$\mathcal{DIFU}_{100}ci@$
Revista de Difusión Científica, Ingeniería y Tecnologías
Vol. 17, No. 1, Enero - Abril 2023
ISSN:2007-3585
Artículo arbitrado
Publicado: 30 de abril de 2023

Classification and detection of powdery mildew damage levels in cucurbits plants through spectral analysis

Clasificación y detección de niveles de daño de cenicilla polvorienta en plantas cucurbitáceas a través de análisis espectral

C. A. Rivera-Romero[*1], **E. R. Palacios-Hernández**[2], **J. U. Muñoz-Minjares**[1], **J. A. Morales-Saldaña**[3], **I. A. Reyes-Portillo**[4], and **M. A. Navarrete-Sánchez**[1]

[1] *Universidad Autónoma de Zacatecas (UAZ), Unidad Académica de Ingeniería Eléctrica Plantel Jalpa,*
Libramiento Jalpa Km., Fraccionamiento Solidaridad, Jalpa, 156+380, 99600, Jalpa, Zacatecas, México.
{c.a.riveraromero, ju.munoz, mnavarrete}@uaz.edu.mx

[2] *Universidad Autónoma de San Luis Potosí (UASLP), Facultad de Ciencias,*
Av. Parque Chapultepec 1570, 78210, San Luis Potosí, S.L.P. México.
epalacios@fciencias.uaslp.mx

[3] *Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí,*
Av. Dr. Manuel Nava No.8 Edificio P, Zona Universitaria, San Luis Potosí, S.L.P., México.
jmorales@uaslp.mx

[4] *Centro de Investigación y Estudios de Posgrado, Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí,*
Av. Dr. Manuel Nava No.8 Edificio P, S. L. P., México,
a318057@alumnos.uaslp.mx

Abstract

In this work, a methodology to detect three damage levels by powdery mildew on cucurbit plants through spectral signatures is proposed. Leaves in fungal germination, leaves with the first symptoms and diseased leaves are considered. A database of spectral signatures with the wavelengths grouped into regions of interest is used. The feature extraction and data reduction is determined with an exploratory analysis and principal component analysis. The wavelength bands with the most descriptive data are between the ranges of $\lambda_{405-679nm}$ for the visible band and $\lambda_{760-988nm}$ for the near infrared band. Then, the detection of damage levels is obtained, with a classification accuracy of 93.2% and a Cohen's kappa value of 0.81. This spectral analysis provides a model that allows features between damage stages for powdery mildew in the cucurbits plants.

Keywords— Spectral signature, feature extraction, principal component, support vector machine

Resumen

Este trabajo, propone una metodología para detectar tres niveles de daño por cenicilla polvorienta en cucurbitáceas a través de firmas espectrales. Se consideran hojas en germinación del hongo, hojas con los primeros síntomas y hojas enfermas. Se utilizó una base de datos de firmas espectrales con las longitudes de onda agrupadas en regiones de interés. Se determinó la extracción de características y la reducción de datos con un análisis exploratorio y un análisis de componentes principales. Las bandas de longitud de onda con los datos más descriptivos están entre los rangos de $\lambda_{405-679nm}$ para la banda visible y $\lambda_{760-988nm}$ para la banda del infrarrojo cercano. Se obtuvo la detección de los niveles de daño, con una precisión de clasificación de $93.2\,\%$ y un valor kappa de 0.81. Este estudio proporciona un modelo que permite establecer características entre las fases de daño de la cenicilla polvorienta en las plantas cucurbitáceas.

Palabras clave— Firma espectral, extracción de características, componentes principales, máquinas de soporte vectorial

[*]Autor de correspondencia

I. Introducción

Los datos espectrales de reflectancia en el espectro-visible (VIS, Visible) y cercano al infrarrojo (NIR, Near Infrared) se han propuesto en un amplio campo de aplicaciones en la agricultura. Estos datos se obtienen a través del muestreo de firmas espectrales en la vegetación, comprendiendo rangos de longitudes de onda desde $\lambda_{250-1100nm}$ y dependiendo del dispositivo utilizado.

Una de los usos principales usos, es la solución de problemas como la identificación de plagas y enfermedades, además, se incluyen en el análisis de diferentes parámetros en las plantas para conocer deficiencias internas. Por lo tanto, la descripción del estado patológico de una planta a través de las propiedades espectrales permite el desarrollo de métodos y algoritmos para un mejor monitoreo de un cultivo [1, 2].

En el caso particular de las hojas de plantas, se han utilizado las firmas espectrales para describir efectos de estrés causados por una enfermedad, plaga o deficiencia de nutrientes, agua, síntomas de marchitamiento, senescencia y cambios en la coloración. Esto es porque las hojas de una planta sana presentan porcentajes de reflectancia variados entre las longitudes de onda del espectro visible (VIS).

Particularmente, las hojas, muestran un aumento en el rango de las longitudes de onda del color verde ($\lambda_{500-550nm}$), debido a la clorofila y las bolsas de aire que se generan en el tejido intermedio. Por lo tanto, se pueden estimar lo niveles de sustancias que forman la estructura interna de la hoja por medio de datos espectrales así como enfermedades.

Por ejemplo, en [3] diferenciaron las fases de infecciones virales en dos variedades de plantas de tabaco (Nicotiana tabacum L.) con reflectancia hiperespectral en las hojas. Las plantas se desarrollaron en condiciones controladas de un invernadero. A las plantas en el estadio de crecimiento que son el desarrollo de hojas y floración, se les introdujeron dos tipos de virus diferentes. Las mediciones de reflectancia se encontraron entre los rangos espectrales de visible y cercano al infrarrojo ($\lambda_{450-850nm}$).

Un prototipo con sensor de reflectancia espectral para seleccionar la aplicación de herbicidas en plantas sin contacto directo fue implementado por [4]. Este trabajo consistió en la toma de medidas de reflectancia espectral de plantas y del suelo. Los datos se obtuvieron por medio de un vehículo de granja que contiene un sensor con un módulo de diodo láser de tres longitudes de onda. La distinción se basó en el cálculo de la pendiente de la respuesta espectral entre los rangos de $\lambda_{635-670nm}$ y de $\lambda_{670-785nm}$ de longitud de onda del láser.

Una aplicación para caracterizar las hojas de la planta de jitomate (Lycopersicon esculentum L.) por minador fue presentada por [5]. Utilizaron espectros de reflectancia para visualizar el rango de longitud de onda de $\lambda_{12500-4000cm^{-1}}$ en dónde se presentó el daño causado por la plaga por medio de un espectrómetro FT-NIR (Fourier Transform – Near Infrared). Entre las longitudes de onda de localización está la de λ_{1450nm} y λ_{1900nm} nm particularmente. Obtuvieron un coeficiente de correlación entre 0.98 y 0.91 entre los niveles del espectro y la infestación.

El contenido de agua en las hojas se analizó por medio de la reflectancia del cercano al infrarrojo y el infrarrojo medio analizando la reflectancia del dosel con respecto a la longitud de onda para analizar la sensibilidad de la reflectancia ante el contenido de agua en las hojas. Entre los resultados muestran que las reflectancias espectrales en $\lambda_{1405\mu m}$, $\lambda_{1875\mu m}$, $\lambda_{2015\mu m}$ y $\lambda_{4375\mu m}$ son las más sensibles al cambio contenido de agua [6].

Además, estudios han especificado por medio de experimentos la importancia de la reflectancia espectral y el dominio en la estructura interna de las hojas [7]. Otro factor interno en las hojas de estudio es el contenido de clorofila, el cual se considera como un indicador importante del crecimiento y la fotosíntesis.

Se ha experimentado la medición espectral del contenido del maíz con tratamiento de inoculación para medir el grado de estrés hídrico [8]. En [9] se probaron espectros de reflectancia ($\lambda_{400-2400nm}$) para distinguir especies y detectar la estructura de la población con rasgos foliares para modelos de clasificación basados en los espectros foliares.

Los modelos identificaron dos especies de arbustos de flor ártico-alpina con una exactitud global del 99,7 %. El análisis espectral ha sido utilizado para el estudio de suelos y la química foliar para especies de coníferas, en donde se encontraron la correlación entre el carbón orgánico y el nitrógeno por los índices espectrales basado en la absorción de la clorofila [10].

Entre otros trabajos en [11], se mostraron resultados de que la reflectancia espectral aumenta en el rango visible e infrarrojo cercano. Esto se debe a los parásitos presentes, considerando el rango $\lambda_{750\sim1400nm}$ sensible de la respuesta espectral. Establecieron un modelo de inversión de clorofila basado en la reflectancia del valle del rojo y como resultado se obtuvieron predicciones del contenido de clorofila bajo diferentes grados de parasitación.

Existe gran diversidad de trabajos y usos de las firmas espectrales para casos muy particulares. Por lo tanto, se propone una nueva metodología en este trabajo. A continuación, se describe la estructura de este estudio en secciones.

En la sección II se describe la base de datos de firmas espectrales que se utilizará y su procesamiento para la

caracterización. Además, se explica como se caracterizan cada nivel de daño y la metodología propuesta, comenzando con un análisis exploratorio de datos.

La sección III describe la parte de clasificación de los niveles de daño. Se inicia con la etapa de extracción de características con el análisis de componentes principales y la selección de componentes. Después, se procede a la clasificación con las máquinas de soporte vectorial y las etapas de entrenamiento y validación de datos para finalizar con la múltiple clasificación.

Para finalizar, en las secciones IV y V, se presentan los resultados de las pruebas de la clasificación de los niveles de daño de la cenicilla polvorienta en plantas cucurbitáceas y la evaluación de desempeño de la clasificación.

II. Base de datos de las firmas espectrales

Se tiene una base de datos formada de firmas espectrales de hojas de plantas cucurbitáceas [12]. Se aplicó un procesamiento a las firmas espectrales con la normalización de la media y el filtro de Saviztky-Golay [13]. Después, se realizó una inspección visual por hoja y por día de muestra.

Para un mejor análisis, se consideró dividir la firma espectral, en rangos de longitudes de onda, para observar los cambios de forma mas detallada entre los porcentajes de reflectancia. Los rangos a considerar fueron las longitudes de onda pertenecientes el espectro del ultravioleta (UV) tomando del $\lambda_{200-400nm}$.

Después se continuó con el rango del visible (VIS), en los cuales los espectros del azul, verde, amarillo y naranja tienen presencia hasta llegar al rojo ($\lambda_{400-700nm}$). Finalmente, se tomo el resto de las longitudes de onda, que conforman la parte del cercano al infrarrojo (NIR), con los rangos del $\lambda_{700-1014nm}$.

En la Fig. 1 se identifican las longitudes de onda que presentan mayor cambio entre los datos en los diferentes días de muestra con la hoja etiquetada como H_5 y los días de muestra (D_1-D_{19}). La finalidad de este análisis es establecer los rangos de longitudes de onda relevantes en el espectro del visible y cercano al infrarrojo que determine características entre los diferentes estados de la enfermedad.

De acuerdo al número de mediciones, se hizo la comparación por hoja y con un rango de entre tres a cinco días para tener mejor visualización de los datos.

La Tabla 1 presenta ejemplos de la inspección visual en las hojas H_5 y H_7 en cuanto a las longitudes de onda óptimas que presentan características en las hojas en diferentes niveles de daño de la enfermedad de cenicilla polvorienta.

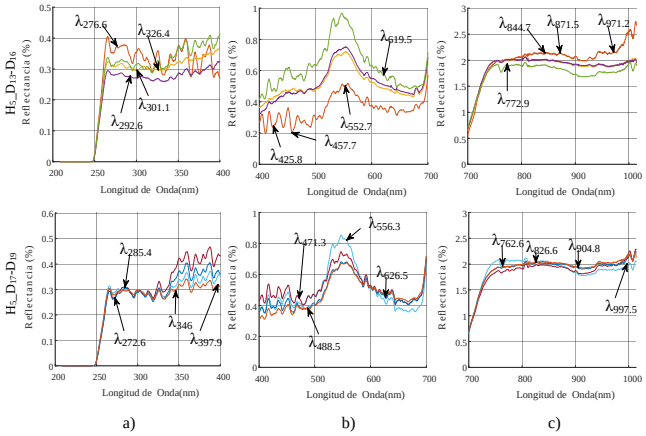

Figura 1: *Inspección visual de las firmas espectrales en la hoja H_5 de diferentes días de muestra (D_1-D_{19}) y dividida en rangos de longitud de onda: a) firma espectral ($\lambda_{250-1014nm}$), b) rango $\lambda_{200-400nm}$, c) rango $\lambda_{400-700nm}$, y d) rango $\lambda_{700-1014nm}$.*

Tabla 1: *Inspección visual en las firmas espectrales considerando rangos de las longitudes de onda (400-700 nm y 700-1000 nm) con diferencias en los porcentajes de reflectancia de las hojas H_5 y H_7.*

Hoja x DM	$\lambda_{400-700nm}$	$\lambda_{700-1000nm}$
H_5x$D_{13} - D_{16}$	$\lambda_{425.8}$, $\lambda_{457.7}$	$\lambda_{772.9}$, $\lambda_{844.7}$
	$\lambda_{552.7}$, $\lambda_{619.5}$	$\lambda_{871.5}$, $\lambda_{971.2}$
H_5x$D_{17} - D_{19}$	$\lambda_{471.3}$, $\lambda_{488.5}$	$\lambda_{762.6}$, $\lambda_{826.6}$
	$\lambda_{556.3}$, $\lambda_{626.5}$	$\lambda_{904.8}$, $\lambda_{997.5}$
H_7x$D_{16} - D_{19}$	$\lambda_{451.2}$, $\lambda_{473.6}$	$\lambda_{760.5}$, $\lambda_{822.1}$
	$\lambda_{506.9}$, $\lambda_{540.9}$	$\lambda_{901.9}$, $\lambda_{937.7}$

II.1. Niveles de daño

Los hojas seleccionadas fueron elegidas considerando cuatro estados de observación: hojas sanas (T_1), hoja sana pero se observaron puntos muy pequeños amarillos visibles que indican que ha llegado la espora del hongo a germinar (T_2), hojas con los primeros síntomas que son pequeñas manchas blancas circulares (T_3) y hojas enfermas cubiertas completamente de esporas blancas provocando un color amarillento por la falta de absorción de la luz (T_4).

Para el grupo T_1 y T_3 se consideraron 240 mediciones, para T_2 y T_4 se tomaron 180 mediciones sobre las cuales se recolectaron los valores de reflectancia en las regiones del ultravioleta, visible y el cercano al infrarrojo (200-980 nm).

En la Fig. 2 se presentan los diferentes estados de las hojas con síntomas. Además, se presenta la firma espectral correspondiente a un promedio de cada estado

de síntomas (T_1-T_4).

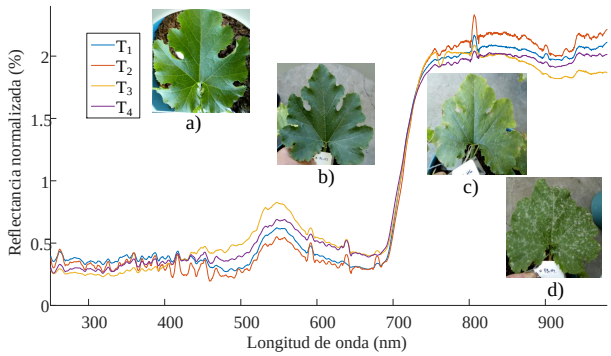

Figura 2: *Niveles de daño de la cenicilla polvorienta: a) T_1, b) T_2, c) T_3 y d) T_4 con imágenes de las hojas de cucurbitáceas y su firma espectral.*

Como en el caso de la condición sana de la hoja (T_1), el color verde es un verde claro de una hoja tierna, en el segundo estado (T_2), se puede apreciar un color verde más oscuro, por ello es difícil determinar un síntoma, considerando el tiempo de germinación del hongo para hacerse visible, por ello las características son muy similares.

Sin embargo, en el estado siguiente (T_3), se pueden observar cambios significativos de color en la hoja no uniformes, sólo son notables en algunas partes, lo que se puede relacionar con los primeros síntomas que son pequeñas manchas blancas grisáceas sobre algunos puntos de la hoja.

Por lo tanto, las condiciones de las hojas continúan siendo similares visiblemente, esto da la dificultad de evaluar una enfermedad, porque la enfermedad no se presenta en todas las hojas y en la misma parte, si no que varía.

Por otro lado, en el estado final (T_4) que se consideró, se toman las diferencias entre el color verde de la hoja y las manchas blancas que son totalmente notables y que se consideran hojas enfermas.

Así, la enfermedad se hace visible y es en dónde se indica la pérdida del verde y la absorción del rojo en el rango del visible. A partir de estos datos sobre las características observadas, se considera realizar el análisis espectral de las firmas espectrales con un total de 840 firmas obtenidas de las hojas.

En la Fig. 3 se propone la metodología para la clasificación de los niveles de daño, partiendo de un análisis exploratorio de las firmas espectrales procesadas[14, 4].

II.2. Análisis exploratorio de las firmas espectrales

A través de un análisis exploratorio de las firmas espectrales agrupadas, es posible encontrar perfiles entre rangos de longitudes de onda que permitan con niveles

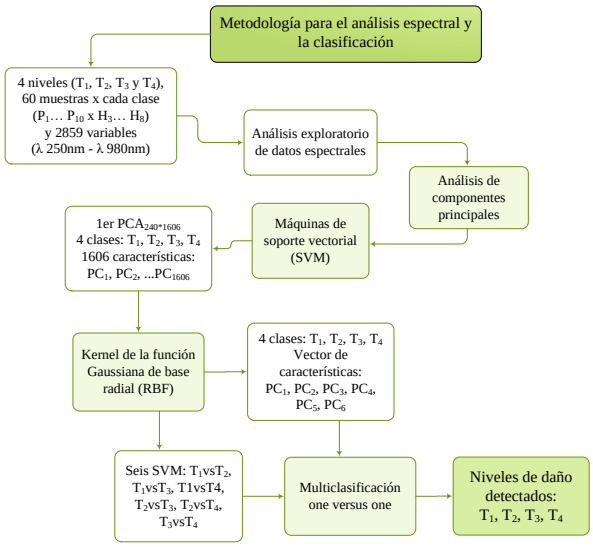

Figura 3: *Metodología propuesta para la clasificación de niveles de daño de la cenicilla polvorienta.*

diferentes en porcentajes de reflectancia que indiquen que algún parámetro interno de la hoja ha sido modificado. Se seleccionaron los datos de seis hojas a partir de la hoja número tres etiquetada como H_3 hasta la hoja número ocho etiquetada como H_8 de diez plantas, que son las hojas (basales).

Estas hojas son en dónde la mayoría de los primeros síntomas de cenicilla se hacen visibles. Inicialmente, se tomaron los datos originales para elaborar los gráficos descriptivos. Para el análisis exploratorio se buscó encontrar perfiles para cada nivel de daño. La Fig. 4 muestra pruebas estadísticas con las firmas espectrales de una hoja sana y una enferma.

Se observa que de acuerdo a los resultados, los valores de la media, la función de densidad y la normalidad, indican que son muchas variables relacionadas con las firmas espectrales considerando su relación con la variante en el espectro del visible y cercano al infrarrojo.

En este caso, los datos son descriptivos para una hoja enferma y una hoja en estado sano y muestra la distribución de los datos en las diferentes condiciones de la hoja.

Como se puede observar, es preferible que los datos estén organizados en rangos para encontrar una mejor descripción del comportamiento de las firmas espectrales en los niveles de reflectancia. Por lo tanto, se trabajó con métodos de estadística descriptiva para el tratamiento de las firmas espectrales con la finalidad de definir características entre las longitudes de onda que diferencien los estados de las hojas.

En la Fig. 5, se muestra el análisis exploratorio de datos con la distribución de las medias en diagramas de cajas por regiones de longitudes de onda y las caras de

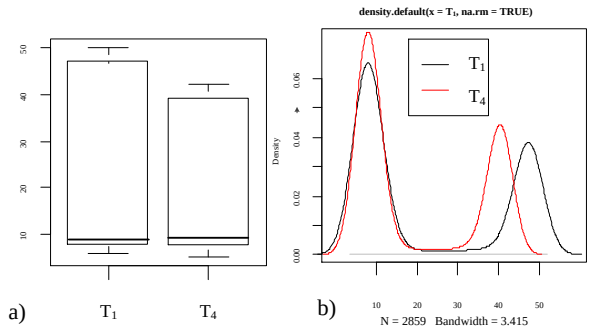

Figura 4: *Resultado de las firmas espectrales completas para describir: a) la media con diagramas de caja, y b) la función de densidad de una hoja sana y una hoja enferma.*

Chernoff. Estas herramientas estadísticas se utilizaron para visualizar datos multivariados.

Las caras de Chernoff tienen forma de un rostro humano, donde las partes individuales (ojos, oídos, boca y nariz) representan los valores de las variables por su forma, tamaño, ubicación y orientación. En este estudio, los rasgos de la cara se describen en la Tabla 2.

Como se observa en la Fig. 5 si existen diferencias entre los grupos: estructura de la cara, el boca y ojos, estableciendo que con los datos considerados si existen diferencias y similitudes entre los grupos.

Tabla 2: *Descripción de rasgos de las caras de Chernoff.*

Rasgo	Longitud de onda	Banda
Altura de la cara	$\lambda_{250-350nm}$	R_1
Ancho de cara	$\lambda_{350-450nm}$	R_2
Estructura de la cara	$\lambda_{450-550nm}$	R_3
Altura de la boca	$\lambda_{550-650nm}$	R_4
Ancho de la boca	$\lambda_{650-750nm}$	R_5
Sonrisa	$\lambda_{750-850nm}$	R_6
Altura de los ojos	$\lambda_{850-950nm}$	R_7

El resultado de este análisis exploratorio con las longitudes de onda divididas en regiones, permitió establecer perfiles entre cada rango. En este caso, entre los diferentes niveles de daño, se observan los cambios entre la estructura de la cara, la forma de la boca y los ojos.

Entre las diferencias más notables mostradas con este análisis, se observan la estructura de la cara, que se relaciona con las longitudes de onda del visible entre la banda del amarillo y verde.

Además, si se consideran la descripción de rangos entre la banda del visible del rojo y cercano al infrarrojo, se observa que la diferencia de la forma de las boca en todos los niveles es diferente. La similitud entre la forma de los ojos se visualiza entre el nivel sano y el nivel enfermo.

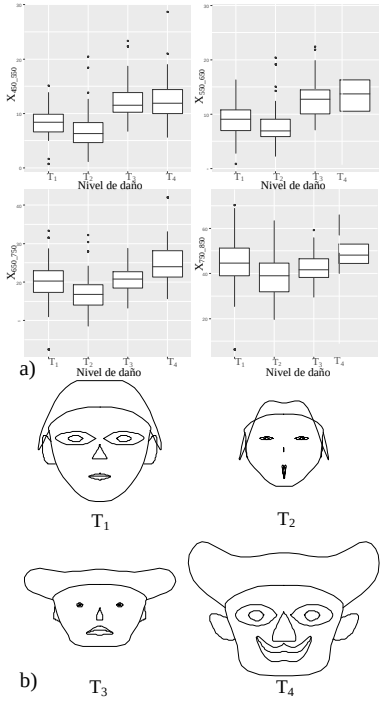

Figura 5: *Análisis exploratorio de los datos espectrales: a) diagramas de caja en los diferentes rangos de longitudes de onda, b) caras de Chernoff para comparar diferencias y similitudes entre rangos de longitudes de onda.*

Sin embargo es notable la similitud entre los estados de germinación y primeros síntomas. Por lo tanto, este análisis nos demuestra que entre las firmas espectrales si existen diferencias que pueden describirse con estadística.

III. Clasificación de los datos espectrales

III.1. Extracción y selección de características

Como primer paso para el proceso de clasificación de los niveles de daño, es necesario iniciar con la etapa de extracción de características. Por lo tanto, se propone el análisis de componentes principales (PCA, principal components analysis).

Este método permite reducir la complejidad de los espacios muestrales con varias dimensiones conservando la información. Si existe una muestra con n datos cada uno con p variables (X_1, X_2, ..., X_p), significa que el espacio muestral tiene n dimensiones.

Este método encuentra un número de factores subyacentes ($z < p$) que describe de forma aproximada lo mismo que las p variables originales. Es decir, en donde antes se necesitaban p valores para caracterizar a cada muestra, se reduce a z valores que reciben el nombre de componente principal.

El método permite reducir la información aportada

por múltiples variables en solo unas componentes [15, 16]. Cada componente principal (Z_i) se obtiene por combinación lineal de las variables originales de la muestra como nuevas variables obtenidas al combinar las variables originales.

La primera componente principal de un grupo de variables (X_1, X_2, ..., X_p) es la combinación lineal normalizada de las variables que tiene mayor varianza:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p \quad (1)$$

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \quad (2)$$

Los términos ϕ_{11}, ..., ϕ_{p1} se conocen como los pesos y su función es definir la componente. ϕ_{11} es el peso de la variable X_1 de la primera componente principal. Los pesos se describen como el peso que tiene cada variable en cada componente y ayudan a definir la información que se centra en cada una de las componentes. Dado un conjunto de datos X con n observaciones y p variables.

El proceso a seguir para calcular la primera componente principal es centrar las variables restando a cada valor la media de la variable a la que pertenece obteniendo que todas las variables tengan media cero. Además, es necesario resolver un problema de optimización para encontrar el valor de los pesos que maximizan la varianza. Para esto se propone el cálculo de valores y vectores propios de la matriz de covarianzas.

Al tener el cálculo de la primera componente (Z_1) se calcula la segunda (Z_2) realizando el mismo proceso, con la restricción de que la combinación lineal no debe estar correlacionada con la primera componente, es decir, que Z_1 y Z_2 tienen que ser perpendiculares. El algoritmo se repite de forma iterativa hasta calcular todas las posibles componentes ($min(n-1, p)$).

Para este caso, debido a la gran cantidad de datos que forman una firma espectral, se consideró realizar la reducción de datos para extracción de características.

En este caso, primero se hizo un análisis de las firmas espectrales reduciendo el total de datos de 2859 valores de longitudes de onda a 1606 valores, considerando los rangos presentados en el análisis exploratorio de datos. Se tomaron las longitudes de onda entre los rangos del $\lambda_{405-679nm}$ para la banda del visible y, $\lambda_{760-988nm}$ para la banda del infrarrojo cercano.

Se procedió a realizar el análisis de componentes principales, resultando un total de 1606 componentes. Se procedió a la selección de características, por lo tanto, para este trabajo sólo se consideraron los seis primeros componentes principales PC_1, PC_2, PC_3, PC_4, PC_5 y PC_6.

Con estos componentes, se formaron los vectores de características. Se considera que el orden de importancia de las componentes es por la magnitud del valor propio relacionado a cada vector propio. Teniendo un total de 840 firmas x 1606 longitudes de onda. En la Fig. 6 se muestran los mapas de características obtenidos del análisis de componentes principales.

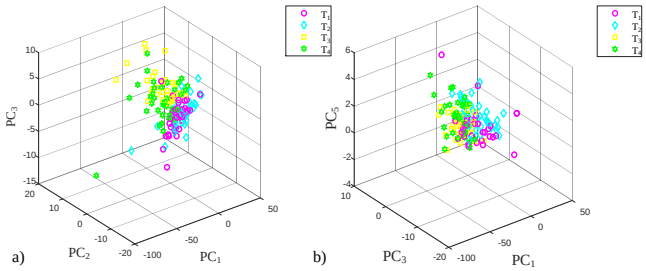

Figura 6: *Mapa de características a partir del análisis de componentes principales: a) características PC_1, PC_2 y PC_3, b) características PC_1, PC_3 y PC_5*

En los mapas de características, se encuentra que los componentes principales que muestran un comportamiento no lineal entre los datos de los diferentes niveles de daño. Al observar el comportamiento de los datos caracterizados, es necesario considerar diferentes alternativas de clasificadores binarios.

Para este caso en particular, se proponen las máquinas de soporte vectorial con los diferentes tipos de núcleos debido a que de acuerdo a los mapas de características, se necesitan definir hiperplanos con funciones de mayor complejidad [17].

III.2. Clasificadores binarios

Las máquinas de soporte vectorial tienen como objetivo generar un modelo de clasificación utilizando un conjunto de muestras de prueba y realizando un entrenamiento y, ser capaz de distinguir la clase a la que pertenece cada una de las muestras que se quieran clasificar.

Este modelo de clasificación se realiza mediante una etapa que se denomina entrenamiento supervisado. En esta fase se introduce al sistema una base de datos que ha sido previamente analizada y clasificada denominada conjunto de entrenamiento.

De esta manera, el clasificador será capaz de realizar una distinción entre clases utilizando los patrones extraídos durante dicho entrenamiento. En general estos algoritmos ayudan a encontrar patrones en datos empíricos (datos de entrenamiento o datos de entrada) con respecto a clases etiquetadas. El modelo resultante se utiliza para hacer una predicción de los datos no etiquetados los cuales se ajustan a datos de entrenamiento.

Una máquina de soporte vectorial (SVM) separa dos clases diferentes a través de un hiperplano. Las máquinas de vectores de soporte son clasificadores binarios con

mayor uso para el aprendizaje automático. Son algoritmos robustos utilizados para generalizar problemas de la vida real. El algoritmo SVM consiste en construir un hiperplano en un espacio de dimensionalidad muy alta que separe clases o grupos de datos.

Esta técnica puede ser utilizada tanto en problemas de clasificación como de regresión. Una buena separación entre las clases permitirá una clasificación correcta. La idea básica es encontrar un hiperplano que separe los datos de d-dimensiones perfectamente en dos clases.

Sin embargo, los datos generalmente no son linealmente separables, así que las SVMs introducen la idea de implementar un núcleo para inducir el espacio característico, el cual emite los datos dentro de un espacio de mayor dimensión en donde los datos sean separables. Por lo general la conversión a un espacio de este tipo causa problemas computacionales [18, 19]. La clave está en que el espacio de dimensión de las SVMs no debe ser tratado directamente.

Dado el conjunto de datos l entrenados $\{\mathbf{x}_i, y_i\}, i = 1, \cdots, l$, donde cada ejemplo tiene d entradas $\mathbf{x}_i \in \mathbf{R}^d$ y una clase etiquetada con uno de dos valores $y_i \in \{-1, 1\}$, todos los hiperplanos en $\mathbf{R}^d$ son parametrizados por un vector $\mathbf{w}$, que es el margen del hiperplano dado y no puede tener un valor de 0 (si $\mathbf{w} = 0$ indica que no existe el margen de separación), y una constante b, expresada en la ecuación siguiente:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{3}$$

donde $\mathbf{w}$ es el vector ortogonal del hiperplano. Dado que un hiperplano ($\mathbf{w}$, b) separa los datos dados por la función,

$$f(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \tag{4}$$

esta función clasifica correctamente los datos de entrenamiento. Sin embargo, un hiperplano dado representado por ($\mathbf{w}$, b) es igualmente expresado por todos los pares $\{\lambda\mathbf{w}, \lambda b\}$ para $\lambda \in \mathbf{R}^+$. Así se define el hiperplano canónico el cual separa los datos del hiperplano a una distancia. Así, para un hiperplano dado, la escala (λ) es un conjunto implícito.

Todos los hiperplanos tienen una distancia funcional ≥ 1. Esto no debe confundirse con el geométrico o la distancia euclidiana también conocida como margen. Para un hiperplano dado ($\mathbf{w}$,b), todos los pares $\{\lambda\mathbf{w}, \lambda b\}$ definen exactamente el mismo hiperplano, pero cada uno tiene una distancia funcional a un punto de datos determinado.

Para obtener la distancia geométrica desde el hiperplano a un punto de los datos, se debe normalizar por la magnitud de $\mathbf{w}$. Esta distancia es simplemente:

$$d\Big((\mathbf{w}, b)\ ,\ \mathbf{x}_i\Big) = \frac{y_i(\mathbf{x_i} \cdot \mathbf{w} + b)}{\parallel \mathbf{w} \parallel} \geq \frac{1}{\parallel \mathbf{w} \parallel} \tag{5}$$

Se requiere que el hiperplano maximice la distancia geométrica que encierra los puntos. De la ecuación (5) se minimiza $\parallel \mathbf{w} \parallel$ para que los puntos queden bien clasificados y quedan sujetos a las restricciones de distancia:

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1, \quad \text{si}: \quad y_i = -1 \tag{6}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq +1, \quad \text{si}: \quad y_i = +1 \tag{7}$$

equivalente a:

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 \qquad \forall i \tag{8}$$

Para lograr la minimización se utilizan los multiplicadores de Lagrange. La matriz se puede definir como $H_{ij} = y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ y se introduce a la notación:

$$\text{minimizar:} \quad \mathbf{w}(\alpha) = -\alpha^T \mathbf{1} + \tfrac{1}{2}\alpha^T H \alpha \tag{9}$$

$$\text{sujeto a:} \quad \alpha^T y = 0 \tag{10}$$

$$\mathbf{0} \leq \alpha \leq C\mathbf{1} \tag{11}$$

Este problema de minimización es conocido como el problema de programación cuadrática (QP, quadratic programming). Muchas técnicas han sido desarrolladas para resolver el problema. El hiperplano óptimo se escribe como:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{12}$$

donde, el vector $\mathbf{w}$ es una combinación lineal de los ejemplos de entrenamiento, y se puede expresar como:

$$\alpha_i(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1) = 0 \quad (\forall i) \tag{13}$$

es decir, que cuando la distancia funcional en un ejemplo es estrictamente más grande que 1 en $y_i((\mathbf{w} \cdot \mathbf{x}_i + b) > 1)$, entonces $\alpha_i = 0$. Así contribuye con los puntos de datos cerrados a $\mathbf{w}$ y son los ejemplos de entrenamiento para los cuales $\alpha_i > 0$, llamados vectores de soporte que se necesitan para definir y encontrar el hiperplano óptimo. Los vectores de soporte son casos límites en la función de decisión.

Se puede considerar α_i como la calificación de dificultad para $\mathbf{x}_i$ y lo importante es que el ejemplo determina el hiperplano. Asumiendo que se tiene el α óptimo del cual se derivó de $\mathbf{w}$, se determina b para completamente especificar el hiperplano. Para hacer esto, se toma cualquier vector de soporte positivo y negativo, $\mathbf{x}^+$ y $\mathbf{x}^-$, para los cuales se sabe que:

$$(\mathbf{w} \cdot \mathbf{x}^+ + b) = +1 \tag{14}$$

$$(\mathbf{w} \cdot \mathbf{x}^- + b) = -1 \tag{15}$$

Entre menos vectores de soporte es mejor la clasificación, además, es una representación más simple del hiperplano y tiene mejor funcionamiento. Ahora, están

los casos cuando los datos no son linealmente separables. Encontrar la curva óptima para ajustar los datos es complicado.

Existe una forma de pre-procesar los datos de tal forma que el problema se transforme en un problema en dónde se encuentre un hiperplano simple. Para obtener esto, se define un mapeo de $\mathbf{z} = \phi(\mathbf{x})$ que transforma las d dimensiones en el vector de entrada $\mathbf{x}$ dentro de un vector de d dimensiones de $\mathbf{z}$. Se elige ϕ como los nuevos datos de entrenamiento $\{\phi(\mathbf{x}_i), y_i\}$ que son los datos separados por un hiperplano.

El hiperplano estará en algún lugar en el espacio característico desconocido. En el espacio de entrada original, los datos se separan por curvas de contorno no continuo [20, 21].

III.3. Múltiple clasificación

Un algoritmo de multiclasificación se puede basar en SVM del tipo uno versus uno, para el cual en este caso se construyeron 6 máquinas de soporte [22]. En la Tabla 3 se muestran los datos obtenidos de las máquinas de soporte que clasifican entre los estados T_1 al T_4 de las hojas. Los datos corresponden primeramente a comparar el estado sano T_1 contra los demás estados y posteriormente tomar T_2 y compararlo con los demás estados. T_1 y T_2 se consideran muy similares en cuanto a las características visibles.

Por lo tanto, las comparaciones entre los demás estados deben ser más variables entre sí, porque se denotan características visibles muy significativas, que es el caso de la enfermedad ya presente en la hoja.

Determinar un clasificador óptimo que genere predicciones correctas para patrones de entradas nuevos significa la selección de un modelo, para lo cual en una SVM no lineal se requiere de un núcleo y valores apropiados para sus parámetros.

Se debe construir el hiperplano que minimize el $h = R^2||\mathbf{w}||^2 + 1$, dónde R es el diámetro de la esfera más pequeña que incluye a todos los datos de entrenamiento y $||\mathbf{w}||$ es la norma euclidiana del vector de pesos. Por lo tanto, para que una SVM clasifique correctamente, se deben escoger los parámetros que minimizen Γ que es un intervalo de confianza.

Se realizaron pruebas con diferentes núcleos de acuerdo a los mapas de características descritos en la Fig. 6, con respecto al comportamiento de los datos que se observan que no son linealmente separables.

Se concluyó que estos datos necesitan de una función que no tiene un comportamiento lineal. Por ello, se consideró usar como núcleo las funciones de base radial, en los cuales el hiperplano óptimo genera curvas entre los datos que permiten la separación de forma más precisa en el hiperplano óptimo de los datos.

En la Tabla 3 presenta los datos obtenidos de las SVM's tomando un núcleo de la función Gaussiana de base radial (RBF, radial base function) [23]. Se entrenaron varios clasificadores binarios, cada uno con dos diferentes clases y valores distintos de σ buscando minimizar Γ en cada SVM. Después de este proceso, continua la etapa de múltiple clasificación, en donde se seleccionan los mejores clasificadores entrenados con los vectores de soporte y los datos de entrenamiento.

La Tabla 3 presenta algunos de los resultados más óptimos del entrenamiento de los clasificadores con parámetros diferentes de σ entre $1.5 - 4$, los cuales resultaron en la minimización del valor de Γ. Por ejemplo, el parámetro $\sigma = 1.5$ resultó ser el óptimo para la máquina de soporte vectorial de las clases T_1 vs T_4 con el resultado minimizado $\Gamma = 2.6748$. Por lo tanto, los vectores de soporte resultantes y los datos de entrenamiento fueron los seleccionados para la siguiente etapa.

En la Fig. 7 se muestra un ejemplo del hiperplano óptimo de separación entre las clases para el caso del clasificador binario con las clases T_1 versus T_2 y las características PC_1 versus PC_2. Se observa que las características no son linealmente separables y por lo tanto el hiperplano óptimo resultante crea los márgenes adecuados para la separación de las clases con base a los vectores de soporte resultantes del entrenamiento.

Tabla 3: *Selección de las máquinas de soporte para clasificar los niveles de T_1, T_2, T_3 y T_4 de las hojas de cucurbitáceas. En el entrenamiento, los casos en negrita, significan que fueron algunos de los clasificadores seleccionados para el proceso de múltiple clasificación.*

SVM	σ	h	Γ
T_1 vs T_4	**1.5**	66.3279	**2.6748**
T_1 vs T_4	2	67.7719	2.7001
T_2 vs T_4	**2**	64.4173	**2.6410**
T_3 vs T_4	2	60.0812	2.5619
T_1 vs T_3	3	50.7323	2.3794
T_2 vs T_3	3	45.7291	2.2735
T_2 vs T_4	3	78.1258	2.8723
T_1 vs T_2	4	1.88e+10	0.0000

IV. Resultados

Las firmas espectrales obtenidas son datos que contienen información relevante de forma muy variada. A partir de la firma espectral de reflectancia, si es posible definir el estado de síntomas de una hoja, pero es importante saber interpretar la información, como en el caso del espectro del visible y sus bandas. Las firmas espectrales presentadas si muestran diferencias y similitudes a través del análisis exploratorio de datos con ayuda de los gráficos y los diagramas.

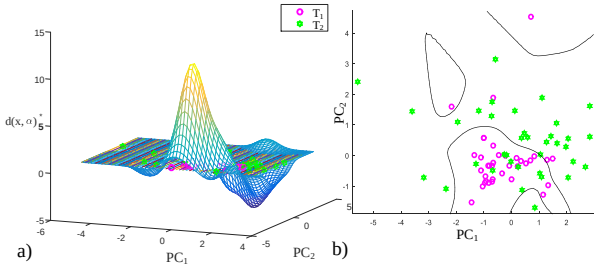

Figura 7: *Resultado del entrenamiento de las máquinas de soporte vectorial en hiperplanos óptimos con los primeros dos componentes principales PC_1 y PC_2: a) hiperplano óptimo en tres dimensiones de T_1 versus T_2, b) hiperplano óptimo en dos dimensiones de T_1 versus T_2.*

Por lo tanto, es posible obtener los perfiles deseados de cada grupo y considerar la reducción, la clasificación y un modelo para definir las características. Las caras de Chernoff son una buena herramienta para establecer similitudes entre clases. En los diagramas de caja, se puede observar una variabilidad entre los grupos como en la longitud de onda del 550 nm al 650 nm y diferencias visuales entre los estados de las hojas.

También se denotan algunos datos fuera del rango como lo son los llamados valores atípicos. Los mapas de características indican que los datos no son linealmente separables, por lo tanto será necesario el uso de métodos no lineales para su clasificación, en este caso se comenzó utilizando las SVM's. Los datos de cada estado de las hojas muestran variabilidad en el porcentaje de reflectancia.

Al ser reducidos por el método de análisis de componentes principales, se observa que los datos son muy similares entre sí y que la relación entre un estado de la hoja sana varia de acuerdo al color que la hoja muestra.

A continuación, se presenta la matriz de confusión resultante de los datos de entrenamiento y la validación de las SVM's de la múltiple clasificación. La Tabla 4 contiene el resultado de las pruebas de los clasificadores binarios con los datos de entrenamiento para obtener los vectores de soporte. Se observa, que todos los datos de entrenamiento fueron clasificados y validados en los diferentes niveles de daño.

Sin embargo, en los datos clasificados se obtuvieron niveles con alto porcentaje de exactitud como en los casos de T_1 con un 88.24 % y T_3 con un valor de 92.16 %. Esto significa que las hojas sanas y las hojas con los primeros síntomas si son distinguibles entre los datos entrenados. Para T_2 y T_3, se obtuvieron valores de 72.55 % y 70.59 %, respectivamente, los cuales se consideran clasificables pero con bajo valor de exactitud.

Se entiende que las características entrenadas tienen problemas para distinguir entre hojas en estado de germinación y las hojas completamente enfermas. La exactitud total de los datos de entrenamiento y validación, resultó en 82.21 %.

Tabla 4: *Matriz de confusión resultante con los datos de entrenamiento y validación.*

SVM	T_1	T_2	T_3	T_4	% Clasificados
T_1	**180**	11	6	7	88.24
T_2	20	**111**	13	9	72.55
T_3	3	4	**188**	9	92.16
T_4	13	12	20	**108**	70.59
% Etiquetados	83.33	80.43	82.82	81.20	**82.21**

El resultado del entrenamiento se considera óptimo para probar los vectores de datos de prueba. Así, con los vectores de soporte obtenidos del entrenamiento y los datos entrenados, se realizó la múltiple clasificación. De los resultados, se obtuvo una exactitud del 86.51 %.

Además, se observó, que al igual que en el entrenamiento, los porcentajes de exactitud resultaron altos para T_1 y T_3 con los valores de 88.89 % y 97.22 %, respectivamente. Se obtuvo un 77.78 % para T_2 y un 77.78 % para T_4. La Fig. 8 muestra en una gráfica los datos que fueron exactamente mal clasificados en conjunto con los verdaderos y falsos positivos (Tabla 5).

Tabla 5: *Matriz de confusión resultante de la clasificación con los datos de prueba.*

SVM	T_1	T_2	T_3	T_4	% Clasificados
T_1	**32**	2	2	0	88.89
T_2	5	**21**	0	1	77.78
T_3	0	0	**35**	1	97.22
T_4	2	1	3	**21**	77.78
% Etiquetados	82.05	87.50	87.50	91.30	**86.51**

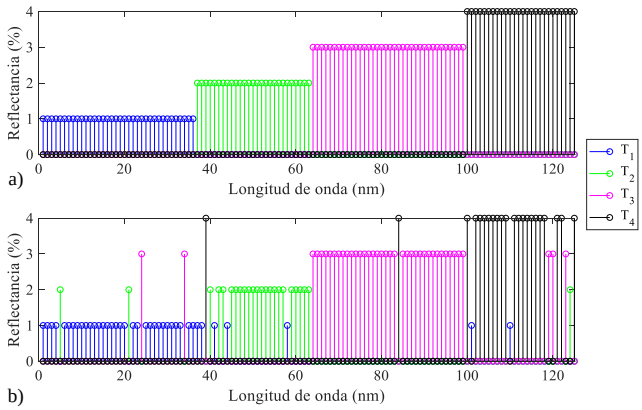

Figura 8: *Resultado de la clasificación con los datos de prueba: a) vector de datos de entrada y b) vector de datos de salida clasificados.*

Un sistema de clasificación debe ser evaluado para conocer su desempeño, por lo tanto, estos resultados fueron sometidos a un análisis de desempeño, el cual consiste en

el cálculo de diferentes parámetros como la sensibilidad y especificidad que se refieren a la probabilidad de que el resultado de las pruebas sean positivos o negativos.

La Tabla 6 contiene los resultados promediados de varias pruebas de clasificación con diferentes vectores formados con datos de prueba. Se observa una exactitud (ACC, accuracy) del 93.25 % y los valores de sensitividad, especificidad y kappa, los cuales son parámetros para medir el total de verdaderos positivos y falsos negativos y determinar el desempeño del clasificador el cual es en este caso, óptimo.

Tabla 6: *Resultado de la evaluación promedio de desempeño de los datos clasificados.*

	$ACC\,\%$	SN	SP	$kappa$
Desempeño	93.25	86.51	95.50	0.8179

V. Conclusiones

En este estudio, se realizó un análisis de los datos espectrales caracterizados para identificar los niveles de daño propuestos T_1-hojas sanas, T_2-hojas en germinación, T_3-hojas con primeros síntomas, y T_4-hojas enfermas de cenicilla polvorienta. El análisis exploratorio permitió establecer similitudes entre clases de acuerdo a las características extraídas de la firma espectral.

Esto permite aplicar métodos estadísticos para la interpretación de los datos. Se observó la variabilidad entre las firmas espectrales y algunos datos extremos y la tendencia lineal entre longitudes de onda continuas.

En este caso, se consideraron las bandas de longitud de onda con mayor cambio en el porcentaje de reflectancia para obtener una clasificación de niveles de daño. Entre estas longitudes de onda, se tienen las pertenecientes a las bandas del verde, rojo y cercano al infrarrojo.

Como se tiene un gran número de longitudes de onda, por firma espectral, el análisis de componentes principales es una técnica que reduce el número de datos en una muestra a un conjunto más pequeño de componentes no correlacionados. Da como resultado datos no separables linealmente, por lo tanto, las máquinas de soporte vectorial son una propuesta para el tratamiento de estos datos en la búsqueda de los perfiles deseados de los síntomas y la clasificación del nivel de daño de la cenicilla polvorienta en las hojas.

Finalmente, se obtuvo la metodología con análisis de datos espectrales para la detección de una enfermedad fúngica en plantas cucurbitáceas y un modelo para definir las características entre clases que puede ser aplicable a otras plantas.

Referencias

[1] Hamed Hamid Muhammed y Anders Larsolle. «Feature Vector Based Analysis of Hyperspectral Crop Reflectance Data for Discrimination and Quantification of Fungal Disease Severity in Wheat». En: *Biosystems Engineering* 86.2 (2003), págs. 125-134. ISSN: 1537-5110. DOI: http://dx.doi.org/10.1016/S1537-5110(03)00090-4. URL: http://www.sciencedirect.com/science/article/pii/S1537511003000904.

[2] Nashwa El-Bendary et al. «Using machine learning techniques for evaluating tomato ripeness». En: *Expert Systems with Applications* 42.4 (2015), págs. 1892-1905. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.09.057. URL: http://www.sciencedirect.com/science/article/pii/S0957417414006186.

[3] Dora Krezhova et al. «Spectral reflectance, chlorophyll fluorescence and virological investigations of tobacco plants (Nicotiana tabacum L.) infected with Tobacco mosaic virus (TMV)». En: (ene. de 2010).

[4] Saman Akbarzadeh et al. «Plant discrimination by Support Vector Machine classifier based on spectral reflectance». En: *Computers and Electronics in Agriculture* 148 (2018), págs. 250-258. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2018.03.026. URL: https://www.sciencedirect.com/science/article/pii/S0168169917310268.

[5] Wondu Garoma Berra. «Visible/Near Infrared Spectroscopic Method for the Prediction of Lycopene in Tomato (Lycopersicon esculentum, Mill.) Fruits». En: *Science, Technology and Arts Research Journal* 1 (dic. de 2013), pág. 17. DOI: 10.4314/star.v1i3.98795.

[6] Ziyang Zhang et al. «Estimation of leaf water content using new vegetation indices combined by near- and middle infrared spectral reflectances». En: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 2017, págs. 4016-4019. DOI: 10.1109/IGARSS.2017.8127881.

[7] Sophie Fabre et al. «Influence of water content on spectral reflectance of leaves in the 3-15-μm domain». En: *IEEE Geoscience and Remote Sensing Letters* 8.1 (2011), págs. 143-147. ISSN: 1545598X. DOI: 10.1109/LGRS.2010.2053518.

[8] Jinhua Sun et al. «Using Spectral Reflectance to Estimate the Leaf Chlorophyll Content of Maize Inoculated With Arbuscular Mycorrhizal Fungi Under Water Stress». En: *Frontiers in Plant Science* 12.May (2021), págs. 1-12. ISSN: 1664462X. DOI: 10.3389/fpls.2021.646173.

[9] Lance Stasinski et al. «Reading light: leaf spectra capture fine-scale diversity of closely related, hybridizing arctic shrubs». En: *New Phytologist* 232.6 (2021), págs. 2283-2294. ISSN: 14698137. DOI: 10.1111/nph.17731.

[10] Jana Albrechtova et al. «Spectral analysis of coniferous foliage and possible links to soil chemistry: Are spectral chlorophyll indices related to forest floor dissolved organic C and N?» En: *Science of the Total Environment* 404.2-3 (2008), págs. 424-432. ISSN: 00489697. DOI: 10.1016/j.scitotenv.2007.11.006.

[11] Jiyou Zhu et al. «The changes of leaf reflectance spectrum and leaf functional traits of osmanthus fragrans are related to the parasitism of cuscuta japonica». En: *Applied Sciences (Switzerland)* 11.4 (2021), págs. 1-15. ISSN: 20763417. DOI: 10.3390/app11041937.

[12] C. A. Rivera-Romero et al. «Visible and near-infrared spectroscopy for detection of powdery mildew in Cucurbita pepo L. leaves». En: *Journal of Applied Remote Sensing* 14.4 (2020), págs. 1-19. DOI: 10.1117/1.JRS.14.044515.

[13] Ian T. Young Alan Victor oppenheim Alan S. Willsky. *Señales y sistemas*. Segunda Edición. Pearson, Prentice Hall, 1998.

[14] Wu Dake y Ma Chengwei. «The Support Vector Machine (SVM) Based Near-Infrared Spectrum Recognition of Leaves Infected by the Leafminers». En: 3 (ago. de 2006), págs. 448-451. DOI: 10.1109/ICICIC.2006.539.

[15] Alvil C. Rencher. *Methods of Multivariate Analysis*. Second. Wiley-Interscience, 2002.

[16] Bryan F. J. Manly. *Multivariate Statistical Methods: A Primer*. Chapman y Hall, 2017.

[17] Joachims Thorsten. «Text categorization with Support Vector Machines: Learning with many relevant features». En: *Machine Learning: ECML-98*. Ed. por Claire Nédellec y Céline Rouveirol. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, págs. 137-142. ISBN: 978-3-540-69781-7.

[18] Lipo Wang. «Support Vector Machines: Theory and Applications». En: Springer, 2005, págs. 1-76, 321-342. ISBN: 978-3-540-24388-5.

[19] N. Cristianini y J. Shawe-Taylor. «An introduction to support vector machines and other kernel-based learning methods. Repr». En: *Introduction to Support Vector Machines and other Kernel-Based Learning Methods* 22.01 (2001). DOI: 10.1017/CBO9780511801389.

[20] Shigeo Abe. «Support Vector Machines for Pattern Classificacion». En: Springer, 2010, págs. 21-350. ISBN: 978-1-4471-2548-8.

[21] J.C. Burges Christopher. «A Tutorial on Support Vector Machines for Pattern Recognition». En: *Data Mining and Knowledge Discovery* 2.6 (1998), págs. 121-167. ISSN: 1573-756X. DOI: 10.1023/A:1009715923555. URL: https://doi.org/10.1023/A:1009715923555.

[22] T. Rumpf et al. «Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance». En: *Computers and Electronics in Agriculture* 74.1 (2010), págs. 91-99. ISSN: 0168-1699. DOI: http://dx.doi.org/10.1016/j.compag.2010.06.009. URL: http://www.sciencedirect.com/science/article/pii/S0168169910001262.

[23] Elham Omrani et al. «Potential of radial basis function-based support vector regression for apple disease detection». En: *Measurement* 55 (2014), págs. 512-519. ISSN: 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2014.05.033. URL: http://www.sciencedirect.com/science/article/pii/S0263224114002541.