Identificación de Maderas Aserradas Propias de la Zona de Santander Mediante el Uso de Herramientas Basadas en Narices Electrónicas

Naren Arley Mantilla Ramírez

Trabajo de Investigación para optar al título de Magíster en Ingeniería de Telecomunicaciones

Director

Alexander Sepúlveda Sepúlveda

Doctor en Ingeniería

Codirectores

Homero Ortega Boada

Doctor en Ingeniería

Luisa Fernanda Ruiz Jiménez

Magíster en Ingeniería de Telecomunicaciones

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2021

A mis padres, modelo y ejemplo en todo sentido; y a toda mi familia. Todos ustedes forman parte de mis raíces y de mi origen, y a todos ustedes les debo mi ser.

A mis amigos, compañeros, colegas, directores y a la comunidad académica que me acompañó. Cada palabra de aliento, cada comentario, cada opinión y cada momento fueron un grano de arena para la construcción de este gran castillo.

A la persona que me acompañó durante los momentos más difíciles, escuchó mis quejas y decepciones, celebró mis éxitos, y me dio fortaleza cuando estaba perdiendo la confianza. Aunque el ciclo terminó, su aporte fue crucial.

A quien fue testigo de la última etapa y todas sus adversidades, incluyendo una pandemia y otras cosas. A quien me acompaña hoy con una gran sonrisa y la mejor actitud.

La lección más importante es que cada persona representa una oportunidad de ser mejor persona, mejor profesional, mejor amigo, ...

Naren

Agradecimientos

Este fue financiado, en parte, en el marco del proyecto "*Plataforma IoT para el desarrollo de servicios inteligentes de apoyo al monitoreo ambiental*, código 1971" −− Vicerrectoría de Investigación y Extensión, Universidad Industrial de Santander; y, la otra parte por la *Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones* (E3T) de la misma universidad.

Agradezco al profesor Iván Porras del Instituto de Proyección Regional y Educación a Distancia de la UIS por sus valiosos comentarios, fundamentales en etapas tempranas del presente trabajo.

Al profesor Franklin, director del trabajo, por su tiempo y dedicación, y por su oportuna y acertada orientación en la formación integral en investigación. También al profesor Homero y Luisa Fernanda, codirectores, por sus importantes aportes.

A todas las personas que me acompañaron en este proceso, desde las dependencias administrativas de la escuela hasta los compañeros investigadores con los que compartí día a día.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

pág.

Resumen

Título: Identificación de Maderas Aserradas Propias de la Zona e Santander Mediante el Uso de Herramientas Basadas en Narices Electrónicas [*]

Autor: Naren Arley Mantilla Ramírez [**]

Palabras Clave: Identificación de Madera, Verificación, Nariz Electrónica, Matriz de Sensores Químicos, Aplicaciones de Aprendizaje Automático, Clasificación de Vectores de Soporte (SVM), Aumento de Datos, GMM-UBM, Modelo de Mezclas gaussianas, Modelo Universal.

Descripción: La deforestación y extracción desordenada de madera ponen en peligro algunas especies maderables vulnerables. Estas especies prohibidas podrían detectarse durante su proceso de transporte si las entidades de vigilancia y control tuvieran los instrumentos de seguimiento adecuados. Si bien en trabajos anteriores se reportan métodos para identificar especies de madera, estos no son aplicables a sitios alejados de las principales ciudades. En el presente trabajo se propone utilizar narices electrónicas (arreglos de sensores químicos) para identificar especies maderables, a partir de los compuestos volátiles que estas emanan. La medición de aromas se realiza mediante el uso de una matriz de 16 sensores químicos, cuyas curvas son la entrada a un procedimiento de estimación de características. Luego, se realiza un análisis de componentes principales, para finalmente aplicar una estrategia de clasificación basada en máquinas de vectores de soporte.

En contraste a trabajos previos, en el presente trabajo se aplica una nueva estrategia al problema de detección de especies: verificación utilizando modelos de mezclas gaussianas y modelos de referencia (*Universal Background Model*). Adicionalmente, en comparación con trabajos reportados recientemente, las condiciones de recolección de muestras son más cercanas a las encontradas en entornos reales para los cuales este trabajo busca resolver el problema; y, el número de muestras es mayor y más variado. De otra parte, debido a que la distribución de muestras recolectadas para las especies no está balanceada, se aplica una técnica de aumento de datos para compensar el desequilibrio en las clases. Al realizar los experimentos se encuentra un desempeño de aproximadamente 80 % en cuanto a identificación de especies. Resultados similares se encuentran en los experimentos de verificación, los cuales mejoran hasta un 88 % cuando se aplica *LASSO* a modo de estrategia de selección de características. A pesar de los resultados prometedores, se deben realizar mayores esfuerzos para obtener un mejor desempeño.

---

[*] Trabajo de Investigación de Maestría

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Alexander Sepúlveda Sepúlveda, Doctor en Ingeniería. Codirector: Homero Ortega Boada, Doctor en Ingeniería. Codirectora: Luisa Fernanda Ruiz Jiménez, Magíster en Ingeniería de Telecomunicaciones.

Abstract

Title: Identification of Sawn Wood own of the Santander Area by Using Electronic Nose-Based Tools [*]

Author: Naren Arley Mantilla Ramírez [**]

Keywords: Wood Identification, Verification, Electronic Nose (E-Nose), Chemical Sensor Arrays, Machine Learning Applications, Support Vector Classification (SVM), Data Augmentation, GMM-UBM, Gaussian Mixture Models, Universal Background Model.

Description: Deforestation and disordered timber extraction endanger some vulnerable timber species. These prohibited species could be detected during their transportation process if surveillance and control entities had adequate monitoring instruments. Although methods for identifying wood species are reported in previous works, they are not applicable to sites far from the main cities. In present work it is proposed to use electronic noses (chemical sensor arrays) to quickly identify wood species, from the volatile compounds their timbers emanate. The measurement of aromas is done by using an array of 16 chemical sensors, whose curves are the input to a feature estimation procedure. Then, principal component analysis is performed, to finally apply a classification strategy based on support vector machines.

In this work, a new approach to the problem of timer species detection is applied: verification by using Gaussian mixture modeling and Universal Background Model (UBM). In addition, in contrast to previous works, the data collecting procedure is closer to those found on real environments for which this work seeks to solve the problem; and the number of samples is larger and more varied. Since the resulting number of samples per species is not balanced, a data augmentation technique is applied to compensate the class imbalance. A performance of approximately 80 % is found in case of identification experiments. Similar results are found for verification task, whose results are improved up to 88 % when applying LASSO as feature selection strategy. Although the promising results, greater efforts should be carried out to obtain a better performance.

---

[*] Master Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Alexander Sepúlveda Sepúlveda, Doctor en Ingeniería. Codirector: Homero Ortega Boada, Doctor en Ingeniería. Codirectora: Luisa Fernanda Ruiz Jiménez, Magíster en Ingeniería de Telecomunicaciones.

Introducción

El abuso de los recursos naturales afecta a un gran número de ecosistemas, genera un desequilibrio ambiental y contribuye negativamente al cambio climático. Por esta razón, se creo la Estrategia Integral de Control a la Deforestación y Gestión de los Bosques "Bosques Territorios de Vida", a nivel nacional, y la Mesa de Bosques de Santander, a nivel regional, con la meta de reducir a cero de la deforestación en el año 2030 (Ministerio de Ambiente y Desarrollo Sostenible, 2017). Como primer paso, se hizo un estudio de la deforestación y sus causas (Arenas et al., 2018), en el que se encontró que los principales motores de deforestación son la expansión de la frontera agropecuaria y de infraestructura, los cultivos ilícitos, los incendios forestales y la extracción de minerales y de madera.

En particular, uno de los principales motores de la deforestación es la extracción insostenible y desordenada de madera que, por escaso conocimiento y/o por tradición, se hace dentro de la ilegalidad y de una manera selectiva, poniendo en peligro algunas especies vulnerables. Este es el caso de maderas como la caoba, el comino crespo, el abarco, el cedro y el roble (Corporación Autónoma Regional de Santander - CAS, 2016; Fondo Mundial para la Naturaleza - WWF, 2017; Diario el Tiempo, 2015; Guerrero Rodríguez, 2017). La explotación y transporte de este tipo de maderas, usualmente se hace en conjunto con maderas ampliamente usadas comercialmente o más comunes como pino, móncoro, urapán, eucalipto y caracolí (Ministerio de Ambiente y Desarrollo Sostenible, 2017), razón por la que las autoridades tienen la necesidad de identificar el tipo de madera que se extrae o transporta, a pesar de no contar los recursos suficientes para hacerlo oportuna y eficazmente. Además, existen especies de madera que se encuentran en diferentes condiciones de vulnerabilidad y por lo tanto su identificación es más crítica.

A pesar de los esfuerzos por proteger los recursos naturales del país, las entidades colombianas aún luchan por combatir la tala ilegal de madera. De hecho, las tasas de deforestación en Colombia siguen siendo notablemente altas, especialmente en los últimos años. Existen campañas de las autoridades y corporaciones ambientales que buscan resolver el problema de la ilegalidad. En esta línea, un procedimiento común, llevado a cabo por la policía, consiste en detener un camión que transporta madera para pedirle al conductor un salvoconducto; luego, se verifica la carga de madera. Sin embargo, no se cuenta con instrumentos de monitoreo apropiados para detectar madera de especies de árboles vulnerables y prohibidas; y por ende, se requieren instrumentos de monitoreo que apoyen los procesos de vigilancia y control. En particular, instrumentos adecuados de monitoreo podrían apoyar el trabajo de la Policía Nacional en la detección de especies de madera vulnerables y prohibidas.

Para las autoridades es difícil actuar frente a la explotación ilegal de madera porque, entre otras razones, las personas que la transportan ilegalmente se valen de la falta de personal capacitado y de la carencia de procesos ágiles de identificación o detección de maderas. Ante esta dificultad, los incautamientos se prolongan y las acciones legales terminan por dilatarse. Por tal motivo, se requiere un método de identificación o detección de maderas ágil y eficiente, que permita realizar acciones de control y vigilancia forestal adecuadas. Este es uno de los programas estratégicos de la Misión Bosques Santander (Mesa de Bosques de Santander, 2018).

En ese orden de ideas, el Centro de Investigación Científica y Tecnológica en Tecnologías de la Información y las Comunicaciones (CentroTIC) de la Universidad Industrial de Santander, desarrolló un proyecto de investigación denominado *"Plataforma IoT para el desarrollo de servicios inteligentes de apoyo al monitoreo ambiental"*, que apoya las iniciativas de la Mesa de Bosques de Santander desde varios ángulos que involucran el uso de drones el procesamiento de imágenes y el uso de narices electrónicas. En este proyecto están vinculadas varias autoridades ambientales y corporaciones que persiguen el objetivo de reducir a cero la deforestación en Santander

(CentroTIC, 2016).

Por todo lo anterior, es importante el desarrollo de herramientas tecnológicas que permitan una rápida detección y que sirvan como método preliminar de identificación o detección. Todas estas necesidades son el origen de este trabajo, con la idea de utilizar narices electrónicas para la detección de especies maderables, pues estos dispositivos pueden ser una herramienta útil y de bajo costo que aporte a la solución de esta problemática.

Descripción del documento

En el Capítulo 2, se hace una recopilación de los antecedentes del Trabajo de investigación, el planteamiento del problema, una breve descripción organoléptica de la madera y un estado del arte de las narices electrónicas dentro del marco de la clasificación de especies maderables. En el Capítulo 3, se describe el sistema de olfato electrónico y las fases de recolección, preparación y toma de muestras utilizadas en el trabajo. Las mediciones de los compuestos volátiles se realizan utilizando una matriz de 16 sensores químicos, cuyas curvas de respuesta ingresan a un sistema basado en el reconocimiento de patrones para la identificación o detección. En el Capítulo 4, se presenta una estrategia de identificación de especies maderables, basada en narices electrónicas y análisis de componentes principales PCA. Se implementa un clasificador por vectores de soporte y se presenta el desempeño obtenido con diferentes subgrupos de parámetros seleccionados como predictores. En el Capítulo 5, se propone un método para la detección de especies de madera a partir de los aromas que emanan las maderas. En este caso, la selección de características se hace a través de tres estrategias: utilizando el Análisis de Componentes Principales (PCA), tomando toda los datos completos como una serie temporal, y aplicando LASSO (*Least Absolute Shrinkage and Selection Operator*, en inglés). La detección se lleva a cabo mediante una mezcla de modelos Gaussianos con el *Universal Background Model*. Finalmente, en el Capítulo 6, se hace un compendio de las conclusiones del trabajo y se hacen recomendaciones para trabajos futuros.

1. Objetivos

Objetivo general

Desarrollar un método de identificación de maderas aserradas propias de la zona de Santander mediante el uso de en narices electrónicas y aprendizaje estadístico.

Objetivos específicos

Generar una base de datos (*dataset*) con los registros de olor obtenidos de la nariz electrónica, para las especies de madera seleccionadas y de interés en Santander; teniendo en cuenta algunas de las condiciones de variabilidad entre muestras que se presentan en el proceso de extracción.

Realizar un estudio de separabilidad entre especies de madera, a partir del uso de técnicas de selección y extracción de características, como el análisis de componentes principales.

Seleccionar un método de reconocimiento de patrones apropiado para el procesamiento de los datos obtenidos y evaluarlo de acuerdo a las condiciones tenidas en cuenta en la generación de la base de datos.

Realizar una validación del desempeño de un primer prototipo IoT de nariz electrónica para para la identificación de maderas de interés en Santander.

2. Marco Teórico

Planteamiento del Problema

Entre las estrategias de identificación de madera, las características macroscópicas como el color, la textura y el olor se destacan porque se pueden usar para establecer rápidamente si una madera está correctamente etiquetada (Wheeler and Baas, 1998), lo que permite que la madera se analice rápidamente y en grandes volúmenes. El uso de estas características macroscópicas, por parte de personal capacitado, es la forma más común de identificación de la madera en Colombia, pero se hace de manera empírica y subjetiva.

También existen métodos precisos basados en análisis taxonómicos y genéticos, en los que se comparan muestras de especies de madera a nivel de secuencias genéticas (Hanssen et al., 2011; Yu et al., 2016). Aunque la confiabilidad de estas pruebas es de casi 100 %, son costosas, demoradas y deben ser realizadas por expertos localizados en centros urbanos alejados del lugar donde se realiza el análisis. Otras técnicas utilizadas involucran diferentes análisis espectroscópicos (Cabral et al., 2012; Rana et al., 2008; Zhao and Cao, 2016; Carballo-Meilán et al., 2016) y de imágenes (Dickson et al., 2017), que siguen requiriendo el apoyo de expertos y toman bastante tiempo. Estos métodos son técnicas efectivas pero aún no cumplen los requisitos para ser aplicados en regiones suburbanas y rurales alejadas de las principales ciudades (Kalaw and Sevilla, 2018). Otra forma de resolverlo es utilizando técnicas basadas en tratamiento de imágenes (FRIM and UTAR, 2018; Agritix, 2016); sin embargo, no todas las especies se pueden identificar de esta manera, especialmente si son de la misma familia (Cordeiro et al., 2016).

Soluciones alternativas proponen analizar los compuestos volátiles emitidos por las especies de madera mediante el uso de estrategias como la cromatografía de gases, que permite iden-

tificar compuestos volátiles específicos dentro de una mezcla (Rinne et al., 2002; Müller et al., 2006; Fedele et al., 2007). Una opción menos costosa y mucho más práctica es el uso de narices electrónicas, que tienen la particularidad de asociar una huella digital de olor a cada muestra, en lugar de identificar individualmente los componentes químicos presentes en la mezcla de volátiles (Kalaw and Sevilla, 2018; Wilson et al., 2005). En Colombia, el uso de narices electrónicas se ha enfocado en la evaluación de calidad de productos agrícolas (Rodríguez et al., 2010; Durán Acevedo et al., 2014) y en detección de explosivos y minas(Gómez Monsalve and Durán Acevedo, 2015). Sin embargo, poco o nada se ha hecho respecto a la identificación de especies maderables propias del territorio nacional usando narices electrónicas.

A nivel global se reporta el uso de narices electrónicas, con altos índices de desempeño, para la identificación de características de diferentes especies maderables a partir de los compuestos volátiles que estas emanan (Garneau et al., 2004; Hamilton et al., 2006; Wilson and Baietto, 2009; Baietto et al., 2010; Wilson, 2012; Baietto et al., 2013; Wilson, 2013; Cordeiro et al., 2016; Kalaw and Sevilla, 2018). Más allá de los buenos resultados reportados en los trabajos mencionados, normalmente, no se tienen en cuenta inconvenientes tales como: dificultad para distinguir un olor específico dentro de una mezcla de olores (Zhang et al., 2018); datos muy específicos en los que se incluyen muy pocas especies maderables, e.g. por pares de especies en el caso de (Cordeiro et al., 2016), o sin tener en cuenta la parte del tronco como albura y duramen (Garneau et al., 2004); y, variabilidad entre las características de muestras de la misma especie tomadas de regiones diferentes de procedencia (Wilson et al., 2005).

Otras posibles interferencias o problemas que se podrían presentar en una situación práctica no fueron tenidas en cuenta. Por ejemplo, en situaciones prácticas no es fácil establecer el origen de procedencia de las muestras de madera, ni el tiempo que ha transcurrido desde que fue cortada o tomada, ni las condiciones de almacenamiento (temperatura, humedad entre otras). En los trabajos mencionados, se tiene bien identificada la zona de origen de las muestras; y, en algunos de ellos

se sigue un riguroso protocolo de almacenamiento, que involucra estrategias como el mantener congeladas las muestras hasta el momento en el que se hace uso de ellas.

En este trabajo se busca estudiar un entorno menos controlado, con condiciones más cercanas a las del funcionamiento de un posible dispositivo final, y teniendo en cuenta una cantidad mayor de muestras, alejando el experimento de condiciones ideales. En particular, utilizamos muestras de madera aserrada en lugar de material recién cortado o sin cortar, después de los procedimientos de transporte. En el caso de la madera recién cortada, los aromas son muy frescos, fuertes y sin mayor interferencia; lo opuesto ocurre con la madera aserrada, seca y transportada. Se utilizaron muestras de madera provenientes de aserraderos en la región del "*Gran Santander*", en Colombia, lo que acercó la recolección de datos a situaciones prácticas. En este sentido, se plantea la siguiente pregunta a responder:

¿Cómo puede ser diseñada una nariz electrónica para que sea capaz de detectar la presencia de madera no autorizadas en un cargamento a partir de muestras de madera tomadas en condiciones similares a las del mismo cargamento?

Además, los los trabajos mencionados realizan tareas de identificación en lugar de verificación. La verificación permite autenticar una muestra de madera individual comparándola con una referencia biométrica específica almacenada en la base de datos, mientras que la identificación la compara con todas las métricas biológicas almacenadas en la base de datos. Hasta donde se sabe, este es el primer trabajo en el que el olor realiza el procedimiento de detección de especies de madera, desde un punto de vista de verificación biométrica.

La madera

La madera es un material orgánico compuesto por fibras de celulosa que resisten fuertes tensiones. Es un tejido estructural fibroso y poroso que se encuentra en los tallos y raíces de los árboles y otras plantas (Peña and Rojas, 2006). La madera ha sido utilizada durante siglos como combustible, para la construcción, como herramienta, para elaboración de muebles y papel, entre

otras aplicaciones. Al ser un producto orgánico, su estructura está sujeta a infinitas variaciones. Así mismo dos especímenes, aún perteneciendo a la misma especie, pueden tener diferentes propiedades. La madera, en general, está compuesta por celulosa, hemicelulosa, lignina y agua. Así mismo, carbono, oxígeno e hidrógeno son los principales elementos que la componen. También se encuentran otros elementos como nitrógeno, calcio, potasio, sodio, magnesio, hierro y manganeso. Adicionalmente la madera también contiene azufre, cloro, silicio, fósforo y otros elementos en pequeñas cantidades. Por supuesto, esta composición varía entre especies (Peña and Rojas, 2006).

Lo primero que se nos ocurre es diferenciar las maderas a simple vista, para esto podemos analizar el color o las vetas formadas por los anillos de crecimiento del tronco del árbol, o el brillo. Algunas especies tienen patrones característicos en las vetas, colores más claros o más oscuros, e incluso su brillo u opacidad puede ser un factor determinante para identificarlas. A diferencia de los componentes estructurales, la composición de los llamados extractos sí varía en amplios rangos entre las especies y depende de muchos factores como factores genéticos, condiciones de crecimiento, el clima y la geografía. Los extractos de madera se presentan por diferentes actividades, algunos de ellos se producen en respuesta a heridas, y algunos de ellos participan en la defensa natural contra insectos y hongos. Estos compuestos contribuyen a diversas propiedades físicas y químicas de la madera, como el color, la fragancia, durabilidad, propiedades acústicas, propiedades higroscópicas, adhesión y secado. De allí que se puedan clasificar diferentes especies, a partir de estas propiedades (SJÖSTRÖM, 1993).

Estado del arte

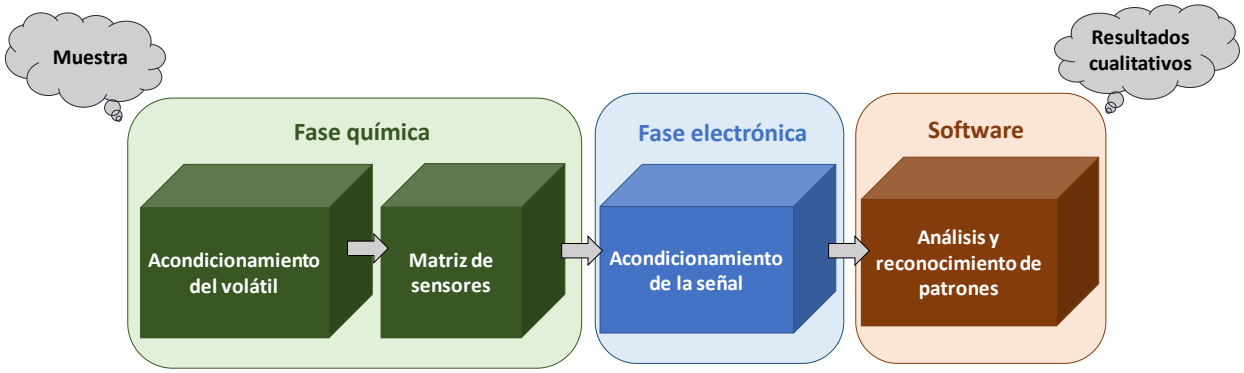
El olfato humano has sido poco atractivo para su emulación electrónica puesto que es mucho más complejo que otros sentidos, como el oído y la vista, debido al enorme número de receptores y neuronas que intervienen en el proceso y su conexión con la corteza cerebral (Gutiérrez and Horrillo, 2014). Sin embargo, y a pesar de su complejidad, esta configuración ha inspirado el desarrollo de un sistema que imita el olfato biológico: la nariz electrónica, el cual es un dispositivo

compuesto de un arreglo de sensores de gas acoplados a un sistema de reconocimiento de patrones con el que se procesa la información obtenida de los sensores (Moreno et al., 2009).

Una nariz electrónica, en general, está constituida por tres fases igual de importantes para su correcto funcionamiento (Figura 1). En la primera fase, se lleva a cabo un proceso químico que inicia con el acondicionamiento de la muestra odorífica y su paso hacia el arreglo de sensores de gas. En la segunda fase, se adquieren las señales eléctricas producidas por la reacción de los sensores mediante un procesamiento electrónico, y se obtiene la información que representa la muestra sensada. Finalmente, estos datos son procesados en la fase de reconocimiento de patrones, a partir del cual se detecta, clasifica e identifica la muestra (Ruiz Jiménez, 2018).

Figura 1

Funcionamiento general de la nariz electrónica.



Nota: Adaptado de Ruiz Jiménez (2018).

Los sistemas de olfato electrónico se han venido usando para un creciente número de aplicaciones. En la industria de alimentos, por ejemplo, se usan narices electrónicas para monitorear la calidad y nivel de maduración de las frutas (Shi et al., 2017). En monitoreo ambiental, entre otros usos, destacan el análisis de la calidad del aire, la calidad del agua, detección y control de polución (Capelli et al., 2014). Así mismo, las narices electrónicas son atractivas en la detección de

explosivos, narcóticos, sustancias peligrosas, perfumería, etc (Guo et al., 2017; Santos and Lozano, 2015).

También se ha extendido el uso de narices electrónicas orientadas hacia la detección, caracterización e identificación de maderas. Existen investigaciones en la detección temprana del decaimiento y pudrición presente en las raíces de algunas especies de árboles por acción de hongos (Baietto et al., 2010), la evaluación de la calidad en un tipo específico de madera (Najib et al., 2012) y la identificación de especies maderables a partir de los compuestos volátiles que estas emanan (Wilson, 2012; Cordeiro et al., 2016; Kalaw and Sevilla, 2018).

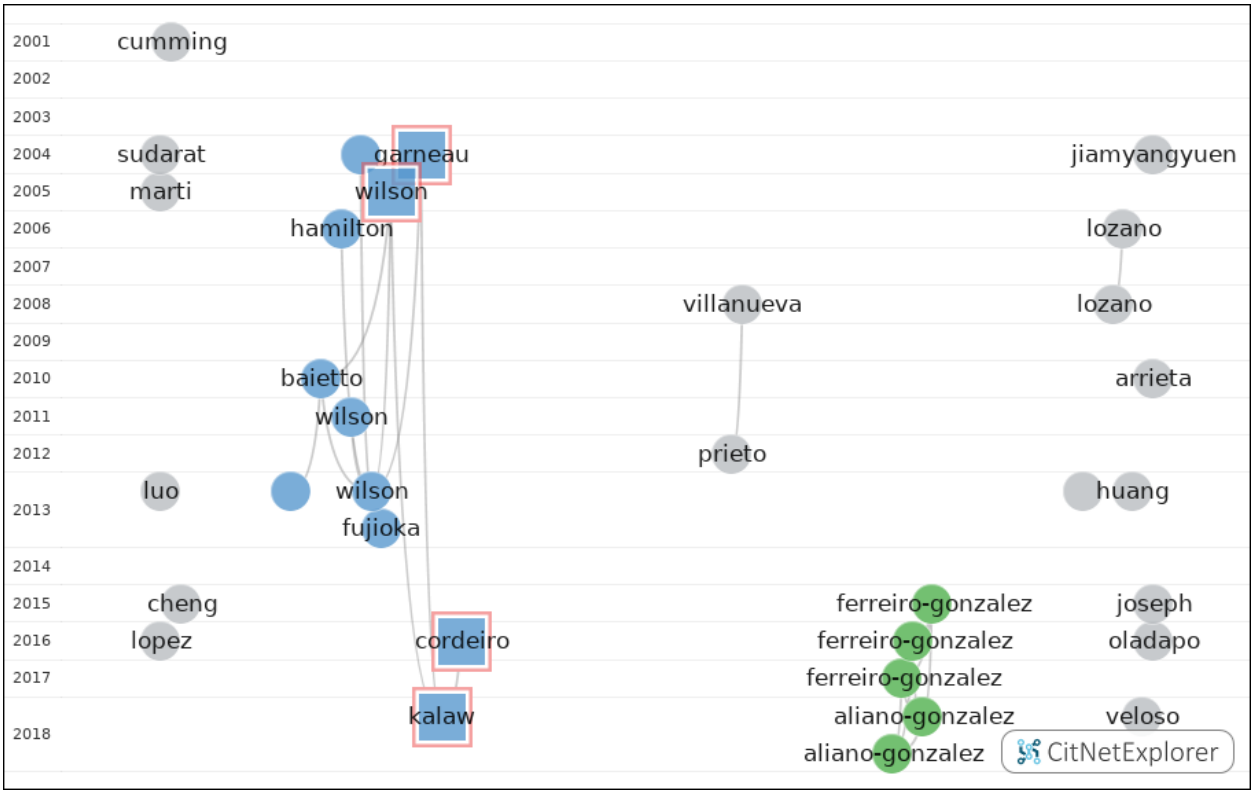
A pesar de los avances, aún quedan algunos inconvenientes específicos por resolver. Por ejemplo, existen dificultades en la selección y calibración de los sensores adecuados para una aplicación debido a su variabilidad (Vergara et al., 2012; Martinelli et al., 2014; Rodriguez-Lujan et al., 2014). Adicionalmente, como se expresa en Zhang et al. (2018), los sensores de gas no son realmente receptores olfativos, y por tanto las narices electrónicas podrían tener problemas al tratar de distinguir un olor específico dentro de una mezcla de olores. Por tanto, según el mismo autor (Zhang et al., 2018), es necesario desarrollar un nuevo tipo de sensores basados en receptores olfativos biológicos además de complejas técnicas de reconocimiento de patrones que se asemejen más al proceso que realiza el olfato humano.

Narices electrónicas y clasificación de madera. Con ayuda de la herramienta CitNetExplorer, se hizo un trabajo de búsqueda y filtrado bibliográfico que se resume en la Figura 2, donde se encuentran diferentes referencias de trabajos que involucran la nariz electrónica en el campo de las maderas, organizados por año y por citaciones. Allí se pueden identificar dos grupos (*clusters*) principales: el uso de la nariz electrónica para la detección de combustibles (incluida la madera), en color verde; y el uso de la nariz electrónica para la identificación de tipos de madera, en color azul. Además, sin agrupar estrictamente pero en líneas cercanas, se encuentran usos de la nariz electrónica para analizar estados de fermentación de vino y cervezas considerando la madera de la que

están hechos los barriles, y su comparación con técnicas como la cromatografía de gases. A partir de los resultados del filtrado de referencias, se realiza un estudio de los trabajos seleccionados para conocer los métodos de clasificación de madera que se utilizan en diferentes contextos.

Figura 2

Referencias sobre la identificación de madera mediante narices electrónicas



Nota: Resultado de la búsqueda, profundización y agrupamiento de referencias sobre la identificación de madera mediante narices electrónicas.

Las narices electrónicas se han empleado exitosamente en la clasificación de especies de madera, bajo condiciones específicas. En 2004, se utilizaron para discriminar entre tres especies diferentes de la familia de las pináceas (*Pinaceae*) a partir de su albura, que es la parte joven de la madera y se encuentra justo debajo de la corteza del árbol, y su duramen, la parte de mayor

edad que está formada por células biológicamente muertas y que es atractiva por ser la parte más resistente del tronco (Garneau et al., 2004). Los autores pudieron establecer diferencias entre estas especies para ambas partes de la madera, pero analizadas por separado.

En 2005, se investigó respecto a las diferencias entre especies de igual familia o género, a partir del uso de narices electrónicas (Wilson et al., 2005). Entre sus resultados y aportes se destacan las observaciones sobre la variabilidad entre las características de muestras de la misma especie, pero tomadas de regiones diferentes, por lo cual se deduce que la procedencia podría afectar la identificación. Esto permite inferir el nivel de importancia del diseño de experimentos la etapa de entrenamiento que debe estar en función de las muestras desconocidas que deberá analizar el sistema en última instancia.

En Brasil también se hizo un trabajo de clasificación de especies maderables con narices electrónicas. En esta oportunidad, existía un particular interés en cuatro especies maderables comúnmente explotadas en ese país y los experimentos buscaban la clasificación en dos escenarios, cada uno para distinguir entre dos pares de especies similares entre sí por su olor. Uno de estos escenarios buscaba clasificar dos especies del mismo género, y se logró con resultados satisfactorios (Cordeiro et al., 2016). Sin embargo, estos experimentos fueron realizados con muestras muy específicas, lo que no garantiza el funcionamiento bajo entornos ligeramente diferentes, tales como condiciones ambientales o geográficas de nuevas muestras.

Los autores Kalaw and Sevilla (2018) le dieron importancia a la rapidez y bajo costo que ofrecen las narices electrónicas y a las ventajas de los sensores de gas. El caso de aplicación es la clasificación de cinco especies maderables en peligro o importantes comercialmente en Filipinas, logrando encontrar grupos (*clusters*) separables a simple vista, mediante el análisis de componentes principales (PCA). No obstante, las muestras tomadas fueron recogidas en una zona específica,

haciéndolas muy poco diversas.

3. Recolección y Toma de Datos

3.1. Arreglo de sensores químicos de la nariz electrónica

Como se mostró anteriormente en la figura 1 un sistema de olfato electrónico, o nariz electrónica, está conformado por tres fases (Ruiz Jiménez, 2018): una fase química, que corresponde a la interacción de los compuestos volátiles con el arreglo de sensores de gas; una fase electrónica, donde se adquieren las señales eléctricas y se acondicionan para obtener una representación matricial temporal de la muestra; y finalmente, una fase de reconocimiento de patrones donde los datos son procesados y se clasifica o detecta el tipo de madera.

Con respecto a las numerosas narices electrónicas disponibles en la literatura, existen narices electrónicas comerciales, cuyos sensores consisten en polímeros orgánicos no conductores. Se han utilizado varios tipos, como la ***Cyranose 320*** (Garneau et al., 2004) y la ***Aromascan A32S*** (Wilson et al., 2005). Sin embargo, este tipo de dispositivos no satisfacen las necesidades de investigación, ya que tiene arreglos de sensores fijos, no modificables ni personalizables. Por otro lado, se han utilizado narices electrónicas personalizadas, por ejemplo: en (Kalaw and Sevilla, 2018), se utilizó una matriz de sensores químicos 8 con principios resistivos basados en nanotubos de carbono; y en (Cordeiro et al., 2016), se utilizó una matriz de 4 sensores de polímero conductivo con principio resistivo. Sin embargo, la fabricación de sensores está fuera del alcance del presente trabajo.

Existen diferentes tipos de sensores comerciales de gas, que varían en tamaño, sensibilidad, aplicación y tecnología utilizada. Los sensores más comunes son aquellos basados en películas semiconductoras de óxido metal, compuestas por cristales de óxido metal tipo *n*, como el dióxido de estaño (SnO_2). En dichos sensores, la sensibilidad ante diferentes gases cambia con la temperatura,

por lo que cuentan con un filamento que se calienta mediante corriente eléctrica con el fin de mantener una temperatura casi constante. Además, antes de ser usados por primera vez, los sensores deben pasar por una etapa de precalentamiento (*pre-heating time*) (Figaro Engineering Inc., 2018). A pesar de estos inconvenientes, estos sensores son preferidos porque presentan características estables a lo largo del tiempo y no demandan procesos de mantenimiento continuo (Ghasemi-Varnamkhasti et al., 2019).

Para este trabajo se utilizó un prototipo de nariz electrónica de laboratorio, que se muestra en la Figura 3; desarrollado en la *Universidad Industrial de Santander (www.uis.edu.co)* bajo la cultura *DIY (Do it yourself, hágalo usted mismo)* Ruiz Jiménez (2018), y que se puede consultar en linea en el catálogo bibliográfico que dispone la universidad (http://tangara.uis.edu.co/biblioweb). Esto permite hacer investigaciones a diferentes escalas y a bajo costo.

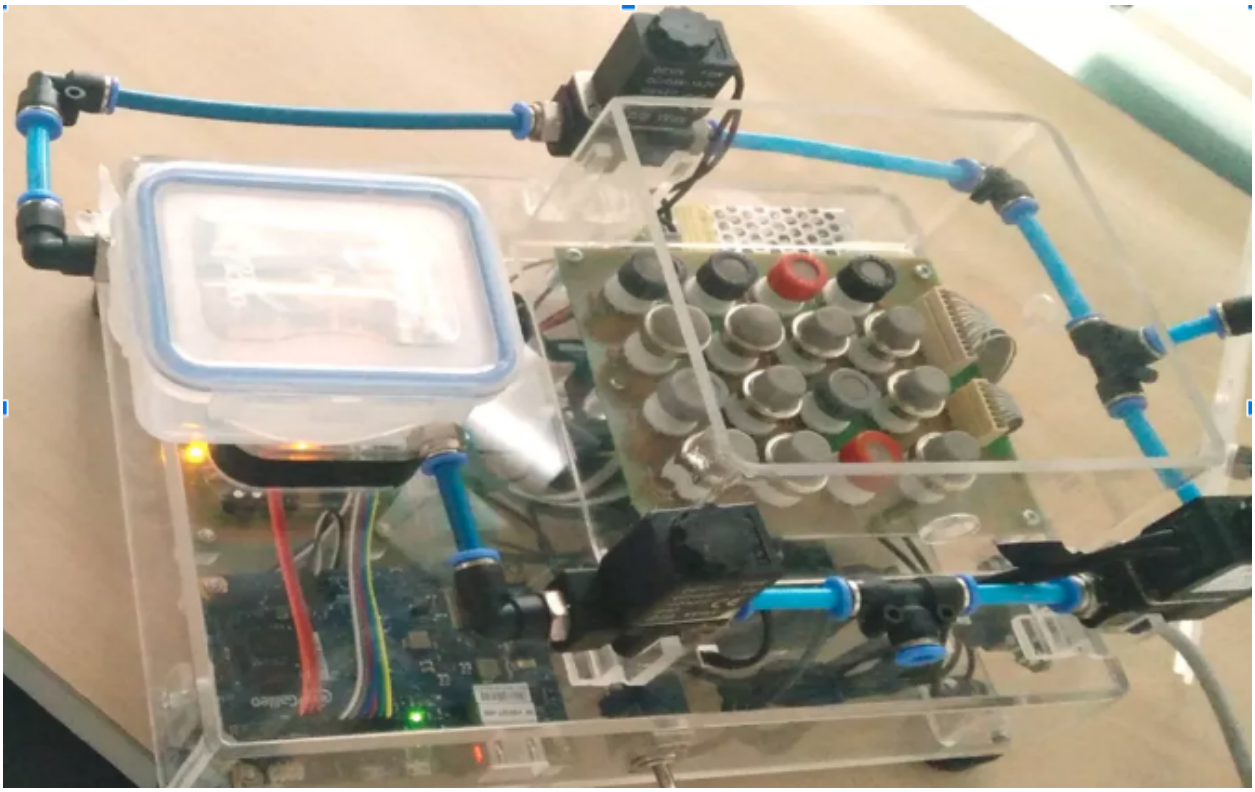
El prototipo está compuesto por una tarjeta de adquisición *Intel Galileo Generation 1* Ruiz Jiménez (2018), para la adquisición y acondicionamiento de las señales de cada sensor (Módulo de adquisición de datos, DAQ). Además, cuenta con una matriz de $4 \times 4 = 16$ sensores semiconductores de óxido-metal para la detección de gases, los cuales varían su resistencia eléctrica debido a la reacción química que ocurre cuando los compuestos volátiles hacen contacto con los sensores. Estos sensores pertenecen a las casas fabricantes *Figaro Engineering* y *Hanwei Electronics*, que se caracterizan por su capacidad para detectar bajas concentraciones de gas y pos su bajo costo. Así mismo, para su conexión, se empleó el mismo circuito sugerido por los fabricantes para el acondicionamiento de la señal (Figaro Engineering Inc., 2018).

El módulo de sensado incluye además una cámara de concentración, una fase móvil, tuberías y electroválvulas que se encargan de acumular los volátiles y transportarlos hasta el módulo de adquisición de datos (DAQ). Sin embargo, del módulo de sensado, solo se utilizan los sensores para tener condiciones cercanas a la realidad.

En la Tabla 1 se listan los sensores utilizados en el prototipo. Las aplicaciones de la mayo-

Figura 3

Prototipo de nariz electrónica desarrollado en la UIS.



Nota: Prototipo de nariz electrónica desarrollado en la UIS Ruiz Jiménez (2018).

ría de estos sensores están orientadas hacia la medición de la calidad del aire, detección de algún gas en particular debido a su toxicidad u otra característica de interés, y, para la detección de hidrocarburos (Ruiz Jiménez, 2018). Por ejemplo, el sensor $MQ-135$, de la marca Hanwei, y el sensor $TGS813$, de la marca Figaro, son utilizados para la detección de gases en general, en aplicaciones relacionadas con la calidad del aire. Su sensibilidad incluye principalmente compuestos como el metano, el hidrógeno, el dióxido de carbono (CO_2) y el monóxido de carbono (CO).

De otra parte, los sensores $MQ4$ $MQ7$, $MQ9$ (Hanwei), $TGS832$ y $TGS826$ (Figaro) son

Tabla 1

Sensores del prototipo de nariz electrónica

SENSOR	MARCA	REF	SENSOR	MARCA	REF
1	HANWEI	MQ-2	9	FIGARO	TGS-832
2	HANWEI	MQ-3	10	HANWEI	MQ-6
3	HANWEI	MQ-4	11	FIGARO	TGS-823
4	HANWEI	MQ-6	12	FIGARO	TGS-816
5	HANWEI	MQ-7	13	FIGARO	TGS-822
6	HANWEI	MQ-8	14	FIGARO	TGS-813
7	HANWEI	MQ-135	15	FIGARO	TGS-826
8	HANWEI	MQ-9	16	HANWEI	MQ-3

Nota: Adaptado de Ruiz Jiménez (2018).

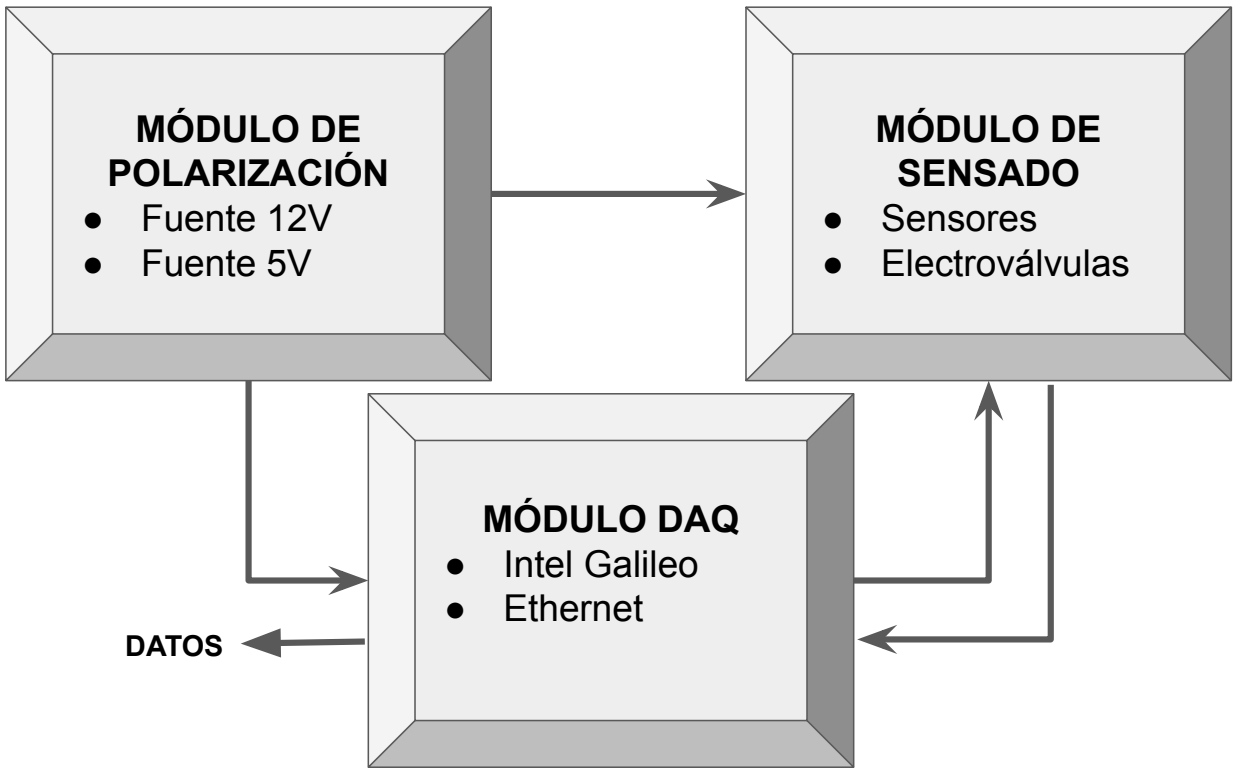
comúnmente utilizados para la detección de un gas o un grupo de gases de interés, como el amoníaco, el CO_2 y gases refrigerantes. Una gama adicional de sensores es utilizada para la detección de alcoholes. Entre ellos se destacan las referencias $MQ3$ y $MQ8$ (Hanwei), $TGS816$ y $TGS823$ (Figaro). Finalmente, en la industria de los hidrocarburos, es importante la detección de gases como metano, butano, propano y gas licuado de petróleo (*LPG* por sus siglas en inglés, *Liquefied petroleum gas*). Para ello, es común el uso de sensores como $MQ2$, $MQ5$ y $MQ6$ (Hanwei), que son especialmente sensibles a estos compuestos.

En el diagrama de la Figura 4, se muestra el esquema de funcionamiento del dispositivo utilizado, descrito en Ruiz Jiménez (2018). El módulo de polarización es el encargado de suministrar la energía a todo el dispositivo; cuenta con dos fuentes: una de 12V que alimenta las electroválvulas y una de 5V que alimenta los demás elementos. El módulo de adquisición de datos funciona como cerebro del dispositivo y es el encargado de controlar el uso de las electroválvulas, además de almacenar los datos de los sensores (módulo de sensado).

En el Apéndice 1, se encuentra información referente al uso de este prototipo de nariz

Figura 4

Diagrama de funcionamiento del prototipo de nariz electrónica desarrollado en la UIS.



Nota: Diagrama de funcionamiento del prototipo de nariz electrónica desarrollado en la UIS (Ruiz Jiménez, 2018).

electrónica como dispositivo IoT.

3.2. Procedimiento de Recolección de Datos

Se tomaron 309 muestras de variados tipos de maderas en diferentes depósitos de madera localizados en poblaciones de la región del Gran Santander−Colombia: Bucaramanga, Lebrija, Socorro, San Gil, Pamplona y Cúcuta. En la Tabla 2 se encuentra un resumen con las muestras obtenidas por cada especie. Para algunas especies, el número de muestras recolectado fue bastante bajo, por lo tanto se organizan en una clase denominada *"otras especies"*.

Cada muestra corresponde a un pequeño bloque de madera que, idealmente, debería tener

Tabla 2

Número de muestras recolectadas por cada clase.

	Nombre científico de la especie	**Número de Bloques**
Cedro	*Cedrela odorata*	84
Móncoro	*Cordia gerascanthus*	47
Pino	*Retrophyllum rospigliosii*	27
Sapán	*Clathrotropis brunnea*	43
Otras	*Tabebuia aurea*, *Zanthoxylum rhoifolium*, *Fraxinus uhdei*, *Anacardium excelsum*, *Simarouba amara*, *Cariniana pyriformis*, *Ficus spp.*, *Quercus humboldtii*, *Guarea guidonia*, *Coffea arabica*, *Alchornea triplinervia*, *Corymbia citriodora*, *Swietenia macrophylla*, y otras desconocidas.	108

unas dimensiones de $10cm \times 10cm \times 10cm$. No obstante, debido a la dificultad de encontrar muestras con tamaños iguales, se flexibilizó esta exigencia y se recolectaron muestras de diferentes tamaños, disponibles en las diferentes carpinterías y depósitos de madera visitados. En las imágenes de la Figura 5, se muestran algunos de los bloques de madera recolectadas y, posteriormente, utilizados como muestras de las respectivas especies.

Para compensar las posibles implicaciones de la diferencia en el tamaño de los bloques de madera, y con el objetivo de estandarizar el procedimiento de la toma de muestras, se optó por utilizar viruta en lugar de los bloques de madera. Para ello, se utilizó un cepillo de madera (tipo garlopa) No. 4 (Figura 6), fácil de adquirir en una ferretería o tienda de materiales para la construcción.

En el desarrollo del experimento para la medición se tienen en cuenta dos tareas previas: primero, como etapa de preparación del dispositivo, se enciende la nariz electrónica durante una hora para que los sensores alcancen su operación de estado estable en el ambiente correspondiente;

Figura 5

Muestras de Madera.

Nota: Bloques de madera utilizados como muestra. A la izquierda, una muestra de pino y, a la derecha, algunas muestras de móncoro.

luego, como etapa previa a la toma de la muestra, se prepara cada muestra (bloque de madera) cepillándolo 20 veces con un cepillo para madera y el material resultante es desechado. Esto con el fin de eliminar posible contaminación por contacto con otras muestras o posibles interferencias con otros elementos.

Después, se realiza el experimento en sí, con un ensayo por cada muestra de madera. En cada ensayo se sigue el procedimiento descrito a continuación:

- Cepillar la muestra 20 veces más y tomar aproximadamente 1 cm^3 de la viruta resultante (Figura 7). El objetivo de cepillar la madera y tomar la viruta es realzar temporalmente la intensidad de los compuestos volátiles, sin recurrir a procedimientos sofisticados. El cepillado de madera puede ser realizado fácilmente por cualquier persona, sin la presencia de un experto.

- Olfatear la muestra (viruta del paso anterior) con la nariz electrónica (Figura 8). El resultado de esta toma de datos es un conjunto de 16 curvas de respuesta, que corresponden a las

Figura 6

Cepillo de madera utilizado para la preparación de las muestras.

variaciones de conductancia relacionadas con cada uno de los 16 sensores y que pueden ser vistas como la huella de olor de la muestra de madera.

- Entre cada ensayo los sensores se dejan reposar un tiempo de, al menos, 5 minutos, permitiendo el ingreso de flujo de aire generado por un ventilador. Al cabo de ese tiempo la nariz electrónica ha regresado a su estado estable y, de esta manera, se busca evitar interferencias de una muestra de un ensayo anterior sobre el ensayo actual.

En cada ensayo, los datos fueron tomados a un periodo de muestreo de 270 *ms*, predefinido en el prototipo usado. Cada curva de respuesta de cada uno de los 16 sensores está formada por 500 muestras divididas en tres fases: lectura base, muestra y recuperación (Figura 9). En la primera fase los sensores reaccionan al aire durante 100 muestras; luego, la viruta de madera correspondiente al respectivo ensayo se pone frente a los sensores durante 300 muestras. Finalmente, la viruta de madera se retira y los sensores se enfrentan solo al aire, se almacenan 100 muestras de esta última fase y no toda la etapa de recuperación del sensor, ya que esta etapa no es considerada para el análisis. Los datos resultantes de este proceso reposan en GitHub: https://github.com/Narenman/WoodSmell. Adicionalmente, con el objetivo de reducir el efecto ruido electrónico en el sistema de adquisición, se realiza un preprocesamiento de los datos que

Figura 7

Viruta resultado de cepillar un bloque de madera.

consiste en el uso de filtro de mediana de orden 5 para cada una de las curvas de respuesta de los sensores.

3.3. Extracción de características

Diferentes características pueden ser estimadas a partir de la curva de respuesta del valor de conductancia de cada uno de los sensores. En particular, en trabajos anteriores se reporta el uso de características relacionados con los valores máximo, mínimo y área bajo la curva de respuesta de cada sensor (Yan et al., 2015). Otra forma es el uso de estrategias que involucran un análisis

Figura 8

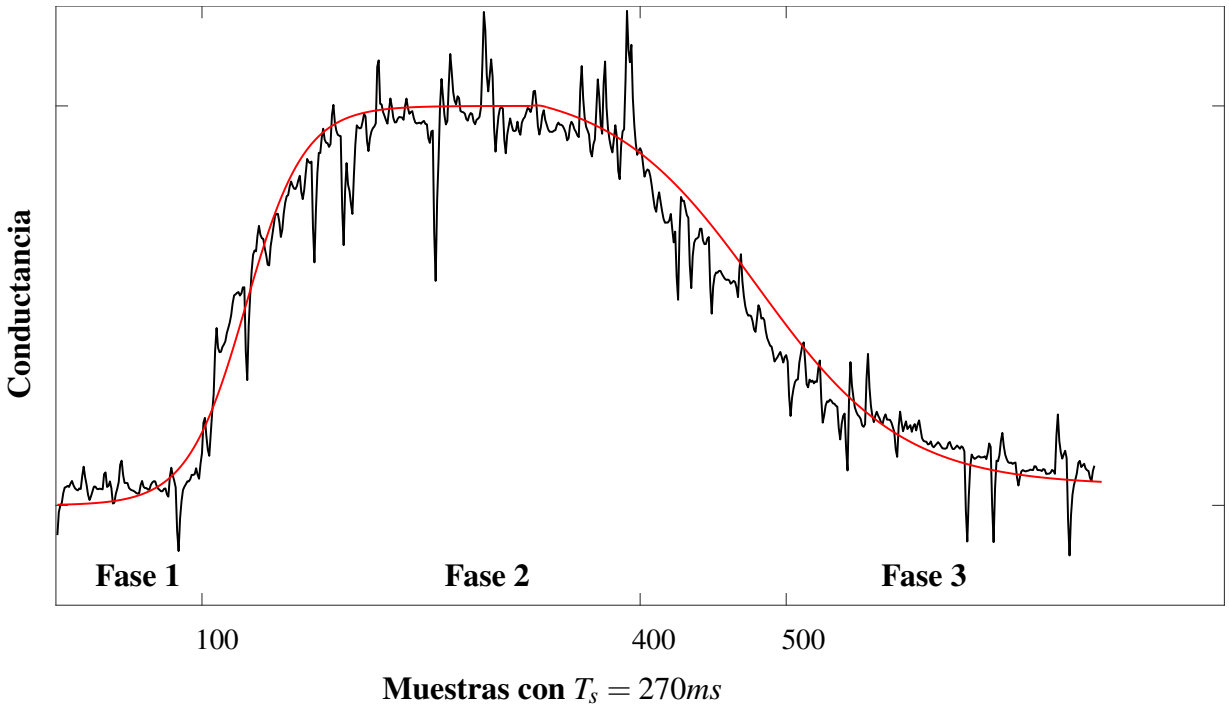
Procedimiento de toma de muestra con la nariz electrónica.

de la respuesta transitoria de los sensores (Rodriguez-Lujan et al., 2014; Rana et al., 2008). Y finalmente existe una tercera forma donde se realiza un ajuste a modelos predefinidos (Carmel et al., 2003). Estas estrategias permiten representar las secuencias de datos de los sensores mediante una cantidad reducida de descriptores, lo cual es deseable en problemas como el nuestro en los que se tienen secuencias de datos medidos de dimensión notablemente más grande que la cantidad de muestras de maderas.

En el presente trabajo, se utilizaron dos estrategias para la etapa de extracción de carac-

Figura 9

Forma de respuesta típica de los sensores del arreglo.



Nota: Forma de respuesta típica de los sensores del arreglo (El eje ordenado corresponde a las muestras tomadas con un periodo de 270*ms*). Se aprecian las diferentes etapas de la respuesta. Fase 1, primeras 100 muestras, donde se hace la lectura base o de referencia. Fase 2, muestras 101 a 400, donde los sensores capturan el olor de la muestra de virutas de madera. Fase 3, muestras 401 a 500, donde los sensores empiezan a retornar a su estado de referencia. En color negro se observa un ejemplo real de respuesta obtenida y en color rojo la curva suavizada esperada después de aplicar el filtro.

terísticas. En la primera, se calculan algunos de los parámetros heurísticos reportados en trabajos recientes, que describen el comportamiento de las curvas de respuesta. En la segunda, para el procedimiento de verificación, se extraen los datos sobre la serie temporal completa.

Para este trabajo se estiman las siguientes características:

- G_0, valor de conductancia inicial, media de las primeras 100 muestras de la respuesta total.
- G_f, valor de conductancia final, media de las últimas 50 muestras de la fase 2 de la respuesta

total.

- G_{max}, valor de conductancia máxima.
- G_{min}, valor de conductancia mínima.
- B, coeficiente de ganancia y A, localización del polos provenientes del ajuste a un modelo auto-regresivo de primer orden (de acuerdo a las respuestas observadas en general para los sensores):

$$H(z) = \frac{Bz^{-1}}{1 + Az^{-1}}. \tag{1}$$

Las primeras 4 características mencionadas, corresponden a parámetros extraídos directamente de cada curva de respuesta. Este tipo de características es uno de los más comunes en los trabajos consultados. Las otras dos características, que corresponden al ajuste a un modelo regresivo, son una manera de intentar representar todo el comportamiento de cada curva de respuesta (esto incluye características transitorias y de estado estable). En estricto orden, las primeras 16 características corresponden al valor G_0 para los 16 sensores; las siguientes corresponden a los 16 valores de G_f, los 16 valores de G_{max}, los valores de G_{min}, los valores de A y, finalmente, los valores de B.

En resumen, se extraen 6 valores por cada una de las 16 curvas de una muestra, con lo que se obtiene un total de 96 características por cada firma o huella odorífica, es decir, una matriz $\mathbf{X}_{309\times96}$ (309 ensayos de dimensión 96). Este conjunto de datos se utiliza para los experimentos de verificación. Para el experimento de identificación, se utilizan muestras de madera provenientes de aquellas especies más comunes. Las especies seleccionadas son: cedro (84 muestras), móncoro (47 muestras), pino (27 muestras) y sapán (43 muestras), para un total 201 muestras, conformando una matriz $\mathbf{X}_{201\times96}$ (201 ensayos de dimensión 96). Estas matrices aumentan de tamaño (número de filas) cuando se aplica una técnica de aumento de datos.

3.3.1. Aumento de Datos. Las dificultades para la recolección de muestras generan dos grandes problemas: el desbalance de las clases y el tamaño reducido del set de datos. El problema de las clases no balanceadas se puede abordar de diferentes maneras, por ejemplo: generando datos sintéticos, haciendo un sobremuestreo de la clase minoritaria o un submuestreo de la clase mayoritaria, o haciendo un ajuste sobre la función de costo para darle una mayor penalidad a la clasificación incorrecta de instancias de la clase minoritaria (Van Hulse et al., 2007; Cieslak et al., 2006; Lusa et al., 2010). SMOTE (Chawla et al., 2002) es una técnica de sobremuestreo que crea muestras sintéticas de la clase minoritaria. Esta técnica permite solventar simultáneamente las dos dificultades mencionadas anteriormente.

SMOTE (*Synthetic Minority Oversampling TEchnique*) resuelve el problema al sintetizar nuevas instancias de la clase minoritaria, entre (en medio de) las existentes (Chawla et al., 2002). Estas nuevas instancias se localizan sobre lineas dibujadas imaginariamente entre las instancias existentes. Para ello, se necesita definir el número de instancias (k) que se toman en cuenta para generar un dato sintético y el número de datos sintéticos generados por cada dato real.

Para generar un dato sintético, se parte de un dato real y sus k vecinos más cercanos. Se traza una linea (de forma imaginaria) desde el dato real hasta cada uno de sus vecinos y, sobre estas lineas, se escoge aleatoriamente un punto que será el dato sintético. Este procedimiento se realiza para cada dato real y se repite cuantas veces sea necesario hasta obtener el número de instancias sintéticas deseado.

La técnica *SMOTE* para aumento de datos se utiliza para lidiar con los problemas de la poca cantidad de datos y el desbalance de las clases. En particular, para los experimentos de identificación, se tomó como referencia la clase mayoritaria (cedro, 84 muestras) y se aumentó el set de datos hasta completar el doble de muestras (168) para las clases con mayor cantidad de muestras (cedro, móncoro, pino, sapán), es decir, un conjunto final de datos de tamaño 672 (4 clases). Con esto, el problema se plantea en torno a una matriz de características de tamaño $\mathbf{X}_{672\times 96}$. Para los

experimentos de verificación, el aumento se hizo hasta completar 216 muestras para las 4 clases mencionadas, y otras 216 para el resto de muestras del *"resto del universo"*; con esto se llega a una matriz de características de tamaño $\mathbf{X}_{1080\times 96}$.

3.3.2. Serie Temporal completa. Para los experimentos experimentos de verificación, también se consideró a cada uno de los $N = 16$ sensores como una característica única, formando un vector de características de longitud N. En otras palabras, cada lectura de cada sensor se convierte en una característica. Además, se incluyen las primeras $s = 400$ muestras de cada curva de respuesta, es decir, 400 lecturas de cada sensor correspondientes a las dos primeras etapas de la curva de respuesta de la Figura 9. Al usar las primeras s muestras, no es necesario reducir la dimensión, porque estamos usando casi toda la información para estimar el modelo GMM, es decir, 400 vectores de longitud de 16 por bloque de madera.

Con esta estrategia, se logra que el set de datos ya no sea una matriz con 309 filas, sino una matriz de $400 \times 309 = 123600$ filas (tamaño 123600×16), que contiene la gran mayoría de la información. En este caso, no se aplica el aumento de datos con *SMOTE*.

4. Identificación de Maderas

4.1. Selección de Características: Análisis de Componentes Principales

Existen diferentes alternativas para la selección de características en problemas de reconocimiento de patrones. Tres de los métodos más utilizados para ello son: selección de un subconjunto, regularización y reducción de dimensión (James et al., 2013). El primer enfoque busca identificar un subconjunto de $q < p$ predictores que se considera están relacionados con la respuesta y así ajustar modelos con variadas configuraciones (subconjuntos reducidos) de variables. De otra parte, la regularización implica ajustar un modelo con todos los p predictores, pero imponiendo restricciones a fin de obtener modelos más conservadores desde el punto de vista de las

entradas. Como consecuencia algunos pesos asociados a las características toman valores cercanos a cero ó llegan a ser cero, reduciendo así la varianza del modelo. Finalmente, la reducción de dimensionalidad implica proyectar los p predictores en un subespacio M-dimensional, donde $M < p$, calculando M diferentes combinaciones, o proyecciones, de las variables. Luego, estas proyecciones M se usan como predictores para ajustar un modelo.

En trabajos reportados previamente, relacionados con clasificación de olores, típicamente se utiliza análisis de componentes principales (PCA, *Principal Component Analysis*) con el fin de reducir la dimensión del problema y así evitar sobreajuste (Akbar et al., 2016; Cordeiro et al., 2016). PCA es una técnica de aprendizaje no supervisado (es decir, que solo utiliza las características de una observación y no la respuesta asociada) que permite reducir la dimensión de observaciones $x_i \in \mathbb{R}^p$ mediante una transformación lineal $\mathbf{V}_q$ que mapea los datos a un nuevo espacio de dimensión $q \leq p$, donde las nuevas variables son no correlacionadas linealmente, al tiempo que conserva la mayoría de la información de los datos originales. Las dimensiones usadas para representar la información son aquellos que resultan tener mayor variabilidad en el nuevo espacio transformado. En particular, la dirección del primer *componente principal* de los datos es aquella a lo largo de la cual las observaciones varían más (mayor varianza) (James et al., 2013).

PCA realiza una transformación lineal y ortogonal de los datos que proyecta una entrada X a una representación Y. Esta nueva representación, que reconstruye los datos de una manera eficiente (con un bajo error cuadrático medio), puede ser de una dimensión menor que la de los datos originales. Por lo tanto, PCA sirve como un método de reducción de dimensionalidad simple y efectivo, conservando la mayor cantidad de información posible (Goodfellow et al., 2016). A continuación se presenta la relación entre la representación de PCA y la representación de datos original X.

Se tiene la matriz de datos $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ que se desea analizar mediante una transformación de la forma $y_i = f(x_i)$. Además, $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$. Se escoge una transformación

$f(\cdot)$ de manera que $cov(\mathbf{Y})$ sea diagonal, indicando que cada nueva variable $y^{(1)}, y^{(2)}, \ldots y^{(q)}$ que forman el vector y_i está no correlacionada con las demás variables.

Se propone reconstruir x_i, a partir de su valor transformado y_i, mediante la transformación de la forma $g(y_i) = \mathbf{V}_q y_i$. $\mathbf{V}_q$ es una matriz de transformación con las restricciones de que las columnas de $\mathbf{V}_q$ sean ortogonales entre sí y que la norma de estas sea unitaria. Por tanto, $x_i \sim g(f(x_i)) = \mathbf{V}_q \mathbf{V}_q^\intercal x_i$. Para hallar $\mathbf{V}_q$ se plantea resolver el problema de minimizar la distancia entre los datos originales y los datos reconstruidos de la forma, (Friedman et al., 2001; Goodfellow et al., 2016)

$$\min_{\mathbf{V}_q} \quad \sum_{i=1}^{N} \|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^\intercal (x_i - \bar{x})\| \tag{2}$$

donde $\bar{x} = E\{x_i\}$. En la práctica es común remover $\bar{x}$ de los datos originales antes de obtener la estimación $\hat{\mathbf{V}}_q$.

Al resolver el problema antes mencionado se encuentra que $\hat{\mathbf{V}}_q$ se construye a partir de los autovectores de $\mathbf{X}^\intercal \mathbf{X}$ correspondientes a los autovalores de mayor magnitud, es decir, aquellos componentes con mayor varianza. Los componentes se ordenan por la cantidad de varianza original que describen, de mayor a menor, concentrando la mayor cantidad de la varianza original en los primeros componentes. De esta manera, tomando unos pocos componentes principales es posible representar los datos originales. Para el caso de los experimentos llevados a cabo en este trabajo, se procedió a usar 90 % como el nivel de varianza mínimo para escoger los p componentes principales con los que se representarán los datos originales.

4.2. Resultados Aplicando Análisis de Componentes Principales (PCA)

Los experimentos de clasificación se llevaron a cabo para las 4 diferentes configuraciones de características de entrada mostradas en la Tabla 3, teniendo en cuenta el set de datos aumentado ($X_{672\times 96}$) que está organizado como en se muestra en la Tabla 4. Los parámetros correspondientes a cada una de las configuraciones, relacionados en la Tabla 3, son estimados para cada sensor.

En particular, para el caso de la configuración 1 se estima un total de $16 \times 6 = 96$ características. Para la segunda configuración, se consideran solo las características extraídas directamente de las curvas, y se estiman $16 \times 4 = 64$ predictores. Para la configuración 3, por el contrario, solo se considera los parámetros de ajuste al modelo auto-regresivo, y se estiman $16 \times 2 = 32$ predictores. Finalmente, en la última configuración, se consideran solo los valores de conductancia máximo y final para cada sensor, y se estiman $16 \times 2 = 32$ predictores. En resumen, se obtienen 4 matrices de características con los siguientes tamaños: $\mathbf{X}_{672\times96}$, para la configuración 1, $\mathbf{X}_{672\times64}$ para la configuración 2, $\mathbf{X}_{672\times32}$ para la configuración 3, y $\mathbf{X}_{672\times32}$ para la configuración 4.

Tabla 3

Combinaciones de diferentes características extraídas

Configuración	Características utilizadas
1	G_0, G_f, G_{max}, G_{min}, A y B
2	G_0, G_f, G_{max} y G_{min}
3	A y B
4	G_f y G_{max}

Tabla 4

Número de muestras por cada clase antes y después del aumento de datos para el experimento de identificación.

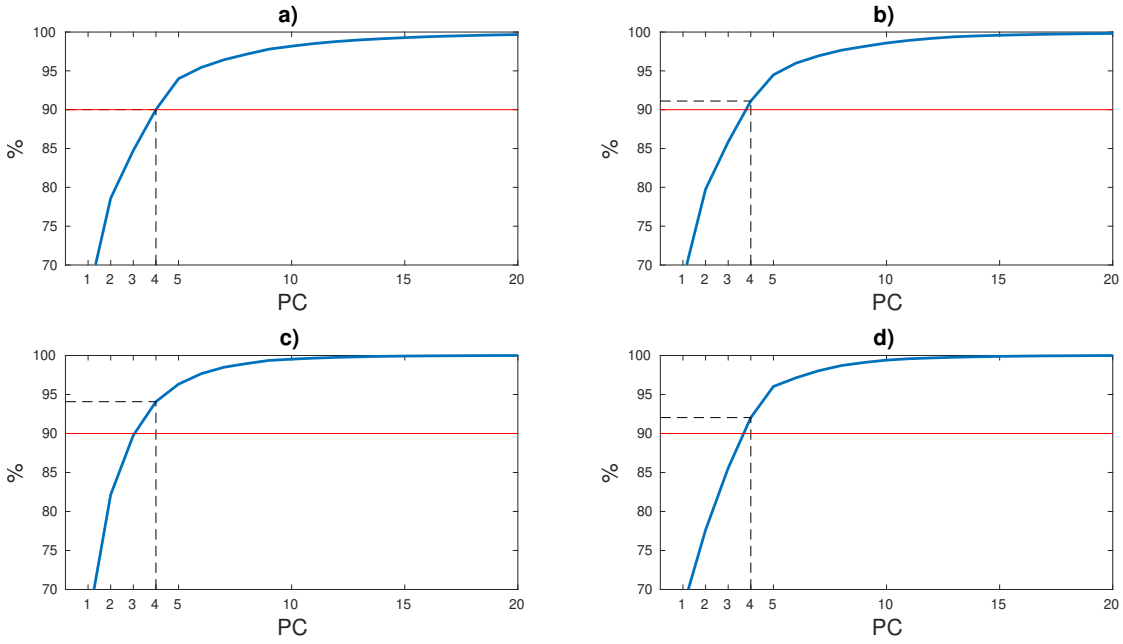
		Número de Bloques	
	Nombre científico de la especie	**Reales**	**SMOTE**
Cedro	*Cedrela odorata*	84	168
Móncoro	*Cordia gerascanthus*	47	168
Pino	*Retrophyllum rospigliosii*	27	168
Sapán	*Clathrotropis brunnea*	43	168

Luego, para la matriz de características de cada configuración se aplicó el análisis de com-

ponentes principales (PCA), y los componentes resultantes se ordenaron según su varianza. En la Figura 10, se muestra la varianza acumulada de los primeros 20 componentes. Del total de componentes se seleccionaron aquellos que, con su varianza acumulada, puedan representar al menos un 90 % de la información. Este número corresponde a los 4 primeros componentes principales, que son los que finalmente ingresan al clasificador por SVM, a modo de características de entrada. De otra parte, con tres componentes principales ya se tiene más del 80 % de la varianza para todas las configuraciones, tal como lo muestra la Figura 10, y también se hace la prueba con 3 componentes principales.

Figura 10

Varianza acumulada en los 20 primeros componentes de PCA.



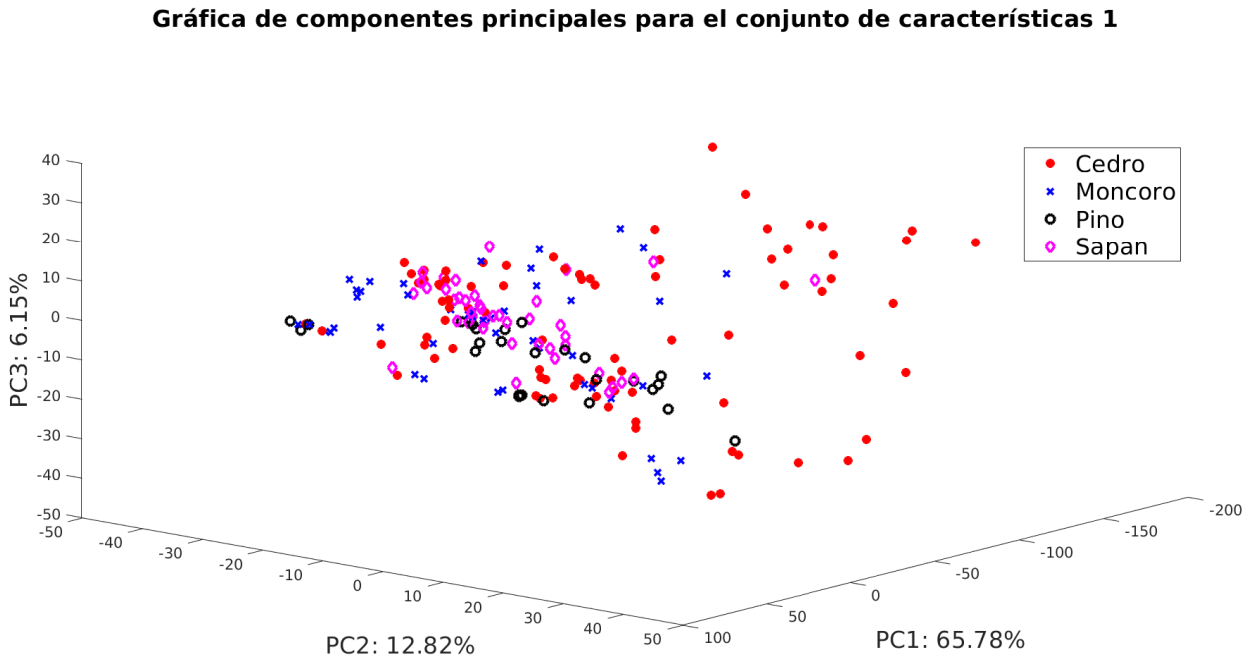
Nota: Varianza acumulada en los 20 primeros componentes principales para la configuración 1 (gráfica **a)**), para la configuración 2 (gráfica **b)**), para la configuración 3 (gráfica **c)**) y para la configuración 4 (gráfica **d)**).

En la Figura 11, se presenta la distribución de los datos utilizando tres componentes principales, resultantes de la configuración 1, para las cuatro clases analizadas en el experimento de

identificación: en color rojo las muestras de cedro, en color azul las muestras de móncoro, en color negro las de pino, y en magenta las muestras de sapán. En la mencionada figura, a primera vista no se observa separabilidad entre las clases analizadas. Se tienen observaciones similares para las restantes 3 configuraciones.

Figura 11

Visualización en 3D de los datos después de aplicar PCA, para la configuración 1.

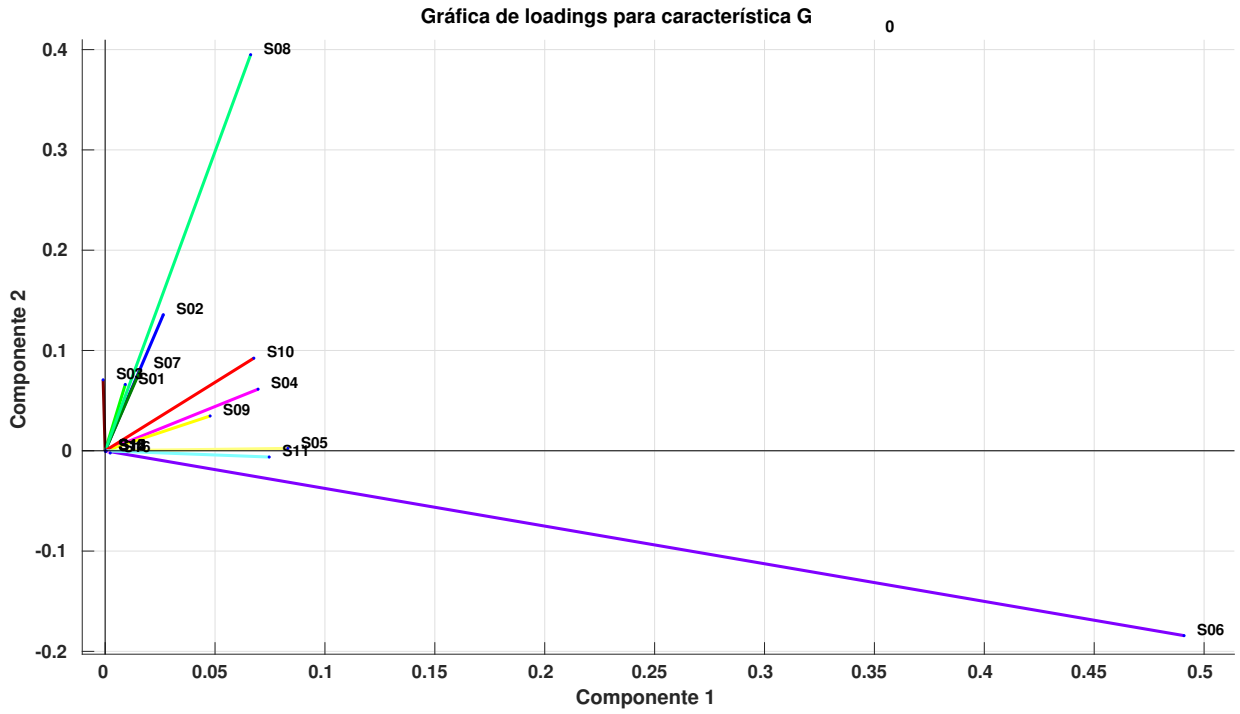


Así mismo, en la Figura 12 se presenta un gráfico de *loadings* para visualizar cuáles sensores son más relevantes para la aplicación de interés. La gráfica de la Figura 12 hace referencia a la primera característica (Conductancia inicial), para la que se observa una predominancia de los sensores 6 y 8. También se realiza un análisis similar para las otras 5 características, observando que los sensores 3, 9 y 11 destacan sobre los demás analizando la conductancia final. Para la conductancia máxima los sensores 5 y 11 son más relevantes; en conductancia mínima los sensores 6 y 8; respecto a la ganancia (A) los sensores 2, 8 y 10; y para la ubicación del polo (B) los sensores

3 y 5 presentan más relevancia.

Figura 12

Gráfica de Loadings para la característica G_0 (conductancia inicial) en los 16 *sensores.*



4.3. Clasificación por máquinas vectores de soporte

Las técnicas descritas anteriormente preparan el conjunto de datos para la etapa de clasificación. Entre los diferentes algoritmos de aprendizaje automático que existen, la clasificación por máquinas vectores de soporte (SVM, *Support Vector Machines*) es uno de los métodos más populares para resolver problemas de clasificación y es ampliamente usado en los problemas de clasificación asociados a las narices electrónicas. La técnica de máquinas vectores de soporte entrega fronteras de tipo no-lineal entre clases o categorías, a partir de construir fronteras lineales, pero en una versión transformada y de mayor dimensión del espacio de características originales y_i (James et al., 2013).

El clasificador de vectores de soporte más sencillo permite la clasificación en una configuración de dos clases con límite lineal. Sin embargo, en la práctica, a veces nos enfrentamos a límites de clase no lineales. Este problema se puede analizar utilizando un grupo más amplio de características, obtenidas a partir de operaciones como funciones polinómicas (cuadráticas, cúbicas e incluso de orden superior) de los predictores originales (James et al., 2013). Esto es, por ejemplo, cambiar el conjunto de p características, $Y_1, Y_2 \dots Y_p$, por un conjunto de $2p$ características, $Y_1, Y_1^2, Y_2, Y_2^2, \dots, Y_p, Y_p^2$.

En el espacio de características extendido, el límite de decisión sigue siendo lineal. Sin embargo, en el espacio de características original, el límite de decisión es de la forma $Q(y) = 0$, donde Q es un polinomio (en este ejemplo, un polinomio cuadrático), y sus soluciones son generalmente no lineales. En particular, se definen los N pares de datos dados por $\{(y_1, z_1), (y_2, z_2), \dots, (y_N, z_N)\}$ donde $y_i \in \mathbb{R}^q$ corresponde a la i-ésima observación en el nuevo espacio de características extendido, de dimensión q, y $z_i \in \{-1, 1\}$ simboliza las etiquetas de las categorías. Se define el hiperplano de clasificación dado por,

$$\{y : \delta(y) = y^\intercal \beta + \beta_0 = 0\} \tag{3}$$

donde β es un vector de magnitud unitaria: $\|\beta\| = 1$. La regla de clasificación para $\delta(y)$ estaría dada por $G(y) = \text{sgn}(y^\intercal \beta + \beta_0)$. Con el fin de obtener esta frontera de decisión $\delta(\cdot)$ se plantea la solución del siguiente problema (Friedman et al., 2001),

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & z_i(y_i^\intercal \beta + \beta_0) \geq 1 - \xi_i, \\ & \xi \geq 0, \end{aligned} \tag{4}$$

donde C es una constante de costo arbitraria. ξ_i, en la restricción $z_i(y_i^\intercal \beta + \beta_0) \geq 1 - \xi_i$, corresponde

a la proporción para la cual la restricción $\delta(y_i) = y_i^\intercal \beta + \beta_0$ ubica a y en el lado contrario de su margen, es decir, en el lado incorrecto. Si se obtiene que $\xi_i > 1$, entonces y_i corresponde a un dato mal clasificado. Al minimizar $\sum_{i=1}^{N} \xi_i$ se minimiza la cantidad total de datos de entrenamiento mal clasificados.

Si buscamos fronteras de tipo no lineal entre clases, en lugar de utilizar y_i como entrada, usamos $h(y_i) = (h_1(y_i), h_2(y_i), \ldots, h_M(y_i))$ para $i = 1, \ldots, N$, que representan una transformación no lineal de las características de entrada. Con ello se produce la función no-lineal de separación $\hat{\delta}(y) = h(y)^\intercal \hat{\beta} + \hat{\beta}_0$, y la función de decisión estaría dada por $\hat{G}(y) = \text{sgn}(\hat{\delta}(y))$. Esta técnica crea hiperplanos que maximizan la separabilidad entre conjunto de datos, a través de funciones kernel $h(\cdot)$.

En particular, en este trabajo se usa a modo de función kernel la función Gaussiana, una función kernel utilizada universalmente. Un kernel gaussiano, generalmente, garantiza la obtención de un predictor óptimo globalmente, minimizando los errores de estimación de un clasificador. Para configurar este clasificador en una aplicación de varias clases (4 clases en este caso), se enfrentan dos clases entre sí y se realiza un entrenamiento progresivo hasta que se logra un ajuste óptimo. Finalmente, se realiza un procedimiento de validación cruzada con k iteraciones (k-folds), separando el conjunto de datos en k subconjuntos iguales, de los cuales $k-1$ se utilizan para entrenar el clasificador y el subconjunto restante para estimar el error de predicción.

4.4. Resultados de la Clasificación por Máquinas Vectores de Soporte

Los experimentos de clasificación se realizaron utilizando los primeros 3 y los primeros 4 componentes principales, extraídos de las matrices de datos correspondientes a cada configuración, después de aplicar la técnica de aumento de datos. La evaluación del modelo se realizó mediante el método de validación cruzada ($k-$fold) donde $k = 10$ y, este proceso se realizó con 100 diferentes ordenaciones aleatorias del conjunto de datos. Los resultados se pueden observar en la Tabla 5 y en la Tabla 6. El error estándar asociado a la estimación del error promedio, se calcula como

$SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n},}$ donde σ es la desviación estándar y n es la cantidad de valores usados para calcular el promedio (James et al., 2013). En este caso, $n = 1000$ debido a que se obtienen 10 valores 100 veces.

Tabla 5

Error de identificación con 3 *componentes principales.*

Set de características	# PC	Error Promedio	Error estándar
Configuración 1	3	27.22 %	0.163 %
Configuración 2	3	25.62 %	0.159 %
Configuración 3	3	26.79 %	0.163 %
Configuración 4	3	24.66 %	0.150 %

Nota: Tasas de error resultado del proceso de validación cruzada ($K-fold$ con $K = 10$) para identificación con diferentes tipos de características de entrada. $PC = 3$.

Tabla 6

Error de identificación con 4 *componentes principales.*

Set de características	# PC	Error Promedio	Error Estándar
Configuración 1	4	23.14 %	0.158 %
Configuración 2	4	21.59 %	0.155 %
Configuración 3	4	22.49 %	0.160 %
Configuración 4	4	20.24 %	0.146 %

Nota: Tasas de error resultado del proceso de validación cruzada ($K-fold$ con $K = 10$) para indentificación con diferentes tipos de características de entrada con el set de datos aumentado. $PC = 4$.

4.5. Discusión de Resultados: Experimento de Identificación

Se aplicó la técnica de análisis de componentes principales a un conjunto de datos aumentado a partir del la información obtenida del olor de cuatro tipos de maderas diferentes, y se observó que con 4 componentes se puede representar un poco más del 90 % de la varianza total de los datos

originales. Sin embargo, esto no es suficiente para obtener una representación con clases visiblemente separables y que facilite la clasificación de los tipos maderas del presente trabajo a partir de arreglos de sensores químicos. En contraste, esta estrategia si ha resultado ser eficiente en otros trabajos (Ruiz Jiménez, 2018; Shi et al., 2017; Capelli et al., 2014; Guo et al., 2017; Santos and Lozano, 2015), aunque para diferentes aplicaciones y en entornos diferentes de toma de datos. En particular, la precisión reportada en el presente trabajo es inferior a la de trabajos previos (Kalaw and Sevilla, 2018; Cordeiro et al., 2016; Wilson, 2012). Sin embargo, las condiciones del presente experimento son más cercanas de las condiciones prácticas. Esto se aprecia en la forma de recolectar los datos, cantidad de muestras, variedad de lugares de procedencia, especies, tratamiento previo y almacenamiento previo de las muestras.

A modo de ejemplo, en Garneau et al. (2004) se realizaron 60 mediciones provenientes de 18 muestras de madera. En Wilson et al. (2005) se tomaron 10 muestras por especie para 23 especies diferentes, lo cual corresponde a un numero reducido por especie. En Kalaw and Sevilla (2018) las muestras fueron recogidas en un lugar muy específico (el campus de la *University of Santo Tomas* en Manila-Filipinas), generando datos con poca variabilidad. En Cordeiro et al. (2016) una muestra es repetida durante varios ciclos de muestreo, generando un conjunto de datos mayor, pero con poca variabilidad.

Respecto al proceso de almacenamiento, trabajos anteriores también han utilizado procedimientos alejados de la práctica. En Garneau et al. (2004); Wilson et al. (2005); Kalaw and Sevilla (2018) las muestras fueron almacenadas herméticamente y congeladas hasta el momento y lugar del experimento; y, en Cordeiro et al. (2016), las muestras fueron almacenadas herméticamente en un laboratorio pero no congeladas. En contraste, para los experimentos realizados en el presente trabajo se buscó trabajar con una cantidad de datos mayor, con muestras de madera que no estaban recién aserradas y en condiciones de almacenamiento no rigurosas ya que decidimos dirigirnos directamente a los depósitos de madera. Con ello se buscó establecer un entorno de trabajo más

cercano al entorno real para el cual se busca resolver el problema.

5. Verificación de Especies Maderables

5.1. Selección de Características mediante LASSO

Otro de los enfoques utilizados para la selección de características es la regularización. La regularización implica ajustar un modelo con todos los p predictores, pero agregando restricciones con el fin de obtener modelos menos complejos; y, en el caso de LASSO, hacerlo menos complejo haciendo que algunos coeficientes asociados a sus respectivas entradas tiendan a cero. Algunos métodos de regularización no eliminan ningún coeficiente, por lo que el modelo final sigue utilizando los p predictores. Otros, como *LASSO*, sí permiten hacer que algunos coeficientes sean exactamente cero (eliminando los predictores correspondientes) (James et al., 2013).

LASSO (*Least Absolute Shrinkage and Selection Operator*) es un método de regularización que permite estimar el conjunto de coeficientes β_i relacionados con los predictores X_i de un modelo de regresión lineal de la forma

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{5}$$

donde Y es la respuesta, y $X_1, X_2, \ldots, X_p$ el conjunto de predictores (James et al., 2013). Tradicionalmente, los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ se estiman mediante un ajuste de mínimos cuadrados, usando los valores que minimizan la suma

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2. \tag{6}$$

LASSO identifica las variables independientes que afectan significativamente la variable

de respuesta agregando un término de regularización con norma $\mathbb{L}^1$ de la forma $\|\mathbf{B}\|_1 = \sum_i \mid \beta_i \mid$ a la regresión (Friedman et al., 2001). Los coeficientes de LASSO minimizan la cantidad

$$\begin{aligned} \hat{\beta}^{lasso} = \min_{\beta} \quad & \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \\ \text{s.t.} \quad & \sum_{j=1}^{p} |\beta_j| \leq t, \end{aligned} \tag{7}$$

donde $t \geq 0 \in \mathbb{R}$.

Reescribiendo el problema de minimización de LASSO la forma equivalente Lagrangiana (Friedman et al., 2001), se tiene

$$\hat{\beta}^{lasso} = \min_{\beta} \quad \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{8}$$

donde $\lambda \geq 0$ es un parámetro de ajuste. LASSO es similar a mínimos cuadrados, pero con una modificación en la función que se minimiza para estimar los coeficientes (James et al., 2013). Para un valor de λ suficientemente pequeño, el término de regularización tiende a cero y el estimador se convierte en “Mínimos Cuadrados ordinarios"; sin embargo, para un valor de λ suficientemente grande, los estimadores β_i son los que tienden a cero (Friedman et al., 2001).

5.2. Resultados de la Selección de Características usando LASSO

Para este análisis, se tuvieron en cuenta los mismos 6 parámetros heurísticos estimados en la fase de extracción de características. De esta manera, se sigue trabajando con 96 características por ensayo, con el set de datos aumentado para el experimento de verificación: $\mathbf{X}_{1080\times96}$. Las primeras 16 características $(1-16)$ corresponden al valor G_0 para los 16 sensores; las siguientes corresponden a los 16 valores de G_f $(17-32)$, los 16 valores de G_{max} $(33-48)$, los valores de G_{min0} $(49-64)$, los valores de A $(65-72)$ y, finalmente, los valores de B $(73-96)$.

En la Tabla 7, se describe el número de individuos por clase, antes y después del aumento de datos. El set aumentado se divide en cinco subgrupos (de 216 muestras cada uno) conformados por 43 (o 44) muestras de cedro y 173 (o 172) muestras de las otras especies. Cuatro de esos grupos se utilizan para la selección de características con *LASSO* y la fase de entrenamiento de la verificación. El grupo restante se utiliza para validar el modelo obtenido en la fase de entrenamiento, basado en las características seleccionadas con *LASSO*. Esto se hace cinco veces, para que cada subgrupo sea usado como conjunto de validación (validación cruzada, *k-fold*, con $k = 5$). Este proceso se realizó 25 veces y, además, se realizaron experimentos similares con las otras tres especies con mayor cantidad de muestras: móncoro, pino y sapán.

Tabla 7

Número de muestras por clase para la verificación.

	Nombre científico de la especie	Número de Bloques	
		Reales	**SMOTE**
Cedro	*Cedrela odorata*	84	216
Móncoro	*Cordia gerascanthus*	47	216
Pino	*Retrophyllum rospigliosii*	27	216
Sapán	*Clathrotropis brunnea*	43	216
Otras	*Tabebuia aurea*, *Zanthoxylum rhoifolium*, *Fraxinus uhdei*, *Anacardium excelsum*, *Simarouba amara*, *Cariniana pyriformis*, *Ficus spp.*, *Quercus humboldtii*, *Guarea guidonia*, *Coffea arabica*, *Alchornea triplinervia*, *Corymbia citriodora*, *Swietenia macrophylla*, and others unknown.	108	216

Nota: Número de muestras por cada clase antes y después del aumento de datos para el experimento de verificación.

Cabe resaltar que algunas de las características son recurrentes en los resultados de cada ensayo y, en consecuencia, son consideradas como las que representan mayor relevancia para la

clasificación. En la Tabla 8, se muestran las características seleccionadas con mayor frecuencia en cada uno de los experimentos de verificación. Además, una observación importante es que las características que representan mayor relevancia, de acuerdo a los resultados de *LASSO*, son aquellas relacionadas con el modelo auto-regresivo (A y B, correspondientes a las características 65 a 96).

Tabla 8

Características seleccionadas con LASSO.

Experimento	Características Seleccionadas
Detección de cedro	1, 2, 8, 10, 13, 18, 21, 43, 46, 58, 63, 66, 67, 68, 71, 74, 75, 78, 79, 80, 81, 86, 88, 89, 90, 92, 93, 94, 95, 96
Detección de móncoro	5, 8, 13, 22, 27, 28, 29, 30, 33, 34, 38, 41, 43, 44, 47, 50, 60, 64, 65, 67, 71, 72, 75, 76, 77, 78, 80, 81, 82, 85, 91, 93, 96
Detección de pino	1, 7, 24, 27, 28, 29, 30, 45, 58, 67, 68, 71, 72, 73, 74, 75, 77, 80, 81, 82, 83, 85, 86, 88, 89, 90, 91, 92, 93
Detección de sapán	1, 2, 3, 11, 15, 25, 41, 45, 46, 59, 60, 62, 65, 66, 71, 74, 76, 78, 79, 80, 81, 83, 87, 88, 89, 90, 91, 93, 94, 95

De los resultados anteriores, se puede inferir que las referencias de los sensores que aportan más información para la clasificación de maderas son:

- Sensores 2 y 16: HANWEI $MQ-3$, que es especialmente sensible a los alcoholes.
- Sensor 8: HANWEI $MQ-8$, utilizado principalmente para la detección de fugas de gas, por su sensibilidad al gas hidrógeno (H_2).
- Sensor 10: HANWEI $MQ-6$, que se utiliza principalmente en la detección de fugas de gas, por su alta sensibilidad a gases licuados de petroleo, propano y butano.
- Sensor 11: FIGARO $TGS-823$, especialmente sensible a vapores de etanol.

- Sensor 12: FIGARO $TGS-816$, sensor de propósito general con sisibilidad a un amplio rango de gases combustibles como metano, propano y butano.
- Sensor 13: FIGARO $TGS-822$, altamente sensible a los vapores de solventes orgánicos, así como a otros vapores volátiles. También tiene sensibilidad a gases combustibles como el monóxido (CO) de carbono, por lo que es un buen sensor de uso general.
- Sensor 14: FIGARO $TGS-813$, altamente sensible al propano, metano y butano, lo que lo hace muy utilizado en aplicaciones relacionadas con detección de fugas de gases.

5.3. Verificación de especies de madera con el enfoque GMM-UBM

Como la idea principal es apoyar a las autoridades en su lucha contra la tala ilegal y selectiva de especies de madera, los algoritmos de clasificación pueden fallar porque eligen dentro de un grupo cerrado de posibilidades. Por otro lado, la verificación de especies puede ser un mejor enfoque desde el punto de vista práctico. Los procedimientos de verificación permiten determinar si una muestra pertenece o no a una clase en particular porque consideran un grupo abierto de posibilidades.

Un *Universal Background Model* (UBM) es un concepto tomado de la biometría que, en este caso, corresponde a un modelo de características que representan, independientemente de la especie de madera, el universo o las condiciones de evaluación general esperadas. Este modelo se compara con un modelo de características específicas de una especie de madera, para tomar una decisión de aceptación (hipótesis nula) o rechazo (hipótesis alternativa). La tarea de verificación se puede resumir al probar si una muestra corresponde a la clase analizada (hipótesis nula) o a una clase desconocida (hipótesis alternativa). En este caso, la hipótesis del impostor (cualquier otra clase) está modelada por el *Universal Background Model* (UBM) (Reynolds et al., 2000; Doddington et al., 2000).

El elemento UBM es básicamente una función de densidad de probabilidad (PDF, *Probability Density Function*) que representa las propiedades de la huella odorífica de la población de especies de referencia. En ese sentido, la huella odorífica *dudosa* (la que se quiere analizar) se compara con respecto al UBM, así como a un modelo PDF de una especie de madera en particular. En ese caso, hay dos modelos: el modelo de una especie de madera (λ_s) y el modelo de referencia UBM (λ_0). Al analizar las observaciones correspondientes a la señal interceptada $\mathscr{X}$, se obtienen dos valores de probabilidad, $p(\mathscr{X}|\lambda_s)$ y $p(\mathscr{X}|\lambda_0)$, con los que se construye el Radio de Verosimilitud (LR, por sus siglas en inglés, *Likelihood Ratio*). Sin embargo, es común usar el *Log Likelihood Ratio* (LLR),

$$\mathscr{L}(\mathscr{X}) = \log p(\mathscr{X}|\lambda_s) - \log p(\mathscr{X}|\lambda_0). \tag{9}$$

A medida que aumenta el valor $\mathscr{L}(\mathscr{X})$, la evidencia de que la huella odorífica dudosa corresponde a la especie que estamos buscando se vuelve más fuerte.

Para el modelado de la PDF, se prefiere el conocido modelo de mezclas gaussianas (GMM, *Gaussian Mixture Models*). El uso de una mezcla de modelos gaussianos está motivado por su capacidad para modelar densidades arbitrarias (Kinnunen and Li, 2010; Reynolds and Rose, 1995). Una GMM se compone de una mezcla finita de componentes gaussianos multivariados y el conjunto de parámetros indicados por λ. Se caracteriza por una combinación lineal ponderada de densidades gaussianas unimodales de C mediante la función:

$$p(o|\lambda) = \sum_{i=1}^{C} \alpha_i \mathscr{N}(o, \mu_i, \sigma_i), \tag{10}$$

donde o es una observación o vector de características de dimensión D, α_i es el peso se mezclado (probabilidad anterior) del i-ésimo componente Gaussiano, y $\mathscr{N}(\cdot)$ es la función de densidad Gaussiana D-variada con vector media μ_i y matriz de covarianza σ_i. El popular algoritmo de *Ex-*

pectation Maximization (EM) se utiliza para maximizar la probabilidad con respecto a un dato dado. El lector interesado se refiere al Apéndice 2 para más detalles.

Como la identificación de maderas es un problema cerrado (hay que escoger entre alguna de N clases posibles conocidas), este enfoque parece insuficiente para la detección de especies de madera. La detección de especies de madera debería abordarse como un problema abierto, es decir, determinando si la muestra corresponde a una especie de interés particular, o, si es alguna otra de identidad desconocida. Para ello se utilizó el enfoque GMM-UBM. En este trabajo, las tareas de verificación se realizan de la siguiente forma: una clase (por ejemplo cedro) vs el modelo universal; los individuos que pertenecen al resto de especies, pero no las mismas muestras, se incluyen en el modelo UBM. Lo mismo sucede con las especies: móncoro, pino y sapán, obteniendo así un experimento de verificación para cada especie.

5.4. Resultados de la Verificación de Especies de Madera con el Enfoque GMM-UBM

Para medir el rendimiento del sistema de detección de especies de madera propuesto, se utilizan las curvas DET (*Detection Error Trade-off*, en inglés) y el valor EER (*Equal Error Rate*, en inglés). Las curvas DET trazan la tasa de falso rechazo (FRR, *False Rejection Rate*) en el eje Y, versus la tasa de falsa aceptación (FAR, *False Acceptance Rate*) en el eje X, donde la curva más cercana a la esquina inferior izquierda del gráfico corresponde al sistema que tiene el mejor actuación. La tasa de falso rechazo (FRR) representa la tasa de muestras que serán rechazadas perteneciendo a la clase de interés (también conocidos como falsos negativos); mientras que, la tasa de falsa aceptación representa las muestras que serán aceptadas aunque, en realidad, o pertenecen a la clase conocida (también llamados falsos positivos) (Martin et al., 1997).

Los experimentos de verificación incluyeron un set de datos aumentado, conformado por 1080 muestras de al menos 18 especies de madera, distribuidos como se muestra en la Tabla 7. Primero, se agruparon 173 muestras de cedro en una clase, y 691 muestras de las otras especies se usaron para adaptar el UBM con una mezcla 4 modelos Gaussianos. El resto de las muestras se

usaron para la validación: 43 muestras de cedro y 173 impostores. Este experimento se denomina ***Detección de cedro***. Se realizaron experimentos similares con las otras tres especies con mayor cantidad de muestras: móncoro, pino y sapán.

Para todos los experimentos, la PDF (*Probability Density Function*, en inglés) de cada clase, que representa una especie de madera particular, está modelada por una mezcla de 4 modelos gaussianos. El valor EER se estima utilizando un procedimiento de validación cruzada de 5 conjuntos; donde el 80 % de las muestras de madera corresponden al conjunto de entrenamiento, mientras que las muestras restantes (20 %) se utilizan para la validación. Cada problema de verificación (detección de cedro, detección de móncoro, detección de pino y detección de sapán) se analizó desde dos puntos de vista: primero, extrayendo características tradicionales (los seis parámetros heurísticos ya mencionados: G_0, G_f, G_{max}, G_{min}, A y B) y aplicando PCA; y segundo, considerando la serie completa de tiempo correspondiente a cada una de las 16 curvas de respuesta.

Además de los experimentos mencionados, bajo las mismas condiciones, se realiza uno adicional en el que se consideran los mismos seis parámetros heurísticos, pero se aplica *LASSO* como estrategia para la selección de características. En este caso, los experimentos de verificación se aplican a partir de las características extraídas con *LASSO*.

5.4.1. Detección utilizando PCA. En los primeros experimentos, se aplicó un procedimiento de extracción de características, y se calcularon seis valores a partir de las 16 curvas. En total, se tienen 96 características para cada muestra, es decir, una matriz de tamaño $\mathbf{X}_{1080\times 96}$. En trabajos previamente reportados, el análisis de componentes principales (PCA) se usa típicamente para reducir la dimensión y evitar un sobre-ajuste (Akbar et al., 2016; Cordeiro et al., 2016). En el caso de esta aplicación, se utilizaron cuatro componentes principales para representar aproximadamente el 90 % de la varianza en los datos originales. Por otro lado, las muestras de validación se comparan con un modelo único de la clase conocida (cedro) y el UBM. De esta forma, se verifica si la muestra analizada pertenece o no a la clase de cedro. Este procedimiento se repite 100 veces,

con diferentes ordenaciones aleatorias del conjunto de datos. La verificación con la clase cedro, aplicando PCA, mostró una tasa de error de clasificación de 37.40 %. En la Figura 13, se muestra una gráfica de la curva DET.

Al igual que con el cedro, experimentos de verificación (utilizando análisis de componentes principales) se realizaron para otras tres especies: móncoro, pino y sapán. La verificación con móncoro, aplicando PCA, mostró una tasa de error de clasificación de 53.64 % y, en la Figura 14, se muestra una gráfica de la curva DET. Para la verificación con pino, el error se estimó en 40.67 % y la curva DET se muestra en la Figura 15. Finalmente, para la verificación de sapán, el error fue de 32.77 % y, en la Figura 16, se muestra una gráfica de la curva DET. En la Tabla 9 se muestra un resumen de los resultados obtenidos usando PCA.

Tabla 9

Valores de EER obtenidos en los experimentos de verificación con PCA.

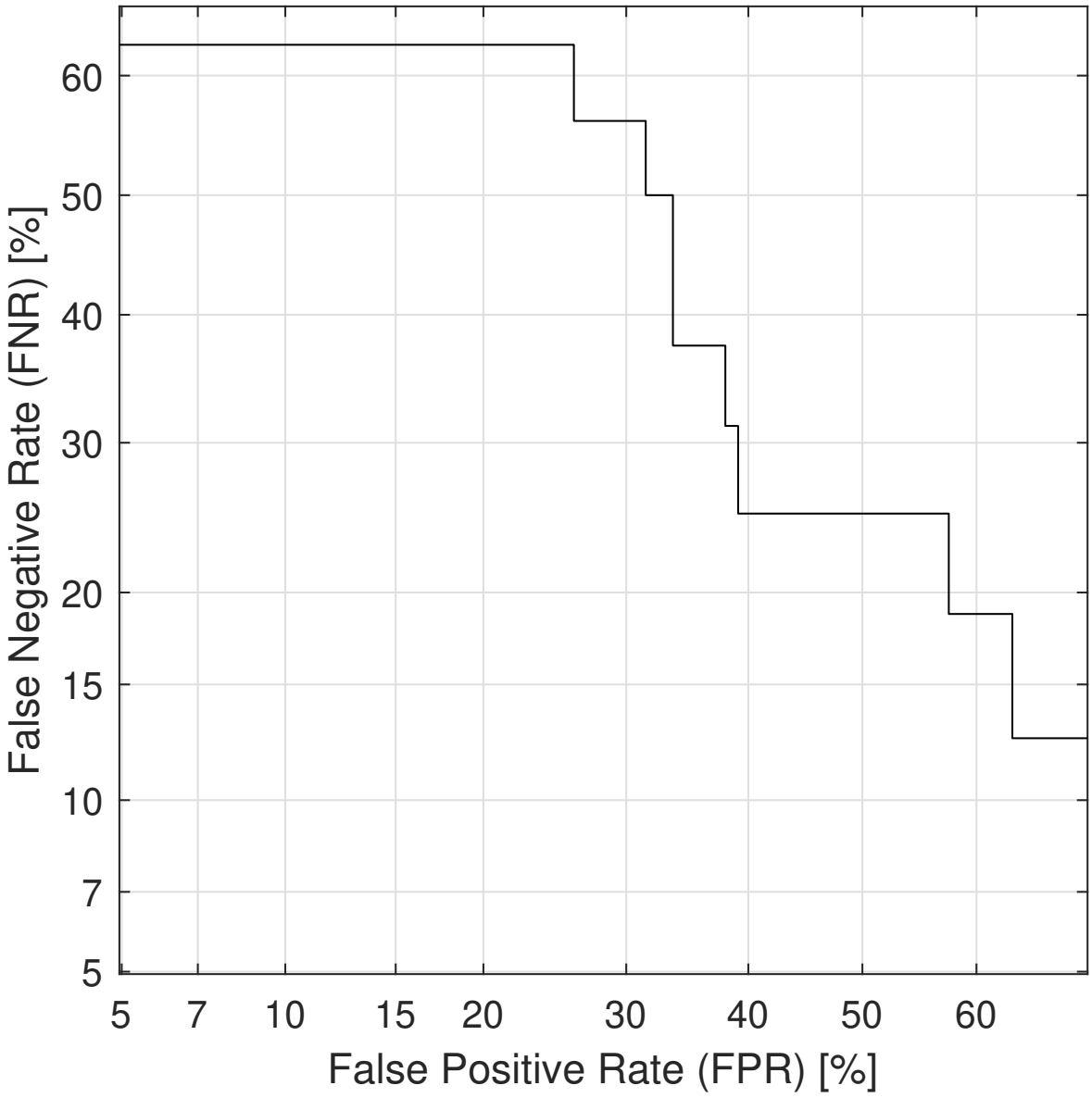
Experimento	Preprocesamiento	EER promedio	Error Estándar
Detección de cedro	PCA	37.40 %	0.207 %
Detección de móncoro	PCA	53.64 %	0.255 %
Detección de pino	PCA	40.67 %	0.253 %
Detección de sapán	PCA	32.77 %	0.221 %

Nota: Resumen de los valores de EER (Equal Error Rate) obtenidos en los experimentos de verificación con Análisis de Componentes Principales (PCA).

5.4.2. Detección usando serie temporal completa. También se aplicó un procedimiento diferente para la extracción de características. Se consideró cada uno de los 16 sensores como una característica, formando un vector de características de longitud 16. Cada una de las 400 muestras obtenidas se considera como un **frame**, por lo tanto, nuestra representación de la huella odorífica de un bloque de madera es una matriz de 16×400. Estos experimentos de verificación se repiten 100 veces, con diferentes ordenaciones aleatorias del conjunto de datos, pero no se utiliza el con-

Figura 13

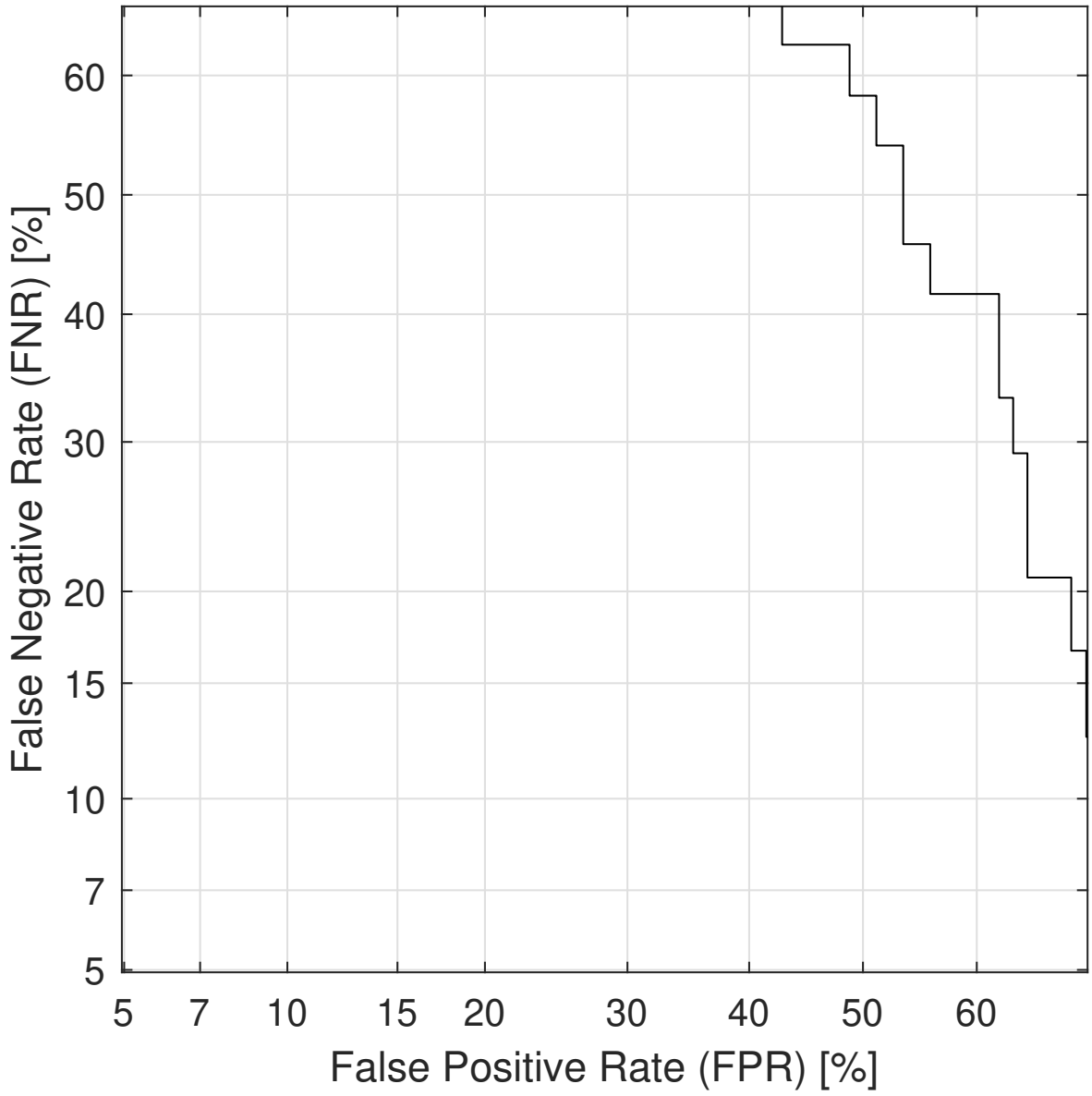
Curva DET para la Detección de cedro con PCA.



Nota: Curva DET (Detection Error Trade-off) para la Detección de cedro con Análisis de componentes principales. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

Figura 14

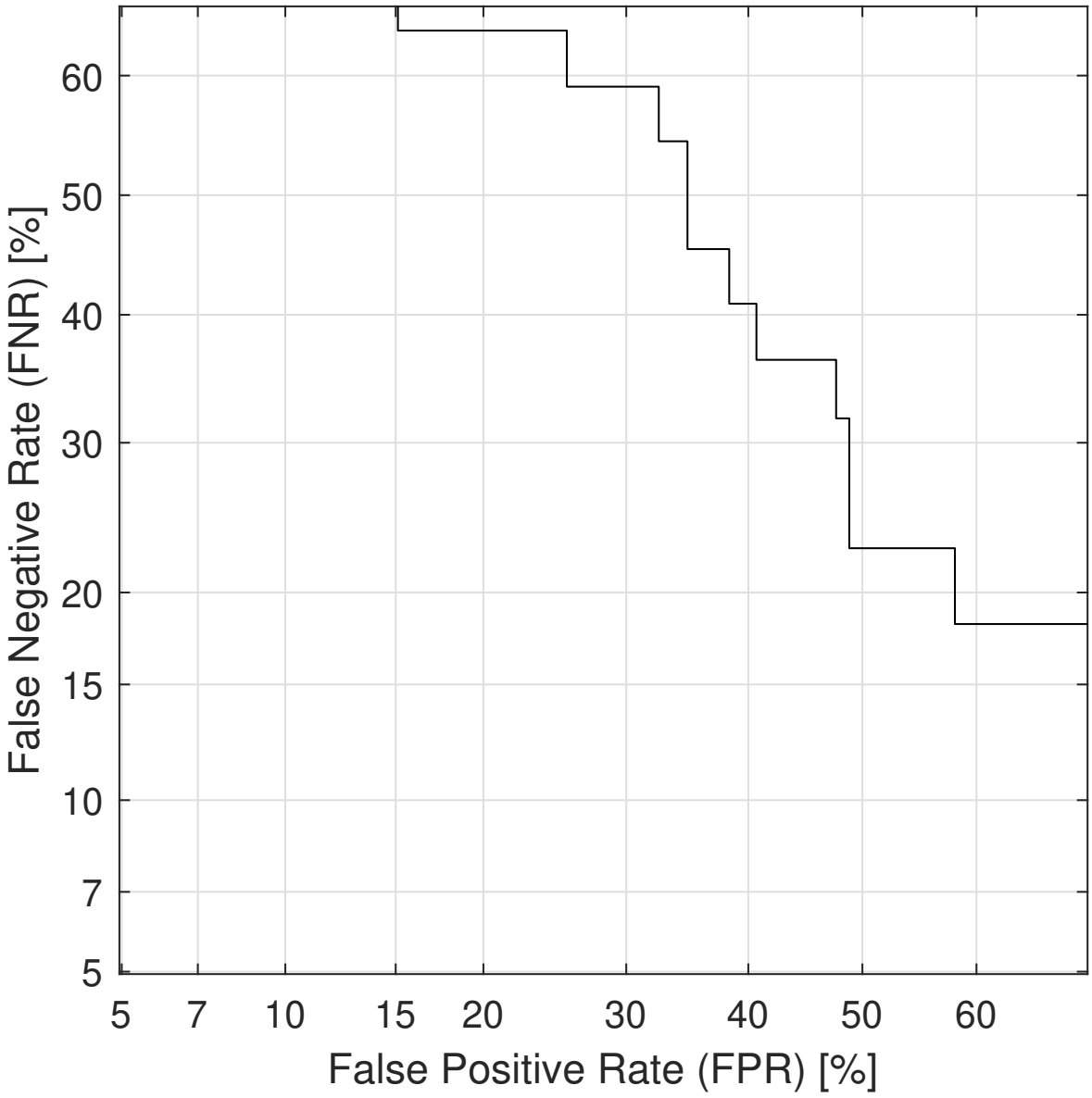
Curva DET para la Detección de móncoro con PCA.



Nota: Curva DET (Detection Error Trade-off) para la Detección de móncoro con Análisis de componentes principales. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

Figura 15

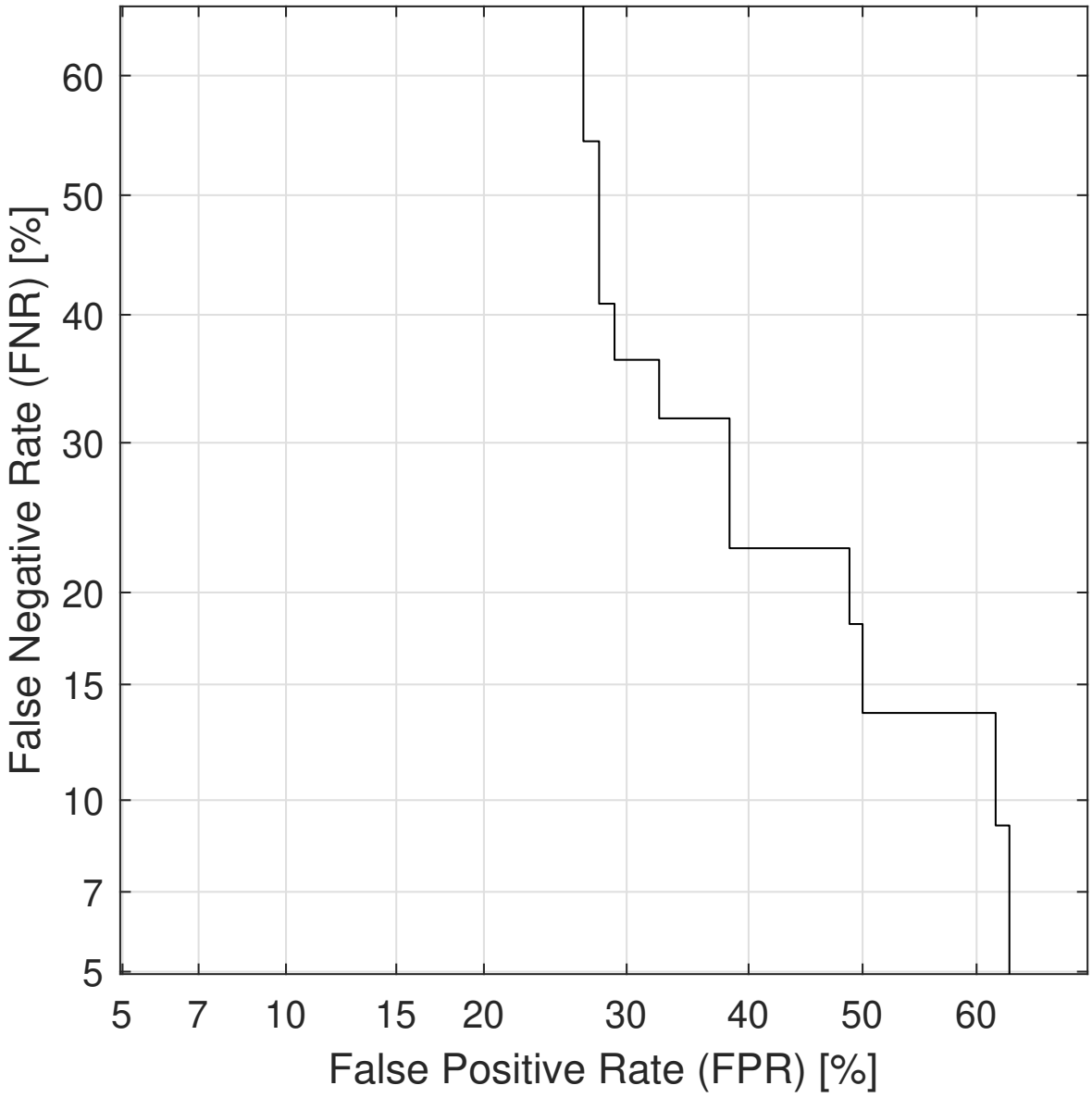
Curva DET para la Detección de pino con PCA.



Nota: Curva DET (Detection Error Trade-off) para la Detección de pino con Análisis de componentes principales. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

Figura 16

Curva DET para la Detección de sapán con PCA.



Nota: Curva DET (Detection Error Trade-off) para la Detección de sapán con Análisis de componentes principales. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

junto de datos aumentado, sino el conjunto de datos original. Los resultados del procedimiento de verificación de la clase de cedro frente al UBM, utilizando la serie de tiempo completa, mostraron una tasa de error de clasificación del 24.18 %. Una gráfica de la curva DET se muestra en la Figura 17.

De la misma manera, se muestran los resultados del procedimiento de verificación de las demás clases (móncoro, pino y sapán) frente al UBM, utilizando la serie de tiempo completa. Para la detección de móncoro, se encontró una tasa de error de clasificación del 34.08 % y una gráfica de la curva DET se muestra en la Figura 18. Para la detección de pino, el error fue del 32.88 % y la curva DET se muestra en la Figura 19. Finalmente, para la detección de sapán, el error fue del 21.38 % y la curva DET se muestra en la Figura 20. En la Tabla 10 se muestra un resumen de los resultados obtenidos usando la serie temporal completa.

Tabla 10

Valores de EER obtenidos en los experimentos de verificación (serie temporal).

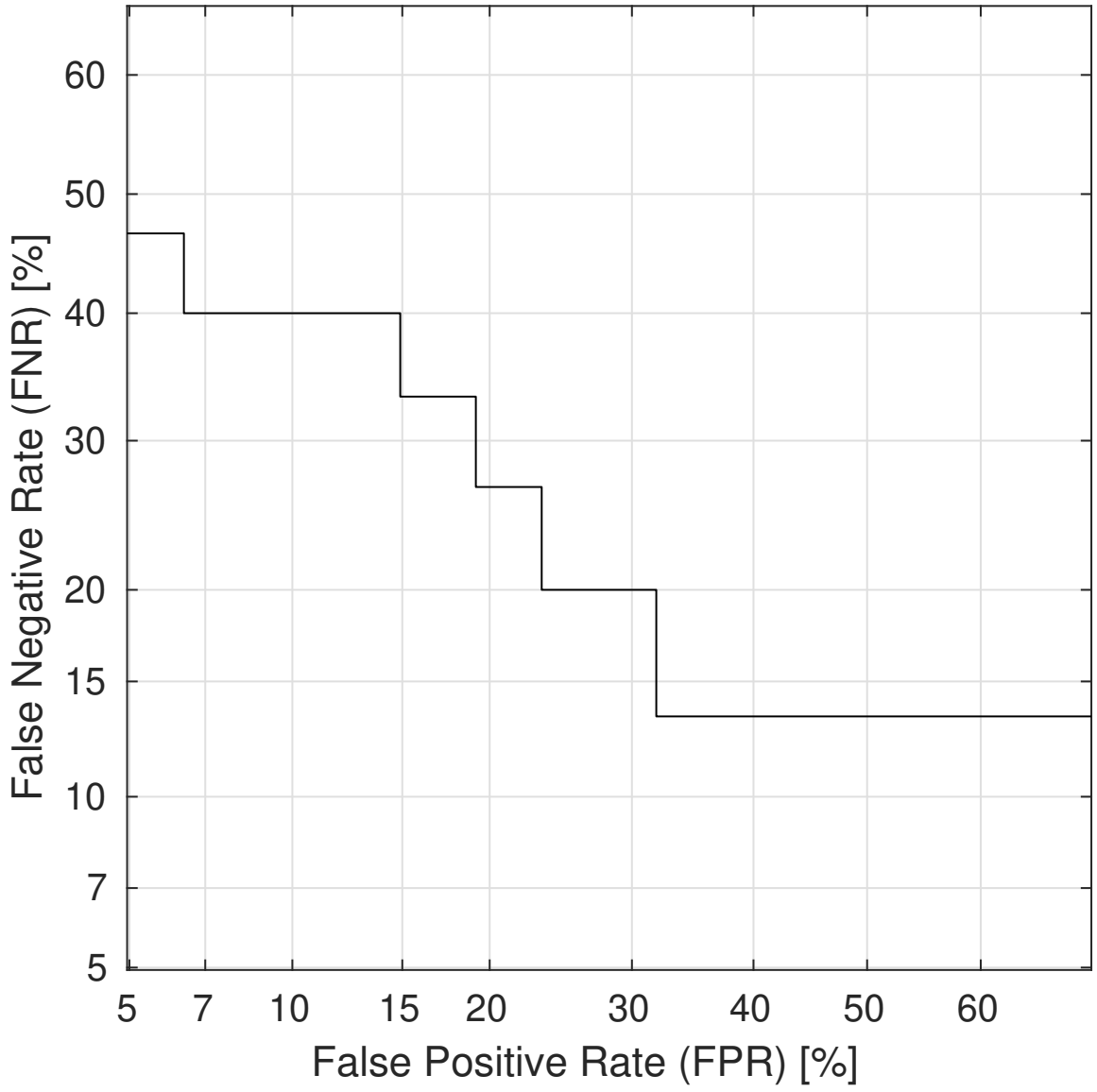
Experimento	Preprocesamiento	EER promedio	Error Estándar
Detección de cedro	Serie Temporal Completa	24.18 %	0.453 %
Detección de móncoro	Serie Temporal Completa	34.08 %	0.669 %
Detección de pino	Serie Temporal Completa	32.88 %	0.648 %
Detección de sapán	Serie Temporal Completa	21.38 %	0.564 %

Nota: Resumen de los valores de EER (Equal Error Rate) obtenidos en los experimentos de verificación con serie temporal completa.

5.4.3. GMM-UBM usando LASSO. En esta fase, se hace la validación del experimento de selección de características utilizando *LASSO*. Para este experimento, se utiliza un procedimiento de validación cruzada de 5 conjuntos; donde el 80 % de las muestras de madera corresponden al conjunto de entrenamiento, mientras que las muestras restantes (20 %) se utilizan para la validación. Dentro de la etapa de entrenamiento, se realiza la selección de características a través de

Figura 17

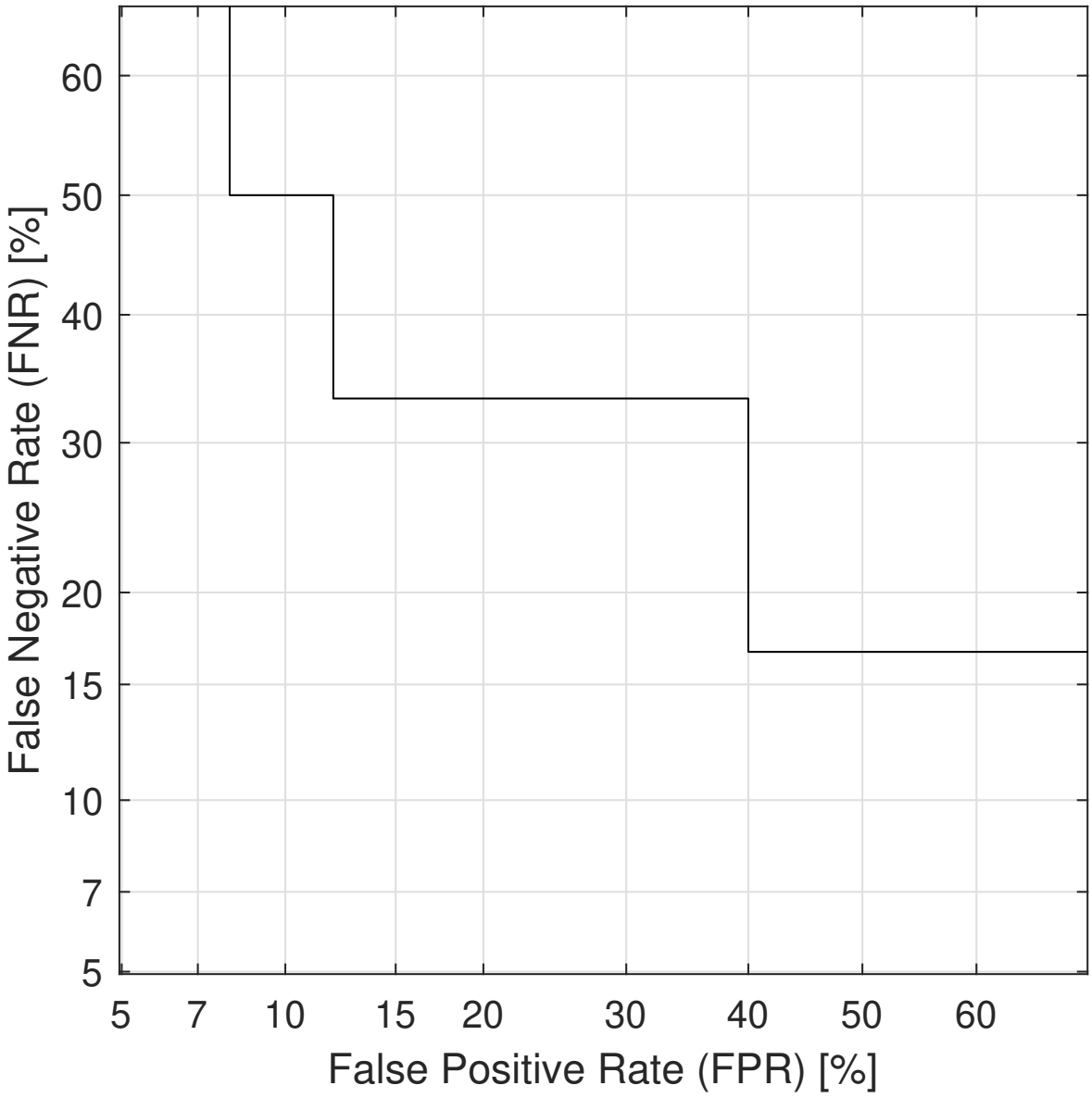
Curva DET para la Detección de cedro con serie temporal completa.



Note: Curva DET (Detection Error Trade-off) para la Detección de cedro con serie temporal completa. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

Figura 18

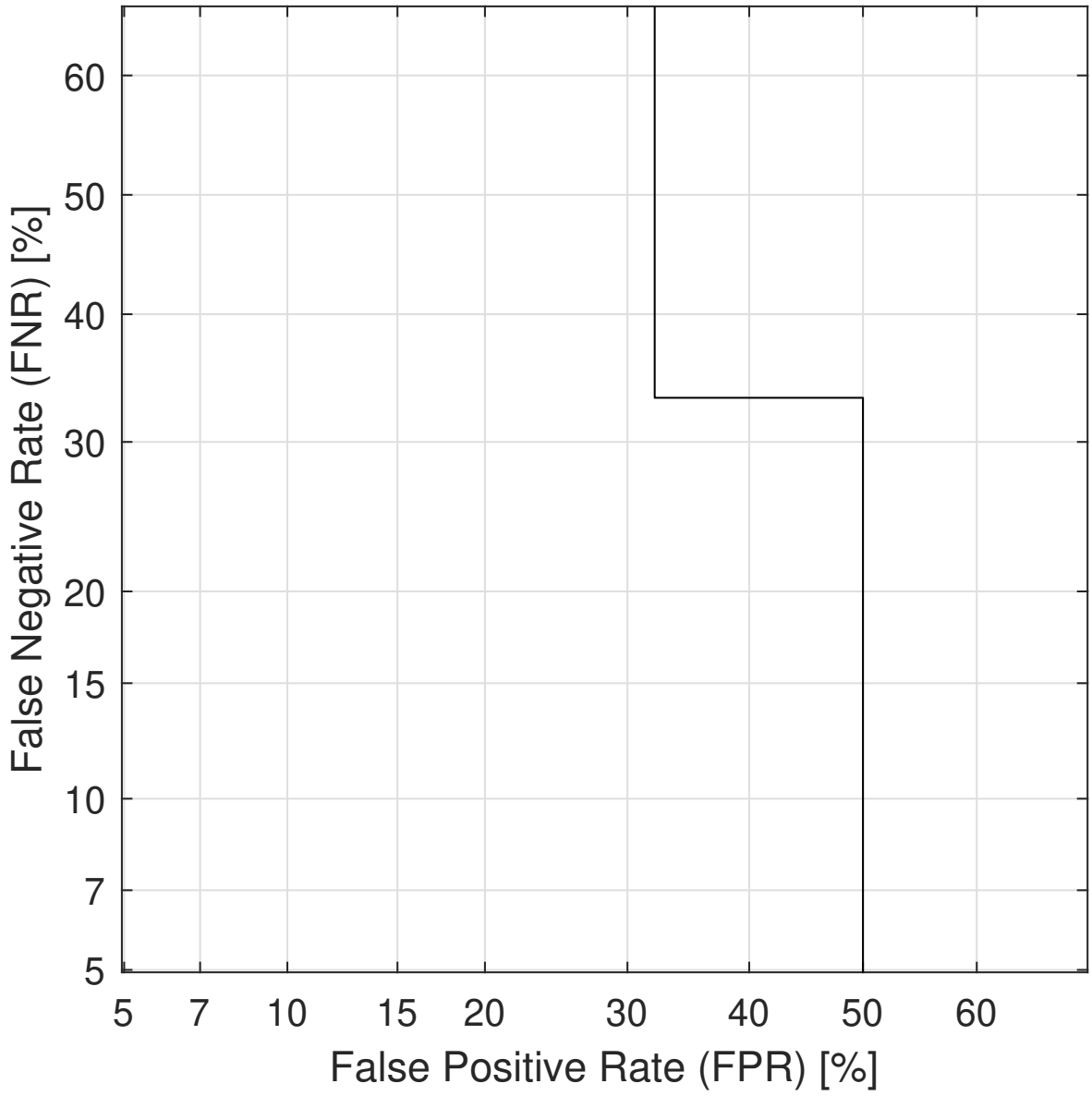
Curva DET para la Detección de móncoro con serie temporal completa.



Nota: Curva DET (Detection Error Trade-off) para la Detección de móncoro con serie temporal completa. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

Figura 19

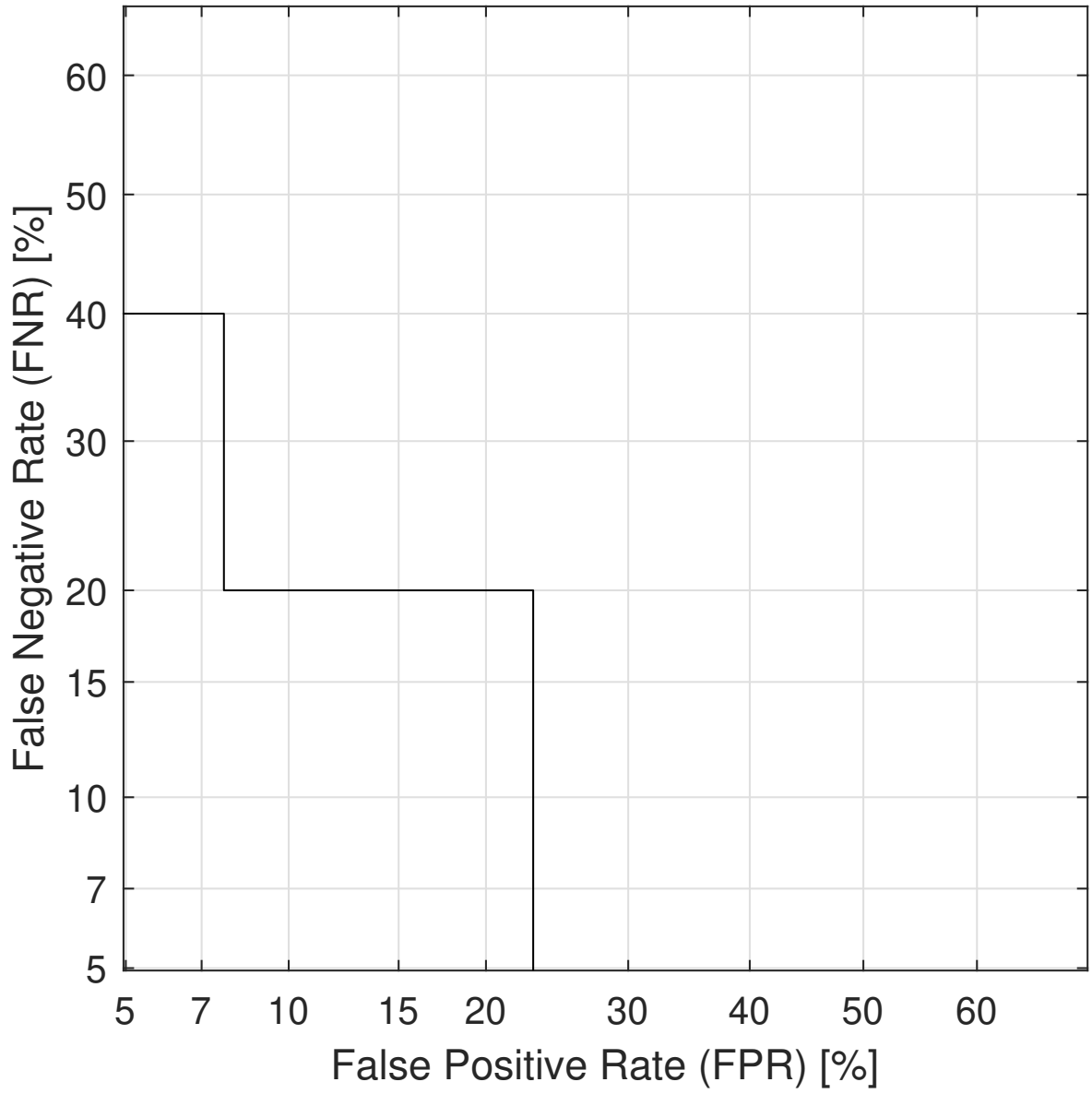
Curva DET para la Detección de pino con serie temporal completa.



Nota: Curva DET (Detection Error Trade-off) para la Detección de pino con serie temporal completa. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

Figura 20

Curva DET para la Detección de sapán con serie temporal completa.



Nota: Curva DET (Detection Error Trade-off) para la Detección de sapán con serie temporal completa. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

LASSO; sobe el conjunto de validación, se utiliza el mismo set de características seleccionadas por *LASSO* para el procedimiento de verificación (mediante el enfoque GMM-UBM). Este experimento se repite 100 veces, con diferentes ordenaciones aleatorias del conjunto de datos. La verificación con la clase cedro, aplicando *LASSO*, mostró una tasa de error de clasificación de 17.83 %. Al igual que con el cedro, el experimento de verificación se realizó para otras tres especies (móncoro, pino y sapán). Un resumen de los resultados se puede ver en la Tabla 11 y, en la Figura 21, se muestra una gráfica de la curva DET para el experimento ***Detección de Sapán***.

Tabla 11

Valores de EER para la verificación utilizado LASSO.

Experimento	Preprocesamiento	EER promedio	Error Estándar
Detección de cedro	*LASSO*	17.83 %	0.366 %
Detección de móncoro	*LASSO*	21.12 %	0.561 %
Detección de pino	*LASSO*	19.05 %	0.172 %
Detección de sapán	*LASSO*	11,92 %	0.330 %

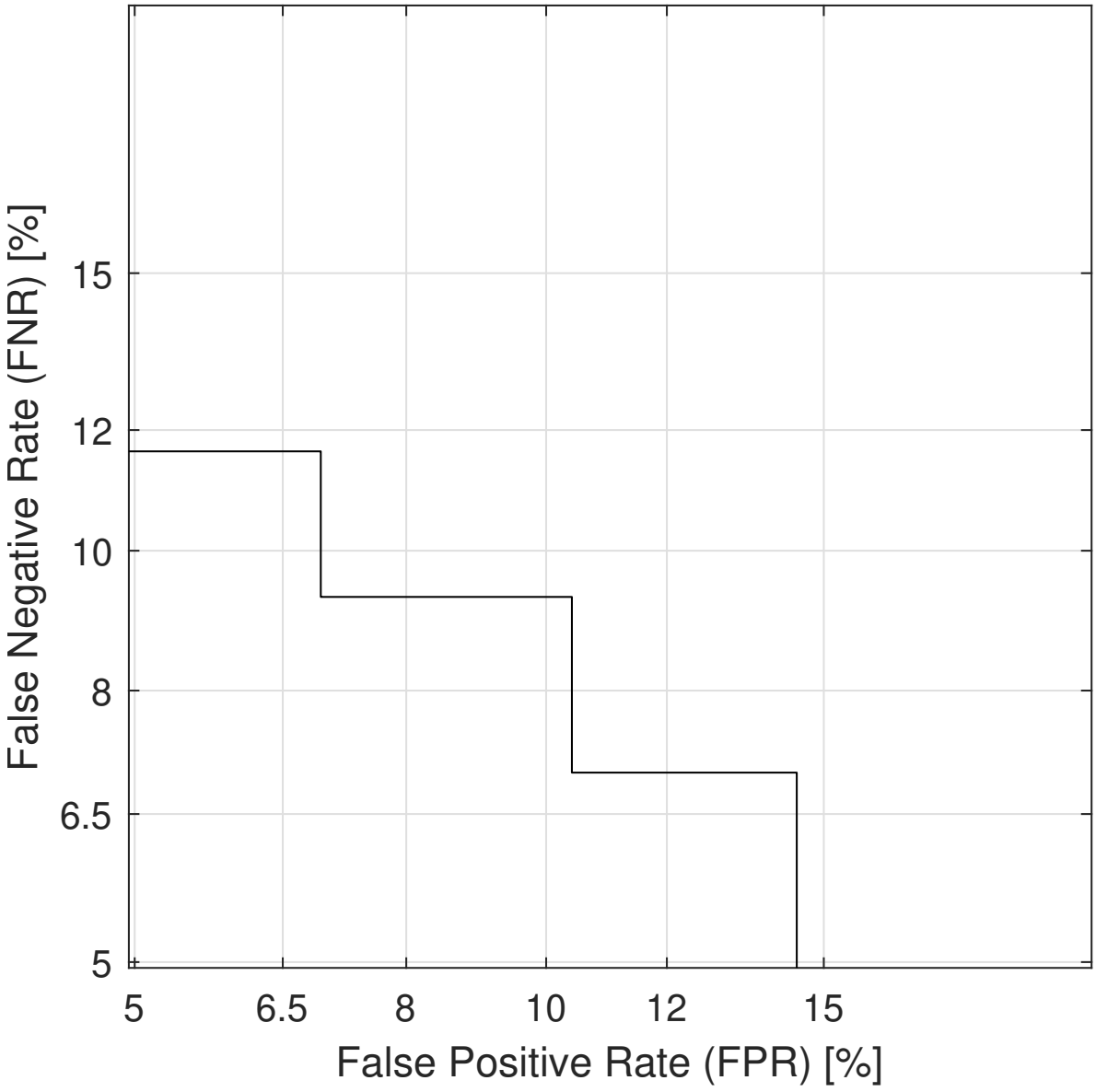
Nota: Resumen de los valores de EER (Equal Error Rate) para la verificación utilizado LASSO.

5.5. Discusión de Resultados del Experimento de Verificación

Al igual que con el experimento de identificación, los errores reportados en este experimento son altos. De otra parte, debido a que al revisar el estado del arte no se encuentran trabajos de detección de madera con narices electrónicas y un enfoque biométrico, los resultados no son directamente comparables a los reportados en (Kalaw and Sevilla, 2018; Cordeiro et al., 2016; Wilson, 2012). Sin embargo, es importante tener en cuenta que las obras mencionadas anteriormente utilizan material de madera recién cortada, cuyos aromas son frescos, fuertes y sin grandes interferencias. Además, esos trabajos reportados usan condiciones controladas. Por el contrario, en este trabajo utilizamos muestras de madera no recién aserrada, que son menos frescas y tienen aromas más débiles. Aunque las muestras de madera no recién aserrada son más difíciles de clasificar, sus

Figura 21

Curva DET para la Detección de sapán, aplicando LASSO.



Nota: Curva DET (Detection Error Trade-off) para la Detección de sapán, aplicando LASSO como selección de características. Tasa de falsos negativos (FNR, por sus siglas en inglés). Tasa de falsos positivos (FPR, por sus siglas en inglés).

condiciones están más cerca de situaciones reales que las que se usan cuando se usan muestras de madera recientemente cortada.

El objetivo de este trabajo es analizar un conjunto de datos mayor, con muestras de madera no recién aserradas y con condiciones de almacenamiento no rigurosas. Esto permite establecer condiciones de entorno menos distantes del entorno real para las cuales se busca resolver el problema. Dicho esto, los resultados dentro del alcance de este estudio son prometedores, ya que muestra que las señales analizadas contienen información importante y discriminatoria para la tarea en cuestión. Sin embargo, aunque se utiliza una mayor cantidad de muestras que en otros trabajos de referencia, el conjunto de datos aún no era lo suficientemente grande. Por ello, y debido a las dificultades para recolectar muestras y ampliar el conjunto de datos, se aplica el procedimiento de aumentado de datos.

Además, en los trabajos referenciados se realizaron tareas de identificación de especies de madera, lo que implica un número limitado y cerrado de especies dentro de las cuales clasificar una muestra. Como alterativa, se propone un procedimiento de verificación, más útil en la práctica, en el que se compara una muestra con un modelo de referencia correspondiente a una especie de interés. Si la muestra de prueba no se parece a la clase objetivo, se dice que pertenece a otra clase; mientras que en los procesos de identificación, se debe asignar una etiqueta de las clases definidas. Los procedimientos de verificación de especies de madera son inusuales, menos aquellos que usan un enfoque basado en el aroma. Dentro de la búsqueda de información llevada a cabo para este trabajo, no se encontraron trabajos previos que apliquen un proceso verificación automática de la especie de madera a partir de mediciones provenientes de narices electrónicas, lo que dificulta comparar los presentes resultados con resultados previos. La propuesta consiste en combinar el uso de narices electrónicas y técnicas de verificación, como GMM-UBM, para determinar rápidamente si una muestra de madera pertenece o no a una especie de interés, según su aroma.

6. Conclusiones y trabajo futuro

En el presente trabajo se desarrollaron dos experimentos: identificación y detección de maderas basadas en su olor; y, en ambos casos, bajo condiciones más exigentes que las que se reportan en los trabajos del estado del arte. Es decir, en el presente trabajo las condiciones son más similares a las del entorno de trabajo donde se pretende usar este tipo de sistemas. Las muestras de madera recién cortada, utilizadas en los trabajos de referencia, son frescas y con un aroma todavía fuerte e intenso. Por el contrario, nosotros utilizamos muestras de madera no recién aserrada que, aunque son más secas y tienen un aroma más débil, representan mejor los escenarios reales de la tala ilegal. Como resultado, las tasas de error son más altas que en otros informes como Cordeiro et al. (2016); Wilson (2012); Kalaw and Sevilla (2018).

Ambos experimentos se hacen mediante el uso de una nariz electrónica de bajo costo, formada por una matriz de 16 sensores químicos de referencias comerciales. Estos sensores se usan como alternativa a propuestas más sofisticadas como: las narices electrónicas ***Cyranose 320*** (Garneau et al., 2004) y ***Aromascan A32S*** (Wilson et al., 2005); arreglo de 8 sensores químicos con principios resistivos basados en nanotubos de carbono (Kalaw and Sevilla, 2018); y, arreglo de 4 sensores de polímero conductor con principio resistivo en (Cordeiro et al., 2016). El uso de narices electrónicas para la identificación o detección de maderas es una área poco explorada y el presente trabajo constituye uno dentro los pocos reportados, aún más si lo acotamos a Colombia.

Al realizar los experimentos de identificación se encuentra un desempeño de alrededor de 79 %, lo cual indica que han de realizarse mayores esfuerzos a fin de obtener un mejor desempeño, similar al obtenido en trabajos previos. Por ejemplo, en Garneau et al. (2004) se reporta un desempeño del 100 % al analizar 30 registros de olor, correspondientes a tres especies diferentes de

la familia de las pináceas (Pinaceae), mediante el uso de narices electrónicas y Análisis de Componentes Principales (PCA). En Wilson et al. (2005), se investigó el uso de redes neuronales para identificar diferencias entre especies pertenecientes a la misma familia y género (dos muestras por árbol, pertenecientes a entre 13 y 30 árboles de 12 especies), con tasas de identificación entre el 94 % y el 99 %. En Cordeiro et al. (2016) se utilizaron narices electrónicas y un análisis de componentes principales (PCA) en dos problemas de clasificación de especies maderables: caoba vs Cedro español, clasificadas con un desempeño del 100 %; y nuez brasileña vs canela negra, especies pertenecientes al mismo genero y clasificadas con 94 % de precisión. Finalmente, en Kalaw and Sevilla (2018) se analizó la separabilidad de clases de cinco especies maderables filipinas (representadas por mediciones entregadas por sensores de gases, a partir de una cantidad reducida de muestras recogidas en una zona específica) generando agrupaciones (clusters) separables a simple vista. Sin embargo, no hay que olvidar que, en el presente trabajo, las condiciones del experimento son mucho menos rigurosas que en los trabajos mencionados.

De otra parte, en el presente trabajo se propuso un enfoque de verificación, el cual se llevó a cabo utilizando el modelo de mezclas gaussiana como modelo de referencia, obteniendo una tasa de desempeño de 78 %. Los procedimientos de verificación podrían ser una mejor opción en escenarios prácticos que los procedimientos de identificación. Hasta donde se sabe, el presente trabajo sería el primero en cuyo enfoque se hizo uso de huellas olfativas desde un punto de vista biométrico para la detección de especies de madera. Al incluir *LASSO* como técnica de selección de características, junto al enfoque de verificación mediante GMM-UBM, la tasa de desempeño llega hasta un 88 %.

De acuerdo a la Tabla 1 y a la información obtenida de los gráficos de *loadings* como el presentado en la Figura 12, se infiere que algunos de los sensores no representan información relevante para el problema planteado. A partir de este análisis se pueden identificar aquellos sensores, dentro de los incluidos en el arreglo, que aportan mayor información desde el punto de vista de

separabilidad de las maderas. Cabe mencionar que el criterio inicial utilizado para la selección de los sensores que forman el arreglo fue una mezcla entre variedad en el tipo de sensores y disponibilidad de los mismos en el mercado, debido al enfoque inicial del trabajo. Con el fin de reducir el número de características y con la idea de identificar los sensores más relevantes para la detección de especies de madera, en el presente trabajo se propuso el uso de la técnica *LASSO*. Se encontró que aquellos tipos de sensores que más aportan son: $MQ-3$, $MQ-6$ y $MQ-8$ de la marca HANWEI, $TGS-813$, $TGS-816$, $TGS-822$ y $TGS-823$ de la marca FIGARO. Fue precisamente bajo este esquema que se obtuvo el menor EER.

Aparte de los resultados de este experimento, no contamos con información acerca de cuáles serían los sensores químicos más apropiados, teniendo en cuenta la composición química de los aromas de las diferentes especies de maderas químicos apropiados para esta aplicación específica. Por lo tanto se sugiere, como trabajo futuro, realizar un análisis de compuestos volátiles que componen los aromas de las diferentes especies de madera. Determinar aquellos sensores óptimos requiere de un análisis químico previo de los aromas presentes en las especies de maderas a trabajar. Por ahora, esa es una tarea está por fuera de nuestro alcance y por tanto se se deja como trabajo futuro. Además de esto, podría ser necesario explorar otras técnicas de extracción de características.

Dado que en este tipo de problemas se cuenta con una cantidad de características de entrada de tamaño comparable con la cantidad de mediciones, se hace necesario implementar estrategias de reducción de dimensionalidad. Aunque PCA ayuda en esta tarea, el aplicar esta técnica no garantiza obtener una mejor representación desde el punto de vista de clasificación y es una técnica de tipo lineal; además, se pierde el sentido físico de cada una de las características, que es importante a la hora de indagar por aquellos tipos de sensores más adecuados para la tarea en cuestión. Adicionalmente, está la opción de utilizar medidas de información mutua para la tarea de selección de características.

Finalmente, se sugiere considerar otras estrategias, como el uso de imágenes, para recolectar información complementaria que mejore el rendimiento del proceso de clasificación o verificación.

Referencias Bibliográficas

Agritix (2016). Xylorix: Macroscopic wood identification system. Recuperado de https://www.xylorix.com.

Akbar, M. A., Ait Si Ali, A., Amira, A., Bensaali, F., Benammar, M., Hassan, M., and Bermak, A. (2016). An empirical study for pca- and lda-based feature reduction for gas identification. *IEEE Sensors Journal*, 16(14):5734–5746.

Arenas, J. J. G., Instituto de Hridrología, M. y. E. A., Álvaro Cubillos Buitrago, Hernández, M. A. C., IDEAM, de Medio Ambiente y Desarrollo Sostenible, M., and Colombia, P. O.-R. (2018). *Principales causas y agentes de la deforestación a nivel nacional*. IDEAM.

Baietto, M., Pozzi, L., Wilson, A. D., and Bassi, D. (2013). Evaluation of a portable MOS electronic nose to detect root rots in shade tree species. *Comput. Electron. Agric.*

Baietto, M., Wilson, A. D., Bassi, D., and Ferrini, F. (2010). *Evaluation of three electronic noses for detecting incipient wood decay*, volume 10. Sensors.

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

Cabral, E. C., Simas, R. C., Santos, V. G., Queiroga, C. L., Da Cunha, V. S., De Sá, G. F., Daroda, R. J., and Eberlin, M. N. (2012). Wood typification by Venturi easy ambient sonic spray ionization mass spectrometry: The case of the endangered Mahogany tree. *Journal of Mass Spectrometry*, 47(1):1–6.

Capelli, L., Sironi, S., and Del Rosso, R. (2014). Electronic Noses for Environmental Monitoring Applications. *Sensors*, 14(11):19979–20007.

Carballo-Meilán, A., Goodman, A. M., Baron, M. G., and Gonzalez-Rodriguez, J. (2016). Application of chemometric analysis to infrared spectroscopy for the identification of wood origin. *Cellulose*, 23(1):901–913.

Carmel, L., Levy, S., Lancet, D., and Harel, D. (2003). A feature extraction method for chemical sensors in electronic noses. *Sensors and Actuators B: Chemical*, 93(1):67 – 76. Proceedings of the Ninth International Meeting on Chemical Sensors.

CentroTIC (2016). Plataforma IoT para el desarrollo de servicios inteligentes de apoyo al monitoreo ambiental.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Cieslak, D. A., Chawla, N. V., and Striegel, A. (2006). Combating imbalance in network intrusion datasets. In *GrC*, pages 732–737.

Cordeiro, J. R., Li, R. W. C., Takahashi, É. S., Rehder, G. P., Ceccantini, G., and Gruber, J. (2016). Wood identification by a portable low-cost polymer-based electronic nose. *RSC Adv.*, 6(111):109945–109949.

Corporación Autónoma Regional de Santander - CAS (2016). En Santander hay 13 especies de flora y 12 de fauna priorizados para su conservación. Recuperado de http://cas.gov.co/index.php/sala-de-prensa/453-en-santander-hay-13-especies-de-flora-y-12-de-fauna-priorizados-para-su-conservacion.html.

Diario el Tiempo (2015). Los cinco árboles maderables con más riesgo de extinción. Recuperado de https://www.eltiempo.com/archivo/documento/CMS-16195856.

Dickson, A., Nanayakkara, B., Sellier, D., Meason, D., Donaldson, L., and Brownlie, R. (2017). Fluorescence imaging of cambial zones to study wood formation in Pinus radiata D. Don. *Trees - Structure and Function*, 31(2):479–490.

Doddington, G. R., Przybocki, M. A., Martin, A. F., and Reynolds, D. A. (2000). The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254.

Durán Acevedo, C. M., Gualdró, Guerrero, O. E., and Hernández Ordoñez, M. (2014). Nariz electrónica para determinar el índice de madurez del tomate de árbol (Cyphomandra Betacea Sendt). *Ingeniería, Investigación y Tecnología*, 15(3):351–362.

Fedele, R., Galbally, I. E., Porter, N., and Weeks, I. A. (2007). Biogenic VOC emissions from fresh leaf mulch and wood chips of Grevillea robusta (Australian Silky Oak). *Atmos. Environ.*, 41(38):8736–8746.

Figaro Engineering Inc. (2018). Operating principle. Recuperado de https://www.figarosensor.com/technicalinfo/principle/mos-type.html.

Fondo Mundial para la Naturaleza - WWF (2017). 15 especies colombianas de árboles amenazados. Recuperado de http://www.wwf.org.co/?uNewsID=299073.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

FRIM, F. R. I. M. and UTAR, U. T. A. R. (2018). My wood: Mobile wood identification for 100 malaysian timbers. Recuperado de http://mywoodid.frim.gov.my.

Garneau, F. X., Riedl, B., Hobbs, S., Pichette, A., and Gagnon, H. (2004). The use of sensor array

technology for rapid differentiation of the sapwood and heartwood of Eastern Canadian spruce, fir and pine. *Holz als Roh - und Werkstoff*, 62(6):470–473.

Ghasemi-Varnamkhasti, M., Mohammad-Razdari, A., Yoosefian, S. H., Izadi, Z., and Rabiei, G. (2019). Selection of an optimized metal oxide semiconductor sensor (mos) array for freshness characterization of strawberry in polymer packages using response surface method (rsm). *Post-harvest Biology and Technology*, 151:53 – 60.

Gómez Monsalve, P. A. and Durán Acevedo, C. M. (2015). Data Acquisition From An Array Of Gas Sensors (E-Nose), Through Xbee Communications Modules. *Revista Colombiana de Tecnologías de Avanzada*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Guerrero Rodríguez, O. C. (2017). El roble colombiano, más débil que nunca. *Revista Semana*. Recuperado de https://sostenibilidad.semana.com/medio-ambiente/articulo/el-roble-colombiano-enfrenta-amenazas-a-su-supervivencia/38691.

Guo, L., Yang, Z., and Dou, X. (2017). Artificial Olfactory System for Trace Identification of Explosive Vapors Realized by Optoelectronic Schottky Sensing. *Adv. Mater.*, 29(5):1–8.

Gutiérrez, J. and Horrillo, M. C. (2014). Advances in artificial olfaction: Sensors and applications. *Talanta*, 124:95–105.

Hamilton, S., Hepher, M. J., and Sommerville, J. (2006). Detection of Serpula lacrymans infestation with a polypyrrole sensor array. *Sens. Actuators B Chem.*

Hanssen, F., Wischnewski, N., Moreth, U., and Magel, E. A. (2011). Molecular identification of Fitzroya cupressoides, Sequoia sempervirens, and Thuja plicata wood using taxon-specific rDNA-ITS primers. *IAWA J.*, 32(2):273–284.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kalaw, J. M. and Sevilla, F. B. (2018). Discrimination of wood species based on a carbon nanotube/polymer composite chemiresistor array. *Holzforschung*, 72(3):215–223.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40.

Lusa, L. et al. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC bioinformatics*, 11(1):523.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The det curve in assessment of detection task performance. Technical report, National Inst of Standards and Technology Gaithersburg MD.

Martinelli, E., Magna, G., Vergara, A., and Di Natale, C. (2014). Cooperative classifiers for reconfigurable sensor arrays. *Sens. Actuators B Chem.*, 199:83–92.

Mesa de Bosques de Santander (2018). Mision Bosques Santander. Technical report, Mesa de Bosques de Santander, Bucaramanga.

Ministerio de Ambiente y Desarrollo Sostenible, I. (2017). Estrategia Integral de Control a la Deforestación y Gestión de Los Bosques en Colombia. Technical report, Ministerio de Ambiente y Desarrollo Sostenible. IDEAM.

Moreno, I., Caballero, R., Galán, R., Matía, F., and Jiménez, A. (2009). La Nariz Electrónica: Estado del Arte. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, 6(3):76–91.

Müller, K., Haferkorn, S., Grabmer, W., Wisthaler, A., Hansel, A., Kreuzwieser, J., Cojocariu, C., Rennenberg, H., and Herrmann, H. (2006). Biogenic carbonyl compounds within and above a coniferous forest in Germany. *Atmos. Environ.*, 40:81–91.

Najib, M. S., Ahmad, M. U., Funk, P., Taib, M. N., and Ali, N. A. M. (2012). Agarwood classification: A case-based reasoning approach based on E-nose. *Proceedings - 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications, CSPA 2012*, pages 120–126.

Peña, S. V. and Rojas, I. M. (2006). *Tecnología de la madera*. Mundi-Prensa Libros.

Rana, R., Müller, G., Naumann, A., and Polle, A. (2008). FTIR spectroscopy in combination with principal component analysis or cluster analysis as a tool to distinguish beech (Fagus sylvatica L.) trees grown at different sites. *Holzforschung*, 62(5):530–538.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41.

Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83.

Rinne, H. J. I., Guenther, A. B., Greenberg, J. P., and Harley, P. C. (2002). Isoprene and monoterpene fluxes measured above Amazonian rainforest and their dependence on light and temperature. *Atmos. Environ.*, 36(14):2421–2426.

Rodríguez, J., Durán, C., and Reyes, A. (2010). Electronic nose for quality control of Colombian coffee through the detection of defects in Cup Tests. *Sensors*, 10(1):36–46.

Rodriguez-Lujan, I., Fonollosa, J., Vergara, A., Homer, M., and Huerta, R. (2014). On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. *Chemometrics Intellig. Lab. Syst.*, 130:123–134.

Ruiz Jiménez, L. F. (2018). Detección de los insectos de la subfamilia Triatominae basado en narices electrónicas. Technical report, Universidad Industrial de Santander.

Santos, J. P. and Lozano, J. (2015). Real time detection of beer defects with a hand held electronic nose. *Proceedings of the 2015 10th Spanish Conference on Electron Devices, CDE 2015*, pages 1–4.

Shi, H., Zhang, M., and Adhikari, B. (2017). Advances of electronic nose and its application in fresh foods: A review. *Crit. Rev. Food Sci. Nutr.*, 8398:1–11.

SJÖSTRÖM, E. (1993). Chapter 5 - extractives. In SJÖSTRÖM, E., editor, *Wood Chemistry (Second Edition)*, pages 90 – 108. Academic Press, San Diego, second edition edition.

Van Hulse, J., Khoshgoftaar, T. M., and Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942.

Vergara, A., Vembu, S., Ayhan, T., Ryan, M. A., Homer, M. L., and Huerta, R. (2012). Chemical gas sensor drift compensation using classifier ensembles. *Sens. Actuators B Chem.*, 166-167:320–329.

Wheeler, E. A. and Baas, P. (1998). Wood identification-a review. *IAWA journal*, 19(3):241–264.

Wilson, A. D. (2012). Application of a Conductive Polymer Electronic-Nose Device to Identify Aged Woody Samples. *The Third International Conference on Sensor Device Technologies and Applications*, pages 77–82.

Wilson, A. D. (2013). Diverse applications of electronic-nose technologies in agriculture and forestry.

Wilson, A. D. and Baietto, M. (2009). Applications and Advances in Electronic-Nose Technologies. *Sensors*, 9(7):5099–5148.

Wilson, A. D., Lester, D. G., and Oberle, C. S. (2005). Application of conductive polymer analysis for wood and woody plant identifications. *For. Ecol. Manage.*, 209(3):207–224.

Yan, J., Guo, X., Duan, S., Jia, P., Wang, L., Peng, C., and Zhang, S. (2015). Electronic nose feature extraction methods: A review. *Sensors*, 15(11):27804–27831.

Yu, M., Liu, K., Zhou, L., Zhao, L., and Liu, S. (2016). Testing three proposed DNA barcodes for the wood identification of Dalbergia odorifera T. Chen and Dalbergia tonkinensis Prain. *Holzforschung*, 70(2):127–136.

Zhang, X., Cheng, J., Wu, L., Mei, Y., Jaffrezic-Renault, N., and Guo, Z. (2018). An overview of an artificial nose system. *Talanta*, 184(January):93–102.

Zhao, P. and Cao, J. (2016). Wood species identification using spectral reflectance feature and optimal illumination radian design. *Journal of Forestry Research*, 27(1):219–224.

Apéndices

Apéndice A. Nariz electrónica como dispositivo IoT

Dentro de las lineas de investigación del Grupo de Investigación RadioGIS, de la UIS, se encuentra la consolidación de la nariz electrónica como un dispositivo IoT y la posterior prestación de servicios TIC, basados en el uso de narices electrónicas. La conformación de una plataforma IoT está sustentada en el desarrollo de un modelo capas que reúnen, en detalle, los componentes que intervienen en la implementación de un servicio TIC (Ruiz Jiménez, 2018). El modelo propuesto por el CentroTIC, que se puede observar en la Figura 22, se compone de seis capas conectadas entre sí: sociedad, tecnologías del usuario, acceso al medio, conectividad, cómputo en la nube (*cloud computing*) y aplicaciones.

En la capa de sociedad, como principales interesados, encontramos a las autoridades y corporaciones ambientales, la UIS y, en general, cualquier entidad que conforme el nodo estratégico denominado *Mesa de Bosques de Santander*. La capa de sociedad interactúa directamente con la capa de tecnologías del usuario, compuesta principalmente por el hardware de la nariz electrónica, presentado en el capítulo Capítulo 3. El hardware, a su vez, interactúa con la siguiente capa (acceso al medio) a través de la tarjeta de adquisición de datos que es capaz de conectarse a internet (capa de conectividad). La capa de cómputo en la nube depende de la orientación funcional del dispositivo y la capa de aplicaciones, que se desarrolla como un complemento al servicio, está orientada hacia los servicios web y necesidades del cliente.

Este trabajo hace parte del proyecto "*Plataforma IoT para el desarrollo de servicios inteligentes de apoyo al monitoreo ambiental*, código 1971" de la Vicerrectoría de Investigación y Extensión de la Universidad Industrial de Santander. Dentro de esa plataforma, existe el espacio

Figura 22

Modelo de Capas de la Plataforma IoT.

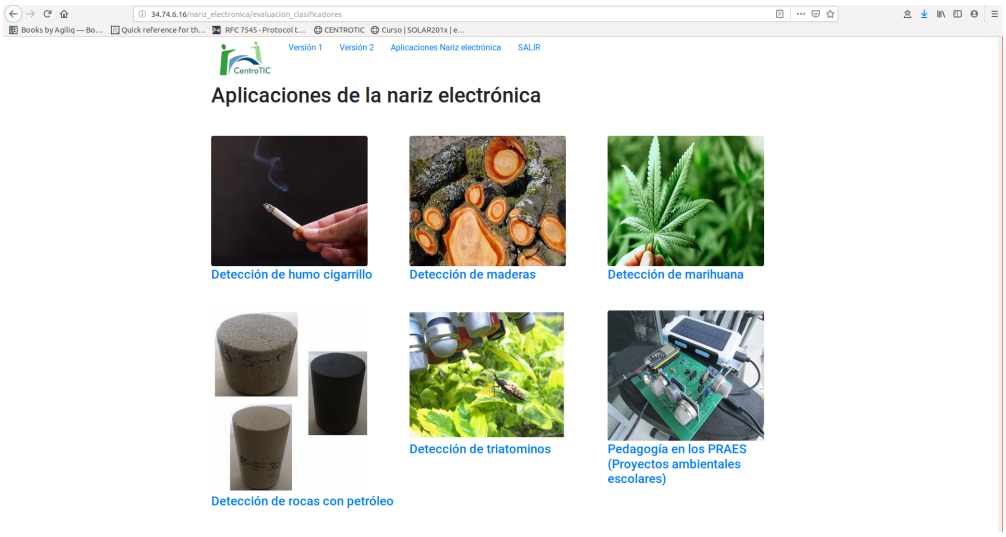
Nota: Modelo de Capas Plataforma IoT. Adaptado de Ruiz Jiménez (2018)

para la aplicación de detección de maderas, como se observa en la Figura 23, donde también se pueden ver otras aplicaciones relacionadas al uso de narices electrónicas.

En 2018, en el marco del evento *U18* organizado por la UIS, se presentó la validación del los uso de la nariz electrónica como dispositivo IoT. En conjunto con el "*Centro Nacional de Investigaciones para la Agroindustrialización de Especies Vegetales Aromáticas y Medicinas Tropicales*" (CENIVAM), se realizó una demostración que incluyó la detección de aceites esenciales extraídos por el mencionado centro de investigación. La demostración fue exitosa y, en este *enlace*, se encuentran algunas de las evidencias del trabajo realizado.

Figura 23

Aplicaciones IoT de la nariz electrónica.



Nota: Aplicaciones IoT de la nariz electrónica.

Apéndice B. Algoritmo de *Expectation Maximization*

El algoritmo de Esperanza Maximización (en inglés *Expectation Maximization*, EM) es un método iterativo que permite encontrar una estimación de los parámetros de un modelo estadístico, con la máxima verosimilitud (o máximo a posteriori, MAP), en casos donde el modelo depende de variables latentes (no observadas). La iteración consiste en alternar entre un paso de esperanza (*Expectation*, E), donde se crea una función para la esperanza de la verosimilitud (*Log-Likelihood*) evaluada con la estimación de parámetros actual, y un paso de maximización (*maximization*, M), donde se calculan los parámetros que maximizan la verosimilitud (*Log-Likelihood*) encontrada en el paso E. Con esta estimación de parámetros, se determina la distribución de las variables latentes en el siguiente paso E (Bishop, 2006).

Dada una distribución de probabilidad conjunta $p(\mathbf{X}, \mathbf{Z}|\theta)$ para un conjunto de variables

observadas $\mathbf{X}$ y no observadas $\mathbf{Z}$, definida por parámetros θ, el objetivo es maximizar (con respecto a θ) la función de verosimilitud dada por

$$p(\mathbf{X}|\theta) = \sum_{Z} p(\mathbf{X}, \mathbf{Z}|\theta), \tag{11}$$

asumiendo que Z es discreta. El algoritmo se resume en:

1. Escoger un conjunto inicial de parámetros θ^{old}.

2. **Paso E:** Evaluar $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.

3. **Paso M:** Evaluar θ^{new}, dado por

$$\theta^{new} = \max_{\theta} \quad \mathscr{Q}(\theta, \theta^{old}) \tag{12}$$

   donde

$$\mathscr{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta). \tag{13}$$

4. Revisar el criterio de convergencia y, si no se satisface, hacer

$$\theta^{old} \leftarrow \theta^{new} \tag{14}$$

   y volver al paso 2.