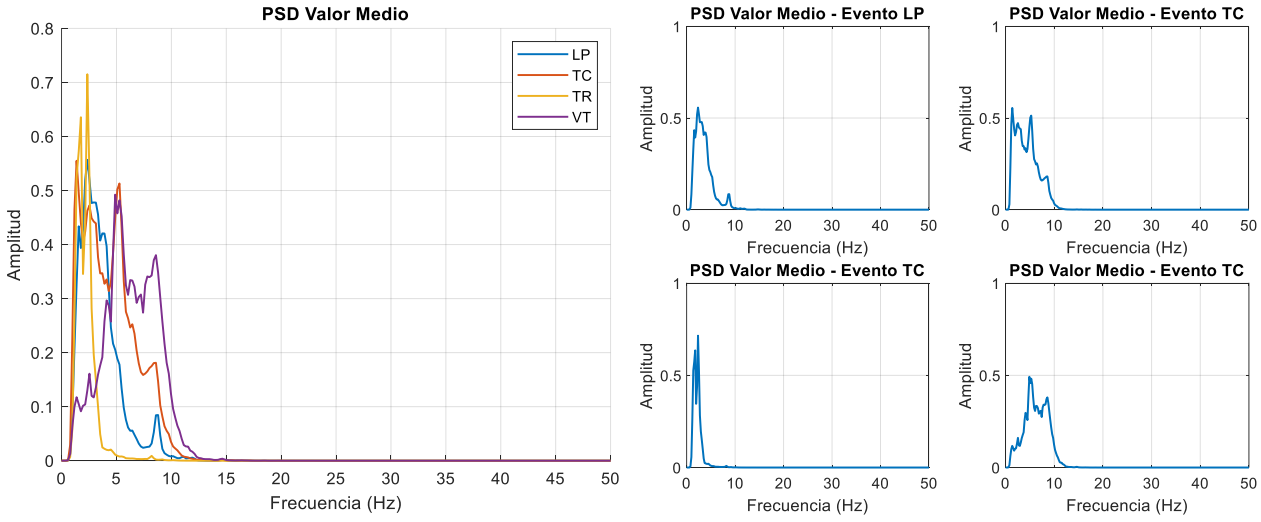
Nota. Este gráfico es una representación bidimensional media de la Figura 36.

PSD Welch promedio

El siguiente objetivo tras identificar las características principales que permiten identificar eventos con metodologías basadas en ML es expresar estas bandas de frecuencia en el espectro de potencias medio de cada evento. Para ello, es necesario obtener la PSD.

Figura 38

PSD media de cada evento.



Nota. Las PSD representativas de cada evento se encuentran normalizadas.

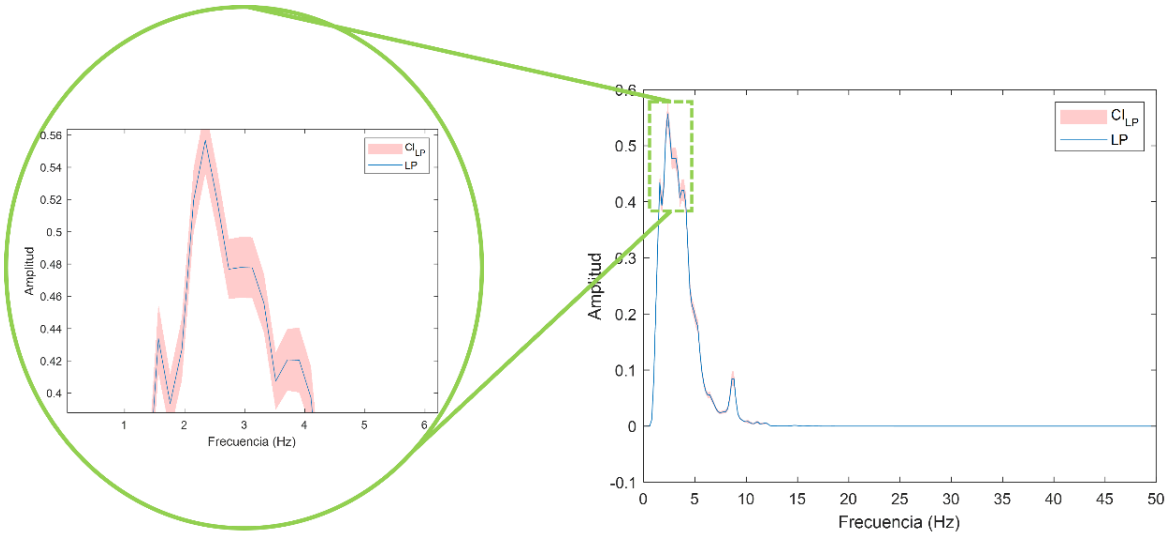
En la Figura 38, se observan las PSD que representan a cada evento, las mismas se encuentran normalizas y expresadas en el rango de frecuencias de 0 a 50 Hz. En este trabajo, se obtiene la PSD para cada evento a partir de las energías calculadas en función de los coeficientes wavelet con MW Daubechies 10. Para ello se emplea el método de Welch con un número de puntos NFFT igual a 512, un valor de solapamiento de 0.5 y con una ventana de tipo Hamming. Finalmente, cada PSD correspondiente a un evento en específico se normaliza con el fin de comparar apropiadamente los resultados obtenidos.

Intervalos de confianza

Para esta sección se emplea la fórmula para el cálculo de los intervalos de confianza para la media. Cada matriz que contiene a las PSD de cada evento tiene una dimensión de $\mathrm{PSD}_{\mathrm{m}\times 257}$, donde m representa el número de muestras por evento. Para cada columna m, se calcula la media y, en función de este valor y el grado de confianza, se obtienen los límites mínimos y máximos que representan la posible ubicación de la media de las energías wavelet.

Figura 39

Intervalos de confianza para los LP con 99% de grado de confianza.

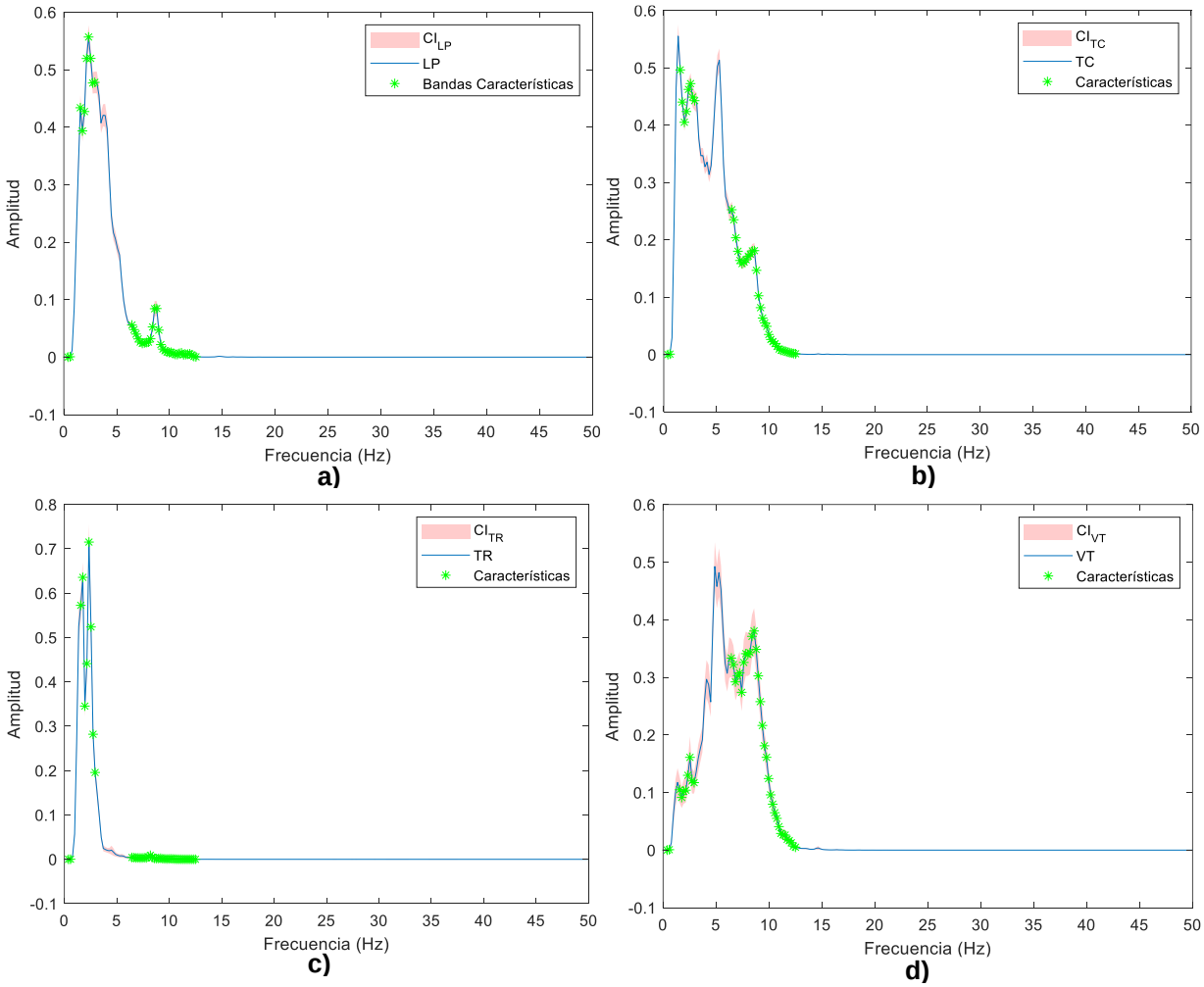


Nota. $\mathrm{CI}_{\mathrm{LP}}$ representa la zona dentro de los intervalos de confianza.

En la Figura 39, se puede observar la gráfica de la PSD en conjunto con los intervalos de confianza calculados para el LP. Se observa que con un valor del 99% como grado de confianza se obtienen límites inferiores y superiores muy estrechos a lo largo de la curva PSD. Esto se traduce a que, la media de la PSD que se obtiene a partir de los coeficientes wavelet de cada señal perteneciente al LP se van a encontrar dentro de la zona de confianza (en rojo).

Figura 40

PSD, intervalos de confianza y características identificadas con modelos DT.



Nota. Los intervalos de confianza y las características seleccionadas se representan mediante sus PSD. En la fila superior: a) LP y b) TC; fila inferior c) TR y d) VT con 99% de grado de confianza para todos los modelos DT. $\mathrm{CI}_{\mathrm{XX}}$ representa la zona de los intervalos de confianza.

En la Figura 40, se expresan de manera gráfica los intervalos de confianza para los cuatro tipos de microsismos con el 99% de grado de confianza, representado en conjunto con las bandas de frecuencia principales obtenidas mediante los modelos DT que se encuentran resumidos en la Tabla 18. Se crea una sola representación para los modelos DT, debido a que los coeficientes cD3, cD5 y cD7 coinciden como las características más importantes en todos los casos.

Tabla 18

Principales características identificadas con DT.

Orden	Característica	Frecuencia
1	cD3	6.25 – 12.5 Hz
2	cD5	1.56 - 3.12 Hz
3	cD7	0.39 - 0.78 Hz

Nota. Las características identificadas son las mismas para los tres métodos de análisis con DT.

En la Figura 41, se representan gráficamente la PSD y los intervalos de confianza para los cuatro tipos de microsismos con el 99% de grado de confianza en conjunto con las bandas de frecuencia principales que se obtienen mediante el modelo SVM, las cuales se encuentran resumidas en la Tabla 19.

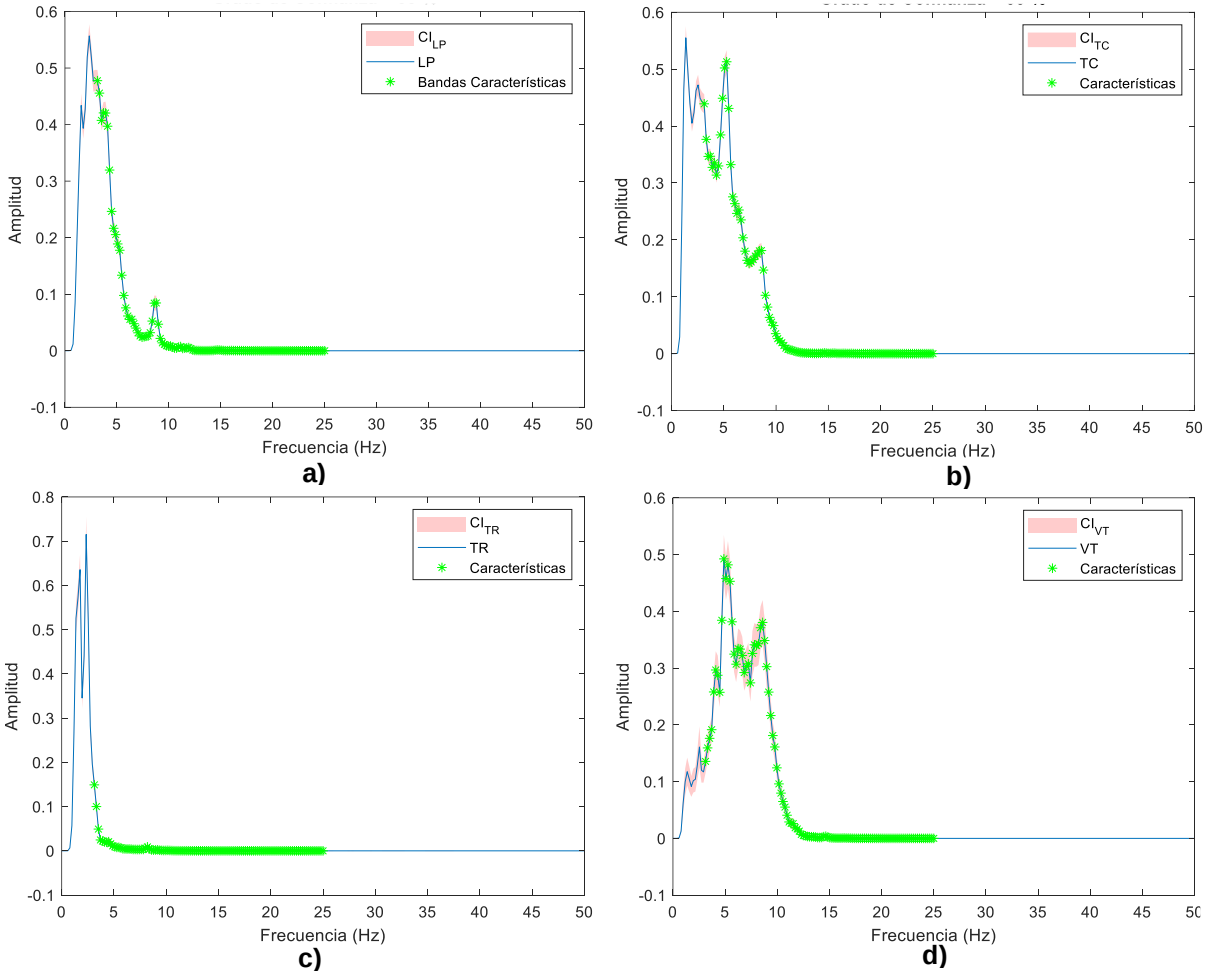
Tabla 19

Principales características identificadas con SVM.

Orden	Característica	Frecuencia
1	cD4	3.12 - 6.25 Hz
2	cD2	12.5 - 25 Hz
3	cD3	6.25 - 12.5 Hz

Figura 41

PSD, intervalos de confianza y características identificadas con modelo SVM.



Nota. Los intervalos de confianza y las características seleccionadas se representan mediante sus PSD. En la fila superior: a) LP y b) TC; fila inferior c) TR y d) VT con 99% de grado de confianza para todos los modelos DT. $\mathrm{CI}_{\mathrm{XX}}$ representa la zona de los intervalos de confianza.

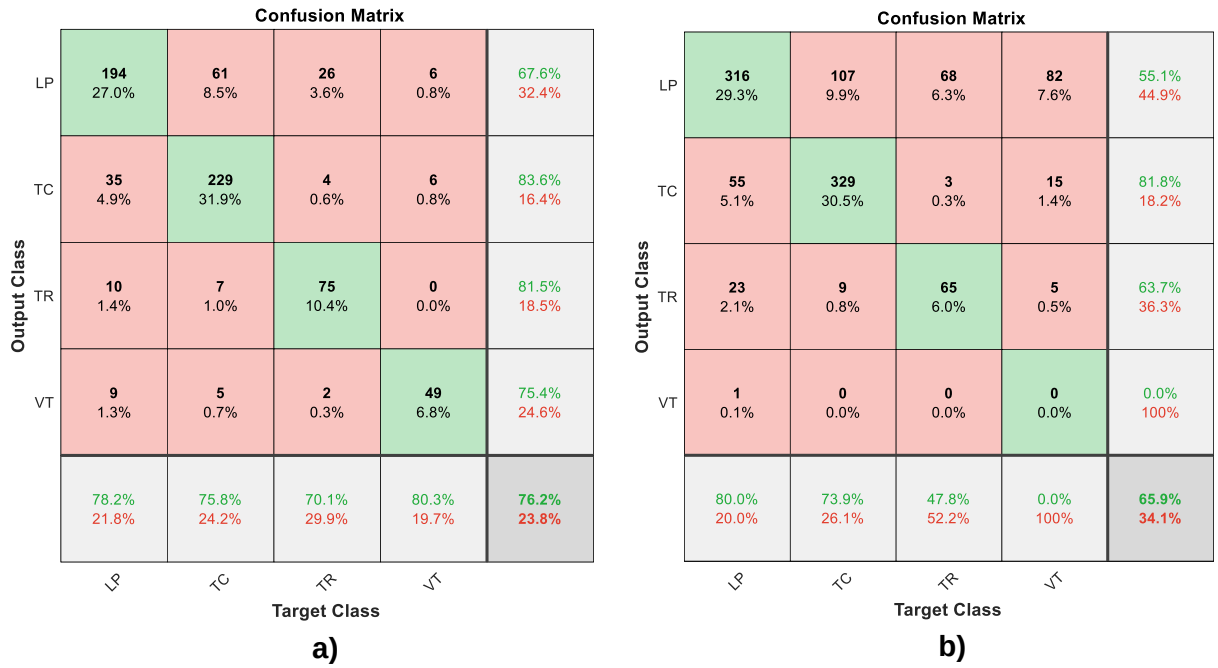
Evaluación de características seleccionadas

Con el fin de encontrar las métricas de rendimiento al usar las características identificadas, se clasifica nuevamente los microsismos con los modelos ML. En este proceso se emplea nuevamente el algoritmo DT y SVM; pero se dispone únicamente de las tres

características más representativas como entrada al modelo de ML. Bajo este contexto, en la Figura 42, se pueden observar los resultados de rendimiento de dicho proceso.

Figura 42

Matrices de Confusión al reentrenar modelos ML.



Nota. Se expresan los reentrenamientos con: a) DT Podamiento y b) SVM

Para los modelos DT, se toma en cuenta únicamente los resultados obtenidos con DT Podamiento, ya que es el modelo que entrega los mejores resultados de entre los tres modelos: configuración automática, validación cruzada y podamiento. De las matrices de confusión que se observan en la Figura 42 se destaca el valor de exactitud que se obtiene en cada modelo. Para DT Podamiento se obtiene una exactitud de clasificación del 76.2% y para el modelo SVM se obtiene una exactitud igual al 65.9%. La precisión de SVM es mucho menor a la que se obtiene con DT y la diferencia que existe entre estas dos métricas es del 10.3%. El principal problema que presenta el modelo SVM es que es incapaz de clasificar cualquier muestra que pertenezca a la clase VT. El modelo la clasifica como falso negativo.

Es importante destacar que esta clasificación se lleva a cabo con únicamente tres características de entrada: cD3 (6.25 - 12.5 Hz), cD5 (1.56 - 3.12 Hz) y cD7 (0.39 - 0.78 Hz) para el modelo DT podamiento; mientras que para el modelo SVM se usan las características cD4 (3.12 - 6.25 Hz), cD2 (12.5 - 25 Hz) y cD3 (6.25 - 12.5 Hz). En la Tabla 20 se brinda un resumen de las métricas que se obtienen a partir de los datos de las matrices de confusión expresadas en la Figura 42. Como se aprecia en dicha tabla, el modelo con mejores resultados en el reentrenamiento es el modelo DT Podamiento.

Tabla 20

Métricas de reentrenamiento de los distintos modelos ML.

Método	A (%)	P (%)	S (%)	R (%)	BER
DT - Defecto	75	75	91	75	0,19
DT - Validación Cruzada	71	72	89	72	0.2
DT - Podamiento	**76**	**77**	**91**	**76**	**0.16**
SVM	66	50	87	50	0.31

Por lo tanto, en base a la información que se expresa en la Tabla 21, las tres características finales que se seleccionan son aquellas que brindan las mejores métricas enfocadas a la exactitud de clasificación. Por consiguiente, las bandas de frecuencia que permiten caracterizar cada evento con un porcentaje de exactitud del 76.2%, al emplear DT con podamiento, son cD3, cD5 y cD7, cuyas frecuencias respectivas se expresan en la Tabla 21.

En base a los rangos de frecuencia establecidos en la Tabla 21, se procede a graficar las tres características que se identifican en el presente trabajo y se las representa en la Figura 43 en conjunto con las PSD medias de cada evento y los intervalos de confianza.

Tabla 21

Características finales identificadas.

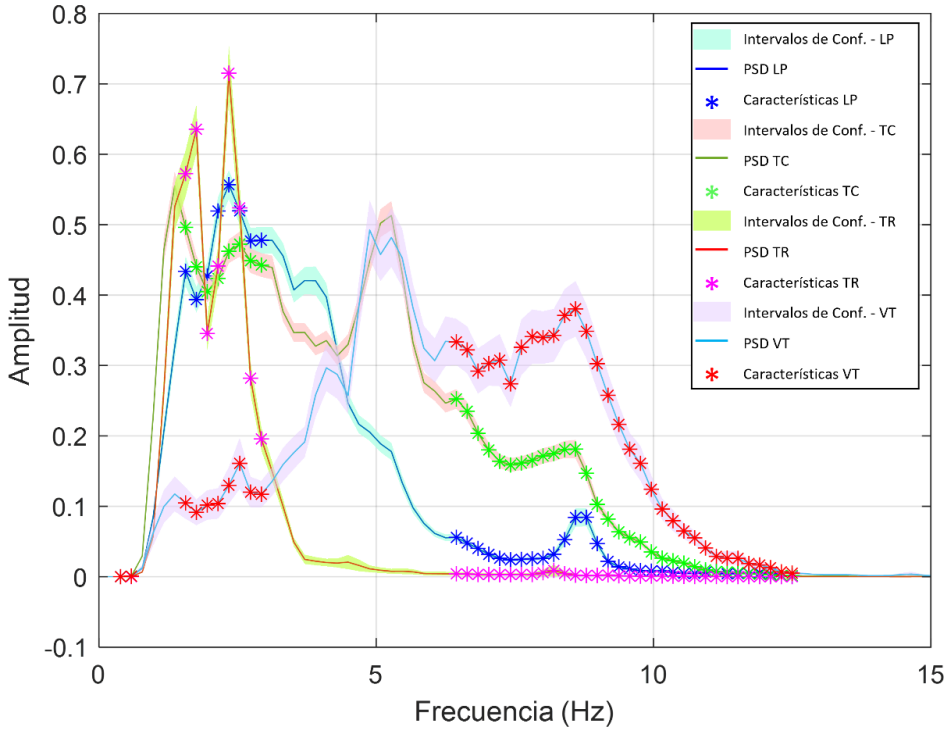
Orden	Característica	Frecuencia
1	cD3	6.25 - 12.5 Hz
2	cD5	1.56 - 3.12 Hz
3	cD7	0.39 - 0.78 Hz

Nota. Estas características identificadas representan el resultado concluyente del presente trabajo de investigación.

En la Figura 43, se logra observar los rangos de frecuencia en los que se aprecian las características de los microsismos, al emplear únicamente 3 características generadas a partir de la WT.

Figura 43

Características finales en conjunto con PSD de cada evento y sus intervalos de confianza.



Nota. Las zonas sombreadas representan los intervalos de confianza.

Para finalizar, se debe destacar la importancia de realizar este reentrenamiento bajo estas condiciones es evaluar cuál es el porcentaje de exactitud de clasificación al usar únicamente las tres características más importantes, obtenidas a partir de la WT en distintos niveles. Para ello, se debe destacar que el hecho de obtener un 76% de exactitud en la clasificación al emplear solo tres características es un logro significativo, ya que se bastaría con aumentar otras características, ya sea en tiempo o frecuencia, para que el modelo sea capaz de apuntar a una exactitud de clasificación cercana al 100%.

Evaluación con Señales Sintéticas

Una vez identificadas las características principales y los intervalos de confianza, se realizan pruebas con señales sintéticas. Estas señales se generan gracias a la colaboración de los equipos de investigación asociados, que trabajan de manera paralela al presente tema de investigación. El método que se emplea para generar dichas señales sintéticas es Bootstrap, con el cual se generan un total de 10 000 señales, divididas en conjuntos de 2 500 correspondientes a los microsismos LP, TC, TR y VT.

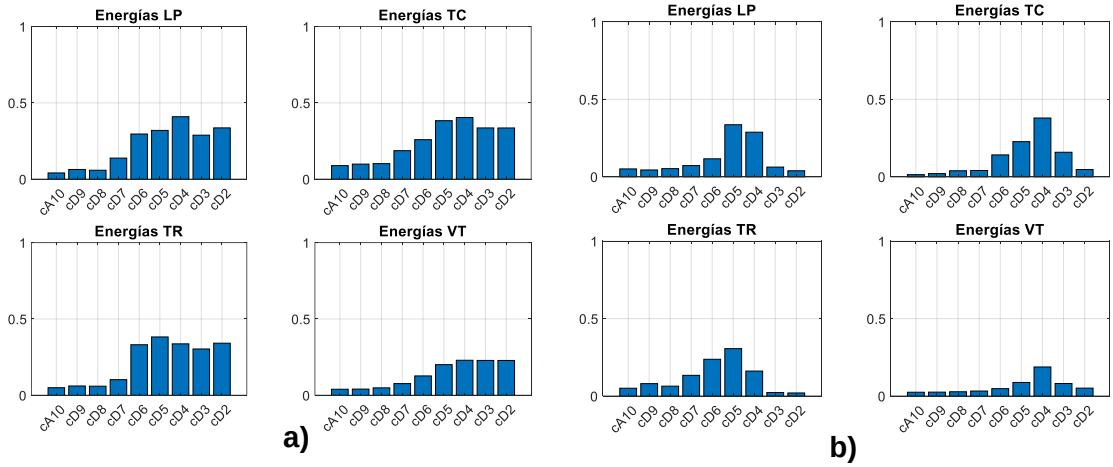
Análisis energético

El primer paso dentro del análisis wavelet, consiste en extraer los niveles medios de energía de los coeficientes cD y cA que se generan de la descomposición de 10 niveles para cada microsismo. En la Figura 44 se representa gráficamente el resultado de esta operación y se puede observar la distribución de los niveles medios de energía de los coeficientes cD y cA, que se obtienen a partir de las señales sintéticas. Tanto para la Figura 44a y como para la Figura 44b, se observa una similitud notable en los valores medios de las bandas de baja frecuencia, como cA10 (0 a 0.0485 Hz), cD9 (0.097 a 0.195 Hz), cD8 (0.195 a 0.39 Hz) y cD7 (0.39 y 0.78 Hz). Sin embargo, esta uniformidad no se refleja en las bandas de alta frecuencia, donde se evidencian diferencias significativas en sus valores medios.

Bajo el enfoque de análisis con señales sintéticas, al observar la Figura 44 se debe destacar que existen características muy singulares que permiten identificar directamente a una señal sintética de las señales originales. En la Figura 44a, los niveles de energía de las bandas de alta frecuencia, cD2 (12.5 a 25 Hz), cD3(6.25 a 12.5 Hz) y cD4 (3.12 a 6.25 Hz), se elevan de manera significativa si se compara con las mismas bandas de las señales originales que se observan en la Figura 44b. Por lo tanto, al comparar directamente las bandas cD2, cD3 y cD4, se puede descubrir si el microsismo a analizar pertenece al grupo de señales sintéticas.

Figura 44

Análisis de energías con coeficientes obtenidos mediante WT de 10 niveles.



Nota. Resultados se generan con WM Daubechies. Donde: a) Valores medios de energía de coeficientes wavelet de señales sintéticas. A la derecha, en b) Valores medios de energía de coeficientes wavelet de señales originales.

Por consiguiente, el resultado de analizar las energías de los coeficientes wavelet cD y cA de las señales sintéticas arrojan una característica única que permite diferenciar entre este tipo de señales y las señales originales.

Entrenamiento de clasificador

El entrenamiento establecido para el presente trabajo de investigación se define por dos enfoques: DT y SVM. Por lo tanto, al tomar los coeficientes de la WT multinivel de las señales sintéticas e ingresarlas como características a estos dos tipos de clasificadores se obtienen los resultados que se observan en la Tabla 22, en donde, se logra apreciar que el modelo con mejor exactitud de clasificación es DT – Auto configuración dentro del enfoque DT. Con el empleo de este modelo se logra obtener una exactitud del 60% de clasificación con un valor de BER del 0.3. Para el enfoque con SVM, se obtiene un valor de exactitud de clasificación del 64% y un BER de 0.2, los cuales son los valores más alto dentro del análisis con señales sintéticas.

Tabla 22

Métricas de entramiento de los distintos modelos ML.

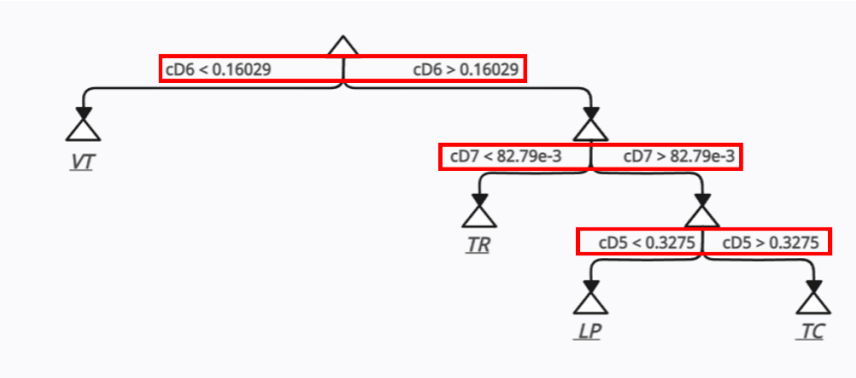
Método	A (%)	P (%)	S (%)	R (%)	BER
DT – Auto Configuración	**60**	**59**	**87**	**60**	**0,3**
DT - Validación Cruzada	57	57	86	57	0,3
DT - Podamiento	59	59	87	59	0,3
SVM	**64**	**64**	**88**	**64**	**0,2**

Selección de características

Tras identificar los modelos ML con más alto rendimiento, se procede a seleccionar las características de los métodos de entrenamiento DT – Auto Configuración y SVM con señales sintéticas. En la Figura 45 se observa la identificación de las características más importantes seleccionadas bajo el enfoque DT. Las características cD6, cD7 y cD5, con frecuencias 0.78 a 1.56 Hz, 0.39 a 0.78 Hz y 1.56 a 3.12 Hz respectivamente, se identifican como características principales bajo el enfoque DT.

Figura 45

Nodos raíz en modelo DT – autoconfiguración.

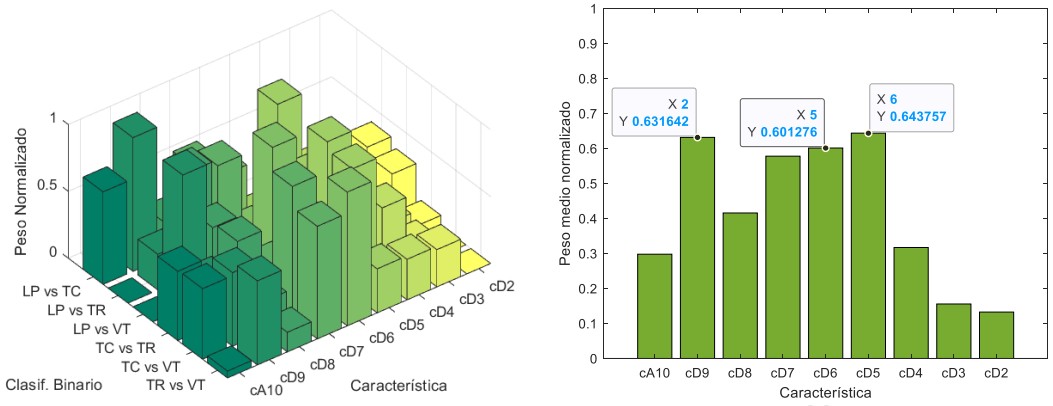


Nota. Con señales sintéticas, las tres características principales identificadas son cD6, cD7 y cD5.

Mediante la Figura 46a, se puede observan los pesos normalizados de cada característica en cada clasificador binario. Se puede observar que no existe una acumulación de pesos en un solo sector; sin embargo, mediante los valores medios expresados en la Figura 46b, se logra identificar los pesos más influyentes dentro del modelo SVM.

Figura 46

Pesos de características con respecto a clasificadores binarios SVM en reentrenamiento



Nota. Pesos de características: a) con respecto a los clasificadores binarios SVM y b) pesos medios con señales sintéticas. Características principales identificadas son cD6, cD7 y cD5.

De esta manera que se detalla anteriormente, se logra identificar a cD9 (0.0975 a 0.195 Hz), cD6 (0.78 a 1.56 Hz) y cD5 (1.56 a 3.12 Hz) como las características principales bajo la clasificación mediante el enfoque SVM.

Reentrenamiento

Por último, con el fin de evaluar el rendimiento de las características seleccionadas se procede a clasificar los microsismos nuevamente. En dicho proceso, se emplean únicamente las tres características más representativas obtenidas mediante el modelo DT y SVM que se detallan en la Tabla 23.

Tabla 23

Principales características identificadas con modelos DT y SVM.

Modelo	Orden	Característica	Frecuencia
DT	**1**	cD6	0.78 – 1.56 Hz
	2	cD7	0.39 – 0.78 Hz
	3	cD5	1.56 – 3.12 Hz
SVM	**1**	cD9	0.0975 - 0.195 Hz
	2	cD6	0.78 - 1.56 Hz
	3	cD5	1.56 – 3.12 Hz

Nota. Las características identificadas son las mismas para los tres métodos de análisis con DT.

Tras reentrenar los modelos con las características seleccionadas que se detallan en la Tabla 23, se observan los resultados de la Tabla 24, en donde destaca el modelo DT con podamiento como aquel que posee la mejor métrica de exactitud de clasificación del 55% con tres características bajo el enfoque de señales sintéticas. Comparado con los resultados

presentados en la Tabla 22 existe una diferencia del 4% con la exactitud de clasificación del modelo DT – Podamiento.

Tabla 24

Métricas de reentrenamiento de los distintos modelos ML.

Método	A (%)	P (%)	S (%)	R (%)	BER
DT – Auto Configuración	54	54	85	54	0,3
DT - Validación Cruzada	46	46	82	46	0,4
DT - Podamiento	**55**	**55**	**85**	**55**	**0,3**
SVM	50	50	84	51	0,3

Capítulo V

Conclusiones y recomendaciones

Conclusiones

- Con el desarrollo del presente trabajo de investigación se ha logrado identificar tres características que permiten definir un microsismo en base a bandas de frecuencia establecidas por la WT, las cuales son: cD3 (3.12 – 6.25 Hz), cD5 (1.56 - 3.12 Hz) y cD7 (0.39 - 0.78 Hz). Esto se obtuvo al emplear métodos de selección de características basados en ML, lo que permitió obtener un porcentaje de exactitud de clasificación del 76% al usar un clasificador DT con podamiento. En consecuencia, se concluye que la relevancia de esto radica en que al usar tan solo tres características de entrada al modelo ML, se obtuvo esta cifra de exactitud y por lo tanto, sería suficiente con aumentar unas cuantas características más, ya sea en tiempo o frecuencia, para que el modelo sea capaz de apuntar a una exactitud de clasificación cercana al 100%.
- Las aplicaciones que tienen las MW son diversas y escoger correctamente la familia de MW adecuada depende del enfoque en el cual se desee aplicar. Es por esto que se considera el principio de similaridad para escoger la MW. En este caso, se han considerado las familias MW Daubechies y MW Symlets ya que son muy similares a las señales micro sísmicas y de entre estas dos, la MW que brinda los mejores resultados es la familia de Daubechies. Pese a que se realizaron pruebas con MW Symlets, se determinó con éxito que mediante la MW Daubechies se obtienen las mejores métricas de rendimiento y por ende, es el principio mediante el cual se obtuvieron las características principales.
- El intervalo de confianza que se establece en el presente trabajo es del 99% y tras analizar el espectro medio de cada evento se concluye que, en promedio, el espectro de cada señal es muy similar a su media, con lo cual se concluye que la base de datos

adquirida es una base de datos académica en donde muy probablemente existió un proceso previo de elección de las mejores señales con el fin de eliminar las señales atípicas para aproximar los datos a una base de datos ideal.

- Para el desarrollo se planteó un desarrollo de manera sucesiva en el cual se procede a descartar los métodos y enfoques que no presentan los mejores resultados. Es así que, en un inicio se planteó un análisis de 6, 8 y 10 niveles de descomposición con WT. La razón de escoger estos niveles de análisis se debe a dos factores: la capacidad de diferenciar un evento de otro mediante los niveles de energía de la wavelet con 6, 8 y 10 niveles y segundo, la reducción del número de características.
- Tal y como se menciona, a medida que se avanzó en el desarrollo del trabajo de investigación, se descartaron varios enfoques de análisis y uno de ellos fue la proporción de división de datos. Mediante procesos de prueba y error, se logró determinar que se obtienen los mejores resultados con una división de datos del 80% para entrenamiento y 20% para prueba.
- Dentro de todo esto, los mejores resultados obtenidos se resumen en el uso de la descomposición con WT de 10 niveles en donde cada coeficiente promedio obtenido es una característica de entrada a un modelo de clasificación basado en ML. La proporción de datos que se emplea es de 80/20 y el modelo ML es DT con podamiento. Al seleccionar las tres características más importantes del modelo ML y reentrenar dicho modelo con estas tres características se obtiene un porcentaje de exactitud del 76%. Bajo estas características es que se obtuvieron los mejores resultados. Al hablar de que se obtuvo un porcentaje de exactitud del 76% al clasificar cuatro eventos de microsismos con un clasificador DT, se destaca la importancia de obtener dicha cifra con apenas tres características de entrada un modelo DT.

Trabajos futuros

- Como trabajos futuros se plantea el desarrollo del mismo proceso bajo un enfoque de datos balanceado, dado a que la base de datos empleado durante el desarrollo de este trabajo de investigación consistió en distintos números de señales para cada evento. Sería apropiado desarrollar este procedimiento planteado para la identificación de características e intervalos de confianza bajo un método de análisis con una base de datos balanceada.
- Dentro de ML existen una gran variedad de técnicas que permiten realizar una clasificación. Además de los enfoques en DT y SVM, se pueden explorar otros algoritmos de ML supervisado a futuro, con el fin de realizar el proceso de selección de características y determinar cuáles son las bandas de frecuencias principales.
- Además de los coeficientes wavelet de detalle (cD) y de aproximación (cA) , podría ser interesante explorar la inclusión de otras características relacionadas al Wavelet. En la Tabla 10 se encuentran datos estadísticos que se obtienen directamente de la WT mediante el diagrama de caja y bigotes. Con estos datos relacionados netamente con la WT, se puede añadir más características y ampliar la información utilizada por los modelos de ML para intentar mejorar su capacidad de distinción entre diferentes tipos de microsismos.
- Otra sugerencia de trabajo a futuro es utilizar una mayor cantidad de datos y ampliar el conjunto de señales micro sísmicas mediante técnicas de generación de señales sintéticas y emplearlas para el entrenamiento y evaluación de los modelos. Esto podría contribuir a una mayor exactitud de clasificación.

Referencias

Alessio, S. M. (2016). *Digital Signal Processing and Spectral Analysis for Scientists: Concepts and Applications.* Switzerland: Springer Cham. doi:https://doi.org/10.1007/978-3-319-25468-5

Altamirano, R. B. (2021). *Sistema de reconocimiento de microterremotos en tiempo real del volcán Cotopaxi aplicando aprendizaje supervisado.* Universidad de las Fuerzas Armadas ESPE, Eléctrica, Electrónica y Telecomunicaciones, Quito. Retrieved from http://repositorio.espe.edu.ec/xmlui/bitstream/handle/21000/23743/T-ESPE-044263.pdf?sequence=1&isAllowed=y

Alvarez, M., Henao, R., & Duque, E. (2007). *Clasificación de eventos sísmicos empleando procesos gaussianos.* Retrieved from Scientia ET Technica: https://doi.org/10.22517/23447214.5385

aprendeIA. (2019). *Métodos de seleccion de características machine learning*. Retrieved from https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/

Arana, C. (2021). *MODELOS DE APRENDIZAJE AUTOMÁTICO MEDIANTE ÁRBOLES DE DECISIÓN.* Universidad del CEMA, Negocios, Buenos Aires. Retrieved from https://ucema.edu.ar/publicaciones/download/documentos/778.pdf

Armijos Sarango, A. E., Palacios Serrano, I. S., & González Martínez, S. (2021). *Evaluación y comparación de algoritmos para la detección automática de eventos sísmicos. Revista Tecnológica - ESPOL, 33(2), 58–74.* Retrieved from https://doi.org/10.37815/rte.v33n2.830.

Barlowe, S. (2011). A visual analytics approach to feature discovery and subspace exploration in protein flexibility matrices. *(Tesis de Doctorado).*

Battaglia, J., & Aki, K. (2003). *Location of seismic events and eruptive fissures on the Piton de la.* doi:10.1029/2002JB002193.

Brownlee , J. (2020, Junio 30). *Rescaling Data for Machine Learning in Python with Scikit-Learn*. Retrieved from Machine Learning Mastery: Making Developers Awesome at Machine Learning: https://machinelearningmastery.com/rescaling-data-for-machine-learning-in-python-with-scikit-learn/

Burrus, S., Gopinath, R., & Guo, H. (1998). Wavelets and Wavelet Transforms. *Recherche, 67*.

Canário, J. P., Fernandes de Mello, R., Curilem, M., Huenupan, F., & Araujo Rios, R. (2020, Junio). Llaima volcano dataset. *ELSEVIER, 30*. doi:https://www.sciencedirect.com/science/article/pii/S2352340920305217

Canário, P., Fernandes, R., Curilem, M., Huenupan, F., & Araujo, R. (2020). Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification. *Journal of Volcanology and Geothermal Research, 401*, 106881.

Cardenas, D., Orozco-Alzate, M., & Castellanos-Dominguez, G. (2013). Selection of time-variant features for earthquake classification at the Nevado-del-Ruiz volcano. *Computers & Geosciences. 51*, 293–304.

Castro, R. (2002). *Análisis de la teoría de ondículas orientada a las aplicaciones en ingeniería eléctrica: Fundamentos. Madrid - España*.

CENAPRED. (2018). *Sistema de reconocimiento automático de señales sísmicas.* Mexico D.F. Retrieved from http://www1.cenapred.unam.mx/DIR_INVESTIGACION/2109/FRACCION_XLI/RV/12_%20HTK_2018.pdf

Cortés , J., & Cano, H. (2007). Del análisis de Fourier a las Wavelets - Transformada continua wavelet (CWT). *Scientia et Technica*(37).

de los Ángeles Linares, M., Ortiz, R., & Marreno, J. M. (n.d.). *Riesgo Volcánico*. Retrieved from Guía didáctica para profesores: https://www.ign.es/web/resources/docs/IGNCnig/VLC-Guia-Riesgo-Volcanico.pdf

Dorado Betancourt, H. A. (2019, Enero). *Wrapper para la construcción de modelos de aprendizaje supervisado basado en arreglos de cobertura que permite la estimación de la importancia de las variables de entrada y la selección de atributos*. Retrieved from http://repositorio.unicauca.edu.co:8080/bitstream/handle/123456789/1299/Wrapper%20para%20la%20construcci%C3%B3n%20de%20modelos%20de%20aprendizaje%20supervisado%20basado%20en%20arreglos%20de%20cobertura.pdf?sequence=1&isAllowed=y#:~:text=Los%20wrappers%20s

Fernández, A. (2007). Estudio de técnicas basadas en la transformada wavelet y optimización de sus parámetros para la clasificación por texturas de imágenes digitales. *(Tesis doctoral).* Universidad Politécnica de Valencia, Valencia.

Franco Marín, L. E. (2019). *Comportamiento eruptivo del Volcán Llaima (2007-2010) e incidencia del terremoto del Maule MW 8.8 en la actividad volcánica y tectónica local.* Retrieved from http://repositorio.udec.cl/jspui/handle/11594/1139

GISI - UNAM. (2018). *Acervo para el mejoramiento del aprendizaje de alumnos de ingeniería e inteligencia artificial.* Retrieved from Transformada Wavelet: https://virtual.cuautitlan.unam.mx/intar/?page_id=1108

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

Ibáñez, J., & Carmona, E. (2000). *SISMICIDAD VOLCÁNICA*.

IBM. (2022). *Árboles de decisión*. Retrieved from https://www.ibm.com/es-es/topics/decision-trees

IGEPN. (2014, Agosto 26). *Actividad sísmica zona de los volcanes Chiles – Cerro Negro*. Retrieved from https://www.igepn.edu.ec/servicios/noticias/content/49-historico?start=8

InteractiveChaos. (2023, 07 11). *Índice de Gini*. Retrieved from https://interactivechaos.com/es/manual/tutorial-de-machine-learning/indice-gini

Kouro, S., & Musalem, R. (2002). Tutorial Introductorio a la Teoría de Wavelet. *Técnicas Modernas en Automáticaa*, 8.

Lara Cueva, R., Paillacho Salazar, V., & Villalba Chaluisa, M. (2017, Marzo). Towards an automatic detection system of signals at Cotopaxi Volcano. *DYNA, 84*, 176-184.

Lara, F., Lara R, Larco, J., Carrera, E., & León, R. (2021). "A deep learning approach for automatic recognition of seismo-volcanic events at the Cotopaxi volcano". *Journal of Volcanology and Geothermal Research*.

Lara-Cueva, R. A., Benítez, D. S., Carrera, E. V., Ruiz, M., & Rojo-Álvarez, J. L. (2016). Feature selection of seismic waveforms for long period event detection at Cotopaxi Volcano. *Journal of Volcanology and Geothermal Research*, 34-49.

Lara-Cueva, R., Bernal, P., Saltos, M. G., Benitez, D. S., & Rojo-Álvarez, J. L. (2014, Febrero). *"Time and Frequency Feature Selection for Seismic Events from Cotopaxi Volcano," 2015 Asia-Pacific Conference on Computer Aided System Engineering, Quito, Ecuador, 2015, pp. 129-134*. Retrieved from doi: 10.1109/APCASE.2015.30.

Lara-Cueva, R., Larco, J. C., Benítez, D. S., Pérez, N., Grijalva, F., & Ruiz, M. (2020). On finding possible frequencies for recognizing microearthquakes at Cotopaxi volcano: A

machine learning based approach. *Journal of Volcanology and Geothermal Research*(407).

Lara-Cueva, R., Paillacho-Salazar, V., & Villalva-Chaluisa, M. (2017). Hacia un sistema de detección automática de señales del volcán Cotopaxi. *Dyna, 84(200)*, 176-184.

Lee, W., & Valdes, C. (1985). *HYPO71PC; a personal computer version of the HYPO71 earthquake location program.* doi:10.3133/ofr85749

Luengas, L., & Toloza, D. (2020, Febrero 14). *Análisis frecuencial y de la densidad espectral de potencia de la estabilidad de sujetos amputados*. Retrieved from https://doi.org/10.22430/22565337.1453

Maisueche, A. C. (2019). *UTILIZACIÓN DEL MACHINE LEARNING EN LA INDUSTRIA 4.0.* Universidad de Valladolid, Eescuela de Ingenierias Industriales, Valladolid. Retrieved from https://core.ac.uk/download/pdf/228074134.pdf

Marrón, S. (2021, Abril 21). *Aprendizaje automático, explicado*. Retrieved from https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Martínez , J., & Castro , R. (2002). Análisis de la teoría ondículas orientada a las aplicaciones en ingeniería eléctrica:Fundamentos. *E.T.D.I. Industriales*, 161.

MathWorks. (2023). *Support Vector Machine (SVM)*. Retrieved from https://la.mathworks.com/discovery/support-vector-machine.html

MathWorks. (n.d.). *MathWorks*. Retrieved from Wavelet Families: https://la.mathworks.com/help/wavelet/ug/wavelet-families-additional-discussion.html

Merry , R. (2005). Wavelet Theory and Applications: a literature study. *DCT Raporten, 2005.053*, 50.

Minango, G. M. (2022). *Clasificación de eventos sismo volcánicos usando características psicoacústicas mediante técnicas de aprendizaje automático supervisado y no supervisado.* Universidad de las Fuerzas Armadas ESPE, Quito. Retrieved from http://repositorio.espe.edu.ec/bitstream/21000/31541/1/T-ESPE-052390.pdf

Mora, M., & Alvarado, G. E. (2001). *¿Qué es un tremor?* Retrieved from Primer taller en actualización de Sismología Volcánica.-Red Sismológica Nacional (UCR-ICE): https://rsn.ucr.ac.cr/documentos/educativos/vulcanologia/5099-que-es-un-tremor

Moreno, H. N. (2009). *Estilos eruptivos 2007-2008 del volcán Llaima, Andes del sur.* Santiago: XII Congreso Geológico Chileno.

Naranjo, J., & Moreno , H. (2011). *Servicio Nacional de Geología y Minería.* Retrieved from Carta Geológica de Chile: https://rnvv.sernageomin.cl/volcan-llaima/

Naranjo, J., & Moreno, H. (2005). *Geología del volcán Llaima, Región de la Araucanía*. Retrieved from Servicio Nacional de Geología y Minería, Carta Geológica de Chile, Serie Geología Básica, No. 88, Santiago-Chile.

Naranjo, J., & Moreno, H. (2011). *Servicio Nacional de Geología y Minería*. Retrieved from Carta Geológica de Chile: https://rnvv.sernageomin.cl/volcan-llaima/

Nieto, N., & Orozco , D. (2008). El uso de la Transformada Wavelet Discreta en la reconstrucción de ondas senosoidales. *Scientia et Technica*(38), 381-386.

OVSICORI. (2023). *Vulcanología*. Retrieved from Universidad Nacional Costa Rica: http://www.ovsicori.una.ac.cr/index.php/faqs/vulcanologia/acerca-de-sismos-asociados-con-actividad-volcanica

Peng, R. L. (2008, Abril). Design of smart sensing component for volcano monitoring. *IET 4th International Conference on Intelligent Environments, 14*, págs. 1-7.

Pérez , O. (2004). *Algoritmos de compresion de imagenes sin movimiento para comunicaciones moviles (3G) utilizando teoria de wavelets.* Puebla-México: Departamento de Ingeniería Electrónica. Escuela de Ingeniería, Universidad de las Américas Puebla.

Pérez Quisaguano, A. S. (2018). *Detección automática de eventos sísmicos en el volcán Cotopaxi mediante técnicas de Aprendizaje de Máquinas:.* Retrieved from https://repositorio.espe.edu.ec/bitstream/21000/13812/1/T-ESPE-057527.pdf

Pérez, N., Benítez, D., Grijalva, F., Lara, R., Ruiz, M., & Aguilar, J. (2020). ESeismic: Towards an Ecuadorian volcano seismic repository. *Journal of Volcanology and Geothermal Research*, ELSEVIER. doi:https://doi.org/10.1016/j.jvolgeores.2020.106855

Probabilidad y Estadística.net. (2023). *Intervalo de confianza para la media*. Retrieved from https://www.probabilidadyestadistica.net/intervalo-de-confianza-para-la-media/#:~:text=El%20intervalo%20de%20confianza%20para%20la%20media%20es%20un%20intervalo,con%20un%20margen%20de%20error.

SERNAGEOMIN. (2023). *Volcán Llaima*. Retrieved from https://rnvv.sernageomin.cl/volcan-llaima/

Sydorenko , I. (2021, March). *Label your data*. Retrieved from Machine Learning and Training Data: What You Need to Know: https://labelyourdata.com/articles/machine-learning-and-training-data

TIBCO. (2023). *¿Qué es el aprendizaje supervisado?* Retrieved from https://www.tibco.com/es/reference-center/what-is-supervised-learning

Tilling, R. I., & Beate, B. (1993). Apuntes para un curso breve sobre los peligros volcánicos. *New México: World Organization of Volcano Observatories.* Santa Fé, Nuevo México, U.S.A.

Vásconez Gómez, F. S. (2020, Junio 30). *Desarrollo de un sistema de clasificación de eventos sismo volcánicos usando librerías de Machine Learning en Python.* Retrieved from http://repositorio.espe.edu.ec/jspui/bitstream/21000/22408/1/T-ESPE-043768.pdf

Wassermann, J. (2011). Volcano seismology. Fürstenfeldbruck. *Geophysikalisches Observatorium der Ludwig-Maximilians Universität München*, 1-67.

Wener-Allen, G., Johnson, J., Ruiz, M., & Lees, J. (2005). *Monitoring volcanic eruptions with a wireless sensor network, Wireless Sensor Networks, 2005. Workshop on Proceeedings of the Second European, pp. 108-120*. Retrieved from DOI: 10.1109/EWSN.2005.1462003

Yugsi, J. P. (2022). *GENERACIÓN DE SEÑALES VOLCÁNICAS ARTIFICIALES DE TIPO LP (LONG-PERIOD) Y VT (VOLCANO-TECTONIC) A PARTIR DE UNA BASE DE DATOS DEL VOLCÁN COTOPAXI USANDO LA TÉCNICA DE BOOTSTRAPPING.* Escuela Politécnica Nacional, Quito. Retrieved from https://bibdigital.epn.edu.ec/bitstream/15000/23276/1/CD%2012691.pdf

Zobin, V. M. (2011). *Introduction to Volcanic Seismology.*

Apéndices