Identificación de especies de maderas locales mediante el uso de nariz electrónica y aprendizaje automático: Un experimento preliminar

Identification of local wood species by using electronic nose and machine learning: A preliminary experiment

DOI: http://doi.org/10.17981/ingecuc.17.1.2021.15

Artículo de Investigación Científica. Fecha de Recepción: 13/07/2019. Fecha de Aceptación: 19/10/2020.

Naren Arley Mantilla Ramírez
Universidad Industrial de Santander. Bucaramanga (Colombia)
naren.mantilla@correo.uis.edu.co

Luisa Fernanda Ruiz Jiménez
Universidad Industrial de Santander. Bucaramanga (Colombia)
luisafernandarj@gmail.com

Homero Ortega Boada
Universidad Industrial de Santander. Bucaramanga (Colombia)
hortegab@uis.edu.co

Alexander Sepúlveda Sepúlveda
Universidad Industrial de Santander. Bucaramanga (Colombia)
alexander.sepulveda@gmail.com

Para citar este artículo:
N. Mantilla Ramírez, L. Ruiz Jiménez, H. Ortega Boada & A. Sepúlveda Sepúlveda, "Identificación de especies de maderas locales mediante el uso de nariz electrónica y aprendizaje automático: un experimento preliminar", *INGECUC*, vol. 17. no. 1, pp. 188–205. DOI: http://doi.org/10.17981/ingecuc.17.1.2021.15

Resumen

Introducción— 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.

Objetivo— 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.

Metodología— 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 las condiciones de recolección de muestras son más cercanas a las encontradas en entornos reales para los cuales este trabajo busca resolver el problema. Además, el número de muestras es mayor y más variado. Sin embargo, el número de muestras recolectadas para cada especie no está balanceado; por lo tanto, se aplica una técnica de aumento de datos para compensar el desequilibrio en las clases.

Resultados— Al realizar los experimentos se encuentra un desempeño de aproximadamente 80%.

Conclusiones— A pesar de los resultados prometedores, se deben realizar mayores esfuerzos para obtener un mejor desempeño.

Palabras clave— Identificación de madera; nariz electrónica; matriz de sensores químicos; aplicaciones de aprendizaje automático; Clasificación de Vectores de Soporte (SVM); aumento de datos

Abstract

Introduction— 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.

Objective— In present work it is proposed to use electronic noses (chemical sensor arrays) in order to quickly identify wood species, from the volatile compounds their timbers emanate.

Methodology— 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 contrast to previous works, in present work the samples collection conditions are closer to those found on real environments for which this work seeks to solve the problem. In addition, the number of samples is larger and more varied. However, the number of samples collected for each species is not balanced; thus, a data augmentation technique is applied to compensate the class imbalance.

Results— When carrying out the experiments, a performance of approximately 80% is found.

Conclusions— Although the promising results, greater efforts must be carried out in order to obtain a better performance

Keywords— Wood identification; Electronic Nose (E-Nose); chemical sensor arrays; machine learning applications; Support Vector Classification (SVM); data augmentation

I. Introducción

La extracción insostenible y desordenada de madera es uno de los motores de la deforestación y el cambio climático tanto en Colombia como en el mundo. Además, por escaso conocimiento y/o por tradición, el aprovechamiento de este recurso se hace dentro de la ilegalidad y de una manera selectiva, poniendo en peligro algunas especies vulnerables. Aunque existen campañas de las autoridades y corporaciones ambientales que buscan resolver el problema de la ilegalidad, hace falta contar con instrumentos de monitoreo que apoyen los procesos de vigilancia y control, para así dotar de herramientas a las autoridades responsables.

Existen diferentes métodos de identificación de maderas, entre los que se destacan aquellos basados en analizar propiedades organolépticas y macroscópicas como el color y el olor [1], lo cual permite que la madera se analice rápidamente y en grandes volúmenes. También existen métodos más precisos basados en análisis taxonómicos y genéticos, en los cuales se toman muestras de las especies de interés para ser comparadas a nivel de sus secuencias genéticas [2], [3]. La confiabilidad de estas pruebas es casi del 100%; sin embargo, son costosas, demoradas y deben ser realizadas por expertos localizados en las principales ciudades. Otras técnicas utilizadas involucran diferentes análisis espectroscópicos [4], [5] y de imágenes [6], las cuales aún requieren el apoyo de expertos y requieren de un tiempo considerable. Aunque son técnicas eficaces, no cumplen con los requisitos necesarios para poder ser aplicadas en regiones sub-urbanas y rurales apartadas de ciudades principales [7].

De manera alternativa, se ha propuesto analizar los compuestos volátiles emitidos por las especies de madera mediante el uso de estrategias como la cromatografía de gases [8]-[10]. Pero aún es una opción costosa. Por el contrario, una opción menos costosa y más práctica es el uso de narices electrónicas [7], [11]. Los sistemas de olfato electrónico se han venido usando en un creciente número de aplicaciones en la industria de alimentos [12], análisis de la calidad del aire [13], detección de explosivos y narcóticos [14], [15], entre otras aplicaciones. También se ha planteado utilizarlas para la identificación de especies maderables a partir de los compuestos volátiles que estas emanan [7], [16], [17]. En lo que se refiere al tipo de sensores, se han utilizado varios tipos: *Cyranose 320* [18]; *Aromascan A32S* [11]; arreglo de 8 sensores químicos con principios resistivos basados en nanotubos de carbon [7]; y arreglo de 4 sensores de polímero conductor con principio resistivo [16].

Entre los trabajos más relevantes del uso de narices electrónicas para la identificación de madera, lo más destacados se mencionan a continuación. En primer lugar, ciertos autores analizaron 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), mostrando diferencias entre estas especies [18]. Posteriormente, se investigó el uso de redes neuronales para identificar diferencias entre especies pertenecientes a la misma familia y género, consiguiendo tasas de identificación desde el 94% hasta el 99% [11]. El set de datos se construyó con dos muestras por árbol, pertenecientes a entre 13 y 30 árboles de 12 especies.

En Brasil se realizó la clasificación de especies maderables con narices electrónicas [16]. El análisis se enfocó en cuatro especies maderables comúnmente explotadas en ese país: caoba vs Cedro español, y, nuez brasileña vs canela negra; mostrando resultados satisfactorios (entre el 94% y el 100%) después de un análisis de componentes principales (PCA). Y finalmente se analizó la separabilidad de clases de cinco especies maderables en Filipinas representadas por mediciones entregadas por sensores de gases [7]. También reportan agrupaciones (clusters) separables a simple vista sobre características obtenidas mediante análisis de componentes principales (PCA, *Principal Component Analysis*). No obstante, las muestras tomadas fueron recogidas en una zona específica, haciéndolas poco diversas, además de utilizar una reducida cantidad de estas.

Más allá de los buenos resultados reportados en los trabajos mencionados, en todos ellos se realizaron experimentos sobre muestras muy específicas, sin tener en cuenta las posibles interferencias o problemas que se podrían presentar en una situación práctica. 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. Para ello, en la sección de metodología, 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, en la sección de resultados, se presenta el desempeño obtenido con diferentes subgrupos de parámetros seleccionados como predictores. Este desempeño se discute en la sección de análisis de resultados, que sirve como antesala a la presentación de conclusiones y el planteamiento de posibles trabajos a futuro.

II. Método

En la Fig. 1 se muestran las principales fases de un sistema de olfato electrónico o nariz electrónica [19]. La fase química corresponde a la interacción de los compuestos volátiles con el arreglo de sensores de gas. En la fase electrónica, se adquieren las señales eléctricas y se acondicionan para obtener una representación matricial temporal de la muestra. Finalmente, estos datos son procesados por algoritmos de reconocimiento de patrones identificar el tipo de madera.

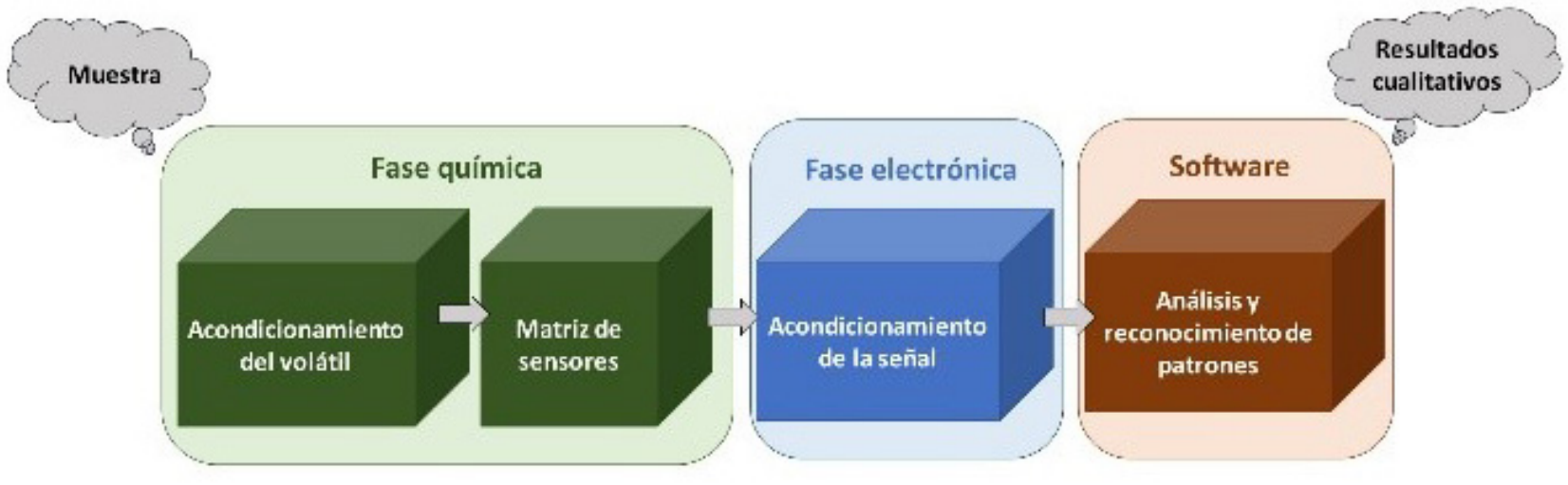

Fig. 1. Esquema general de una nariz electrónica típica.
Fuente: Adaptado de [19].

A. *Arreglo de Sensores Químicos de la Nariz Electrónica*

Existen diferentes tipos de sensores 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*) [20]. 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.

Para este trabajo se utilizó un prototipo de nariz electrónica de laboratorio, que se muestra en la Fig. 2; desarrollado en la Universidad Industrial de Santander bajo la cultura *DIY* (*Do it yourself, hágalo usted mismo*) [19], y que se puede consultar en línea 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.

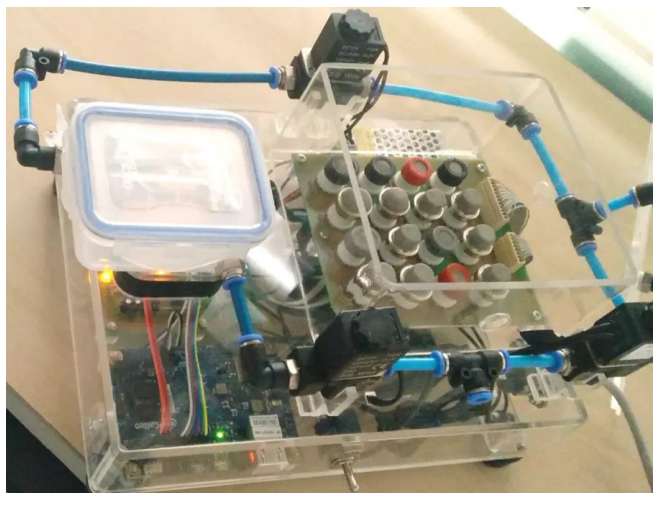

Fig. 2. Prototipo de nariz electrónica desarrollado en la UIS.
Fuente: [19].

El prototipo está compuesto por una tarjeta de adquisición *Intel Galileo Generation* 1 [19], para la adquisición y acondicionamiento de las señales de cada sensor (Módulo de adquisición de datos, DAQ). Además, cuenta con una matriz de 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 [20].

Los sensores conforman el módulo de sensado que, además, incluye 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 mayoría 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 [19]. Por ejemplo, los sensores MQ6 y MQ135 de la marca Hanwei, y el sensor TGS813 de la marca Figaro, son utilizados para la detección de gases en general para estimar 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).

Tabla 1. Sensores del prototipo de nariz electrónica desarrollado por la UIS.

Sensor	Marca	Referencia
1	Hanwei Electronics	MQ 2
2	Hanwei Electronics	MQ 3
3	Hanwei Electronics	MQ 4
4	Hanwei Electronics	MQ 6
5	Hanwei Electronics	MQ 7
6	Hanwei Electronics	MQ 8
7	Hanwei Electronics	MQ 135
8	Hanwei Electronics	MQ 9
9	Figaro Engineering	TGS 832
10	Hanwei Electronics	MQ 6
11	Figaro Engineering	TGS 823
12	Figaro Engineering	TGS 816
13	Figaro Engineering	TGS 822
14	Figaro Engineering	TGS 813
15	Figaro Engineering	TGS 826
16	Hanwei Electronics	MQ 3

Fuente: [19].

De otra parte, los sensores MQ7 (Hanwei), TGS832 y TGS826 (Figaro) son comúnmente utilizados para la detección de un gas o un grupo de gases de interés, como el amoníaco, el CO_2 o gases refrigerantes. Además, otra gama de sensores es utilizada para la detección de alcoholes, debido a su sensibilidad a estos compuestos. 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 MQ2, MQ4 y MQ9 (Hanwei), que son especialmente sensibles a estos compuestos.

En el diagrama de la Fig. 3, se muestra el esquema de funcionamiento del dispositivo utilizado descrito en [19]. 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, es el encargado de controlar el uso de las electroválvulas y de almacenar los datos de los sensores (módulo de sensado).

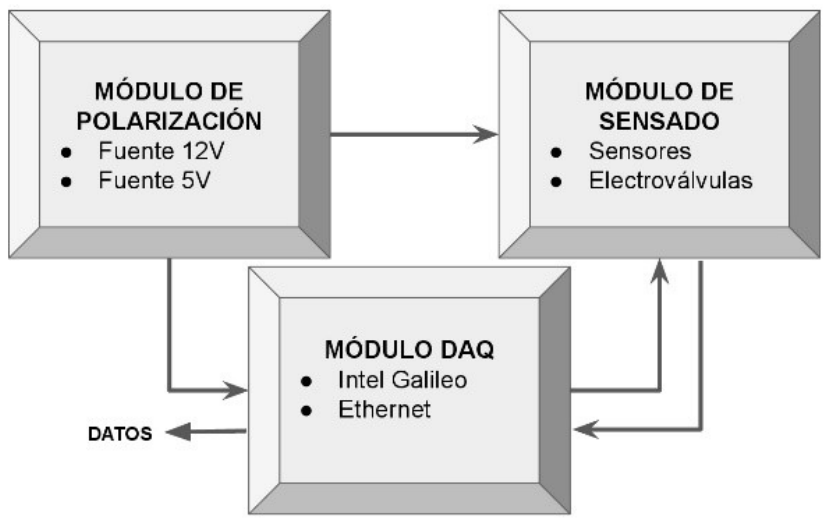

Fig. 3. Diagrama de funcionamiento del prototipo de nariz electrónica desarrollado en la UIS.
Fuente: Autores.

B. *Procedimiento de toma 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. Sin embargo, para esta investigación exploratoria se han utilizado muestras de madera provenientes de aquellas especies más comunes. Estas son: cedro (84 muestras), móncoro (47 muestras), pino (27 muestras) y sapan (43 muestras), para un total de 201 muestras.

El desarrollo del experimento para la medición tiene dos tareas previas: primero se enciende la nariz electrónica durante una hora para que los sensores alcancen su operación de estado estable en el ambiente correspondiente; luego, 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.

Luego se realiza el experimento en sí, con un ensayo por cada muestra de madera. En cada ensayo se ha seguido el procedimiento descrito a continuación: se cepilla la muestra otras 20 veces; se toma aproximadamente 1 cm^3 de la viruta de madera resultante y se ingresa a la nariz electrónica. El objetivo 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. El resultado de esta toma de datos es un conjunto de 16 curvas de respuesta, que corresponden a las 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. La curva de respuesta de cada uno de los 16 sensores se divide en tres fases: lectura base, muestra y recuperación (Fig. 4). En la primera fase, los sensores reaccionan al aire durante 100 muestras; luego, las virutas de madera correspondientes se colocan 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

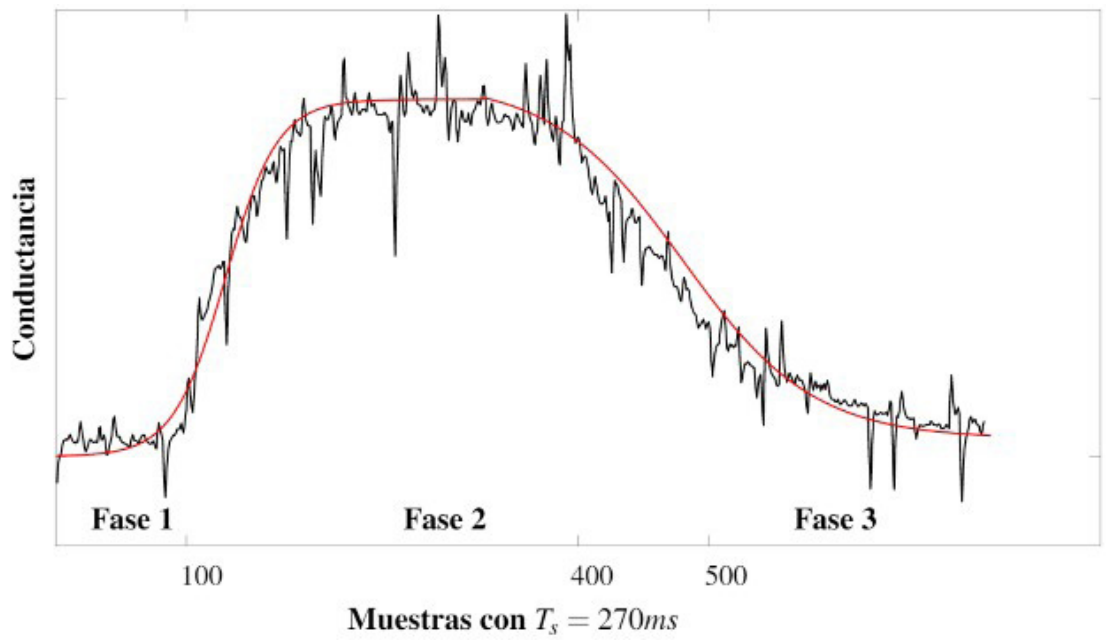

Fig. 4. Forma de respuesta típica de los sensores del arreglo. En color negro se observa un ejemplo de respuesta obtenida y en color rojo la respuesta esperada.
Fuente: Autores.

C. *Extracción de características*

Con el objetivo de reducir el efecto ruido electrónico en el sistema de adquisición, se realiza un pre-procesamiento de los datos, que consiste en el uso de filtro de mediana de orden 5 para cada una de las curvas de respuesta de los sensores. Diferentes características pueden ser estimadas a partir de la curva de respuesta del valor de conductancia de cada sensor. En particular, en trabajos anteriores se reporta el uso de características relacionados con los valores máximo, mínimo y área bajo la curva [21]. Otra forma es utilizando estrategias que involucran un análisis de la respuesta transitoria de los sensores [5], [22]. Y finalmente existe una tercera forma donde se realiza un ajuste a modelos predefinidos [23]. 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 tiene secuencias de datos medidos de dimensión notablemente más grande que la cantidad de muestras de maderas. En el presente trabajo se estiman las siguientes características:

- G_0, valor de conductancia inicial resultado de promediar las primeras 100 muestras de la respuesta total.
- G_F, valor de la conductancia final resultado de promediar las últimas 50 muestras de la fase 2 de la respuesta total.
- G_{MAX}, valores de conductancia máxima.
- G_{MIN}, valores de conductancia mínima.
- B, coeficientes de ganancia; y A, ubicación del polo, correspondientes al modelo auto-regresivo de primer orden ajustado (1):

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

Las primeras 4 características extraídas, 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. 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 X_{201X96} (201 ensayos de dimensión 96). Esta matriz aumenta de tamaño (número de filas) cuando se aplica la técnica de aumento de datos.

D. *Aumento de Datos*

Las dificultades para la recolección de muestras generan dos grandes problemas: desbalance de las clases y tamaño reducido del set de datos. El problema de las clases no balanceadas se puede abordar de diferentes maneras, como: generando datos sintéticos, haciendo un sobre-muestreo de la clase minoritaria o un sub-muestreo 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 [24]-[26]. SMOTE [27] es una técnica de sobre-muestreo 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 [27]. Estas nuevas instancias se localizan sobre líneas 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 línea (de forma imaginaria) desde el dato real hasta cada uno de sus vecinos y, sobre estas líneas, 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 en este trabajo para lidiar con los problemas de la poca cantidad de datos y el desbalance de las clases. En particular, 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 cada clase, 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 matriz $\mathbf{X}_{672X96.}$

E. *Análisis de Componentes Principales*

En trabajos reportados previamente, típicamente se utiliza análisis de componentes principales (PCA, *Principal Component Analysis*) con el fin de reducir la dimensión del problema y evitar sobreajuste [16], [28]. Aunque la forma de extracción de características previamente expuesta logra reducir notablemente la dimensionalidad de los datos de entrada, ello no resulta suficiente; por tanto, se recurre a la técnica de PCA a fin de lograr una reducción adicional. El análisis de componentes principales permite reducir la dimensión de observaciones $x_i \in R^p$ mediante una transformación lineal V_q que mapea los datos a un nuevo espacio de dimensión $q \leq p$ donde las nuevas variables son no correlacionadas, al tiempo que conserva la mayoría de la variabilidad de los datos originales [29], [30].

Las columnas de la nueva matriz transformada, denominados 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 esta aplicación, se procedió a usar máximo $q = 4$ componentes principales, con los cuales se representa aproximadamente el 90% de la varianza de los datos originales.

F. *Clasificación por 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, se escogió la

clasificación por vectores de soporte (SVM, *Support Vector Machines*), un método popular para resolver problemas de clasificación. La técnica de 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 [31].

Si buscamos fronteras de tipo no lineal entre clases, en lugar de utilizar y_i como entrada usamos $h(y_i) = (h_1(y_i), h_2(y_i), ..., h_M(y_i))$ para $i = 1, ...N$. Con ello se produce la función no lineal de separación $\hat{\delta}(y) = h(y)^T \hat{\beta} + \hat{\beta_0}$, y la función de decisión estaría dada por $\hat{G}(y) = sgn(\hat{\delta}(y))$. Esta técnica crea hiperplanos que maximizan la separabilidad entre conjunto de datos, a través de funciones *kernel* [30]. En este trabajo se usa a modo de función *kernel* la función Gaussiana. Para configurar este clasificador en una aplicación de varias clases (4 clases para este problema), se enfrentan dos clases entre sí y se realiza un entrenamiento progresivo hasta que se logra un ajuste óptimo, con un rendimiento verificado a través de la matriz de confusión.

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.

III. Resultados

Los experimentos de clasificación se llevaron a cabo para las 4 diferentes configuraciones de características de entrada, tal como se muestra en la Tabla 2. Los parámetros relacionados en cada una de las configuraciones son estimados para cada sensor. En particular, para el caso de la configuración 1 se estima un total de 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 predictores. En resumen, se obtienen 4 matrices de características con los siguientes tamaños: $\mathbf{X}_{672X96}$, para la configuración 1, $\mathbf{X}_{672X64}$ para la configuración 2, $\mathbf{X}_{672X32}$ para la configuración 3, y $\mathbf{X}_{672X32}$. para la configuración 4.

Tabla 2. Combinación de diferentes características extraídas.

Configuración	Características Utilizadas
Configuración 1	G0, GF, GMAX, GMIN, A, y B
Configuración 2	G0, GF, GMAX, y GMIN
Configuración 3	A, y B
Configuración 4	GF y GMAX

Fuente: Autores.

Luego, para cada matriz de características de cada configuración se aplicó el análisis de componentes principales (PCA), y los componentes resultantes se ordenaron según su varianza. En la Fig. 5, 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. También se hace la prueba con 3 componentes principales.

En la Fig. 6 se presenta la distribución de datos para tres componentes principales resultantes de la configuración 1, para las tres clases analizadas: 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.

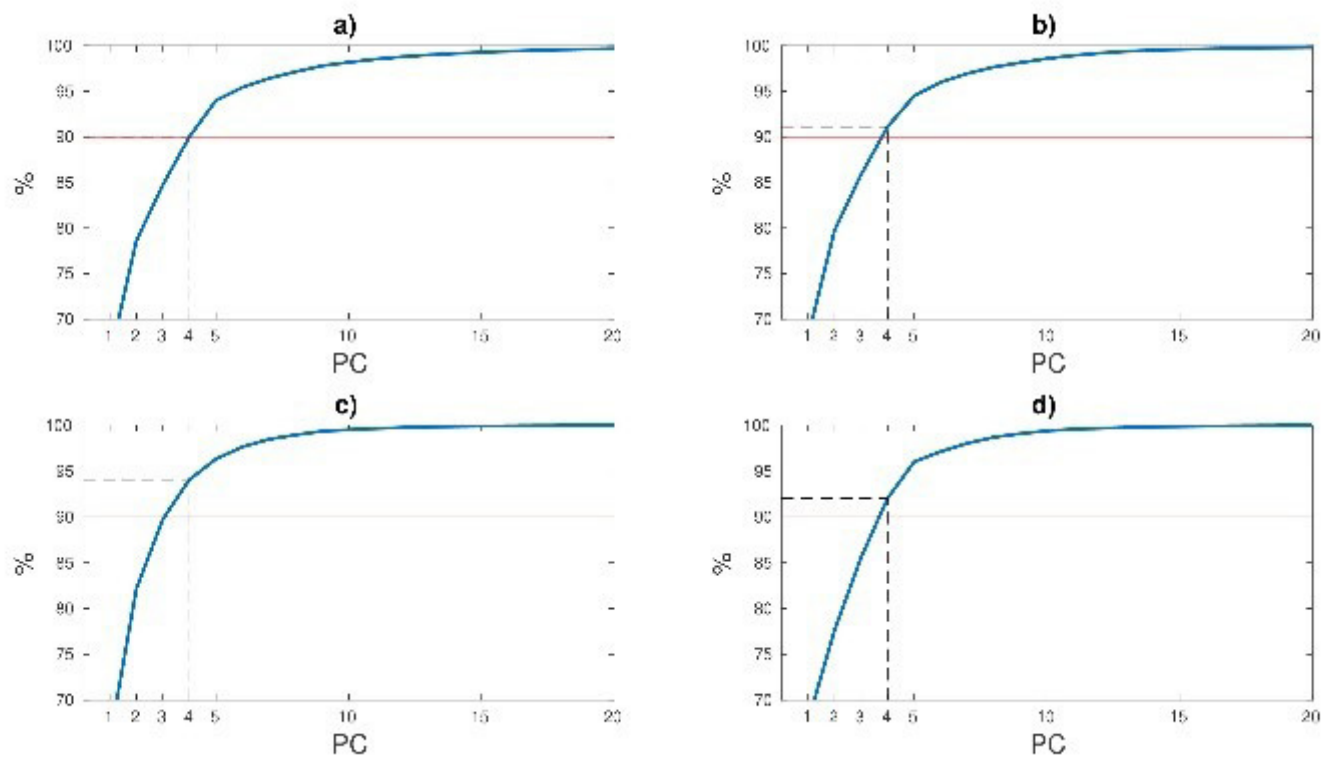

Fig. 5. Varianza acumulada en los 20 primeros componentes principales (PC) para la configuración 1 (a), para la configuración 2 (b), para la configuración 3 (c) y para la configuración 4 (d).
Fuente: Autores.

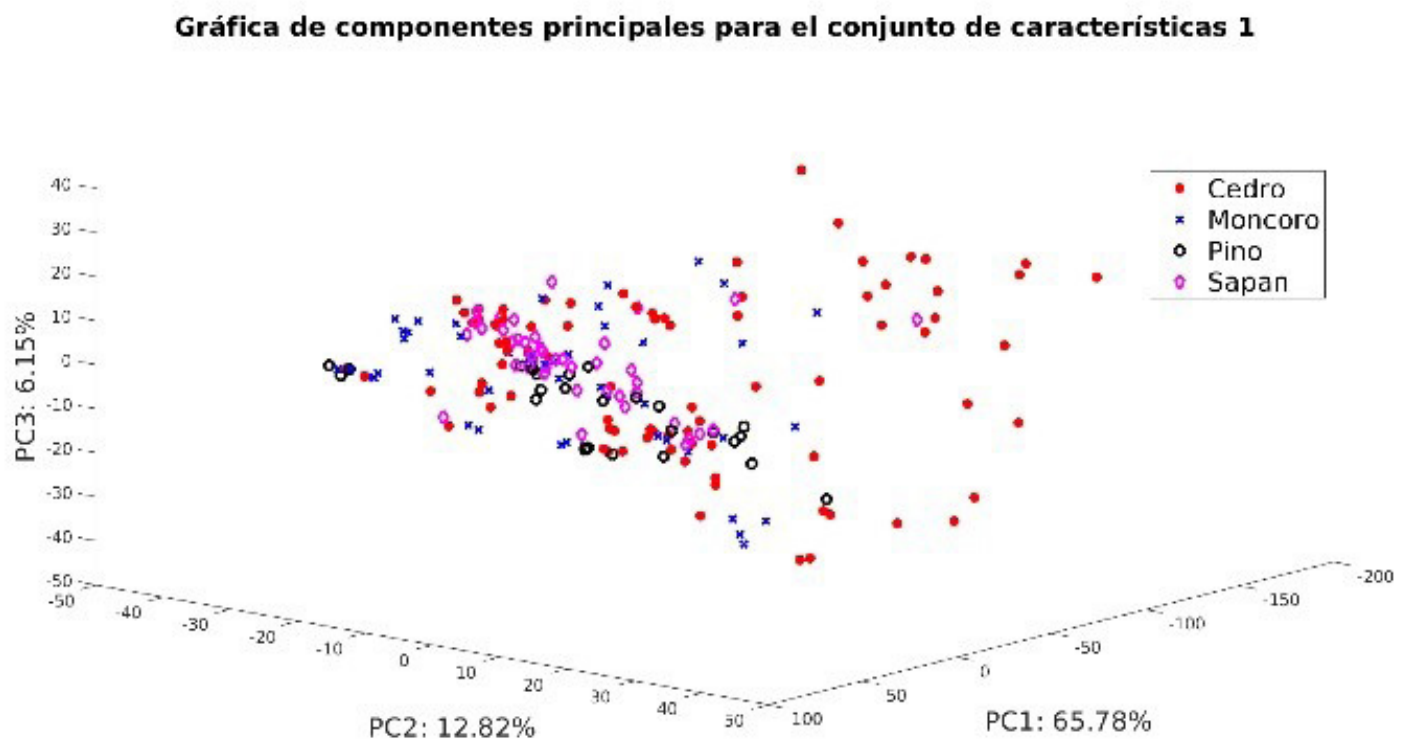

Fig. 6. Visualización 3D de los datos después de aplicar el análisis de componentes principales (PCA), para la configuración 1.
Fuente: Autores.

De otra parte, con tres componentes principales ya se tiene más del 80% de la varianza para todas las configuraciones, a primera vista no se observa claramente separabilidad entre las tres clases analizadas, tal como lo muestra la Fig. 5. Se tienen observaciones similares para las restantes 3 configuraciones.

Así mismo, en la Fig. 7, se presenta un gráfico de loadings para visualizar cuáles sensores son más relevantes para la aplicación de interés. La gráfica 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 la ubicación del polo B los sensores 3 y 5.

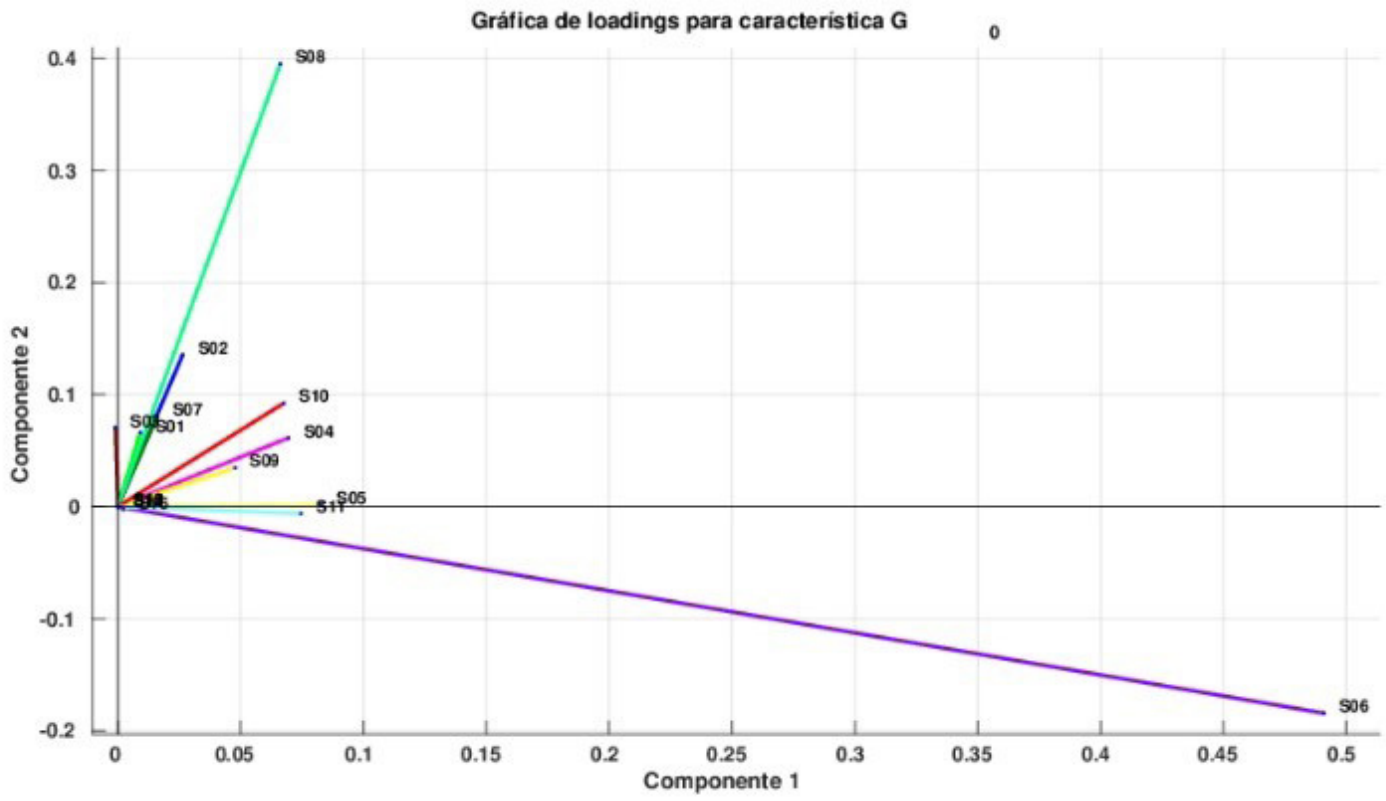

Fig. 7. Gráfica de Loadings para la característica G0 (conductancia inicial) en los 16 sensores.
Fuente: Autores.

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-folds) donde k = 10 y con *boostraping* de 100 veces. Los resultados se pueden observar en la Tabla 3 y Tabla 4.

Tabla 3. Tasas de error resultado del proceso de validación cruzada (k-folds, con K = 10) para diferentes tipos de características de entrada con el set de datos aumentado. PC = 3.

Configuración	# PC	Error Promedio	Desviación estándar
Configuración 1	3	27.22%	5.16%
Configuración 2	3	25.62%	5.02%
Configuración 3	3	26.79%	5.15%
Configuración 4	3	24.66%	4.75%

Fuente: Autores.

Tabla 4. Tasas de error resultado del proceso de validación cruzada (k-folds, con K = 10) para diferentes tipos de características de entrada con el set de datos aumentado. PC = 4.

Configuración	# PC	Error Promedio	Desviación estándar
Configuración 1	4	23.14%	4.99%
Configuración 2	4	21.59%	4.91%
Configuración 3	4	22.49%	5.08%
Configuración 4	4	20.24%	4.62%

Fuente: Autores.

IV. Discusión de Resultados

Se aplicó la técnica de análisis de componentes principales a un conjunto de datos aumentado a partir de 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 [12]-[15], [19], 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 [7], [16], [17]. Sin embargo,

las condiciones de nuestro experimento son más cercanas de las condiciones prácticas. Esto se aprecia en la forma de recolectar los datos, cantidad de muestras, variedad de lugares procedencia, especies, tratamiento previo y almacenamiento previo de las muestras.

A modo de ejemplo, en [18] se realizaron 60 mediciones provenientes de 18 muestras de madera. En [11] se tomaron 10 muestras por especie para 23 especies diferentes, lo cual corresponde a un número reducido por especie. En [7] 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 [16] 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 [7], [11], [18] las muestras fueron almacenadas herméticamente y congeladas hasta el momento y lugar del experimento; y, en [16], las muestras fueron almacenadas herméticamente en un laboratorio pero no congeladas. En contraste, para 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.

V. Conclusiones y Trabajo Futuro

En el presente trabajo se desarrolló un sistema de identificación de maderas basado en su olor que, a diferencia de trabajos anteriores, se desarrolló bajo condiciones más exigentes y más cercanas a las de los entornos de trabajo en los cuales se pretende usar este tipo de sistemas. Al realizar los experimentos se encuentra un desempeño de alrededor de 79.65%, lo cual indica que han de realizarse mayores esfuerzos a fin de obtener un mejor desempeño. El uso de narices electrónicas para la identificación y detección de maderas es un área poco explorada y el presente trabajo constituye uno dentro los pocos reportados, aún más si lo acotamos a Colombia.

De acuerdo a la Tabla 1 y a la información obtenida de los gráficos de *loadings* como el presentado en la Fig. 7, 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 que, al responder a determinados compuestos volátiles, aportan mayor información desde el punto de vista de separabilidad de las maderas. Cabe mencionar que el criterio utilizado para la selección de los sensores fue una mezcla entre variedad en el tipo de sensores y disponibilidad de los mismos en el mercado debido al enfoque inicial del trabajo. Sin embargo, 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 deja como trabajo futuro.

También, a modo de trabajo futuro, se plantea avanzar en la solución de interrogantes relacionados con la selección de características, tipos de sensores y estrategias alternas de aprendizaje automático. Dado que 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.

De otra parte, es importante tener en cuenta que a futuro lo que se busca es poder detectar (verificar) especies maderables vulnerables y prohibidas, lo cual es un problema de tipo abierto, en contraste a la identificación. En particular, la identificación de maderas es un problema de tipo cerrado en el sentido de que el dispositivo ha de escoger entre alguna de N clases posibles conocidas, 4 en el presente caso. Sin embargo, la detección de maderas debería abordarse como un problema de tipo abierto, en analogía con los sistemas biométricos. Es decir, el clasificador ha de informar si la muestra corresponde a una especie protegida en particular, o, si es alguna otra de identidad desconocida que podría llegar incluso a no ser parte del conjunto de datos de entrenamiento.

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

Referencias

[1] E. A. Wheeler & P. Baas, "Wood identification-a review," *IAWA J*, vol. 19, no. 3, pp. 241–264, 1998. Available: https://brill.com/view/journals/iawa/19/3/article-p241_2.xml?language=en

[2] F. Hanssen, N. Wischnewski, U. Moreth & E. A. Magel, "Molecular identification of Fitzroya cupressoides, Sequoia sempervirens, and Thuja plicata wood using taxon-specific rDNA-ITS primers," *IAWA J*, vol. 32, no. 2, pp. 273–283, 2011. https://doi.org/10.1163/22941932-90000057

[3] M. Yu, K. Liu, L. Zhou, L. Zhao & S. Liu, "Testing three proposed DNA barcodes for the wood identification of Dalbergia odorifera T. Chen and Dalbergia tonkinensis Prain," *Holzforschung*, vol. 70, no. 2, pp. 127–136, 2016. https://doi.org/10.1515/hf-2014-0234

[4] E. C. Cabral, R. C. Simas, V. G. Santos, C. L. Queiroga, V. S. da Cunha, G. F. de Sá, R. J. Daroda & M. N. Eberlin, "Wood typification by Venturi easy ambient sonic spray ionization mass spectrometry: The case of the endangered Mahogany tree," *J. Mass Spectrom*, vol. 47, no. 1, pp. 1–6, 2012. https://doi.org/10.1002/jms.2016

[5] R. Rana, G. Müller, A. Naumann & A. Polle, "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*, vol. 62, no. 5, pp. 530–538, 2008. https://doi.org/10.1515/HF.2008.104

[6] A. Dickson, B. Nanayakkara, D. Sellier, D. Meason, L. Donaldson & R. Brownlie, "Fluorescence imaging of cambial zones to study wood formation in Pinus radiata D. Don," *Trees - Struct Funct*, vol. 31, no. 2, pp. 479–490, 2017. https://doi.org/10.1007/s00468-016-1469-3

[7] J. M. Kalaw & F. B. Sevilla, "Discrimination of wood species based on a carbon nanotube/polymer composite chemiresistor array," *Holzforschung*, vol. 72, no. 3, pp. 215–223, 2018. https://doi.org/10.1515/hf-2017-0097

[8] R. Fedele, I. E. Galbally, N. Porter, and I. A. Weeks, "Biogenic VOC emissions from fresh leaf mulch and wood chips of Grevillea robusta (Australian Silky Oak)," *Atmos Environ*, vol. 41, no. 38, pp. 8736–8746. Dec. 2007. https://doi.org/10.1016/j.atmosenv.2007.07.037

[9] K. Müller, S. Haferkorna, W. Grabmer, A. Wisthaler, A. Hansel, J. Kreuzwieser, C. Cojocariu, H. Rennenberg & H. Herrmanna, "Biogenic carbonyl compounds within and above a coniferous forest in Germany," *Atmos Environ*, vol. 40, No. 1, pp. 81–91, 2006. https://doi.org/10.1016/j.atmosenv.2005.10.070

[10] H. J. I. Rinne, A. B. Guenther, J. P. Greenberg & P. C. Harley, "Isoprene and monoterpene fluxes measured above Amazonian rainforest and their dependence on light and temperature," *Atmos Environ*, vol. 36, no. 14, pp. 2421–2426, May. 2002. https://doi.org/10.1016/S1352-2310(01)00523-4

[11] A. D. Wilson, D. G. Lester & C. S. Oberle, "Application of conductive polymer analysis for wood and woody plant identifications," *For Ecol Manage*, vol. 209, no. 3, pp. 207–224, May. 2005. https://doi.org/10.1016/j.foreco.2005.01.030

[12] H. Shi, M. Zhang & B. Adhikari, "Advances of electronic nose and its application in fresh foods: A review," *Crit Rev Food Sci Nutr*, vol. 58, no. 16, pp. 1–11, 2017. https://doi.org/10.1080/10408398.2017.1327419

[13] L. Capelli, S. Sironi & R. Del Rosso, "Electronic Noses for Environmental Monitoring Applications," *Sensors*, vol. 14, no. 11, pp. 19979–20007, 2014. https://doi.org/10.3390/s141119979

[14] L. Guo, Z. Yang & X. Dou, "Artificial Olfactory System for Trace Identification of Explosive Vapors Realized by Optoelectronic Schottky Sensing," *Adv Mater*, vol. 29, no. 5, pp. 1–8, 2017. https://doi.org/10.1002/adma.201604528

[15] J. P. Santos & J. Lozano, "Real time detection of beer defects with a hand held electronic nose ," presented at *10th Spanish Conference on Electron Devices*, CDE, MD, ES, pp. 1–4, 11-13 Feb. 20015. https://doi.org/10.1109/CDE.2015.7087492

[16] J. R. Cordeiro, R. W. C. Li, É. S. Takahashi, G. P. Rehder, G. Ceccantini & J. Gruber, "Wood identification by a portable low-cost polymer-based electronic nose," *RSC Adv*, vol. 6, no. 111, pp. 109945–109949, 2016. https://doi.org/10.1039/c6ra22246c

[17] A. D. Wilson, "Application of a Conductive Polymer Electronic-Nose Device to Identify Aged Woody Samples," *3 IARIA*, Xpert Publishing, RO, IT, pp. 77–82, 2012. Available: https://www.fs.usda.gov/treesearch/pubs/45153

[18] F. X. Garneau, B. Riedl, S. Hobbs, A. Pichette & H. Gagnon, "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 Werkst*, vol. 62, no. 6, pp. 470–473, 2003. https://doi.org/10.1007/s00107-004-0508-8

[19] L. F. Ruiz, "Detección de los insectos de la subfamilia Triatominae basado en narices electrónicas," *tesis maestría*, UIS, BGA, CO, 2018.

[20] Figaro Engineering Inc, "Operating principle," *figaro Engineering*, 2018. Available: https://www.figaro-sensor.com/technicalinfo/principle/mos-type.html

[21] Jia Yan, X. Guo, S. Duan, P. Jia, L. Wang, C Peng & S. Zhang, "Electronic Nose Feature Extraction Methods: A Review," *Sensors*, vol. 15, no. 11, pp. 27804–27831, Nov. 2015. https://doi.org/10.3390/s151127804

[22] I. Rodriguez-Lujan, J. Fonollosa, A. Vergara, M. Homer & R. Huerta, "On the calibration of sensor arrays for pattern recognition using the minimal number of experiments," *Chemom Intell Lab Syst*, vol. 130, pp. 123–134, Jan. 2014. https://doi.org/10.1016/j.chemolab.2013.10.012
[23] L. Carmel, S. Levy, D. Lancet & D. Harel, "A feature extraction method for chemical sensors in electronic noses," *Sens Actuators B:Chem*, vol. 93, no. 1-3, pp. 67–76, Aug. 2003. https://doi.org/10.1016/S0925-4005(03)00247-8
[24] J. Van Hulse, T. M. Khoshgoftaar & A. Napolitano, "Experimental perspectives on learning from imbalanced data," presented at *Proceedings of the 24th international conference on Machine learnin*, ICML, NY, USA., pp. 935–942, Jun. 20, 2007. https://doi.org/10.1145/1273496.1273614
[25] D. A. Cieslak, N. V Chawla & A. Striegel, "Combating imbalance in network intrusion datasets," *IEEE International Conference on Granular Computing*, GRC, ATL, USA, pp. 732–737, 2006. https://doi.org/10.1109/GRC.2006.1635905
[26] R. Blagus & L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC Bioinf*, vol. 11, no. 1, pp. 1–17, 2010. https://doi.org/10.1186/1471-2105-11-523
[27] N. V Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J Artif Intell Res*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953
[28] M. A. Akbar, A. Ait Si Ali, A. Amira, F. Bensaali, M. Benammar, M. Hassan & A. Bermak, "An Empirical Study for PCA and LDA-Based Feature Reduction for Gas Identification," *IEEE Sens J*, vol. 16, no. 14, pp. 5734–5746, 2016. https://doi.org/10.1109/JSEN.2016.2565721
[29] I. Goodfellow, Y. Bengio & A. Courville, *Deep Learning*, CBG: MIT Press, 2016.
[30] J. Friedman, T. Hastie & R. Tibshirani, *The elements of statistical learning*, NY, USA: Springer, 2001.
[31] G. James, D. Witten, T. Hastie & R. Tibshirani, *An introduction to statistical learning*. NY, USA: Springer, 2013.

Naren Mantilla Ramírez es Ingeniero Electrónico de la Universidad Industrial de Santander (Colombia) y actualmente cursa estudios de Maestría en Ingeniería de Telecomunicaciones en la misma Universidad. Sus áreas de interés corresponden a Sistemas de control y aplicaciones de aprendizaje automático. https://orcid.org/0000-0002-3185-7387

Luisa Fernanda Ruiz es Ingeniera Electrónica y Magíster en Ingeniería de Telecomunicaciones de la Universidad Industrial de Santander (Colombia). Actualmente es profesora de la Universidad Manuela Beltrán (Colombia) en el área de Ingeniería Biomédica. Sus áreas de interés corresponden al tratamiento de señales y aplicaciones del aprendizaje automático en sistemas sensoriales. https://orcid.org/0000-0002-0205-7815

Homero Ortega Boada es Doctor en Ciencias de la Ingeniería de la Universidad Internacional de Aviación Civil de Kiev (Ucrania) y conjuga un cúmulo de experiencias en la industria, la regulación del sector de TIC y en investigación científica en temas de comunicaciones. Lo cual se derivan de su paso por la empresa Ericsson, la Universidad Industrial de Santander (UIS) y la Agencia Nacional del Espectro, donde se desempeñó como Asesor de la Dirección General de la Agencia Nacional del Espectro y como subdirector Interino. En la UIS ha sido líder de proyectos, docente investigador, director del Grupo de investigación RadioGis y del CentroTIC. https://orcid.org/0000-0003-0343-862

Alexander Sepúlveda Sepúlveda es Ingeniero Electrónico y Magíster en Automatización Industrial de la Universidad Nacional de Colombia. Con título de doctor en Ingeniería-Automática por parte la misma universidad. Actualmente es profesor asociado de la Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones de la Universidad Industrial de Santander. Sus áreas de interés corresponden al Tratamiento de Señales del Habla y aplicaciones del Aprendizaje Automático. https://orcid.org/0000-0002-9643-5193