

UNIVERSITAT POLITÈCNICA DE CATALUNYA  
DEPARTAMENTO DE INGENIERÍA AGROALIMENTARIA Y BIOTECNOLOGÍA  
PROGRAMA DE DOCTORADOS INDUSTRIALES  
DOCTORADO EN TECNOLOGÍA AGROALIMENTARIA Y BIOTECNOLOGÍA

# **Desarrollo de modelos para la identificación de variedades y fitopatologías mediante técnicas fotónicas**

Tesis presentada por  
Sergio Borraz Martínez

Tesis realizada en el Departamento de Ingeniería Agroalimentaria y Biotecnología bajo la dirección de:

Directora: D. Anna Gras Moreu  
Departamento de Ingeniería Agroalimentaria y Biotecnología (DEAB)  
Co-director: D. Joan Simó Cruanyes  
Fundació Miquel Agustí y DEAB

Colaboración: Ricard Boqué Martí  
Universitat Rovira i Virgili  
Departamento de Química Analítica y Orgánica

Colaboración: Francesc Tarrés Ruiz  
Universitat Politècnica de Catalunya  
Departamento de Teoría de la señal y comunicaciones

**Castelldefels, 2021**



<<Nada en la vida es para ser temido,  
sino para ser entendido.  
Ahora es el momento de entender más,  
de modo que podamos temer menos>>

Marie Curie



A mi familia,  
porque siempre me ha apoyado



# Agradecimientos





## Agradecimientos

---

Este proyecto no hubiera sido posible sin no hubiera contado con la ayuda y el apoyo de muchas personas, las cuales me han ayudado a mantenerme erguido ante la adversidad, y me han dado energías para avanzar. Por tanto, en las siguientes líneas, me gustaría agradecerles su ayuda durante el doctorado.

A Agromillora Iberia por haberme escogido para llevar a cabo este proyecto, brindándome la oportunidad de poder crecer profesionalmente. En especial a mi tutora de tesis en la empresa, Mariàngela Mestre, y a Joan Torrents, por apostar por las nuevas tecnologías. Y como no, agradecer a todos mis compañeros/as del Centro de Materiales Iniciales, por aguantarme durante estos 3 años. También a Mireia Bordas, por su esfuerzo y apoyo, además de por todo lo que me enseñó del cultivo de plantas *in vitro*. A Natalia, Nora e Inma, con las que compartí mis preocupaciones, en ocasiones frustraciones, y sobre todo grandes momentos, los cuales guardo con gran cariño en mi memoria.

Agradecer también a Agromillora Sur su gran hospitalidad durante los tres meses de mi estancia en Chile. A Ely Chernilo, por haberme ofrecido la oportunidad de vivir esa gran experiencia. A Natalia Torres y el resto del equipo de QA y de Producción, por todo ese trabajo (“la pega”) que realizamos juntos, el cual, entre risas y momentos para enmarcar, conseguisteis que se me hiciera mucho más ameno. A mi amigo y compañero de piso Román Toro, con el cual aprendí muchísimo de la cultura chilena, mediante nuestras extensas charlas acompañadas con un vaso de Coca-cola. Gracias por hacerme sentir como en casa, aun estando a más de 11.000 km de ella. Llegados a este

punto, no me puedo olvidar de Rebeca, juntos vivimos grandes aventuras, desde descubrir paisajes impresionantes de la cordillera de los Andes, hasta vernos atrapados dentro de un supermercado asediado por las protestas. Estas vivencias son las que forjan amistades que, pese a la distancia, son fáciles de conservar.

A mis directores de tesis, Anna Gras y Joan Simó. Muchas gracias por haber sido durante estos años mis guías en la ardua, en muchas ocasiones difícil, y siempre sorprendente, senda de los proyectos doctorales. Muchas gracias nuevamente, por toda vuestra ayuda y apoyo.

A Ricard Boqué (Universitat Rovira i Virgili) i a Francesc Tarrés (Universitat Politècnica de Catalunya). Aunque oficialmente habéis sido colaboradores de este trabajo, para mí también habéis sido tutores. Muchas gracias por todo el esfuerzo, tiempo, dedicación y paciencia, además de por todos los grandes conocimientos que me habéis transmitido en los campos de la espectroscopía y quimiometría, Ricard; y en el ámbito de la visión artificial y deep learning, Francesc. Sin vuestra ayuda nada de esto hubiera sido posible.

A Doctorats Industrials por el apoyo económico ofrecido a lo largo de la tesis, sin el cual no hubiera sido posible realizar este proyecto.

A todos/as mis amigos/as. Con especial mención a Anna Mur, Antonio Luque, Albert Ribes, Mariona Bessa, Jesús Clemente y Laura López. Porque sé que puedo contar con vosotros siempre que lo necesite.

A Cristina Medina, por muchas palabras y líneas que pudiera escribir, jamás alcanzaría a plasmar ni una milésima parte de todo el esfuerzo, apoyo y energía que me has transmitido durante estos años.

A mi familia, que siempre me ha apoyado en todo lo que me he propuesto. A mis padres, Silvia y Jaime, a los que quiero muchísimo y que siempre se han preocupado por que fuera feliz. A mi hermano Jorge. Nuestras escapadas de montaña han sido el mejor remedio contra el agotamiento, la saturación y el estrés.

A todos ellos, muchas gracias.



## **Resumen / Abstract**



## Resumen

---

En el cultivo de árboles frutales lo más habitual es trabajar con plantas injertadas. De esta forma, la variedad de árbol frutal injertada ofrece unas cualidades de fruto específicas y el portainjerto ofrece unas características agronómicas y de adaptación determinadas. Los injertos que provienen de variedades de una misma especie suelen presentar una gran proximidad genética, lo que genera que sean muy difíciles de identificar mediante caracteres agromorfológicos. Por esta razón, es necesario la implementación de métodos más complejos para poder distinguirlos. En la actualidad, los métodos más utilizados para la identificación varietal se basan en el análisis del ADN, los cuales permiten obtener perfiles moleculares. Sin embargo, estas técnicas biomoleculares requieren de instrumentos e instalaciones sofisticadas, no son rápidas de ejecutar y presentan un coste elevado. Estos inconvenientes hacen inviable la implementación de las técnicas moleculares como métodos de análisis rutinario de grandes volúmenes de muestras. En este contexto, el sector viverístico demanda nuevos métodos que seas rápidos, no destructivos, eficaces y con un coste por análisis bajo. El objetivo principal de esta tesis es desarrollar un método que permita identificar y diferenciar variedades a nivel de planta de vivero junto con la capacidad de evaluar el estado fisiológico de las mismas. La tesis se articula en 4 artículos publicados en revistas indexadas junto con otros ensayos expuestos en el presente trabajo. Los principales resultados son: a) Se han desarrollado protocolos para el uso de técnicas de espectroscopia de infrarrojo cercano, cámara hiperespectral y visión artificial acoplado a *deep learning* en esquejes de frutales y portainjertos de cerezo; b) La espectroscopia de infrarrojo cercano demostró ser una técnica capaz de diferenciar, con un grado muy elevado de especificidad, seis variedades de almendro (*Prunus dulcis*), pero presentó limitaciones para su implementación en una línea de producción; c) El uso de la cámara hiperespectrales ha demostrado adecuado para la diferenciación de tres variedades de portainjerto y el análisis del estado fisiológico, incluyendo la detección prematura de *Pseudomonas*

*syringae*; d) El uso de la visión artificial acoplada a *deep learning* ha demostrado ser eficaz en la diferenciación de variedades de almendro; e) El análisis conjunto de los resultados indican que la mejor solución para ser implementada en una línea de producción pasaría por la implementación de la cámara espectral para el control del estado fisiológico de las plantas y la visión artificial acoplada a *deep learning* para la identificación varietal.

**Palabras clave:** Espectroscopia de infrarrojo cercano, Cámara hiperespectral, Visión artificial, Deep learning, Quimiometría, Control varietal, Detección de fitopatologías, *Prunus dulcis*.



## Abstract

---

In the cultivation of fruit trees, the most common is to work with cuttings. In this way, the grafted fruit tree variety offers specific fruit qualities and the rootstock offers certain agronomic and adaptation characteristics. The grafts that come from varieties of the same species usually present a great genetic proximity, which makes them very difficult to identify by means of agro-morphological characters. For this reason, it is necessary to implement more complex methods to be able to distinguish them. Nowadays, the most widely used methods for varietal identification are based on DNA analysis, which allow obtaining molecular profiles. However, these biomolecular techniques require sophisticated instruments and facilities, time consuming, and are expensive. These drawbacks make the implementation of molecular techniques unfeasible as methods for the routine analysis of large volumes of samples. In this context, the nursery plant sector demands a rapid, no-destructive, accurate, and economic methods. The main objective of this thesis is to develop a method that allows identifying and differentiating varieties at the nursery plant level, together with the ability to assess their physiological status. The thesis is articulated in 4 articles published in indexed journals together with essays presented in the present work. The main results are to a) Protocols have been developed for the use of near infrared spectroscopy techniques, hyperspectral camera and artificial vision coupled to deep learning in fruit cuttings and cherry rootstocks; b) Near-infrared spectroscopy proved to be a technique capable of differentiating, with a very high degree of specificity, six varieties of almond (*Prunus dulcis*), but it presented limitations for its implementation in a production line; c) The use of the hyperspectral camera has proven adequate for the differentiation of three rootstock varieties and the analysis of the physiological state, including the premature detection of *Pseudomonas syringae*; d) The use of artificial vision coupled with deep learning has proven to be effective in the differentiation of almond varieties; e) The joint analysis of the results indicates that the best solution to be implemented in a production line

would go through the implementation of the spectral camera for the control of the physiological state of the plants and the artificial vision coupled to deep learning for the varietal identification.

**Keywords:** Near-infrared spectroscopy, Hyperspectral camera, Computer vision, Deep learning, Chemometric, Varietal control, Phytopathology detection, *Prunus dulcis*.

## Abreviaciones

---

ASCA: Analysis of variance – simultaneous component analysis

CNN: Convolutional neural network

CV: Computer vision

CV-DL: Computer vision coupled deep learning

CPs: Componentes principales

DI: Doctorado industrial

DL: Deep learning

ELISA: Enzyme-Linked ImmunoSorbent Assay

FAO: Food and Agriculture Organization

HSV: Valor de saturación / *Hue, Saturation, Value* en inglés

IA: Inteligencia artificial

IoT: Internet de las cosas / *Internet of the things* en inglés

IOA: Índice de orientación agrícola

IRTA: Institut de Recerca i Tecnologia Agroalimentàries

ML: Machine learning

MSC: Multiplicative Scatter Correction

NIR: Near infrared

NIRS: Near-infrared spectroscopy

NW: Norris-Williams

PCA: Principal Component Analysis

PCR: polymerase chain reaction

PIB: Producto interior bruto

PLS: Partial Least Squares

PLS-DA: Partial Least Squares Discriminant Analysis

PLS-R: Partial Least Squares Regression

RGB: Red-green-blue

RMESCV: Root Mean Square Error of Cross-Validation

SG: Savitzky-Golay

SGD: Stochastic Gradient Descent

SIMCA: Soft independent modelling by class analogy

SNP: Single nucleotide polymorphisms

SNV: Standard Normal Variate

UFC: Unidades formadoras de colonias

VL: Variable latente

VNIR: Visible – infrarrojo cercano

# Índice



# Índice

---

Contexto .....	27
1.1 Empresa .....	29
1.2 Motivación.....	31
Introducción .....	33
2.1 Innovación en la agricultura .....	35
2.2 Fundamentos de las técnicas fotónicas .....	36
2.3 Fundamentos de la espectroscopia de infrarrojo cercano (NIRS).....	40
2.3.1 Conceptos de la vibración molecular en el infrarrojo cercano .....	41
2.3.2 Espectro NIR .....	45
2.4 Cámara espectral.....	48
2.5 Visión artificial junto con <i>Deep learning</i> (CV-DL) .....	53
2.5.1 Inteligencia artificial y Machine learning .....	53
2.5.2 Conceptos del aprendizaje profundo o <i>Deep learning</i> .....	56
2.5.3 Arquitectura de las redes neuronales computacionales.....	59
2.5.4 Entrenamiento de una red .....	64
2.5.5 Transfer learning .....	70
2.5.6 Aumento de datos ( <i>data augmentation</i> ) .....	70
2.5.7 Arquitectura red neuronal convolucional VGG16 .....	71
2.6 Análisis de datos multivariantes.....	72
2.6.1 Métodos quimiométricos de análisis .....	74
Objetivos .....	85
Capítulo 1. Development of a Methodology to Analyze Leaves from <i>Prunus dulcis</i> Varieties Using Near-Infrared Spectroscopy .....	89
Capítulo 2. Multivariate Classification of <i>Prunus dulcis</i> Varieties Using Leaves of Nursery Plants and Near-Infrared Spectroscopy.....	101
Capítulo 3. Diferenciación varietal y control del estado fisiológico mediante cámara hiperespectral .....	113
Capítulo 4. Varietal Quality Control in the Nursery Plant Industry Using Computer Vision and Deep Learning Techniques .....	129

Capítulo 5. Combining Computer Vision and Deep Learning to Classify Varieties of <i>Prunus dulcis</i> for the Nursery Plant Industry .....	143
Discusión.....	155
4.1 Espectroscopia de infrarrojo cercano.....	157
4.2 Cámara hiperspectral .....	161
4.3 Visión artificial junto <i>Deep learning</i> (CV-DL) .....	165
4.4 Aplicación de los métodos estudiados .....	168
4.5 Nivel de maduración tecnológica .....	170
Conclusiones.....	173
Bibliografía.....	177



**Contexto**



## Contexto

---

El presente proyecto se enmarca dentro del *Pla de Doctorats Industrials*, el cual se trata de una iniciativa de la Generalitat de Catalunya, en colaboración con el sistema universitario y de investigación para promocionar la incorporación de doctores en las empresas.

El programa de Doctorado Industrial (DI) [1] se basa en un proyecto de investigación estratégico de una empresa que se desarrolla en colaboración con una universidad o centro de investigación, el cual se convertirá en una línea doctoral. Estos proyectos permiten desarrollar la investigación dentro de un entorno dual, constituido por la empresa y el mundo académico.

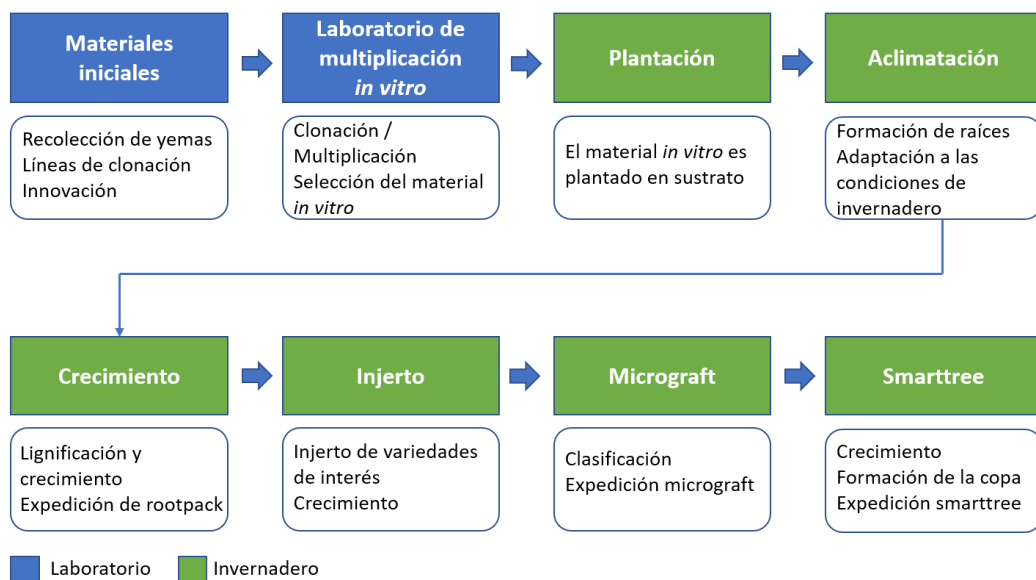
Esta iniciativa surge con la finalidad de dar respuesta al reto de transferir la tecnología y el conocimiento que el sistema universitario e investigador es capaz de generar, hacia el sector productivo, con el fin de poder favorecer un desarrollo económico y social.

Dentro del programa de *doctorats industrials* del presente proyecto, se encuentran adheridos la Generalitat de Catalunya, como entidad financiadora del proyecto de investigación; la Universitat Politècnica de Catalunya [2], en colaboración con la Universitat Rovira i Virgili [3] como centros académicos; además de la empresa Agromillora Iberia S.L.U. [4], en la cual se desarrolló el proyecto.

### 1.1 Empresa

Agromillora Group es una empresa multinacional líder en el sector viverístico y un referente en la producción y comercialización de árboles frutales y portainjertos, siendo el propagador número uno de especies vegetales leñosas y semi leñosas para el sector agrícola. Se encuentra presente en los cinco continentes gracias a sus once filiales, las cuales representa la puerta de entrada a mercados de 25 países.

Agromillora se caracteriza por ser la única plataforma con presencia global y ser líderes mundiales en multiplicación vegetal *in-vitro*. Su infraestructura consta de 62,8 ha de invernaderos con capacidad de entregar alrededor de 80 millones de plantas al año. La misión y la visión de Agromillora es impulsar la creación de plantas más productivas, sanas y resistentes y trabajar para implementar nuevas soluciones agronómicas, todo ello, con la finalidad de incrementar la eficiencia y rentabilidad de los cultivos. Los clientes principales son otras empresas del sector viverístico y productores.



**Figura 1.** Esquema del proceso de producción de frutales en Agromillora Iberia.

El proceso de producción de frutales se inicia con la clonación de los portainjertos. Para ello se recolectan yemas de árboles madre certificados, los cuales se encuentran, bajo estrictas condiciones de vigilancia y control, en el Centro de Materiales Iniciales (CMI). A partir de estas yemas se crean clones mediante técnicas de cultivo *in vitro*. A medida que aumenta el número de clones y se desarrollan, estos son extraídos de los botes de multiplicación *in vitro* y pasan a la fase de plantación y aclimatación. Cuando los portainjertos alcanzan el tamaño adecuado, estos pueden

ser expedidos a otros viveros o seguir el proceso de producción de Agromillora, encaminándose hacia la fase de injerto. En este punto, se injertan yemas de las variedades comerciales de interés en el portainjerto. Dichas yemas también provienen de campos de planta madre. Durante el crecimiento y desarrollo de las plantas injertadas, se realizan tratamientos de saneamiento, colocación de un tutor y diferentes ciclos de corte del ápice apical, todo ello con el objetivo de que el producto final sea una planta con las cualidades óptimas para ser plantada en el campo. Lo anteriormente explicado se muestra representado esquemáticamente en la [Figura 1](#).

## 1.2 Motivación

En el marco de diversas reuniones entre responsables de Agromillora Iberia e investigadores de la Escola d'Enginyeria Agroalimentària i Biosistemes de Barcelona (EEABB), perteneciente a la Universitat Politècnica de Catalunya (UPC), se trataron una serie de puntos que según la empresa eran claves para mejorar su sistema productivo. Los representantes de la empresa expusieron la necesidad de disponer de un sistema de control de la pureza varietal y la detección de fitopatologías. En aquel entonces, se realizaban análisis moleculares para el control de fitopatologías y había instaurado un sistema de trazabilidad para el control varietal. Pese a estos esfuerzos, estos no eran suficientes. Por esta razón, una de sus principales preocupaciones era desarrollar controles más rápidos, más eficientes y baratos para incrementar el número de muestras analizadas en tiempo real. Se planteó, como óptimo, el control de la totalidad de las plantas en uno o varios puntos de la cadena de producción.

Para el desarrollo de la agricultura es importante la obtención de nuevos materiales vegetales, a menudo híbridos interespecíficos, que sean de interés agrícola y permitan impulsar y mejorar el rendimiento de los cultivos. La diversidad de especies y variedades en el sector de producción de plantas de vivero es realmente

## Contexto

grande y requiere la incorporación de sistemas de control fiables, tanto para la detección de patologías como para la identificación de material vegetal. Por ello, el control varietal en la fase de producción es un gran reto. La capacidad para discriminar rápidamente entre las variedades vegetales es de gran interés, no sólo para Agromillora, sino que también para todo el sector viverístico.

En los últimos años, las tecnologías basadas en la amplificación del ácido desoxirribonucleico (ADN) han sido las más usadas para la identificación de especies y variedades vegetales. Aun así, se tratan de técnicas demasiado caras y lentas como para analizar grandes volúmenes de muestras. Según datos proporcionados por el *Institut de Recerca i Tecnologia Agroalimentàries (IRTA)* [5], los análisis mediante *single nucleotide polymorphisms* (SNPs) [6,7] de un lote de entre 300-500 plantas tiene un coste de 42 euros por muestra, lo cual equivale a un coste total que oscilaría entre los 12600 – 21000 euros. Por lo que respecta a la detección de fitopatologías, métodos de análisis como las pruebas *Enzyme-Linked ImmunoSorbent Assay* (ELISA)[8–10] o también técnicas basadas en ADN, como la reacción en cadena de la polimerasa (PCR)[11–13], presentan limitaciones similares. En este contexto, las tecnologías basadas en espectroscopia, visión artificial e inteligencia artificial (IA), parecen aportar algunas soluciones al problema del coste económico y de rapidez, además de permitir en muchos casos la realización de análisis no destructivos.

# Introducción





## Introducción

---

### 2.1 Innovación en la agricultura

La Food and Agriculture Organization (FAO) [14] considera de vital importancia la inversión en innovación en el sector agrícola para intensificar la producción agrícola de manera sostenible, además de una pieza clave para lograr acabar con el hambre y la malnutrición en el mundo. Es tal la importancia de la investigación, desarrollo e implementación de nuevos sistemas productivos que permitan mejorar el rendimiento y la eficiencia del sector agrícola, que hace que estos formen parte de los objetivos sobre el desarrollo sostenible incluidos en la agenda 2030 [15]. Por lo tanto, para poder afrontar los retos, presentes y de futuro, es indispensable la continua actualización del sistema productivo agrícola.

Actualmente, se está produciendo una modernización del sector agrícola, impulsada por la innovación y la incorporación de tecnologías emergentes. Tal es así, que se habla de la cuarta revolución agrícola o de la agricultura 4.0 [16]. Este cambio viene impulsado por la introducción en el sector de tecnologías emergentes como el Internet de las cosas (IoT), la computación en la nube, la robótica, la inteligencia artificial (IA) o el *Big Data* [17–20]. Además, este desembarco tecnológico se ve favorecido por encontrar, en el sector agrícola, un fértil campo donde poder extenderse y echar raíces, ya que posibilidades y nichos de implantación no le faltan. Agrupados bajo el paraguas de la agricultura inteligente o agricultura digital, están surgiendo aplicaciones en áreas tan diversas como en el control de plagas [21], en la optimización en el uso de pesticidas y herbicidas [22], en la elección de la fecha óptima de siembra [23], la monitorización de cultivos mediante el uso de vehículos aéreos no tripulados (agricultura de precisión) [24–26], la implementación del *big data* en agricultura [27,28]. También se están abriendo un hueco las ciencias robóticas mediante la utilización de robots para cosechar productos que hasta ahora

solo podían ser cosechados a mano [29,30]. La mayor parte de estas tecnologías se basan en el continuo monitoreo de los cultivos con el fin de obtener una gran cantidad y variedad de datos que permitan tener un conocimiento más profundo del estado del cultivo. Además, el análisis de dichos datos se emplea para interpretar el presente y, de esta forma, intentar predecir el futuro, y así poder tomar decisiones más precisas y oportunas [31].

A pesar de los grandes beneficios de la innovación, es importante tener en cuenta que también puede acarrear efectos negativos. De esta forma, la tecnología es un arma de doble filo porque tiene el potencial de causar daño, además de proporcionar beneficios [32]. La mayor parte de las consecuencias negativas hacen referencia a cuestiones sociales y medioambientales [18]. Por ello, es importante apostar por una innovación sostenible.

Un aspecto que puede poner en peligro la revolución de la agricultura 4.0 es la reducción en la financiación pública. Tal y como se muestra en el informe *Seguimiento de los progresos relativos a los indicadores de los ODS relacionados con la alimentación y la agricultura correspondientes a 2020* [33] realizado por la FAO, el gasto público en agricultura ha disminuido en la mayoría de las regiones del mundo desde el año 2000. El índice de orientación agrícola (IOA) para los gastos públicos, el cual compara la contribución a la agricultura por parte del país con la contribución al PIB del sector, muestra que del 2001 a 2018, el IOA para los gastos públicos cayó del 0,42 al 0,28 en todo el mundo [33]. Por lo tanto, para una plena consolidación de la agricultura 4.0 en todo el sector, es necesario revertir esta situación.

## 2.2 Fundamentos de las técnicas fotónicas

En el presente estudio, se han aplicado diferentes técnicas que van desde la espectroscopia de infrarrojo cercano (*Near infrared spectroscopy*, NIRS) hasta la visión artificial (*Computer vision*, CV), incluyendo la utilización de imágenes hiperespectrales. Pese a que se tratan de técnicas distintas, pertenecientes a campos

de la ciencia diferentes, todas ellas tienen en común los mismos principios de interacción entre la radiación electromagnética y la materia. En lo que realmente se diferencian es en la manera en la que se estudia y se observa dicha interacción. En la espectroscopia se obtiene un espectro, en la visión artificial se trabaja con una imagen y mediante las cámaras espectrales se obtiene un cubo de datos que contiene información de la imagen y del espectro al mismo tiempo. Para poder entender mejor los fundamentos teóricos de las técnicas empleadas, es importante tener presente los conceptos básicos que se exponen a continuación, los cuales son comunes para todas ellas.

### Radiación electromagnética

La radiación electromagnética es una forma de energía que se compone de ondas electromagnéticas. Según la teoría electromagnética de Maxwell [34], las ondas electromagnéticas surgen de la combinación de un campo eléctrico y un campo magnético que oscilan en planos individuales dispuestos perpendicularmente el uno del otro. Ambos campos oscilan en fase y se propagan como una onda sinusoidal. Por otra parte, la mecánica cuántica describe las ondas electromagnéticas como una combinación de onda y partícula (fotón) [35], ya que, en base a esta rama de la física, las ondas electromagnéticas presentan un comportamiento dual [36].

Las propiedades de la radiación electromagnética vienen definidas por tres factores: La frecuencia ( $\nu$ ), la longitud de onda ( $\lambda$ ) y la energía de un fotón ( $E$ ) [37].

- La frecuencia es el número de ciclos de onda por unidad de tiempo.

$$\nu = \frac{c}{\lambda} \quad (1)$$

- La longitud de onda es la distancia entre dos máximos consecutivos de la onda.

$$\lambda = \frac{c}{f} \quad (2)$$

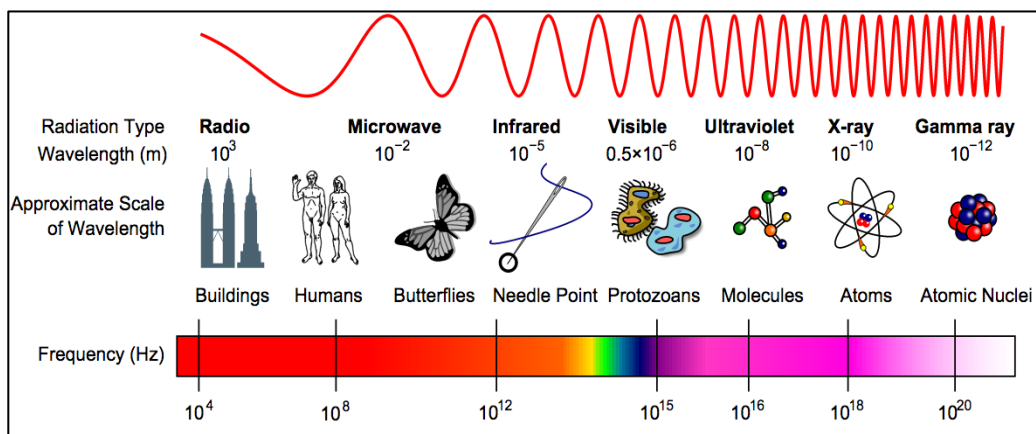
## Introducción

- La energía de un fotón depende de la frecuencia de la onda electromagnética.

$$E = \frac{hc}{\lambda} = h\nu \quad (3)$$

donde  $c$  es la velocidad de la luz en el vacío (299,792,458 m/s) y  $h$  es la constante de Planck ( $6,62607015 \cdot 10^{-34}$  J·s).

Las ondas electromagnéticas se clasifican y agrupan en base a su frecuencia o longitud de onda, formando lo que se conoce como el espectro electromagnético. La longitud de onda (*wavelength* en inglés) es una unidad de distancia, que normalmente se expresa en múltiplos o submúltiplos del metro. No obstante, otra unidad comúnmente utilizada es el número de onda (*wavenumber* en inglés), que consiste en el número de ondas en una longitud de un centímetro y se expresa como  $\text{cm}^{-1}$ . La **Figura 2** muestra el espectro electromagnético, el cual va desde los rayos gamma, las ondas con mayor frecuencia, menor longitud de onda y mayor energía, hasta las ondas de radio, las cuales tienen poca frecuencia, gran longitud de onda y poca energía.

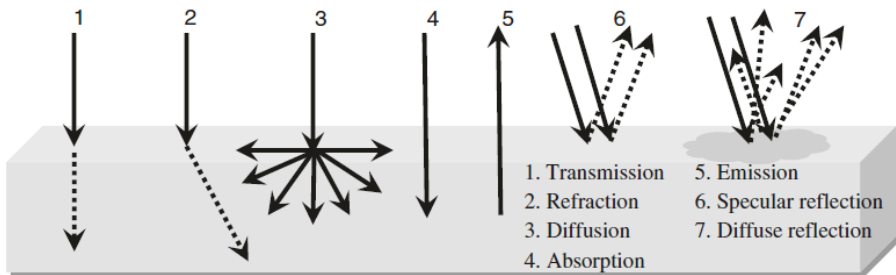


**Figura 2.** Espectro electromagnético. Fuente: Imagen de internet libre de licencia.

Se conocen múltiples formas de interacción entre la radiación y la materia, la mayoría de ellas se encuentran descritas gráficamente en la **Figura 3** [37]. De todas

ellas, las únicas dos interacciones que generan un cambio en el estado energético de la materia son la absorción y la emisión de radiación, ya que en ambas se produce un fenómeno de transferencia de energía. Por otro lado, la suma del conjunto de interacciones radiación – materia, dentro del rango electromagnético del visible, es lo que da lugar a la imagen.

Todas las interacciones aportan información sobre la composición y estructura de la materia [38]. Dependiendo de la técnica empleada se estudian una o varias de estas interacciones.



**Figura 3.** Posibles interacciones entre un objeto y una onda electromagnética [37].

En el presente estudio, las tecnologías de espectroscopia de infrarrojo cercano y la cámara espectral se basan en la detección de la reflectancia, ya que la fuente de radiación y el detector se encuentran en el mismo lado del objeto investigado (geometría de reflexión). En el caso de la espectroscopia de infrarrojo cercano, a partir del espectro de reflectancia se calculó el de absorbancia (4), con el cual se trabajó, mediante la siguiente transformación:

$$abs = \log \left( \frac{1}{Reflectancia} \right) \quad (4)$$

Por lo que respecta a la cámara espectral, se empleó directamente el espectro de reflectancia. En cambio, mediante la cámara digital empleada en la visión artificial, se trabajó con la transmitancia, debido a que el detector se encontraba en el lado opuesto al de la fuente de luz (geometría de transmisión).

## 2.3 Fundamentos de la espectroscopia de infrarrojo cercano (NIRS)

La espectroscopia estudia la interacción entre la radiación electromagnética y la materia con la finalidad de poder analizar y explicar dichas interacciones. En el caso de la espectroscopia de infrarrojo cercano se emplea el rango de longitudes de onda que van desde el final del espectro visible (Vis), 780 nm ( $12500\text{ cm}^{-1}$ ), hasta el inicio de la región del infrarrojo medio (MIR), 2500 nm ( $4000\text{ cm}^{-1}$ ). Dicha región recibe el nombre de infrarrojo cercano (NIR). La espectroscopia NIR se caracteriza por ser una técnica rápida, económica y no destructiva. Aunque en ocasiones puede ser aconsejable, no requiere el procesado de la muestra ni la adición de ningún tipo de sustancia [39–41].

La región del infrarrojo fue descubierta en el año 1800 por Fredrick William Herschel mediante un sencillo experimento, en el cual se generaba la difracción de la luz mediante un prisma [42]. Herschel midió la temperatura de las diferentes bandas del espectro visible, observando que esta iba en aumento a medida que se desplazaba de la región del azul hacia la banda del rojo. Al observar este incremento de temperatura, prosiguió con las mediciones. De esta forma, descubrió que la medición de la temperatura alcanzaba su máximo en una región superior a la banda del rojo, fuera del espectro visible. Herschel acababa de descubrir que la energía del sol no sólo se limitaba al espectro visible (lo que podemos ver) [43]. Esto no sólo supuso el descubrimiento de la región del infrarrojo, sino que además marcó el punto de partida para el descubrimiento del resto de regiones del espectro electromagnético [43]. No obstante, no fue hasta llegada la segunda mitad del siglo XX, cuando la espectroscopia de infrarrojo cercano experimentó un gran impulso, gracias especialmente a los trabajos desarrollados por Karl Norris, juntamente con las importantes contribuciones de Williams [44], McClure [45] y Shenk [43,46,47]. Actualmente, la espectroscopia de infrarrojo cercano se encuentra presente en

numerosos sectores, en los que se emplea para diversos usos. Por ejemplo, en la industria farmacéutica [48–50] y petroquímica [51,52], en el sector agrícola [53–55] y, también, en el sector sanitario [56–58], entre muchos otros. Asimismo, hoy en día existen en el mercado una amplia variedad de equipos NIR, tanto instrumentos más precisos de uso restringido en el laboratorio, como equipos portátiles o de uso *on-site* y *on-line* con capacidad para ser utilizados en ambientes más extremos [59], lo que facilita su implementación en múltiples ambientes.

La absorción de radiación electromagnética dentro de este rango genera un cambio en el estado energético de la molécula, el cual se traduce en un cambio en el estado vibracional de los enlaces atómicos que constituyen dicha molécula [60]. De esta forma, se define una relación entre el estado energético y el estado vibracional.

Para entender los fundamentos de la espectroscopia de infrarrojo cercano y, de esta forma, poder realizar una correcta interpretación de los espectros resultantes, es necesario comprender los elementos o conceptos básicos de la espectroscopia vibracional.

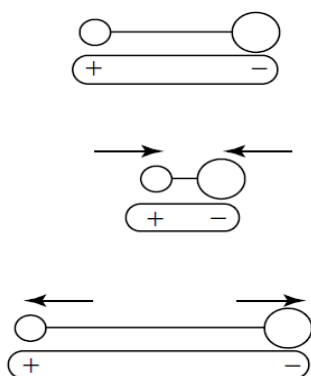
### 2.3.1 Conceptos de la vibración molecular en el infrarrojo cercano

Se denomina vibración molecular al movimiento de los átomos que conforman la molécula, dentro de un rango definido por las particularidades de los átomos que constituyen los enlaces covalentes que dan forma a dicha molécula. En este tipo de enlaces se produce la compartición de uno o más electrones entre los dos átomos que conforman el enlace. Además, hay átomos que tienen la capacidad de atraer con más fuerza hacia su órbita a los electrones compartidos en el enlace covalente, los cuales reciben el nombre de átomos electronegativos. Por el contrario, los átomos que tienen menor capacidad de atracción de los electrones compartidos reciben el nombre de elementos electropositivos. Este fenómeno genera la formación de un potencial de carga en el enlace, dando lugar a un dipolo. De esta forma, se genera un ambiente más electronegativo alrededor del átomo más electronegativo, y un

## Introducción

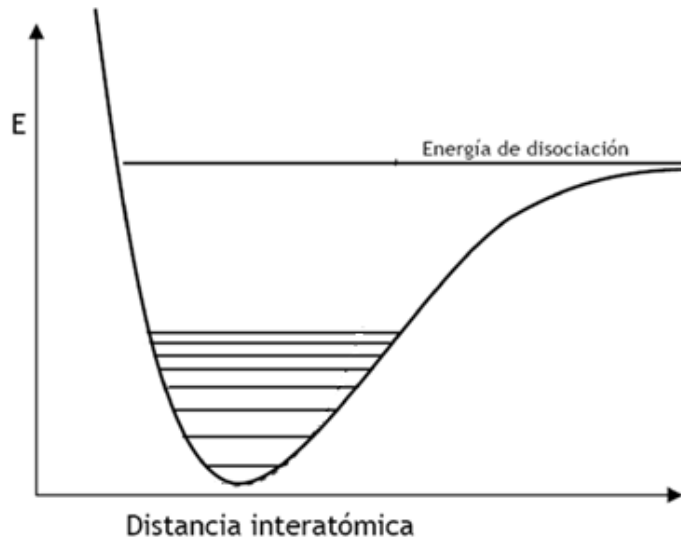
ambiente más electropositivo en el entorno del átomo electropositivo. La intensidad o magnitud del potencial del dipolo recibe el nombre de momento dipolar, el cual se encuentra en continuo cambio debido a las vibraciones de los enlaces químicos. Cuando el dipolo se forma a nivel de enlace se habla de dipolo local, en cambio a escala molecular se habla de molécula polar (momento dipolar netos).

Durante la vibración, los átomos de un enlace se aproximan y se alejan el uno al otro de manera cíclica. En la **Figura 4** se ilustra este movimiento. Esto genera que la energía potencial cambie continuamente, haciéndose máxima cuando los átomos se encuentran en el punto de mayor proximidad debido a las fuerzas de repulsión, y también en el punto de mayor distanciamiento entre ambos como consecuencia de las fuerzas de atracción. En cambio, se alcanza un punto de energía próximo a cero cuando los átomos se encuentran a la distancia de equilibrio [61]. De esta forma, el movimiento de los átomos en la molécula está confinado dentro de un pozo de potencial descrito gráficamente en la **Figura 5**.



**Figura 4.** Movimiento de un enlace covalente en una molécula diatómica heteronuclear [60].





**Figura 5.** Energía potencial en el modelo del oscilador no armónico. Modificación [61]

La frecuencia vibracional de las moléculas depende del estado energético de las mismas y, además, no cambia de manera continua, sino todo lo contrario. La variación se da de forma discontinua, saltando de una frecuencia vibracional discreta a otra (cuanto o *quantum*), sin pasar por las frecuencias intermedias, dando lugar a lo que se conoce como estados vibracionales. Estos estados vibracionales se definen como estados energéticos cuánticos [60,62].

La forma más precisa de describir el movimiento vibracional de un enlace es mediante el modelo de oscilador no armónico [63] ya que, como se observa en la **Figura 5**, los potenciales de energía no son simétricos a ambos lados del punto de equilibrio (punto de menor energía del enlace) ni la distancia energética entre estados vibracionales contiguos es siempre la misma (a mayor energía los estados vibracionales se encuentran más juntos). Esto se debe a la existencia de fuerzas que influyen en la vibración, como la repulsión coulombica entre los núcleos de los átomos y la disociación del enlace, ambas contempladas en el modelo de oscilador no armónico [61].

Los estados vibracionales vienen definidos por los átomos que conforman los enlaces que constituyen la molécula. Por ello, la absorbancia de energía necesaria para que se produzca un cambio en el estado vibracional es específica para cada enlace químico. Aunque también se ve influenciada por la disposición de los átomos que conforman los enlaces contiguos, además de las moléculas adyacentes y la composición de la matriz química donde se encuentra la molécula [59,64,65]. Todo ello define en gran medida el comportamiento de las moléculas en el espectro del infrarrojo cercano y, por consiguiente, el espectro de la muestra.

Para que se produzca transferencia de energía entre la radiación incidente y una molécula, esta última debe experimentar un cambio neto en el momento dipolar, debido al movimiento vibracional y rotacional de sus enlaces [61]. Cuando esto ocurre, el campo eléctrico asociado a la radiación electromagnética puede interactuar con el campo eléctrico generado por la fluctuación del momento dipolar. Cuando la frecuencia de la radiación iguala a la frecuencia de vibración de la molécula se produce una transferencia de energía que genera un cambio en la amplitud del movimiento vibracional de la molécula [60–62]. Es importante destacar que la energía transportada por la radiación electromagnética ha de ser suficiente como para salvar la distancia energética entre los diferentes estados vibracionales.

Debido a estos requisitos, no todas las moléculas son susceptibles de absorber energía en la región del infrarrojo cercano. Un ejemplo de moléculas no activas en la región del NIR son las especies homonucleares ( $H_2$ ,  $O_2$ ,  $N_2$ , etc.), pues al estar formadas por átomos de un mismo elemento, no hay una atracción diferencial de los electrones hacia los átomos, con lo cual, el momento dipolar se mantiene en todo momento a cero. Además, debido a las particularidades de la espectroscopia NIR, las principales moléculas que son activas en esta región espectral son aquellas en cuyos enlaces está presente el átomo de hidrogeno asociado a otros átomos ligeros, como son los enlaces C-H, N-H, O-H y S-H. Esto se debe a que el átomo de hidrogeno experimenta oscilaciones muy acusadas debido a su bajo peso atómico, lo cual

genera bandas de absorción de mayor intensidad que otros enlaces en los que no está presente. Por ejemplo, las bandas correspondientes a enlaces como C=O, C-C y el C-Cl son muy débiles o, incluso, ausentes en la región del NIR [62,66]. Aunque, si estos enlaces se localizan adyacentes a enlaces C-H, pueden afectar a la vibración de este y, por consiguiente, a su huella espectral. Tal es así, que es posible diferenciar hidrocarburos saturados de insaturados mediante la espectroscopia NIR [60].

Estos conceptos son clave para reconocer qué compuestos son sensibles a la radiación infrarroja y, por tanto, influirán en el espectro que se obtenga, y que compuestos, pese a estar presentes en la muestra, no serán visibles en la huella espectral en el infrarrojo cercano.

### 2.3.2 Espectro NIR

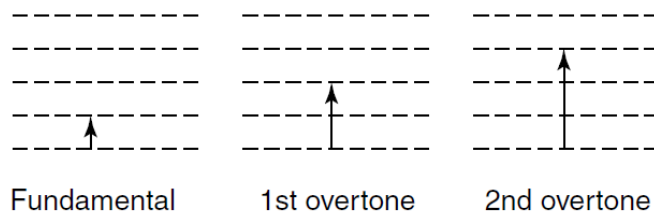
La espectroscopia NIR se caracteriza por una baja absorción y una alta reflectividad, también denominada *scattering* [59]. Los espectros presentan bandas de absorción anchas, poco definidas y de baja intensidad que las observables en el resto de las regiones del infrarrojo (IR). Su origen está asociado a sobretonos o armónicos (*overtones* en inglés), además de bandas de combinación de las vibraciones fundamentales [59,64,65]. En la espectroscopia NIR no son visibles las transiciones entre estados vibracionales fundamentales.

Los sobretonos son transiciones energéticas entre estados vibracionales no contiguos ( $\Delta v = \pm 2, \pm 3, \dots$ ), lo cual implica que para salvar la distancia energética, estas transiciones requieren de una mayor aportación de energía en comparación con las transiciones vibracionales fundamentales ( $\Delta v = \pm 1$ ). Esto hace que la probabilidad de que suceda una transición de primer sobretono ( $\Delta v = \pm 2$ ) sea, dependiendo del enlace, entre 10 y 100 veces más baja que la de las transiciones entre estados vibracionales contiguos. Por esta razón, únicamente son observables el primer y segundo sobretonos [61]. Todo ello genera que las transiciones de tipo *overtone* se den en un número inferior al de las transiciones fundamentales, lo que explica porque

las bandas de absorción, en la región del infrarrojo cercano, son de baja intensidad. Las bandas de los sobretonos aparecen entre 780 y 2000 nm dependiendo del orden del sobretono, la naturaleza y fuerza del enlace químico, cubriendo de esta forma buena parte de la región del NIR. Por otro lado, las bandas de combinación de las vibraciones fundamentales únicamente se dan en moléculas poliatómicas cuando se producen cambios simultáneos en la energía de dos o más modos de vibración. Las frecuencias de las bandas de combinación son combinaciones lineales de las frecuencias fundamentales que interaccionan [61].

$$f_{comb} = m_1 f_1 \pm m_2 f_2 \pm m_3 f_3 \pm \dots \quad (5)$$

donde  $m_i$  son números enteros y  $f_i$  las frecuencias de las transiciones que constituyen la banda de combinación. Estas bandas aparecen en el rango del NIR entre 1900 y 2500 nm. La **Figura 6** muestra unos ejemplos de transiciones energéticas correspondientes a bandas fundamentales y sobretonos.



**Figura 6.** Niveles energéticos para las bandas fundamentales y sobretonos [60].

La anchura de las bandas en el NIR se debe principalmente al solapamiento de sobretonos y bandas de combinación. Esto es debido a que una misma molécula puede dar lugar a múltiples combinaciones de sobretonos y bandas de combinación a lo largo de la región del NIR (mayor orden del sobretono menor intensidad de banda) [59,64]. Esta característica de superposición de bandas dificulta la identificación de los componentes de una muestra. Pese a ello, sí que se pueden relacionar longitudes de onda con la absorción de radiación por parte de algunos enlaces concretos para la cuantificación de algunos parámetros de la muestra como

el contenido en agua, proteínas o materia grasa, entre otros [67–69]. También permite definir la huella espectral del objeto de interés, para luego crear modelos de discriminación [70–72]. Además, en la bibliografía es posible encontrar tablas y figuras que nos pueden ayudar a la hora de identificar los enlaces químicos predominantes en el espectro [63], tal y como se muestra en la **Tabla 1**. Es importante remarcar que la espectroscopia de infrarrojo cercano no se suele utilizar para el estudio de la composición de una muestra (obtención de los componentes específicos), sino más bien para el estudio de características más cualitativas y, en algunos casos, cuantitativas (proteína, humedad, materia grasa, etc.).

**Tabla 1:** Bandas más comunes de compuestos orgánicos en el infrarrojo cercano. Modificación [60].

Longitud de onda (nm)	Asignación
2200–2450	Combinación C–H stretching
2000–2200	Combinación N–H stretching, combinación O–H stretching
1650–1800	Primer sobretono C–H stretching
1400–1500	Primer sobretono N–H stretching, primer sobretono O–H stretching
1300–1420	Combinación C–H stretching
1100–1225	Segundo sobretono C–H stretching
950–1100	Segundo sobretono N–H stretching, segundo sobretono O–H stretching
850–950	Primer sobretono C–H stretching
775–850	Tercer sobretono N–H stretching

En general, la reflectancia en el NIR tiene dos componentes, la radiación reflejada de forma difusa y la reflejada de forma especular (reflejos en forma de espejo). Los reflejos especulares normalmente se minimizan mediante el diseño del instrumento y la geometría de muestreo, ya que no contienen ninguna información química. La luz reflejada de forma difusa, que es aquella que se refleja en una amplia gama de

direcciones, es la principal fuente de información en los espectros NIR [73]. Esta contiene información sobre la composición química de la muestra (absorción) y sobre su microestructura (dispersión), donde se incluye el tamaño de partícula, presencia de rugosidades, células, fibras, fluctuaciones de densidad, etc. Las dos principales formas de dispersión de la radiación (que no incluyen la transferencia de energía con la muestra) son la de Rayleigh y la de Lorentz-Mie.

La dispersión de Rayleigh depende en gran medida de la longitud de onda. Se produce cuando las partículas tienen un diámetro mucho menor que la longitud de onda de la radiación electromagnética [73]. En cambio, la dispersión de Lorentz-Mie aparece cuando las partículas presentan tamaños mayores al de la longitud de onda, como ocurre generalmente en la espectroscopia NIR. Esta depende de las formas de las partículas y no depende en gran medida de la longitud de onda [73].

Por lo general, las propiedades de dispersión de la radiación asociadas a la estructura, como es el caso del tamaño de partícula, suelen ser minimizadas mediante técnicas de procesamiento de espectros, ya que causan un aumento del ruido de los datos (variabilidad espectral). Este aspecto es tratado en mayor profundidad en el apartado 2.6.1.

## 2.4 Cámara espectral

La cámara espectral es una herramienta que combina espectroscopia e imagen [74]. Por un lado, la imagen proporciona información sobre la intensidad de la luz reflejada por la muestra en cada uno de los píxeles que la forman. Por otro lado, la espectroscopia aporta información espectral. Una cámara espectral fusiona ambas técnicas, ofreciendo como resultado un espectro por cada píxel [74]. De esta manera, es posible aprovechar la gran información aportada por los espectros, junto con las ventajas de análisis que ofrecen las imágenes.

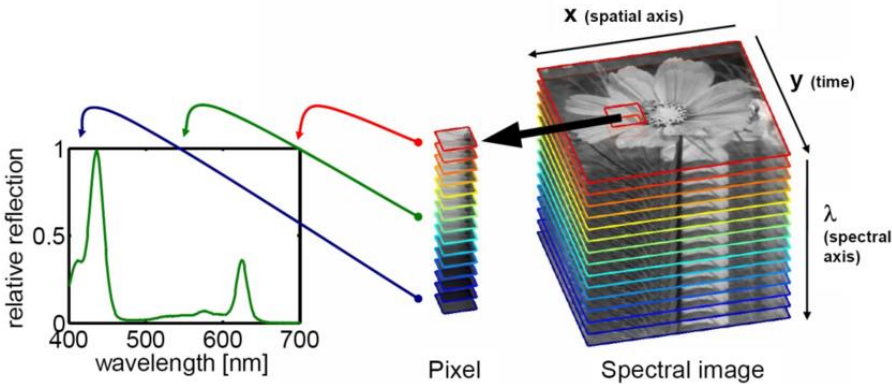
La espectroscopia de imagen surgió en la década de 1970, cuando se empezaron a realizar mediciones de campo en apoyo al análisis de datos del satélite Landsat-1 (1972), el cual portaba un escáner multispectral [75,76]. El término “imagen hiperespectral” apareció por primera vez en una publicación científica en el año 1985 [77]. Debido a que sus primeras aplicaciones se centraron en el área de la teledetección, la propia NASA participó en su desarrollo. En la actualidad, las cámaras espectrales han demostrado ser una herramienta eficaz en la industria alimentaria para el control de adulteraciones [78,79] y para la evaluación de la calidad y seguridad de productos alimentarios [80,81], también en el ámbito sanitario [82,83] y en el estudio de la Tierra y medioambiente [84,85]. Otro sector en el que las cámaras espectrales destacan especialmente es el agrícola, donde se emplean, por ejemplo, en la agricultura de precisión [86,87], control de fitopatologías [88,89] y discriminación de especies o variedades [90,91],

Con un equipo de espectroscopia que no incluye imagen se obtienen espectros sin definición espacial [76,92]. Además, el espectro resultante es el promedio espectral de toda la región analizada, la cual suele ser pequeña. Por el contrario, gracias a la imagen (mapa espacial) que proporciona la cámara espectral es posible analizar regiones específicas de la muestra de manera independiente, eliminar u obviar aquellas que no son de interés, buscar regiones de la imagen que presenten unas características espectrales específicas, además de ofrecer una mayor área de análisis.

Las cámaras espectrales captan la radiación electromagnética reflejada por la muestra en unas bandas espectrales específicas, obteniéndose una imagen para cada una de estas bandas [74]. Las múltiples imágenes adquiridas se apilan unas encima de otras en un orden correlacionado con la distribución de las bandas en el espectro electromagnético. A continuación, se alinean los píxeles de todas ellas, dando lugar a lo que se conoce como cubo de datos de imagen espectral o hipercubo espectral, tal como se muestra en la **Figura 7**. El hipercubo espectral presenta tres dimensiones,

## Introducción

dos de ellas son espaciales y conforman el plano de la imagen. La tercera dimensión se genera por el propio apilamiento de las imágenes y es la que contiene la información espectral [93].



**Figura 7.** Ejemplo de un cubo de datos de una imagen hiperespectral. Cada píxel consta de un espectro de reflectancia completo en su posición [177].

Para obtener el espectro de una región particular de la imagen (muestra analizada), lo que se hace es graficar la intensidad de los píxeles de la región seleccionada en cada una de las imágenes que conforma el hipercubo, o lo que es lo mismo, en cada una de las bandas espectrales. De esta forma, se obtiene un gráfico en dos dimensiones del espectro de reflectancia de la región seleccionada, donde el eje de ordenadas representa la intensidad de píxel y el de abscisas la longitud de onda.

Se pueden distinguir dos tipos de cámaras espectrales en función del número de bandas en las que trabajen:

- **Cámaras multiespectrales:** miden como máximo en 10 bandas del espectro. En este tipo de cámaras, por lo general, las bandas son anchas (varias longitudes de onda) y el rango de longitudes de onda puede ser discontinuo [93].



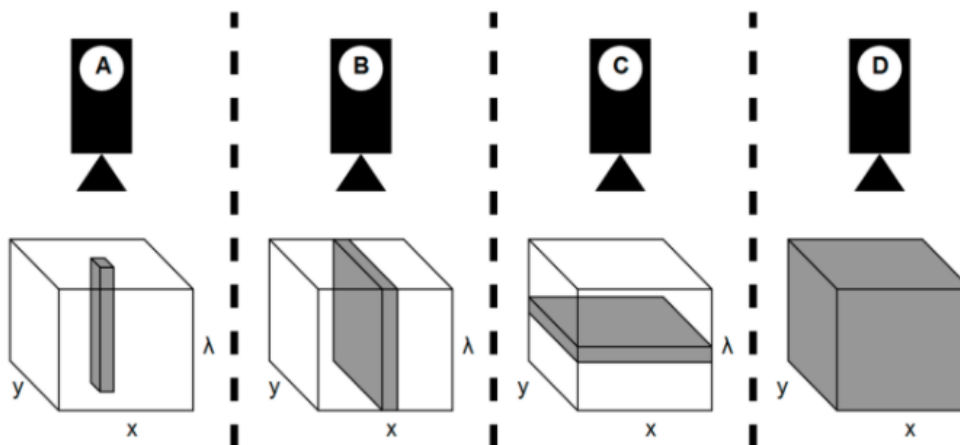
- Cámaras hiperspectrales: miden en más de 10 bandas del espectro. Pueden trabajar en cientos de bandas estrechas que cubren un rango continuo de longitudes de onda.

También se distinguen varios tipos de cámaras espectrales dependiendo de la forma en la que se adquieren los datos [76,94] (**Figura 8**):

- **Whisk-broom** o **barrido de puntos**: adquiere todas las bandas a la vez píxel por píxel moviendo el detector en el plano  $xy$ . El inconveniente más destacado es la lentitud en la adquisición de datos.
- **Push-broom** o **barrido de líneas**: en lugar de un escaneo basado en píxeles, se adquiere una secuencia de píxeles que forman una línea. La adquisición de las bandas espectrales se realiza al unísono. El escaneo de la muestra se realiza por el desplazamiento de la cámara por encima de la muestra o viceversa a lo largo de una dimensión espacial ( $x$  en el ejemplo de la **Figura 8**). Este método se caracteriza por ser más simple y proporcionar una mayor relación señal/ruido. El principal inconveniente es que debe usarse un tiempo de exposición lo suficientemente corto para evitar el riesgo de inconsistencias a nivel de las bandas espectrales (saturación o subexposición).
- **Barrido de planos** o **escaneo de longitud de onda**: Mantiene fijo el campo de visión de la imagen, es decir, no hay desplazamiento ni de la cámara ni de la muestra. Para construir el cubo espectral se adquiere una imagen monocromática de dos dimensiones espaciales ( $x, y$ ) del campo de visión completo en una sola longitud de onda a la vez. Este proceso se repite sucesivamente para el resto de las longitudes de onda. Cabe destacar que este método no es adecuado para entornos en movimiento.
- **Full-Field** o **disparo único**: este método adquiere todos los datos espaciales y espectrales a la vez. Es el más moderno de todos y se encuentra aún en desarrollo. Su principal desventaja es la baja resolución espacial.

## Introducción

En el presente estudio se utilizó una cámara hiperespectral VNIR (visible e infrarrojo cercano) de tipo *push-broom* capaz de medir 120 bandas dentro del rango de 400 a 1000 nm, con lo cual incluye toda la región del visible y la primera parte de la del infrarrojo cercano.



**Figura 8.** Modos de adquisición de datos hiperespectrales: (A) representa el modo *whisk-broom* o barrido de punto; (B) representa el modo *push-broom* o de barrido de línea; (C) corresponde al modo de barrido de planos; (D) representa el modo disparo único [76].

La absorción de energía en la región del espectro visible causa cambios energéticos a nivel de electrones, a diferencia del infrarrojo cercano, en el cual los cambios son a nivel de estados vibracionales. El proceso es similar al explicado anteriormente para la espectroscopia vibracional. Cuando una molécula absorbe radiación en el espectro visible, pasa de un estado energético basal a uno de mayor energía, llamado estado de excitación. Esto sucede porque la energía absorbida genera que un electrón pase a un estado de valencia superior. Los diferentes estados energéticos son característicos de cada molécula, lo cual define el espectro de absorción. Es importante recordar que para que se produzca un cambio en el estado energético, la radiación incidente debe aportar una energía equivalente a la distancia energética entre los diferentes estados. Una vez la molécula se encuentra excitada,

esta retorna a su estado energético basal mediante la emisión o transmisión de energía.

Al trabajar con material vegetal, los espectros resultantes dependen en gran medida de la interacción de la radiación incidente con los pigmentos de la planta, aunque también influyen otros aspectos, como la estructura de la propia hoja.

## 2.5 Visión artificial junto con *Deep learning* (CV-DL)

La visión artificial, también denominada visión por computador, consiste en adquirir, procesar y analizar, mediante ordenadores, imágenes del mundo real con la finalidad de que estos puedan interpretar dichas imágenes y ejecuten una respuesta autónomamente en base a dicha interpretación. Los datos que emplea la visión artificial son imágenes, ya sean fotografías, imágenes en vídeo, escáneres, etc. La visión artificial conjuga varios procesos concatenados, siendo las fases de análisis e interpretación de los datos donde entra en juego la inteligencia artificial.

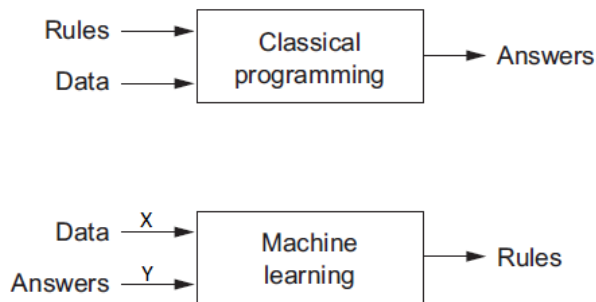
### 2.5.1 Inteligencia artificial y Machine learning

La Inteligencia artificial (IA) surgió en la década de 1950 y se puede definir como el esfuerzo de automatizar tareas intelectuales normalmente realizadas por humanos [95]. No obstante, en la década del 1940 ya existían las primeras técnicas basadas en redes neuronales (*neural networks*, NN), las cuales fueron abandonadas a finales de la década de 1960. Desde su aparición, la inteligencia artificial ha experimentado varias oleadas de interés y optimismo, seguidas de decepción y abandono [96].

Hasta finales de la década de 1980, la IA se basaba en reglas creadas por programadores (programas), mediante las cuales, el programador definía cual debía ser la respuesta del sistema para cada situación [97]. Es lo que se conoce como programación clásica o *Symbolic AI*. Se creía que con la creación de un número suficiente de reglas se podía simular el pensamiento humano. Este planteamiento era

adecuado para resolver problemas lógicos bien definidos, pero era insuficiente para afrontar problemas más complejos, como el reconocimiento de voz o la clasificación de imágenes. Esto favoreció la aparición, en la década de 1990, de un nuevo enfoque denominado *machine learning* [95].

El *machine learning* (ML) se basa en introducir datos, junto con la respuesta que se espera, de manera que sea el propio sistema el que defina como procesar la información (las reglas) para lograr el resultado deseado. Posteriormente, dichas reglas pueden ser aplicadas a nuevos datos. De esta forma el sistema es entrenado en lugar de programado [95,97]. La **Figura 9** muestra gráficamente las diferencias principales entre la programación clásica y el *machine learning*. Por esta razón, es a partir de este punto cuando ya se habla de aprendizaje automático.



**Figura 9.** Comparación conceptual entre la programación clásica y el *machine learning* [95].

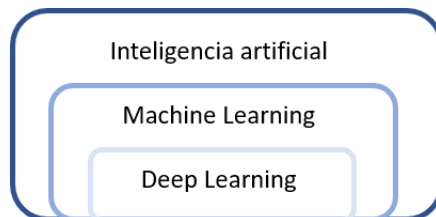
Para lograr un correcto entrenamiento del sistema, este debe recibir mucha información en forma de ejemplos de entradas y de salidas, de manera que se encuentre una estructura estadística en estos ejemplos que finalmente permita que el sistema genere reglas para automatizar la tarea. Para poder crear modelos de *machine learning*, son necesarios tres elementos [95]:

- Datos de entrada: son los datos de ejemplo que se van a emplear para entrenar el modelo, los cuales conforman la matriz X.

- Ejemplos de las salidas esperadas: la respuesta esperada a los datos de entrada. Corresponde a la matriz  $Y$ .
- Una manera de medir si el modelo está trabajando bien

La clave del *machine learning* es que el sistema “aprenda” a transformar o procesar los datos de entrada con el objetivo de obtener representaciones de los mismos (datos transformados) que ayuden a encontrar una relación con la salida esperada. Una representación es una forma diferente de ver los datos (representarlos o codificarlos) [95]. Por ejemplo, una imagen se puede representar en formato rojo-verde-azul (RGB, por sus siglas en inglés) o en formato de matiz, saturación e intensidad (HSV, por sus siglas en inglés). Dependiendo de la tarea que se deba desarrollar, una representación puede ser más útil que otra. Los modelos de *machine learning* tienen que encontrar las representaciones de los datos más adecuadas para el caso de estudio. En el contexto del *machine learning* la palabra “aprender” (*learning*) describe el proceso automático de búsqueda de las representaciones más adecuadas. Esta búsqueda se realiza dentro de un conjunto de operaciones predeterminadas (proyecciones lineales, translaciones, operaciones no lineales, etc.) denominadas espacio de hipótesis [95].

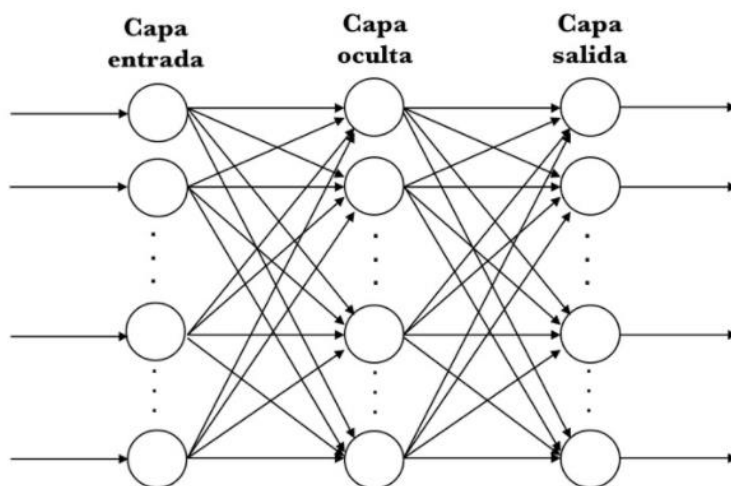
A partir de este punto, han ido surgiendo nuevos métodos de aprendizaje automático, todos ellos basados en los conceptos desarrollados en el *machine learning*. Uno de los más importantes es el aprendizaje profundo o *Deep learning*. La **Figura 10** ilustra la relación entre inteligencia artificial (IA), *machine learning* (ML) y *deep learning* (DL).



**Figura 10.** Relación entre Inteligencia artificial (IA), *Machine learning* (ML) y *Deep learning* (DL) [96].

## 2.5.2 Conceptos del aprendizaje profundo o *Deep learning*

El *deep learning* es un subcampo específico dentro del *machine learning*, de forma que todo lo anteriormente explicado sobre el *machine learning* es totalmente aplicable al *deep learning*. El *deep learning* fue desarrollado en la década del 2010, gracias principalmente a las aportaciones de Geoffrey E. Hinton, Yoshua Bengio y Yann LeCun [98], con la finalidad de incrementar el potencial de las representaciones de los datos originales. Por ello, se caracteriza por presentar una arquitectura de red neuronal computacional, la cual consiste en capas sucesivas e interconectadas de neuronas computacionales en las que se realizan representaciones de los datos de entrada. Tal y como se puede observar gráficamente en la **Figura 11**, cada capa se nutre de las transformaciones realizadas por la capa anterior, esto implica que a medida que nos adentramos en las sucesivas capas de la red neuronal, se alcanza un nivel de abstracción superior (mayor transformación de los datos de entrada).

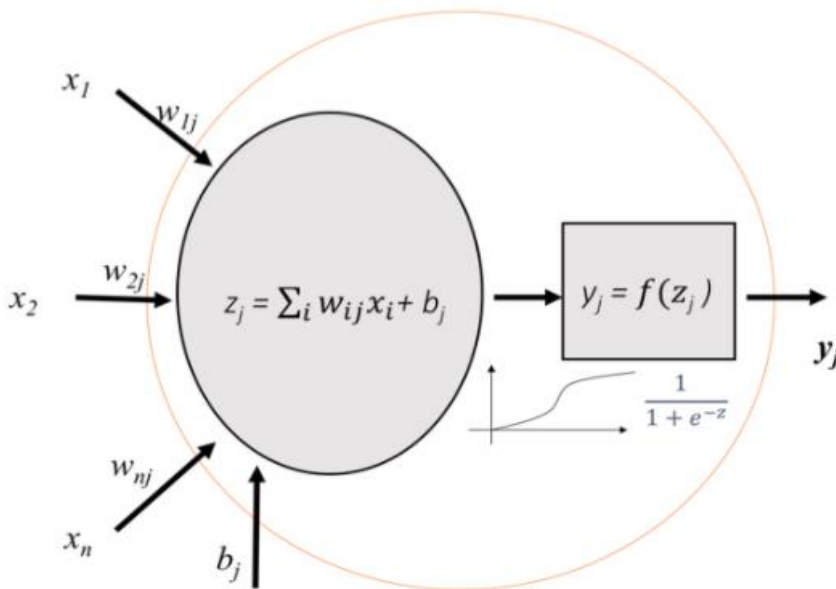


**Figura 11.** Ejemplo de red neuronal computacional [96].

Los datos son introducidos dentro de la red neuronal a través de la capa de entrada. La capa de salida es la que reporta la respuesta. Las capas del centro reciben

el nombre de capas ocultas y son las encargadas de procesar los datos de entrada. En la actualidad, no existe un consenso claro por lo que respecta al número de capas que una red neuronal ha de tener. Si bien es verdad que al incrementar el número de capas se aumenta el nivel de procesamiento de los datos, lo que es favorable para las tareas más complejas. Esto a su vez puede ser negativo debido a que aumenta el número de parámetros dificultando la correcta convergencia de la red.

Los datos de entrada son transformados por las neuronas computacionales, las cuales se tratan de funciones no lineales que reciben el nombre de funciones de activación. La función de activación recibe unas entradas ( $x_i$ ), cada una de ellas es multiplicada por un peso específico ( $w_i$ ). A la suma de los productos resultantes ( $z_j = \sum x_{ij} w_i$ ) se le añade un sesgo o *bias* ( $b$ ). Esto da como resultado la salida de la función, también denominada activación ( $y_j$ ) [97]. Esta es transmitida, como entrada, al siguiente grupo de neuronas a las que está conectada. La **Figura 12** muestra un ejemplo de neurona computacional, también llamada Perceptrón.



**Figura 12.** Ejemplo de neurona computacional [96].

Durante la fase de entrenamiento, se denomina aprendizaje del modelo a la búsqueda de los valores para los parámetros, pesos y sesgo, que dan lugar a las representaciones más útiles. Por consiguiente, el valor de dichos parámetros representa lo que ha aprendido la neurona hasta ese momento. Una vez se tiene un modelo entrenado, si se analizan las neuronas por separado, se puede ver que cada neurona realiza una parte de la tarea global realizada por la red neuronal. Además, cada neurona aprende de manera autónoma a realizar esa tarea.

Los tipos de función de activación más utilizados son la sigmoidea, Tanh y la ReLU. Incluso existe un tipo de función de activación que está diseñada específicamente para ser colocada en la capa de salida de la red neuronal, la softmax.

- Sigmoidea: permite reducir valores extremos, los cuales se relacionan con valores atípicos (outliers), pero sin eliminarlos. Además, hace que la salida se encuentre entre 0 y 1.

$$S(x) = \frac{1}{1+e^{-x}} \quad (6)$$

donde  $x$  es el valor de entrada.

- Tangente hiperbólica o Tanh: es similar a una sigmoidea, con la diferencia que esta función se encuentra comprendida entre 1 y -1. La ventaja de Tanh respecto a la sigmoidea, es que puede trabajar más fácilmente con valores negativos.

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} \quad (7)$$

donde  $x$  es el valor de entrada,  $\sinh(x)$  y  $\cosh(x)$  son el seno y el coseno hiperbólicos, respectivamente, de la entrada  $x$ .

- Rectified Linear Unit o ReLU: la particularidad de esta función es que activa el nodo si la entrada supera cierto umbral. Este umbral se suele establecer en



0. De esta forma, si la señal es inferior a cero, no se activa, pero si es superior a cero genera una salida. Actualmente, es una función muy utilizada.

$$ReLU(x) = \max(0, x) = \begin{cases} 0 & \text{para } x < 0 \\ x & \text{para } x \geq 0 \end{cases} \quad (8)$$

donde  $x$  es el valor de entrada

- *Softmax*: Esta función está acotada entre 0 y 1, e indica la probabilidad con la que ha realizado la predicción.

$$Softmax(y)_j = \frac{e^{y_j}}{\sum_{k=1}^K (e^{y_k})} \quad (9)$$

donde  $y_j$  es una de las salidas de la capa de salida sobre la cual queremos calcular el porcentaje,  $y_k$  son todas las salidas de la capa de salida y  $K$  es el número de clases.

Es importante destacar que la función de activación  $f(z)$  no puede ser lineal porque de ser así se limitaría la capacidad de procesamiento de la red, debido a que, si se encadenan y se suman múltiples regresiones lineales, el resultado final es una regresión lineal, de manera que esto equivaldría a que nuestra red neuronal colapsara mostrando el comportamiento de una sola capa de neuronas.

### 2.5.3 Arquitectura de las redes neuronales computacionales

A nivel de arquitectura de las redes neuronales, es conveniente hablar de los dos tipos arquitectura que han sido empleadas en el presente estudio. Se tratan de las redes neuronales densamente conectadas (*Densely connected networks*) y de las redes neuronales convolucionales (*Convolutional neural networks o CNNs*).

#### Redes densamente conectadas (*Densely connected networks*)

Las redes densamente conectadas presentan una estructura clásica en donde cada neurona de una capa se encuentra conectada con todas las neuronas de la

siguiente capa, lo que recibe el nombre de *fully connected*. Este tipo de red opera sobre tensores de 1D, es decir, toda la información acaba condensada en un único vector. Intuitivamente es la arquitectura más simple, pero presenta varios inconvenientes. Por un lado, esta organización genera que haya un gran número de parámetros que deben ser ajustados durante el proceso de entrenamiento. El número de parámetros en este tipo de red depende del tamaño de los vectores de entrada. Por otro lado, este tipo de red aprende patrones globales en un espacio global de entrada [96]. Es decir, la red neuronal aprende una característica de la imagen en una posición determinada dentro de esta, si esa característica aparece en una nueva localización dentro de la imagen, la red debe aprenderla nuevamente.

Pese a estas limitaciones, las redes densamente conectadas siguen siendo utilizadas en la actualidad, ya sea para resolver problemas sencillos o acopladas al final de una red convolucional dando lugar a un clasificador.

### Redes convolucionales (*Convolutional neural network, CNNs*)

Cada capa convolucional está constituida por varios filtros o *kernels*, cada uno de los cuales realiza una transformación específica de los datos de entrada, dando lugar a los mapas de características (*features maps* en inglés). De esta forma, una capa convolucional se compone de múltiples representaciones de la imagen de entrada o *features*, una por cada filtro que contenga la capa. Las capas convolucionales operan sobre tensores de tres dimensiones (3D), dos ejes espaciales (altura y anchura) y un tercer eje denominado canal. Para una imagen en blanco y negro, el eje canal únicamente tiene una dimensión (escala de grises), mientras que si la imagen es en color, el eje canal tiene tres dimensiones (rojo, verde y azul) [96]. En el contexto de redes convolucionales, el término filtro hace referencia a la matriz de pesos y sesgo por la cual se multiplica la información procedente de la capa anterior, dando lugar a una transformación de la imagen. De esta forma, el valor de cada punto de la capa convolucional corresponde al producto escalar entre el filtro y las neuronas de la capa

de entrada a las cuales esté conectada. Funcionan de forma similar a los filtros que se usan para retocar imágenes, aunque en este caso se emplean para buscar características locales [96].

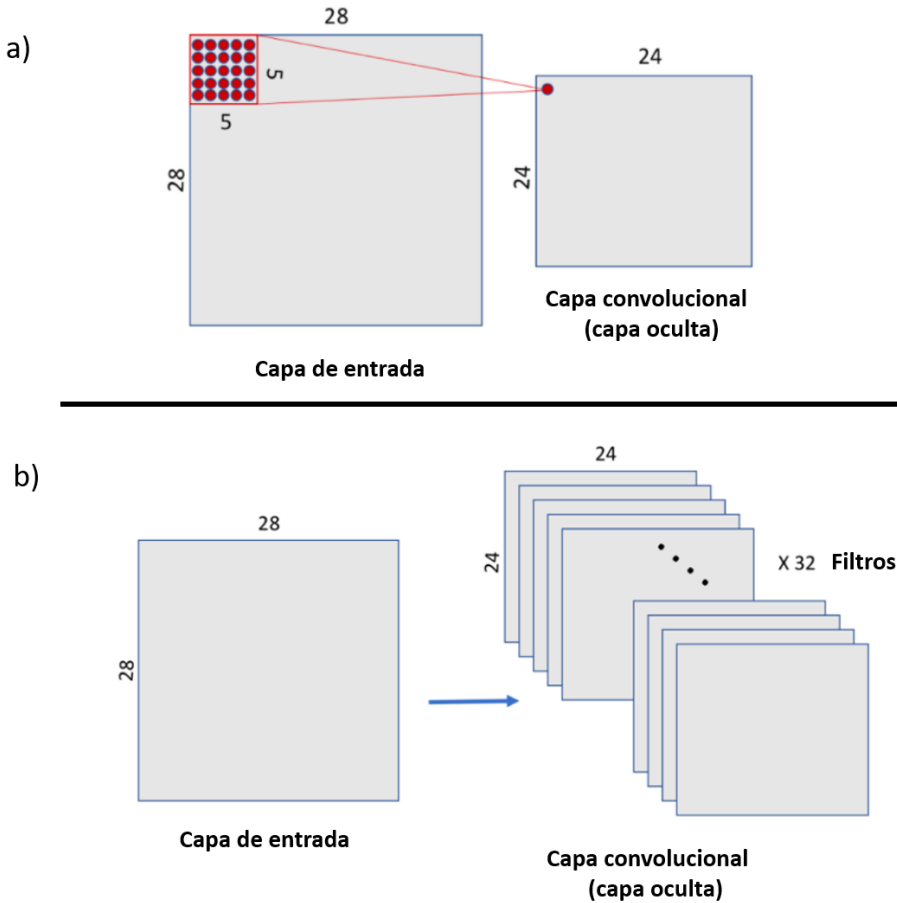
En las redes convolucionales no se conectan todas las neuronas de la capa de entrada con todas las neuronas de la primera capa oculta, como ocurre en las redes densamente conectadas. En esta tipología de red, la conexión se realiza a través de una ventana de conexión de tres dimensiones (altura, anchura y canal), la cual tiene un tamaño definido. Dicha ventana va recorriendo toda la capa anterior con un desplazamiento o paso de avance (*stride* en inglés) también determinado. Debido a las características en el desplazamiento de la ventana de conexión, se pueden dar solapamientos entre las diferentes posiciones de la ventana a medida que esta se desplaza por toda la imagen.

Esta ventana de conexión, en realidad, es un filtro, el cual a medida que va recorriendo la imagen va realizando la transformación de la misma, dando lugar de esta forma a un nuevo *feature*. Dicha transformación, es llevada a cabo por una única neurona, denominada neurona convolucional. Como se ha comentado, un filtro de una capa convolucional transforma los datos de entrada de una manera determinada, ya que emplea una misma matriz de pesos y sesgo. De forma que solo permite detectar una característica. Para poder obtener más *features* de la imagen original, se emplean varios filtros en cada una de las capas convolucionales, uno para cada característica que se desee detectar. La **Figura 13** muestra gráficamente el proceso de convolución para una capa convolucional con 32 filtros.

Al haber una única neurona por filtro, supone una gran ventaja, ya que permite reducir notablemente el número de parámetros de la red convolucional respecto a los de una red densamente conectada. Por ejemplo, si se parte de una capa de entrada con un tamaño de  $28 \times 28 \times 3$  y una ventana de conexión de tamaño  $5 \times 5$ , la cual tiene un desplazamiento de una posición. Cada una de las neuronas de la capa

## Introducción

convolucional tendrá  $5 \times 5 \times 3 + 1$  parámetros [96]. Por el contrario, en una capa densa serían 43200 parámetros ( $5 \times 5 \times 24 \times 24 \times 3$ ) más sesgos.

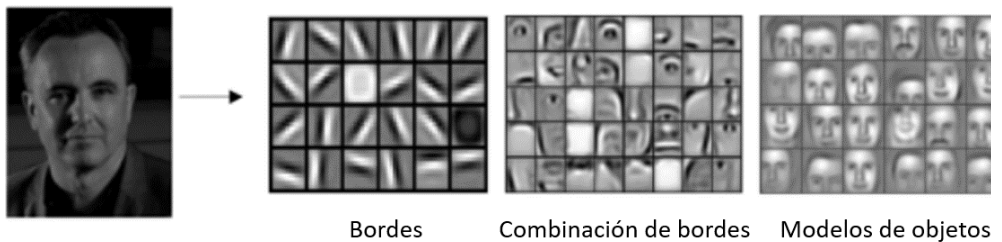


**Figura 13.** Proceso de convolución en una capa convolucional; a) Representa la condensación de información de la capa de entrada en la capa convolucional; b) Representa la construcción de 32 filtros en la capa convolucional, el conjunto de los cuales constituye la dimensión profundidad de la mencionada capa [96].

A diferencia de las redes densamente conectadas, las CNN aprenden patrones locales en pequeñas ventanas de tres dimensiones. Además, esta arquitectura permite que una vez se ha aprendido una característica, esta pueda ser reconocida en cualquier punto de la imagen. Otra característica diferencial de las redes

convolucionales respecto a las densamente conectadas, es que en las convolucionales el número de parámetros no depende del tamaño de la imagen, sino que este viene determinado por la arquitectura de la red (número de capas, número de filtros, etc.). Tal como se muestra en el ejemplo de la

**Figura 14**, conforme la información va pasando de una capa hacia otra más profunda, la red emplea los patrones básicos aprendidos en las capas inferiores para aprender patrones más complejos y abstractos.



**Figura 14.** Ejemplo del procesamiento de la información por parte de una red neuronal convolucional (CNN). Imagen modificada [96].

Este tipo de redes computacionales, a parte de las capas convolucionales, suelen presentar otro tipo de capas llamadas capas de *pooling*, las cuales siempre se encuentran detrás de una capa de convolución. La función de estas capas es simplificar la información de salida de las capas convolucionales creando una versión condensada de la misma [96]. De esta forma se reduce el tamaño de los mapas de características a procesar [95]. Su funcionamiento se basa en una ventana 2D de pequeño tamaño que recorre la salida de la capa convolucional. La información contenida en cada posición de la ventana se condensa en una neurona en la capa de *pooling*. Además, no hay una matriz de pesos con la que multiplicar la entrada, ya que se trata de un proceso que no es susceptible de ser modificado por el entrenamiento de la red (no suma parámetros a la red). Para poder entenderlo mejor, la **Figura 15** muestra gráficamente el proceso de pooling. Esta reducción de la resolución se puede realizar de dos maneras principalmente:

- *Max-pooling*: transfiere a la capa de *pooling* el valor más alto de los contenidos en la ventana;
- *Average-pooling*: realiza una media de los valores contenidos en la ventana.

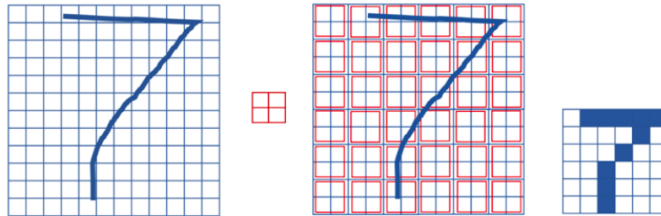


Figura 15. Representación del proceso de *pooling* [96].

## 2.5.4 Entrenamiento de una red

Antes de entrenar una red neuronal, se debe dividir el conjunto de datos en tres subconjuntos: de entrenamiento, de validación y de test. El de entrenamiento se emplea para que el modelo pueda ajustar sus parámetros. El conjunto de validación se emplea para comprobar lo bien o mal que está aprendiendo nuestro modelo. Una vez entrenado el modelo, se evalúa su efectividad mediante el conjunto de test, con el cual, se pretende simular una prueba real.

Para entrenar una red neuronal se deben definir una serie de argumentos y de hiperparámetros, los cuales van a determinar cómo se va a realizar el entrenamiento. Los hiperparámetros son variables de configuración que son especificadas por el programador ya que no pueden ser estimadas a partir de los datos de entrada. Los argumentos e hiperparámetros más importantes son los siguientes [95–97]:

### 1. Argumentos

- *Función de loss*: esta función cuantifica, durante el entrenamiento, la cercanía entre la predicción realizada por la red y la salida esperada, es decir, el error. A medida que se van realizando ajustes en los parámetros del

modelo, nos vamos desplazando por la función de *loss* en busca del mínimo absoluto, o lo que es lo mismo, el mínimo error. Existen diversas funciones de *loss* o, dicho de otra forma, diversas maneras de calcular el error, como el error absoluto, el error cuadrático o la entropía cruzada. Dependiendo de la tarea a resolver se escoge una función de *loss* u otra.

- Optimizador: es el algoritmo que permite a la red neuronal calcular los pesos de los parámetros a partir de los datos de entrada y de la función de *loss*, indicando la dirección y la magnitud del ajuste de los parámetros del modelo. La finalidad del optimizador es lograr el mejor ajuste de los parámetros para reducir al máximo el error. Existe una amplia variedad de optimizadores. La mayoría ellos se basan en el calcular el gradiente de la función de *loss* (magnitud) en la que se encuentra el modelo y ajustar los parámetros para descender por dicho gradiente (dirección). Algunos de los optimizadores más utilizados son *Adam*, *RMSprop* y *Stochastic Gradient Descent (SGD)*.
- Métrica: define la manera de monitorizar y visualizar el proceso de aprendizaje de la red neuronal. Por lo general se suele utilizar la exactitud o precisión (*accuracy* en inglés), la cual indica la fracción de imágenes correctamente clasificadas. Cuando se emplea la precisión como métrica, indirectamente se está empleando también la tasa de error, ya que ambas son complementarias (precisión = 1 – error).

## 2. Hiperparámetros

- Número de epochs: indica el número de veces que se usarán todos los datos de entrenamiento y validación durante el proceso de aprendizaje. Si el número de *epochs* es muy grande significa que realizaremos muchos ciclos de entrenamiento con los mismos datos, lo cual incrementa el riesgo de que aparezca *overfitting* o sobreajuste.
- Batch size: indica el número de datos que usaremos para cada actualización de los parámetros del modelo.

- *Learning rate*: es un valor que ayuda a ajustar la velocidad a la que el modelo se desplaza por la función de *loss*, o lo que es lo mismo, la magnitud del ajuste de los parámetros del modelo. Para ello, se multiplica el valor del gradiente por el valor de la *learning rate*, dando lugar a la magnitud de ajuste de los parámetros. Si la *learning rate* es grande, el proceso de aprendizaje se acelera, pero se corre el riesgo de que, al aproximarnos a un mínimo, el modelo salte al otro lado del mismo, entrando en un bucle de saltos alrededor del mínimo sin que este sea alcanzado o incluso provocando la divergencia del algoritmo.
- *Momentum*: su función es intentar evitar que el entrenamiento del modelo quede atrapado en un mínimo local. Para ello, calcula el promedio ponderado de los pasos anteriores para obtener así cierta inercia que le permita salir y no quedarse atascado en pequeños mínimos locales.

Mediante esta serie de argumentos se define cómo va a ser el entrenamiento, lo cual posteriormente tiene su repercusión en el éxito del mismo y, por tanto, del modelo.

El aprendizaje de una red neuronal se puede definir como un proceso iterativo, del cual podemos distinguir tres fases [95,96]:

1. *Forwardpropagation*: se introducen los datos a través de la capa de entrada. Luego se comparan las predicciones del modelo con las salidas esperadas. En base a esto se calcula el error de predicción mediante la función de *loss* (Figura 16). La función de *loss* es uno de los parámetros requeridos para cuantificar lo cercano que está una determinada red neuronal de su ideal durante el proceso de entrenamiento. Mediante el proceso de entrenamiento se busca reducir ese error y aproximar lo más posible a cero (mínimo de la función de *loss*). Para lograrlo es necesario conocer donde se ha equivocado la red para poder reducir el error, el cual se promedia usando todas las muestras de un *batch*.



2. *Backpropagation*: consiste en propagar, a través de la red, el valor de la *loss* en dirección opuesta al recorrido realizado por las muestras. Es decir, se empieza por la capa de salida y se avanza hacia la capa de entrada. Hay que tener en cuenta que debido a que una capa trabaja con las salidas de la capa anterior, el error se debe ir asignando en base a la contribución relativa que haya aportado cada neurona a la salida original. De esta forma a cada neurona de la red se le asigna un valor de *loss* que describe su contribución relativa al error total.
  
3. *Ajuste de los parámetros de la red*: Una vez se conoce la contribución de cada neurona al error total, se realiza el ajuste de los parámetros de cada neurona en base a la *loss* asignada mediante la acción del algoritmo de optimización. En el caso de las capas convolucionales, este ajuste se da a nivel de filtro (todas las neuronas de un mismo filtro comparten los mismos parámetros). De esta forma, aquellos filtros que hayan contribuido más al error global de la red neuronal experimentarán un mayor ajuste de sus parámetros. El funcionamiento del optimizador se basa en calcular la derivada de la función de *loss* en el punto que nos encontramos, lo que nos da el gradiente de la función (**Figura 16**). A continuación, se realiza un ajuste de los parámetros en la dirección que nos permita ir bajando por la función de *loss* hasta alcanzar el mínimo local más cercano, es decir, mínimo error. Es preferible hablar de mínimo local en lugar de mínimo absoluto porque la función de *loss* es muy compleja y no se puede calcular en su totalidad, sino que únicamente podemos saber su pendiente en el punto en el que nos encontramos. Por lo tanto, nunca se sabe con total seguridad si se trata de un mínimo local o global. Este proceso se realiza de forma iterativa hasta lograr los mejores resultados posibles. La **Figura 17** resume el proceso de entrenamiento anteriormente descrito.

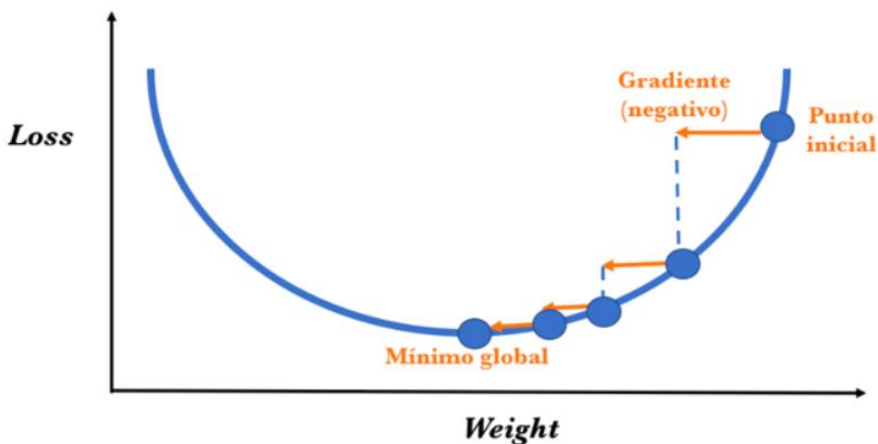


Figura 16. Proceso de entrenamiento representado sobre una función de *loss* ideal [96].

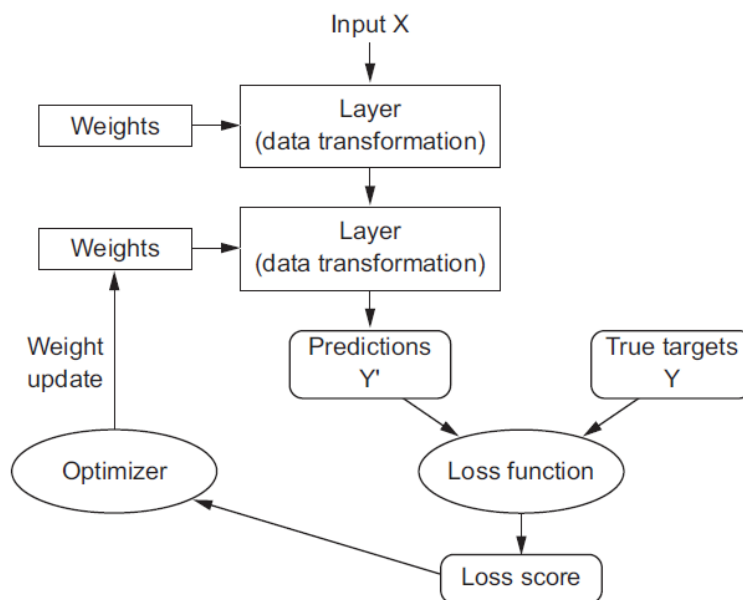
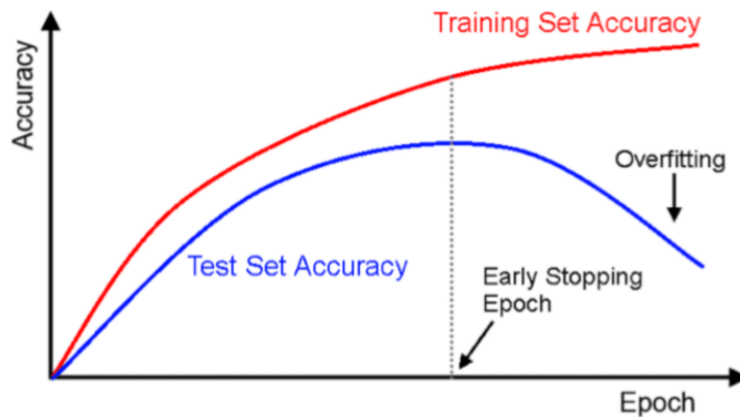


Figura 17. Esquema del entrenamiento de una red neuronal [95]

Se debe evitar realizar un entrenamiento excesivo de la red neuronal. Ya que esto podría ocasionar que los parámetros se ajustaran excesivamente bien a las muestras de entrenamiento, llegando aprender aspectos muy específicos de las mismas, pero

poco relevantes para la caracterización de la clase. Como consecuencia de ello, se reduciría la capacidad de clasificar correctamente nuevas muestras. Es lo que se denomina *overfitting* o sobreajuste [99], el cual se encuentra representado gráficamente en la **Figura 18**.



**Figura 18.** Representación gráfica de la aparición de *overfitting* durante el entrenamiento de una red neuronal computacional [95].

La aparición del *overfitting* causa un incremento repentino en el error de la validación durante el entrenamiento. Existen una serie de técnicas llamadas regularizadores, las cuales ayudan a evitar el *overfitting*. Una de ellas es el conocido como *dropout* [100], el cual consisten en eliminar aleatoriamente (establecer a cero) una serie de características de la salida de una capa neuronal durante el entrenamiento. Por ejemplo, si una capa presenta un vector de salida [0.2, 0.5, 1.3, 0.8, 1.1] para una imagen determinada, al aplicar *dropout* varios de los valores del vector escogidos aleatoriamente se transforman en ceros, [0, 0.5, 1.3, 0, 1.1]. La tasa de *dropout* hace referencia a la fracción de valores que se transforman en ceros. Por lo general esta se establece entre 0.2 y 0.5 [95]. Esta asignación aleatoria de ceros se modifica al final de cada *batch*.

## 2.5.5 Transfer learning

Al empezar a entrenar una red neuronal desde cero (*scratch*), los parámetros suelen tener un valor aleatorio, de manera que su optimización requiere de un largo y arduo proceso iterativo. Además, es necesario disponer de un gran número de datos. En muchos casos, no es posible cumplir ambos requisitos. Para estos casos existe un método llamado *transfer learning* [99,101], el cual nos permite descargar una red neuronal junto con sus parámetros ya entrenados. Luego, únicamente es necesario realizar pequeños ajustes de estos para adaptarlos a la nueva tarea que se quiera realizar. Además, este método permite mejorar los resultados del modelo [102]. Dentro del *transfer learning* existen dos modalidades:

- *Feature extraction*: Únicamente se entrenan las últimas capas densas de la red neuronal, también denominadas clasificador.
- *Fine tuning*: se entrena el clasificador junto con alguna de las capas predecesoras. Permite lograr un ajuste más fino del modelo.

## 2.5.6 Aumento de datos (*data augmentation*)

Como se ha comentado, el número de imágenes es un factor limitante a la hora de entrenar una red neuronal. Como más compleja sea la tarea a realizar y más clases se deban caracterizar, mayor número de imágenes serán necesarias. En ocasiones, adquirir un gran número de imágenes es francamente complicado. Por ello, se suele recurrir a una técnica denominada *data augmentation* [99,103]. A través de este método es posible incrementar el número de imágenes de entrenamiento. Para lograrlo, esta técnica lo que hace es modificar aleatoriamente las imágenes originales, generando de esta forma, nuevas imágenes que podrían pasar por originales. El objetivo es que el modelo nunca vea exactamente las mismas imágenes en las diferentes epochs de entrenamiento [99], lo cual también ayuda a evitar la aparición del indeseado overfitting. Estas modificaciones suelen consistir en rotar,

desplazar, cambios de zoom, cortes aleatorios, etc. Pese a sus ventajas, se debe andar con cuidado con las modificaciones específicas que se apliquen al conjunto de datos, para no generar imágenes que nunca podrían encontrarse en la realidad. De lo contrario, se estaría empeorando la eficacia del modelo [99]. Otro punto importante, es que el *data augmentation* únicamente se debe aplicar sobre el conjunto de entrenamiento, nunca sobre el de test.

### 2.5.7 Arquitectura red neuronal convolucional VGG16

La arquitectura de la red neuronal convolucional VGG (*Visual Geometry Group*) fue creada por Karen Simonyan y Andrew Zisserman en el 2014 [104] y fue empleada en la competición de *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC) [105,106]. Los autores diseñaron dos arquitecturas VGG, una compuesta por 19 capas y otra por 16 capas. Esta última es la que se ha empleado en el presente estudio.

La base convolucional de la red VGG16 está constituida por 16 capas, que están ordenadas en cinco bloques. Los bloques 1 y 2 contienen dos capas convolucionales cada uno con 64 y 128 filtros o *kernels*, respectivamente. Los otros tres bloques restantes (bloques 3, 4 y 5) están formados por tres capas convolucionales con 256, 512 y 512 filtros por bloque, respectivamente. Cada capa convolucional usa filtros con una ventana de conexión o campo receptivo de  $3 \times 3$ . El desplazamiento o paso de avance del filtro es de  $1 \times 1$  píxel, y el relleno espacial de la entrada de la capa convolucional es tal que la resolución espacial se conserva después de la convolución [104]. Además, al final de cada uno de los cinco bloques convolucionales hay una capa *max\_pooling* que trabaja sobre una ventana de  $2 \times 2$  píxeles. A continuación de la base convolucional, la red VGG16 contiene un clasificador que está compuesto por tres capas completamente conectadas (*Fully connected* en inglés). La **Figura 19** muestra un esquema de la arquitectura Vgg16.

En el presente estudio se empleó una versión modificada de la red neuronal convolucional VGG16. La red VGG16 empleada conserva toda la arquitectura del

bloque convolucional de la VGG16 original a excepción del clasificador, el cual fue sustituido por otro creado específicamente para la situación experimental de interés. Los detalles aparecen descritos en el Capítulo 4.

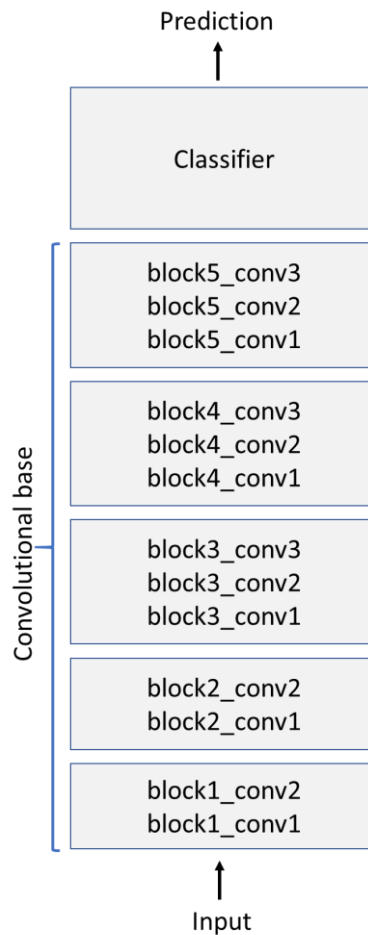


Figura 19. Arquitectura de la red neuronal convolucional VGG16. Creación propia.

## 2.6 Análisis de datos multivariantes

Se habla de datos multivariantes cuando se estudian tres o más características (variables) por muestra. Un ejemplo de datos multivariantes son los obtenidos mediante técnicas espectroscópicas, en las cuales, una misma muestra se analiza a

diferentes longitudes de onda, siendo considerada cada una de ellas una variable [107].

Para poder trabajar los datos multivariantes, estos deben ser organizados formando matrices, donde las filas son las muestras ( $i=1\dots m$ ) y las columnas son las medidas o variables de estudio ( $j=1\dots n$ ). Los datos pueden agruparse en una sola matriz o en varias matrices. Por lo general, se suele trabajar con dos matrices. Por un lado, tenemos la matriz de datos ( $X$ ), la cual contiene las variables independientes o predictoras. Esta matriz agrupa la información adquirida durante el desarrollo del ensayo. Por ejemplo, el caso de la espectroscopia serían los datos espectrales. Por otro lado, está la matriz de referencia ( $Y$ ), la cual puede contener datos cuantitativos, cualitativos o definir las clases de interés. Las variables presentes en la matriz  $Y$  reciben el nombre de variables dependientes o respuesta.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{pmatrix} \text{ Orden } m \cdot n \quad (10)$$

Una vez estructurados los datos en matrices se puede empezar a estudiar el comportamiento de las variables. Es en este punto donde entran en juego los métodos de análisis multivariante. Estos se emplean principalmente para la reducción de la dimensionalidad de los datos y la búsqueda de relaciones entre varios grupos de variables pertenecientes a una misma matriz o a varias matrices (matriz  $X$  y matriz  $Y$ ). De esta forma, el análisis multivariante ayuda al estudio conjunto de los datos, facilitando la interpretación y comprensión de la información [107]. Dentro del análisis de datos multivariantes existe un amplio repertorio de técnicas y métodos. Los más relevantes, para datos espectrales, se muestran en la [Figura 20](#).

Algunos de los métodos de análisis multivariante más utilizados son los de reducción de variables, como el análisis de componentes principales (PCA). Además

de métodos en los que se buscan relaciones entre las matrices X e Y, como es el caso de la regresión de mínimos cuadrados parciales (PLS) junto con el análisis discriminante de mínimos cuadrados parciales (PLS-DA). Algunos de estos métodos de análisis multivariante son comúnmente empleados dentro del campo de la quimiometría

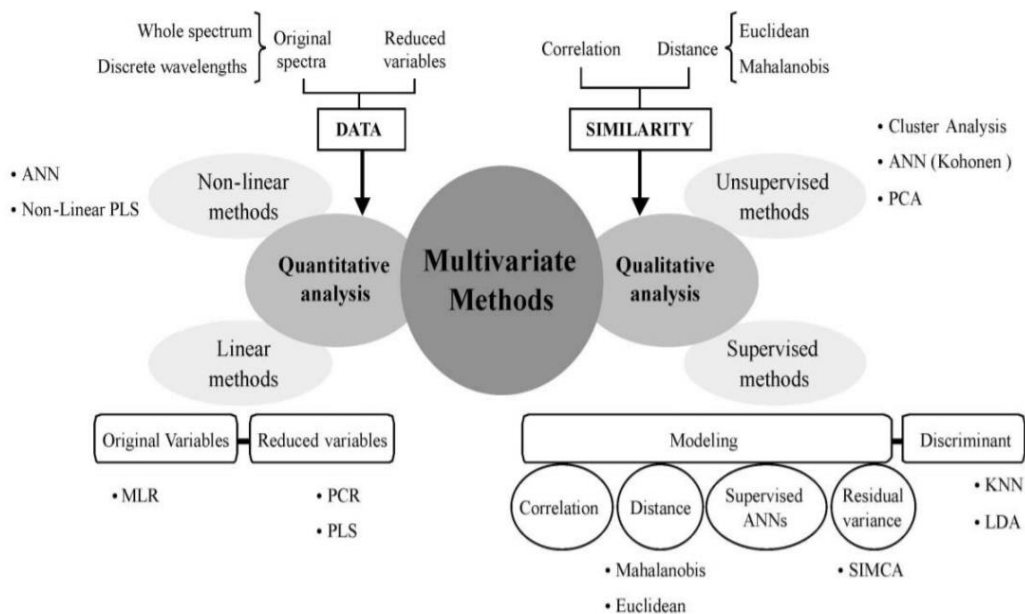


Figura 20. Esquema donde se muestran los principales métodos de análisis multivariante aplicado a la espectroscopia NIR [64].

### 2.6.1 Métodos quimiométricos de análisis

Massart (1998) definió la quimiometría como «una disciplina química que utiliza las matemáticas, la estadística y la lógica formal para a) diseñar o seleccionar procedimientos experimentales óptimos; b) proporcionar la máxima información química de relevancia mediante el análisis de datos químicos; c) obtener conocimientos sobre sistemas químicos» [108]. De esta forma, la quimiometría ayuda a extraer el máximo de información útil de los datos.



El origen de la quimiometría se remonta a los trabajos realizados por Kowalski, Isenhour y Jurs en el año 1969 [109–113]. A partir de esas primeras publicaciones, otros grupos de investigación contribuyeron con sus trabajos al desarrollo de lo que hoy se conoce como quimiometría. Uno de esos grupos de investigación fue el liderado por Svante Wold, el cual, aparte de trabajar en algoritmos de clasificación como el *Soft independent modelling by class analogy* (SIMCA), fue el primero en referirse a este nuevo conjunto de herramientas con el nombre de quimiometría en el año 1974 [114].

La quimiometría incluye en primer lugar el tratamiento de los datos, lo que nos permite resaltar la señal y reducir el ruido. Y en segundo lugar la modelización y la calibración, a través de las cuales se busca una relación cuantitativa o cualitativa entre la señal analítica y alguna propiedad de la muestra.

#### Pretratamiento de la señal

Las técnicas quimiométricas se emplean frecuentemente para mejorar la señal de los espectros obtenidos mediante espectroscopia de infrarrojo cercano [73,115–118], donde los espectros pueden verse significativamente influenciados por efectos no deseados derivados de la dispersión de la radiación, como el desplazamiento de la línea de base (efecto aditivo) y otros fenómenos agrupados bajo el término de no linealidades (*non-linearities*) [73,119]. Debido a que estos fenómenos se encuentran estrechamente ligados a la presencia de partículas de diferentes tamaños, son muy comunes cuando se trabaja con muestras trituradas o en polvo [115]. Además, las características propias de los espectros NIR, como son la presencia de bandas anchas y solapadas o la baja intensidad de pico, hacen indispensable la utilización de técnicas de filtrado de señal, las cuales en el ámbito de la espectroscopia reciben el nombre de métodos de pretratamiento de espectros. De esta forma, las técnicas quimiométricas de tratamiento de señal tienen como objetivo resaltar la componente que depende de las propiedades de la muestra y eliminar o reducir los

efectos indeseados. Ello permite mejorar el análisis exploratorio de los datos, la calibración lineal del modelo (forzar a que los datos cumplan la ley de Lambert-Beer) y la construcción de modelos de clasificación [73]. No obstante, la quimiometría no solo se aplica en el campo de la espectroscopia, sino que también está presente en otros ámbitos como es la cromatografía [120–122] y la espectrometría de masas [123,124].

Las técnicas de pre-procesamiento de señal consisten en aplicar algoritmos matemáticos. El orden en el que se ejecutan estas modificaciones matemáticas es muy importante, ya que cada tratamiento matemático de la señal se aplica sobre el resultado del tratamiento anterior. A continuación, se muestran las técnicas de preprocesamiento de espectros más comunes ordenadas conforme deberían ser aplicadas.

- Para la corrección de artefactos derivados de la dispersión del haz de luz (*scatter effects*) se suelen utilizar las técnicas *Standard Normal Variate (SNV)* [125], *Multiplicative Scatter Correction (MSC)* [126] y *Normalization*. Estas técnicas están diseñadas para reducir la variabilidad (física) entre las muestras debida a la dispersión, y además ajustan cambios en la línea de base (*base line*) entre las muestras [73,127].
- Las derivadas espectrales. Dentro de este grupo destacan las técnicas *Norris-Williams (NW) derivation* [128] y, especialmente, *Savitzky-Golay (SG) derivation* [129]. Las derivadas son una forma de filtro de paso alto y se usan cuando las características de baja frecuencia (donde la frecuencia se define como el cambio de variable a variable), tales como los efectos aditivos y multiplicativos, producen interferencias, siendo las características de mayor frecuencia las que contienen la señal de interés [127]. La primera derivada elimina cualquier desplazamiento de la línea de base (efecto aditivo) mientras que la segunda derivada acentúa aún más las características de frecuencia más alta, eliminando de esta forma tanto interferencias por desplazamiento de la línea de base como tendencias lineales

(efecto multiplicativo) [127]. Debido a que las derivadas restan énfasis a las frecuencias más bajas y enfatizan las frecuencias más altas, tienden a acentuar el ruido (señal de alta frecuencia) y, por consiguiente, reducen la relación señal/ruido. Con el objetivo de solventar este inconveniente, ambas técnicas realizan un proceso de suavizado de los espectros (*smoothing*) previo al cálculo de la derivada, y de esta forma se eliminan las pequeñas oscilaciones (ruido) del espectro. Por esta razón, se desaconseja la aplicación de derivadas sin aplicar un suavizado previo [73,115,127].

- Al final del preprocesamiento de espectros se suele aplicar algún algoritmo de centrado y/o escalado. Se deben aplicar al final porque el resto de las técnicas están pensadas para eliminar las interferencias y potenciar la señal. En cambio, el objetivo del centrado y escalado es destacar las diferencias entre las muestras. Para ello, asumen que todo lo que resta en el espectro después de los pretratamientos anteriores es información de interés. La técnica más usada es el centrado en la media (*Mean-centering*), el cual calcula la media de cada columna y la resta a cada valor de esta (a cada valor le resta el valor promedio de su columna).

$$x_c = x_{i,j} - \bar{x}_j \quad (11)$$

donde  $x_c$  es el valor centrado,  $x_{i,j}$  es el valor de la fila  $i$  y columna  $j$  antes del centrado y  $\bar{x}_j$  es la media de la columna  $j$ . De esta forma el valor resultante nos informa de cómo cada fila (cada muestra) difiere de la muestra promedio de la matriz de datos [127]. Por otro lado, el autoescalado (*Autoscale*) asume que la magnitud de la medida es proporcional a su importancia y que el nivel de ruido es similar en todas las variables [127]. Este consiste en dividir el valor resultante del centrado por la desviación estándar de la columna.

$$x_e = \frac{x_{i,j} - \bar{x}_j}{\sigma_j} \quad (12)$$

donde  $x_e$  es el valor autoescalado,  $x_{i,j} - \bar{x}_j$  es el centrado y  $\sigma_j$  es la desviación estándar de la columna  $j$ .

A pesar de la amplia utilización de estas técnicas de tratamiento de señal y de la extensa bibliografía, no existe un acuerdo claro de cuando es mejor utilizar unas u otras. Al final la responsabilidad recae en el criterio del investigador. Esto hace del proceso de pretratamiento de espectros uno de los pasos más complejos y con mayor controversia.

Una vez se ha realizado el pretratamiento de espectros, se inicia el proceso de análisis y modelización de los datos.

### Análisis de Componentes Principales (*Principal Component Analysis*, PCA)

Un problema del análisis multivariante es que el volumen bruto de los datos puede dificultar el reconocimiento de pautas y relaciones (*pattern recognition* en inglés) [107]. Además, es frecuente encontrar cierta correlación entre las variables que conforman los datos, de manera que parte de la información resulta redundante. Por ello, el objetivo de muchos métodos de análisis multivariantes es la reducción de la dimensionalidad de los datos [64,107]. El Análisis de Componentes Principales [130] permite reducir el número de variables cuando estas presentan correlación. Para ello, PCA busca componentes principales (CP)  $Z_1, Z_2, \dots, Z_n$  que sean combinaciones lineales de las variables originales  $X_1, X_2, \dots, X_n$  que describen cada muestra. Esto da lugar a un nuevo conjunto de variables, denominadas variables latentes (VL) o componentes principales (Ecuación 8), las cuales constituyen unos nuevos ejes que definen el espacio en el cual se representan las muestras.

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \dots + a_{1n}X_n \\ Z_2 &= a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \dots + a_{2n}X_n \\ Z_3 &= a_{i1}X_1 + a_{i2}X_2 + a_{i3}X_3 + \dots + a_{in}X_n \end{aligned} \tag{13}$$

Los coeficientes  $a_{11}$ ,  $a_{12}$ ,  $a_{13}$ , etc., denominados *loadings*, se escogen de manera que las nuevas variables, a diferencia de las variables originales, no se encuentren correlacionadas las unas con las otras [107]. De esta forma, cada componente principal es ortogonal respecto a los componentes principales precedentes. Además, el orden de los CPs viene dado por el porcentaje de variación que explica cada uno de ellos, siendo el primer componente principal (CP1) el que explica más variación, seguido del segundo, etc. A menudo los dos primeros CP aglutinan la mayor parte de la variación, lo que permite representar los datos empleando únicamente dos dimensiones en lugar de las  $n$  originales.

El PCA se suele utilizar como primera etapa para visualizar los datos antes de aplicar métodos de calibración o clasificación. Además, forma parte de las técnicas de reconocimiento de pautas no supervisado, ya que únicamente trabaja a nivel de la matriz de datos ( $X$ ), permitiendo incrementar el conocimiento estructural de los datos y buscar relaciones entre muestras.

### Regresión de mínimos cuadrados parciales (*Partial least Square Regression, PLS-R*)

El objetivo del método PLS-R es definir un modelo mediante el cual se puedan predecir las variables respuesta/dependientes (matriz  $Y$ ) a partir de las variables predictoras/independientes (matriz  $X$ ). Dicho proceso se denomina calibración. Como se trata de datos multivariantes, cada variable respuesta depende de múltiples variables predictoras. Esto suele generar que el número de variables predictoras sea superior al de muestras de calibración, lo cual se conoce como alta dimensionalidad de la matriz  $X$ . Por esta razón, la técnica PLS se basa en la reducción del número de variables predictoras (matriz  $X$ ) en unos pocos componentes importantes llamados factores PLS (variables latentes). La representación gráfica de los datos mediante los factores PLS, en forma de *scores* y *loadings*, permite visualizar y comprender los patrones de los datos. Los *loadings* indican la influencia o peso de cada variable original en cada variable latente, o lo que es lo mismo, las variables originales con

mayor covariancia con las variables respuesta. Mientras que los *scores* representan las coordenadas de las muestras en el nuevo espacio creado por la proyección de las variables latentes (nuevo espacio de coordenadas donde cada VL es un eje) [131].

A diferencia del PCA, donde se busca la máxima variancia entre las variables que componen la matriz de datos X, el algoritmo PLS busca la máxima covarianza entre las variables predictoras (matriz X) y las variables respuesta (matriz Y) [132]. El proceso se asemeja a realizar dos análisis de componentes principales simultáneos, uno para la matriz X y otro para la matriz Y, en el cual se emplea la varianza de la matriz Y como guía para la descomposición de la matriz X [132]. De esta forma el algoritmo PLS conecta la matriz X y la matriz Y en un mismo espacio. Además, a través de este método se logran resaltar las variables predictoras ( $x_n$ ) que más se correlacionan con las variables respuesta y reducir la influencia de aquellas variables que son menos importantes a la hora de predecir las variables y.

Una vez calculados los factores PLS, se debe escoger el número adecuado de los mismos. Contra más factores PLS se escojan para crear el modelo, mayor porcentaje de información se dispondrá, pero a su vez, aumentará la complejidad del modelo. Para encontrar el punto óptimo entre información y complejidad, los modelos PLS-R se deben validar con un conjunto externo de datos (*test-set*), es decir, un conjunto de muestras que no se hayan utilizado para la calibración del modelo. Cuando no es posible disponer de un conjunto externo independiente de muestras, porque el número de muestras total es limitado, se puede optar por lo que se conoce como validación cruzada (*cross-validation* en inglés). La validación cruzada consiste en dividir los datos en dos subconjuntos, el de entrenamiento (mayor número de muestras), el cual se emplea para calibrar el modelo, y el de validación (menor número de muestras), cuya finalidad es evaluar su funcionamiento (validarlo). Debido a que la evaluación depende en gran medida de cómo se haya realizado la división de los datos, es necesario repetir el proceso realizando diferentes particiones de los datos, para luego calcular la media aritmética de los resultados de evaluación.

En todo caso, el proceso de validación consiste en predecir la variable respuesta con el modelo y calcular el residual, es decir, la diferencia entre el valor de  $y$  predicho y el valor de  $y$  conocido. Ello da lugar al error medio de validación, que se grafica en función del número de factores PLS del modelo. En el caso de la validación cruzada, el estadístico que se obtiene es el *Root Mean Square Error of Cross-Validation* (RMSECV) [133].

$$RMSECV = \sqrt{\frac{\sum_{i=1}^n (Y_i^{cv} - Y_i^{ref})^2}{n}} \quad (14)$$

donde  $n$  es el número de muestras que se ha utilizado para validar el modelo,  $Y_i^{cv}$  es el valor predicho por el modelo en la validación cruzada e  $Y_i^{ref}$  es el valor de referencia de la muestra  $i$ . Por lo general, el error del modelo se reduce a medida que se incrementan el número de factores PLS, hasta que se alcanza un punto en el cual esta tendencia se rompe. Es este punto el que determina el número de factores PLS óptimo. Si se sobrepasa el número óptimo de factores se produce un sobreajuste del modelo (*overfitting* en inglés), que provoca que el modelo se ajuste muy bien a los datos de calibración, pero no prediga correctamente los datos de validación.

El modelo de regresión resultante relaciona las variables predictoras ( $x_n$ ) con las variables respuesta ( $y_n$ ). Dicho modelo puede ser utilizado en análisis posteriores para predecir el valor de la variable respuesta a partir de las variables predictoras.

#### Análisis discriminante de mínimos cuadrados parciales (*Partial Least Square Discriminant Analysis*, PLS-DA)

El objetivo del análisis discriminante es utilizar una serie de muestras, de las cuales se conoce la clase o categoría a la que pertenecen, para encontrar una regla que permita la asignación de nuevas muestras a alguna de las clases ya definidas

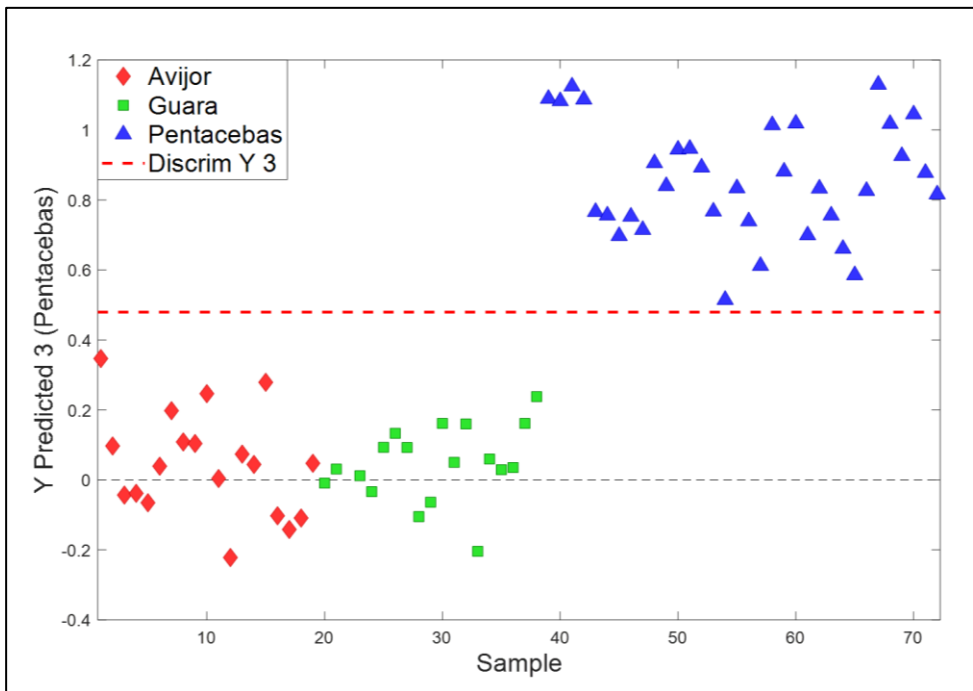
[107]. El PLS-DA es un método de clasificación lineal que combina las propiedades del algoritmo PLS junto con el potencial de discriminación de las técnicas de clasificación [131]. El método PLS-DA consta de dos pasos: 1) búsqueda de la relación entre la matriz de datos X y la matriz de referencia Y mediante el algoritmo PLS; 2) Construcción de la regla de discriminación.

PLS-DA emplea el algoritmo PLS para buscar variables latentes que presenten máxima covarianza con las variables Y [131]. Pero a diferencia del PLS, en un modelo PLS-DA la matriz de referencia Y contiene información cualitativa, es decir, informa de la clase a la que pertenece cada muestra de la matriz X. La información de la matriz Y se transforma en valores numéricos (*dummy variables* en inglés), los cuales suelen ser 0 y 1. En el caso de una clasificación binaria, el valor 1 se asigna a la clase de interés y el valor cero a la otra clase. No obstante, cuando se construye un modelo real no se suelen obtener valores enteros, sino que los valores predichos se distribuyen alrededor de 0 o de 1 [134]. Cuando se trabaja con dos clases, únicamente es necesaria una variable y (una sola columna en la matriz Y). Cuando hay más de dos clases, se debe crear una variable y en la matriz de referencia para cada una de ellas (una columna por clase) [131]. El procedimiento sería el mismo, 1 para la clase seleccionada y 0 para el resto de las clases. A continuación, y en base a la distribución de los datos en el nuevo espacio constituido por los factores PLS (VL), se define un discriminador o *threshold* para cada una de las variables y (clases), basado en el Teorema de Bayes [135]. El discriminador crea un umbral dentro del rango 0-1 por encima del cual la muestra es asignada a la clase correspondiente o, de lo contrario, es rechazada. En el proceso de cálculo (y optimización) del discriminador se realiza una estimación de la sensibilidad (proporción de positivos correctamente asignados) y la especificidad (proporción de negativos correctamente asignados) en base a los datos observados y asumiendo que la distribución de los mismos es representativa para la clase. Una vez hecho esto, se sitúa el discriminador en el punto de corte entre ambas curvas. De esta forma se intenta minimizar el número de falsos positivos y



falsos negativos en futuras predicciones. El cálculo del discriminador varía en función del número de factores PLS seleccionados. En este sentido, al igual que en el modelo PLS, en el modelo PLS-DA se debe escoger el número de variables latentes a través de un proceso de validación. Todo esto se muestra gráficamente en las [Figura 21 y 22](#).

De esta forma, el resultado del modelo PLS-DA son una serie de valores de predicción, uno para cada clase, los cuales son comparados con los diferentes discriminadores para realizar la asignación de la muestra a alguna de las clases definidas.



**Figura 21.** Discriminador o *threshold* (línea discontinua roja) para la variedad Pentacebas en un modelo PLS-DA con tres clases.

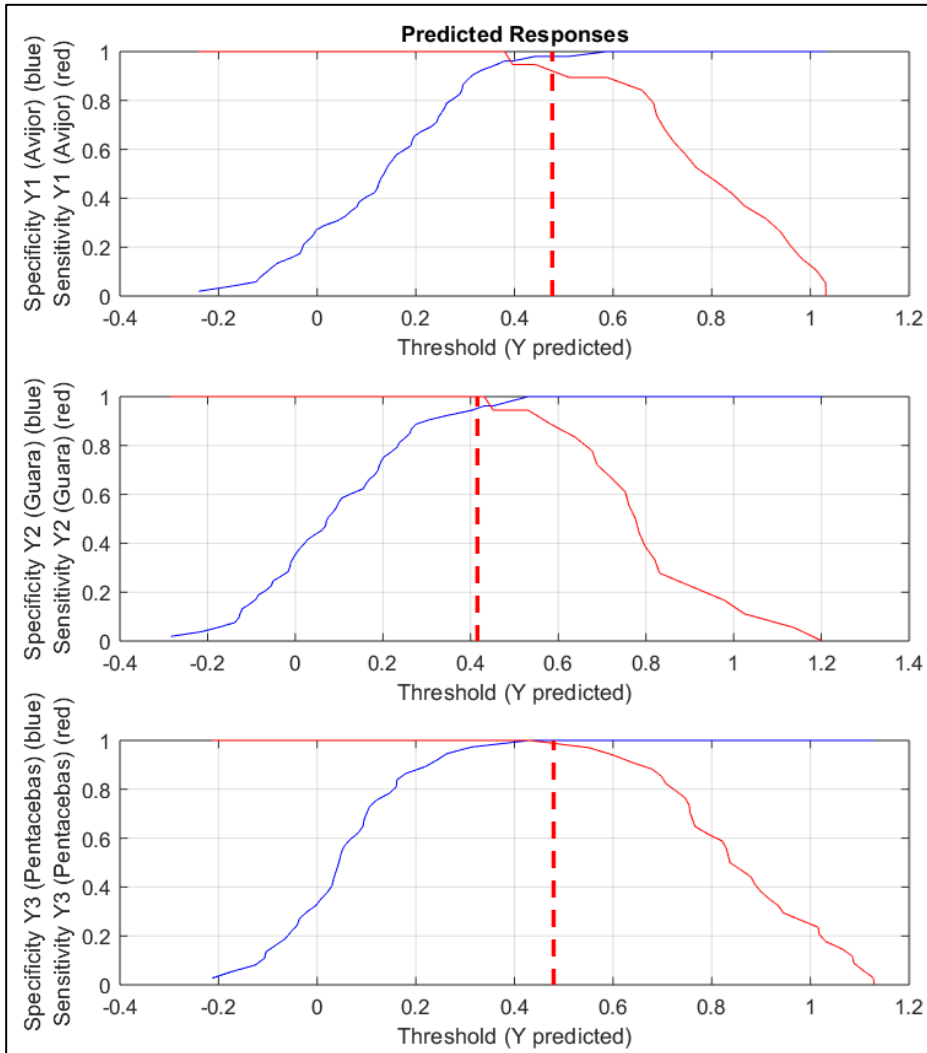


Figura 22. Estimación del discriminador en el punto de corte entre la sensibilidad y la especificidad para cada clase.

# Objetivos



## Objetivos

---

### **Objetivo general**

El objetivo general es desarrollar un método que permita identificar y diferenciar variedades a nivel de planta de vivero. Además, se plantea la capacidad de evaluar el estado fisiológico de las mismas. En este sentido, es importante que la técnica desarrollada sea rápida y de bajo coste.

### **Objetivos específicos**

Para lograr el objetivo general, la tesis doctoral queda estructurada en tres objetivos específicos:

- 1) Definir un protocolo de muestreo del material vegetal en el campo y en el vivero, además de definir el tratamiento que se les debe aplicar a las muestras de acuerdo con los resultados obtenidos, así como la determinación y puesta a punto del instrumento técnico.
- 2) Desarrollar modelos matemáticos con la utilización de software específicos que nos permitan:
  - a. Discriminación varietal para un rápido fenotipado
  - b. Detección de la presencia o ausencia de fitopatologías con antelación a la aparición de síntomas visuales.
- 3) Alcanzar un nivel 5 de madurez de la tecnología en la escala “*Technological Readiness Level (TRL)*”, además de estudiar la viabilidad para la implantación del instrumento en la cadena de producción de la empresa.

Los objetivos de la tesis doctoral se acordaron en base a necesidades expuestas por Agromillora.



**Capítulo 1. Development of a  
Methodology to Analyze Leaves  
from *Prunus dulcis* Varieties  
Using Near-Infrared  
Spectroscopy**







## Development of a methodology to analyze leaves from *Prunus dulcis* varieties using near infrared spectroscopy

Sergio Borraz-Martínez<sup>a,c,\*</sup>, Ricard Boqué<sup>b</sup>, Joan Simó<sup>a,d</sup>, Mariàngela Mestre<sup>c</sup>, Anna Gras<sup>a</sup>

<sup>a</sup> Universitat Politècnica de Catalunya, Department of Agri-Food Engineering and Biotechnology, Esteve Terrades 8, 08860, Castelldefels, Spain

<sup>b</sup> Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, 43007, Tarragona, Spain

<sup>c</sup> Agromillora Iberia S.L.U, Center of Initial Materials, Ctra. BV-2247 km. 3, 08770, Sant Sadurní d'Anoia, Spain

<sup>d</sup> Fundació Miquel Agustí, Esteve Terrades 8, 08860, Castelldefels, Spain

### ARTICLE INFO

#### Keywords:

Optimization  
Almond trees  
Leaf analysis  
Varietal purity  
NIR  
PLS-DA

### ABSTRACT

Near-infrared spectroscopy (NIRS) can be a faster and more economical alternative to traditional methods for screening varietal mixtures of nursery plants during the propagation process to ensure varietal purity and to avoid errors in the dispatch batches. The global objective of this work was to develop and optimize a NIR spectral collection method for construction of robust multivariate discrimination models. Three different varieties of *Prunus dulcis* (*Avijor*, *Guara*, and *Pentacebas*) of agricultural interest were used for this study. Sources of variation were investigated, including the position of the leaves on the trees, differences among trees of the same variety, and differences at the varietal level. Three types of processed samples were investigated. Fresh leaves, dried leaves, and dried leaves in powder form were included in each analysis. A study of spectral pre-treatment methods was also performed, and multivariate methods were applied to analyze the influence of different factors on classification. These included principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and ANOVA simultaneous component analysis (ASCA). The results indicated that variety was the most important factor for classification. The spectral pre-treatment that provided the best results was a combination of standard normal variate (SNV), Savitzky-Golay first derivative, and mean-centering methods. With regard to the type of processed sample, the highest percentages of correct classifications were obtained with fresh and dried powdered leaves at both the training set and test set validation levels. This study represents the first step towards the consolidation of NIRS as a method to identify *Prunus dulcis* varieties.

### 1. Introduction

Rapid discrimination between vegetal varieties is a key requirement for all nursery plant production. The huge diversity of vegetal materials necessitates the incorporation of new control systems along the nursery plant production chain to avoid mixing varieties and to ensure varietal purity in the dispatch batches.

Nowadays, the most extensively used methods for varietal identification are based on DNA analysis. These techniques include DNA amplification by the polymerase chain reaction (PCR) [1] followed by analysis of genetic variations, such as single nucleotide polymorphisms (SNPs) [2]. However, these biomolecular techniques are very expensive for routine analysis of a large number of samples. In this context, the use of spectroscopic analysis combined with chemometrics has recently increased. This combination comprises a rapid, accurate, and non-destructive methodology for the classification and authentication of

agricultural products [3].

Near-infrared (NIR) spectroscopy has proved to be a powerful analytical tool and has been widely used in various sectors, including the petrochemical [4] and pharmaceutical industries [5]. It has also become a well-established technique for the quantitative and qualitative analysis of agricultural products [6]. Several recent studies have employed spectroscopic techniques for species discrimination [7,8], or differentiation of varieties within a species, such as tomato [9], rice [10] and lettuce [11]. For these reasons, NIRS can be considered a potential candidate for the differentiation of *Prunus dulcis* varieties.

Despite recent studies, there is a lack of knowledge regarding the best methodology for accurate sampling of leaves. Most of the published works on species discrimination do not consider factors derived from the nature of the samples, which are potential sources of variance. For example, mature trees have a heterogeneous canopy composed of leaves in different phenological stages. It is thus important to take the

\* Corresponding author.

E-mail address: [sergio.borraz@upc.edu](mailto:sergio.borraz@upc.edu) (S. Borraz-Martínez).

<https://doi.org/10.1016/j.talanta.2019.05.105>

Received 22 March 2019; Received in revised form 21 May 2019; Accepted 27 May 2019

Available online 05 June 2019

0039-9140/ © 2019 Elsevier B.V. All rights reserved.

sampling procedure into account, especially when an analysis is performed with whole leaves. Improper sampling may generate invalid data, the use of which could lead to incorrect conclusions [12]. To perform correct sampling, it is important to recognize sources of variation and to control for factors from which variation originates. Therefore, it is necessary to first develop a sampling protocol and to select the best material for use.

Another analytically relevant aspect is the study of sample processing methods, which may considerably alter the vibrational spectrum of a sample compared to that collected with the sample in its native state. Due to economic and time constraints, it is generally best to avoid any type of sample processing. Moreover, modifying the native architecture of biological tissues can result in the loss of information. Thus, performing analyses *in vivo* is preferred whenever possible [13]. Occasionally, however, sample processing is an indispensable step. In any case, the option that best accomplishes the objective of the study must be selected.

The aim of this work is to determine how sampling of vegetal material affects the collection of NIR spectra for the construction of a multivariate discriminant model for *Prunus dulcis* varietal classification. The specific objectives are to 1) determine whether there are differences among the analyzed regions of the leaves or between their upper and lower surfaces; 2) to determine whether differences exist due to the age of the leaves; 3) identify the best sampling procedure for varietal discrimination of almond trees; and 4) study pre-treatment of spectral data and to identify the pre-treatment that leads to the best classification model.

## 2. Material and methods

### 2.1. Experimental design

#### 2.1.1. Assay one

The first assay was performed to obtain information about the analyzed regions of fresh and dried leaves. Specifically, the NIR spectra differences resulting from including or excluding the primary veins of the leaves were examined (Fig. 1) together with analysis of the upper

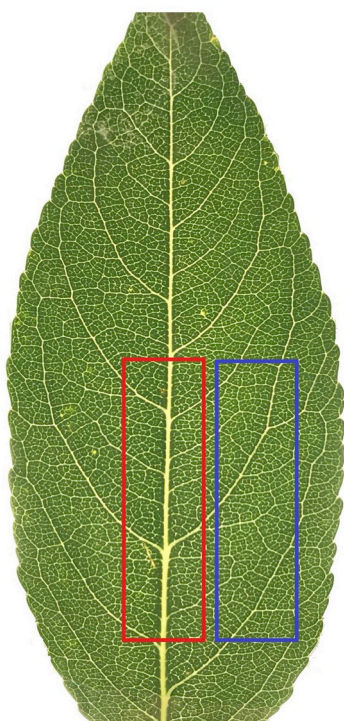


Fig. 1. Image of an almond leaf showing the two studied regions.

**Table 1**

Summary of the samples used in the study.

	Varieties	Fresh samples	Dried samples	Dried-powdered samples
Assay one	Guara	10	10	not used
	Pentacebas	10	10	not used
Assay two	Avijor	80	80	80
	Guara	80	80	80
Assay three	Pentacebas	80	80	80
	Avijor	80	80	80
	Guara	80	80	80
	Pentacebas	80	80	80

(adaxial) and lower (abaxial) leaf surfaces. Twenty samples from two varieties of almond trees, *Guara* and *Pentacebas*, were used for each experiment (Table 1). Results were evaluated by using PCA and PLS-DA models. This assay focused on aspects that affected only fresh and dried leaves. The information obtained in this assay was used for the development of the next two assays.

#### 2.1.2. Assay two

The second assay was designed to study the NIR spectra differences between young leaves and adult leaves and among samples from different trees of the same variety. The assay was performed on the *Guara* and *Pentacebas* varieties and on a third variety, *Avijor*. Four trees per variety were sampled, twelve in total. Twenty leaves were collected from each tree. Ten of the leaves were collected from the upper part of the branch (apex), which corresponded to young leaves, while the other ten were adult leaves that were collected from the lower part of the branch. Two hundred forty leaves were sampled in total (Table 1). Results were evaluated by using PCA and ASCA-ANOVA models.

#### 2.1.3. Assay three

Three different leaf processing methodologies were studied in the third assay; one for fresh leaves, one for dried leaves, and the other for dried powdered leaves. The aim was to determine whether the water content and macrostructures of the leaves had any influence on discrimination results. It is important to note that every sample was processed with each of the three methods in order to increase the comparative robustness. We also identified the most suitable pre-treatment method for NIR spectral analysis. The applicability of NIRS for discriminating between *Prunus dulcis* varieties was evaluated by mean partial least squares discriminant analysis (PLS-DA). The available data were randomly divided into calibration (70%) and validation (30%) sets, but both sets contained the same proportion of each variety to prevent unbalanced representation of the almond tree classes. To improve the robustness of comparing results from the three sample processing methods, the same samples included in the three sample processing datasets were used for both cross validation and test set validation. All of the samples used for assay two were also used for this assay (Table 1).

### 2.2. Description of the sampling field

Vegetal material used in this study came from almond trees located at the mother plant field from the Center of Initial Materials of Agromillora Iberia, S.L.U. in Sant Sadurn d'Anoia (Catalonia, Spain). These trees are under a strict control in order to prevent the appearance of diseases and to ensure the sanitary quality of nursery plants. The use of molecular biology techniques to assess the traceability of the varieties was not necessary in this case because the almond trees were previously certificated by the company.

The samples were stored in a plastic bag after collection, assigned identifiers, and stored at 4 °C until analysis.

### 2.3. Sample pre-processing

Samples were analyzed either as fresh leaves without processing, as dried leaves, or as dried powdered leaves. To obtain dried leaves, fresh leaves were heated in an oven at 65 °C for 48 h. A weight was placed on the leaves to keep them flat and to facilitate their posterior analysis. Once dried, the leaves were pulverized to a homogeneous powder with a grinder. Once samples were dried, they were stored in a desiccator with silica gel to prevent any influence from moisture. Only one leaf was used per experiment. Each sample was analyzed in the three ways. First, they were analyzed in fresh, second in dried and finally in powdered. In all the experiments each sample was composed of one leaf only.

### 2.4. Acquisition of NIR spectra

Samples were scanned in reflectance mode using an Antaris II FT-NIR analyzer (Thermo Scientific, USA) equipped with an integrating sphere module. Samples were measured in the spectral range of 12000–3800  $\text{cm}^{-1}$  (833–2630 nm). For each spectrum, 32 scans were averaged with a resolution of 4  $\text{cm}^{-1}$ . Each sample was analyzed in triplicate. Fresh leaves and dried leaves were placed directly over the sphere and covered to prevent interference from environmental light. The powdered leaf samples were measured in a standard sample cup that came with the instrument. A background spectrum was collected every 20 min. All spectra were recorded as  $\log(1/R)$ , where R was the reflectance. Room temperature was maintained at  $\sim 25$  °C, and the humidity remained constant throughout the spectral acquisition process.

### 2.5. Spectral data pre-treatment

This was an important step, because although different pre-treatments have been reported on extensively [14–16], there is still no clear consensus regarding the best pre-treatment or a guideline to follow. As can be seen in Fig. 2, the spectra contained very little noise. The raw spectra had to be corrected for additive and multiplicative effects that were probably due to light scattering.

A basic pre-treatment was performed in assays one and two, which consisted of the standard normal variate (SNV) method with mean centering. In the assay three, four different pre-treatments were applied and compared to identify the combination that provided the best results

in the PLS-DA model. The combinations used were: SNV method with mean centering; SNV method with Savitzky-Golay (SG) first derivative and mean centering; and finally, SNV method followed by de-trending and mean centering. Spectral pre-treatments were performed using PLS\_Toolbox (Eigenvector Research Incorporated, Manson, WA) with MATLAB R2017b (MathWorks, Natick, MA).

SNV is a normalization procedure for spectral light scattering correction. It is used to correct additive and multiplicative effects in the spectra due to particle size variation. SNV calculates the standard deviation of all the variables in a given sample spectrum. The entire data set is then normalized by this value, which yields a unit standard deviation ( $s = 1$ ) for the sample spectrum [17]. De-trending is sometimes used to remove constant, linear, or curved offsets and is often used in conjunction with SNV. With this method, the mean value or linear trend is subtracted from a vector or matrix. To achieve this, a polynomial of a given order is fitted to the entire data set, and the polynomial is simply subtracted. This algorithm fits all points in the baseline and the signal [17]. SG first derivative was applied to remove baseline drift and to enhance small spectral differences. The SG derivative method includes a smoothing step, the Savitzky-Golay algorithm, which corrects for the increased noise due to application of the derivative. The SG derivatization algorithm requires selection of the filter width, which is the size of the window, the order of the polynomial, and the order of the derivative [18]. In this work, we selected a 15-point window and applied a second order polynomial. Mean centering is one of the most common pre-processing methods, in which the mean value of each column is calculated and subtracted from each individual value in the column. After mean centering, the mean of each column equals zero, and each row of mean-centered data reflects only how it differs from the average sample in the original data matrix [16].

### 2.6. Principal component analysis (PCA)

PCA captures the largest amount of variance in the data and reduces the dimensionality of the original dataset through calculation of a new set of variables called principal components (PCs). The PCs are linear combinations of the original variables. Samples and variables are projected onto the new PCs in the calculated PCA space. Samples are defined by their scores, and variables are defined by their loadings. Inspection of the scores and loading plots can lead to a better understanding of the different sources of variation in the data. As a data reduction technique, PCA is frequently the first step in the analysis of a

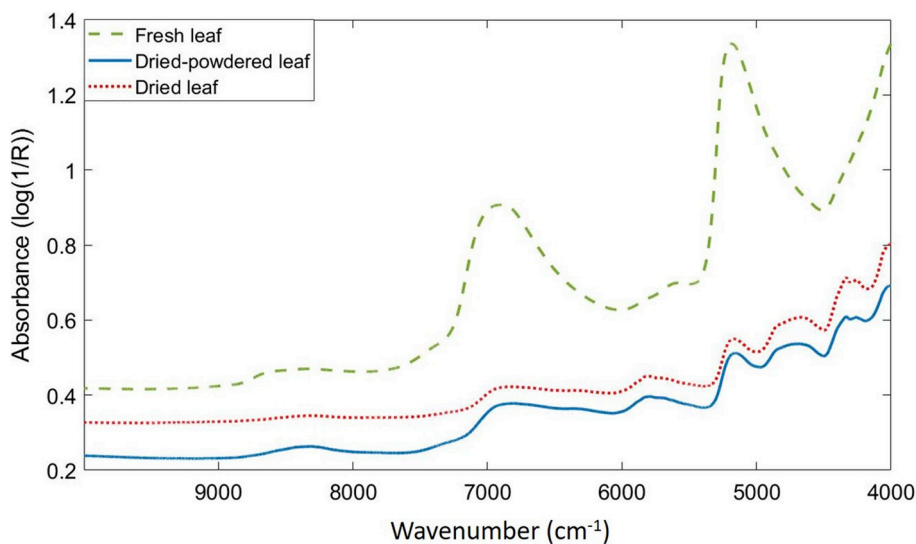


Fig. 2. Mean raw spectra from the three processed sample types. Fresh leaf (green dashed line); dried powdered leaf (blue solid line); and dried leaf (red dotted line). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

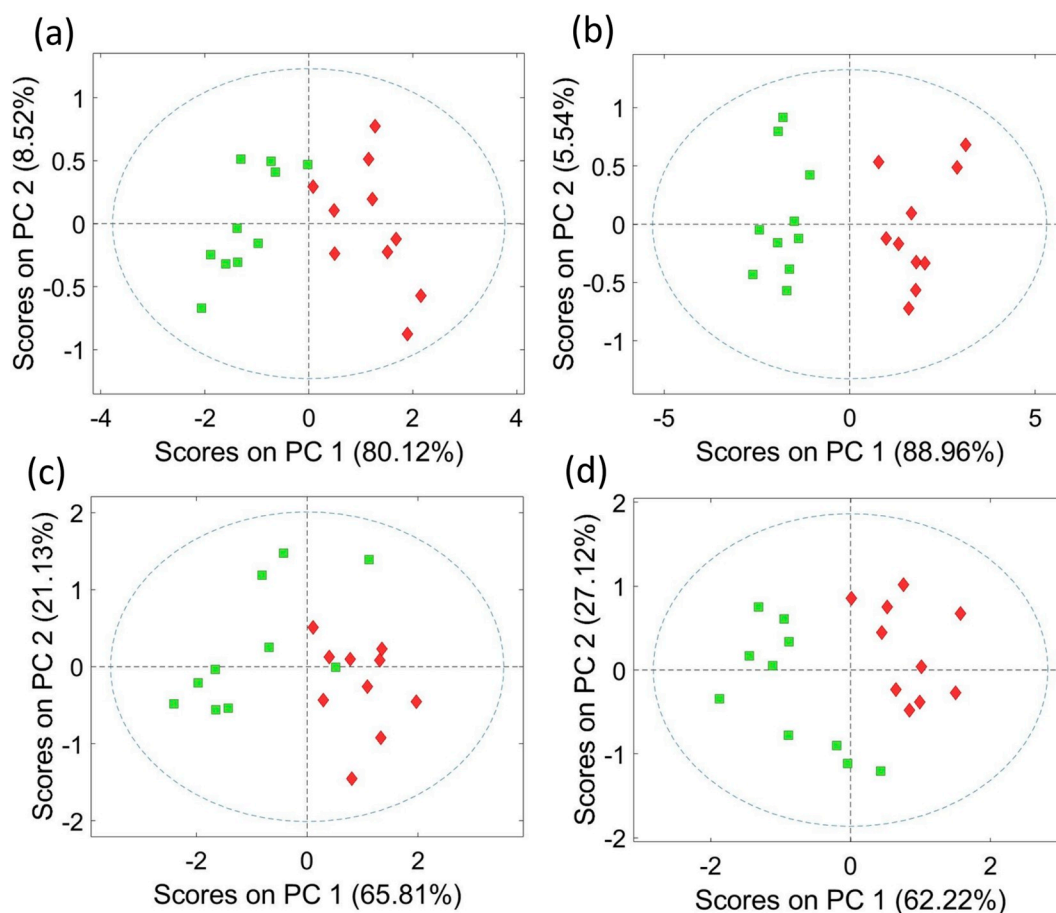


Fig. 3. PCA results from the Guara and Pentacebas varieties with and without inclusion of the midvein. The presence of the midvein is indicated by red diamonds, and absence of the midvein is indicated by green squares. a) Dried leaf of the Guara variety; b) dried leaf of the Pentacebas variety; c) fresh leaf of the Guara variety; d) fresh leaf of the Pentacebas variety. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

high-dimensional data set. It can then be followed by classification, clustering, or other multivariate techniques [19].

### 2.7. Partial least squares discriminant analysis (PLS-DA)

PLS-DA is a classification technique widely used in research studies concerning both varietal classification and authentication of geographical origin [10,20]. PLS-DA is based on the PLS regression algorithm, which searches for linear combinations of the original variables (latent variables) that display maximum covariance with the Y-variables (classes). A discriminator, or threshold, is created that separates the different classes [21]. This technique allows determination of whether or not a given sample belongs in a specific predefined class [22]. The optimal number of factors or latent variables (LVs) for the PLS-DA models was estimated with a cross-validation procedure, and the number yielding the minimum classification error was selected. Venetian blinds cross validation was used for the calibration with a data split of 10 and one sample per blind (thickness).

### 2.8. ASCA-ANOVA

Designed experiments with a single dependent variable are typically analyzed with ANOVA [23]. Problems occur when hundreds or thousands of variables are measured simultaneously, which is the case in spectroscopic analysis. ANOVA is thus not useful for analyzing multivariate data. Multivariate ANOVA (MANOVA) [24], the natural multivariate extension of ANOVA, breaks down when the number of measurements is smaller than the number of variables [25].

ANOVA-simultaneous component analysis (ASCA) [26] is a method

used to determine which factors in a fixed-effect experimental design are significant relative to the residual error. ASCA allows an ANOVA-like analysis, even when there are more variables than samples. Two matrices are used to perform the procedure. The X-matrix contains the experimental data, while the F-matrix represents the experimental design. PCA of each factor in the effect (X) matrix reduces the number of variables to a smaller number of principal components. In this way, the parameter estimation functionality of ANOVA is merged with PCA, and the presence of more variables than samples is no longer problematic [27]. Due to the hierarchy of factors analyzed in the present study, a nested design referred to as multi-level simultaneous component analysis (MLSCA) [28] was applied. Hence, the leaf age factor was nested within the tree factor, which in turn was nested within the variety factor.

## 3. Results and discussion

### 3.1. Assay one

#### 3.1.1. Comparison of leaf midvein and lamina

Whether differences exist within the same leaf is a question that frequently arises. For this reason, spectra were collected in different areas of healthy leaves. The two regions of the leaves used for comparison are shown in Fig. 1. PCA was performed with two of the almond tree varieties, *Guara* and *Pentacebas*, to identify possible differences between the measurement areas on fresh and dried samples. These results are shown in Fig. 3.

Differences when including or not the primary vein were detected. The data clouds with and without midvein form separate clusters in

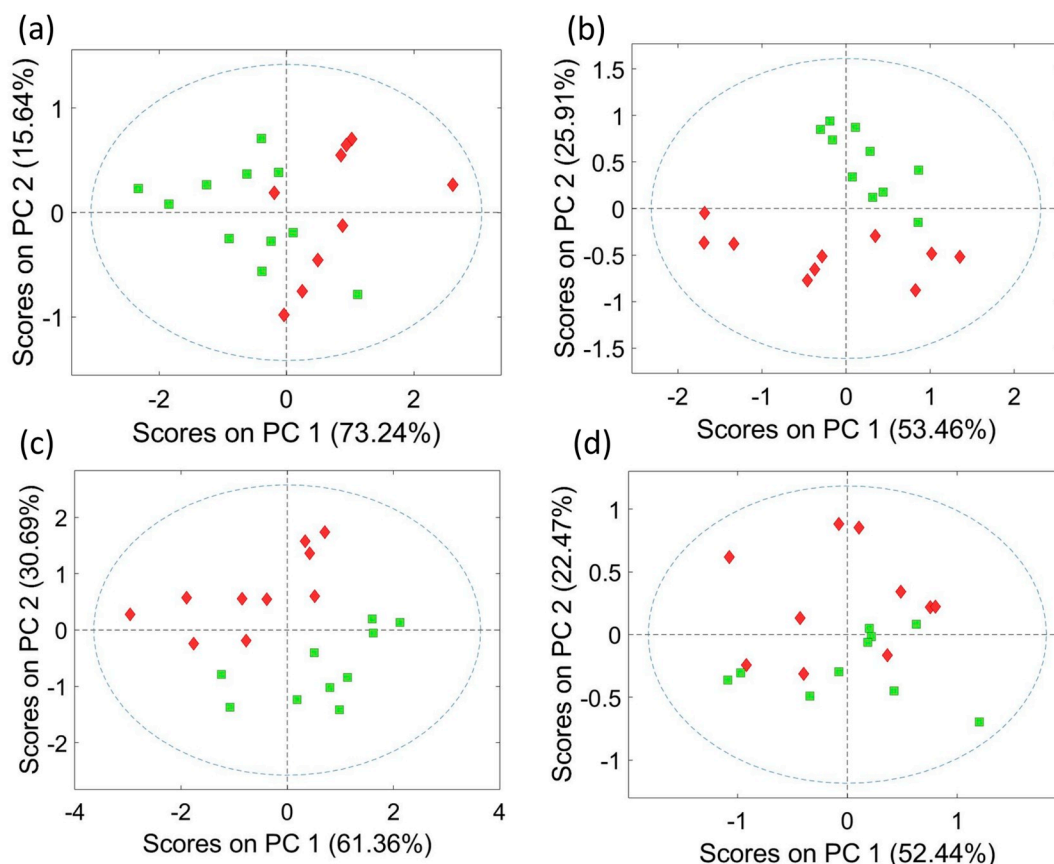


Fig. 4. PCA results showing the differences between the adaxial (red diamonds) and abaxial (green squares) leaf surfaces. a) Dried leaf of the Guara variety; b) dried leaf of the Pentacebas variety; c) fresh leaf of the Guara variety; d) fresh leaf of the Pentacebas variety. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2  
PLS-DA results from the comparison of adaxial and abaxial leaf surfaces.

Leaf surface	Real class	Data set	Fresh samples		Dried samples	
			Assigned class		Assigned class	
			Guara	Pentacebas	Guara	Pentacebas
Adaxial	Guara	Cross-validation	100%	100%	100%	100%
	Pentacebas		100%	100%	100%	100%
Abaxial	Guara	Cross-validation	100%	100%	100%	100%
	Pentacebas		100%	100%	100%	100%

both kinds of pre-processed samples. This cluster separation can be observed in both varieties, although the separation is clearer for the Pentacebas variety. Differences were detected whether or not the primary vein was included. The data clouds with and without the midvein formed separate clusters for both processed sample types. This cluster separation was observed with both varieties, although the separation was more pronounced in the results from the *Pentacebas* variety. Considering the macrostructures and compositions of the analysis regions were not equivalent, which was reflected in their spectral signatures, these differences were justifiable. When the primary vein was scanned, the reflectance spectra of both the primary vein and the laminar regions located on either side of the primary vein were collected. Taking into account that secondary veins were present in the laminar regions, identifying differences between these regions indicated the primary vein had a profound influence on the spectra.

The apical region and a region adjacent to the leaf margin showed more damage and decay than the central region of the leaves. Consequently, the central region was usually more stable. The leaf size

could make it difficult to completely exclude the primary vein during measurement of the laminar region. Collecting spectra in the central region, including the primary vein, could therefore provide a standardized measure.

### 3.1.2. Comparison of adaxial and abaxial surfaces

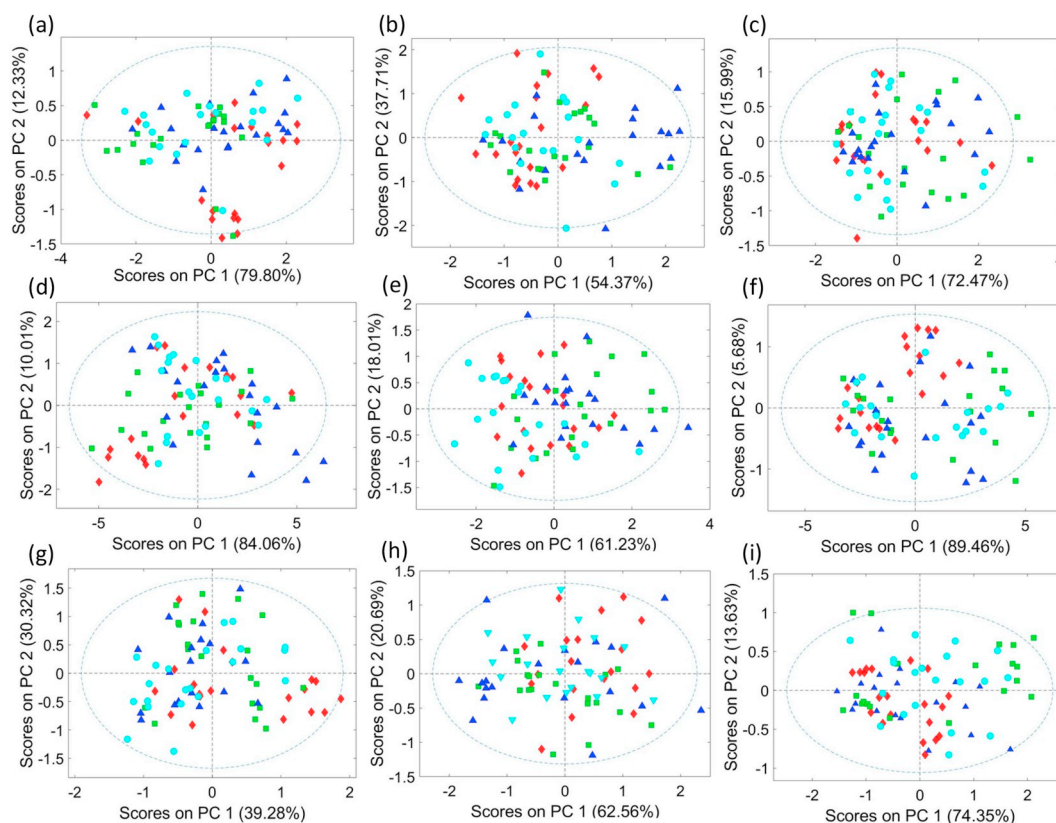
Differences between the upper and lower surfaces of the leaves were also investigated. These results are shown in Fig. 4. In both fresh and dried samples, results of PCA revealed differences between the spectra obtained from the upper and lower leaf surfaces. However, this difference was not as clear in fresh leaves of the *Pentacebas* variety. The upper and lower surfaces of leaves in all plants are different. In addition, the stomas are usually present on the abaxial surface together with trichomes and others surface features. The differences between these two surfaces could be the cause for separation of their spectra in the PCA plots.

A PLS-DA model was built to determine which surfaces were most suitable for discriminating between two almond tree varieties using fresh or dried samples. The classification results are shown in Table 2. The PLS-DA model had a classification score of 100% for both types of processed samples when the upper leaf surface was analyzed. Perfect discrimination was obtained using the lower leaf surface as well. Based on these results, the differences identified by PCA did not affect the discrimination results with either surface.

## 3.2. Assay two

### 3.2.1. Variability between trees of the same variety

Differences among trees of the same variety are important to consider when building a classification model. This source of variation



**Fig. 5.** PCA results from the study of differences between trees of the same variety. Each tree is represented by a different symbol (triangle, circle, diamond, and square). a) Fresh leaf of the Avijor variety; b) fresh leaf of the Guara variety; c) fresh leaf of the Pentacebas variety; d) dried leaf of the Avijor variety; e) dried leaf of the Guara variety; f) dried leaf of the Pentacebas variety; g) dried powdered leaf of the Avijor variety; h) dried powdered leaf of the Guara variety; i) dried powdered leaf of the Pentacebas variety.

determines the number of trees of each variety that must be sampled for development of the final model. If the variance is very large, it could affect the model's discrimination capability. The PCA results from assay 2 are shown in Fig. 5.

No differences were identified among the four trees studied within each variety. This was the case for fresh, dried, and dried powdered leaves. This was remarkable, because if significant differences were found, it would have been more difficult to build a good classification model. Also noteworthy was that the same results were obtained with samples processed with the three different methods, and with samples of different varieties. Such similar behavior in all cases is a positive indicator when creating a classification model. A more exhaustive study of the variability between trees was performed using the ASCA-ANOVA method, which is discussed in section 3.2.3.

### 3.2.2. Variability between leaves of the same variety

Since differences among almond trees of the same variety were not detected at the PCA level, we decided to include all samples of the same variety in a single PCA model. This made it easier to study the variability among samples within each variety while increasing the robustness of the model with more samples. The results of PCA modelling are shown in Fig. 6.

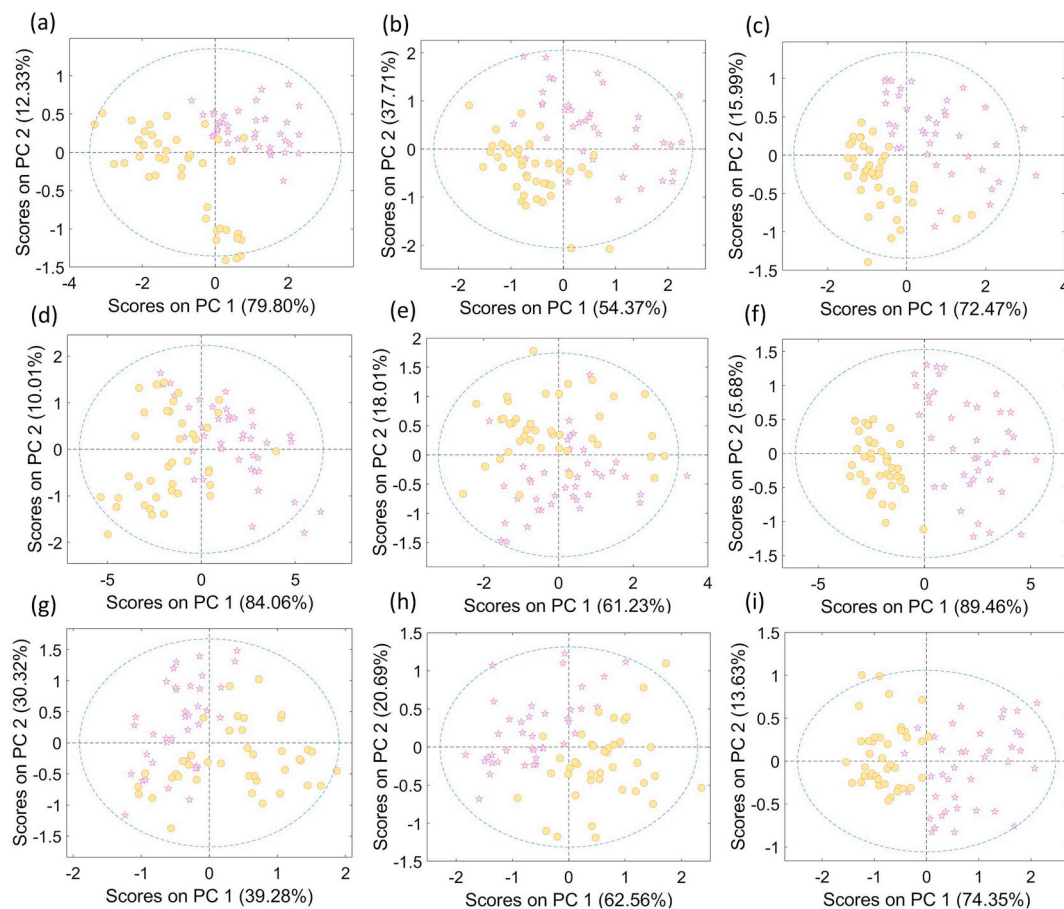
Two clusters could be distinguished using only the first two principal components. This separation was very clear in some cases, such as the dried processed samples of the *Pentacebas* variety, for which the two clusters were completely separated (Fig. 6f). The results of all of the PCA models were similar, regardless of the sample processing method or the variety studied. However, overlap between the two data clusters was observed in some cases, such as dried samples of the *Avijor* variety (Fig. 6d). The overlap could be explained by the presence of leaves in a phenological stadium intermediate between young and adult. It was

possible to observe the progressive growth of the leaves, although this was not the goal of the assay. In any case, the results indicated there were differences between young and adult leaves at the spectral level. This difference should be considered at the time of sampling.

### 3.2.3. ASCA-ANOVA analysis

To study variability between *Prunus dulcis* varieties more deeply, an ASCA-ANOVA model was constructed for young and adult leaves from trees of the same variety. The modelling results are shown in Table 3. The raw spectra pre-treatment used to develop the model, SNV with mean centering, was the same as that used for the PCA models.

Tree variety was the most influential factor for variance among fresh and dried powdered leaves and accounted for 30.26% and 24.99%, respectively, of the total effect in these samples. Despite explaining 19.25% of the effect for dried leaves, tree variety was not the factor that accounted for the majority of variance. For two of the three processing methods, the variety factor had the greatest effect, which indicated that differences between varieties were important. The tree factor explained little of the variance for the three processing methods, which was in agreement with the PCA results shown in Fig. 5. This indicated strong homogeneity between trees of the same variety, an aspect that could be key for effective discrimination between varieties. For fresh and dried powdered leaves, the age (young/adult) factor explained a higher percentage of variance than the tree factor, but it accounted for less of the variance than tree variety. In dried powdered leaves, the difference between the age and tree factors was not large. The age factor accounted for 6.68% of the variance, while the tree factor explained 1.87%. The difference was more notable for fresh leaves, as the age factor accounted for 19.11% of the explained variance. The age factor was most significant for dried leaves, accounting for 24.18% of the explained variance. Therefore, the age factor had a



**Fig. 6.** PCA results from the study of differences between young (yellow circles) and adult (pink stars) leaves. a) Fresh leaf of *Avijor* variety; b) fresh leaf of *Guara* variety; c) fresh leaf of *Pentacebas* variety; d) dried leaf of *Avijor* variety; e) dried leaf of *Guara* variety; f) dried leaf of *Pentacebas* variety; g) dried powdered leaf of *Avijor* variety; h) dried powdered leaf of *Guara* variety; i) dried powdered leaf of *Pentacebas* variety. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

**Table 3**  
Results of ASCA-ANOVA modelling to study variance of the factors.

Factor	Fresh leaves		Dried leaves		Dried-powdered leaves	
	Principal components	Effect %	Principal components	Effect %	Principal components	Effect %
Variety	2	30.26	2	19.25	2	24.99
Tree	3	1.83	3	4.69	3	1.87
Young/adult	1	19.11	1	24.18	1	6.68
Residual	6	48.80	3	51.88	3	66.46

greater effect in non-powder samples. These results also correlated with the results of the PCA (Fig. 6), in which differences due to leaf age were observed, but overlap of the cluster regions was detected.

All of the variance not explained by the studied factors accumulated in the residual term. In the three types of processed samples, the residual accounted for a high percentage of the variance. Fresh leaves had a lower residual than either the dried or dried powdered leaves. It was thought that the main source of uncontrolled variance was the physiological state of the leaves, which included damage to the leaves and climatologic agents. The combination of these abiotic factors with biotic factors influences plant physiology [29,30]. It is important to note that the leaves used in this study came from trees located in an outdoor field.

### 3.3. Assay three

#### 3.3.1. Spectral pre-treatment study

Table 4 shows the results of the PLS-DA modelling using different spectral pre-treatments. The best classification results for the three types of samples were obtained with the SNV pre-treatment and application of the SG first derivative and mean centering. This was curious, because although modelling was performed for one material (almond tree leaves), the samples analyzed were completely different in terms of their macrostructures and dry compositions. With this spectral pre-treatment, 100% classification accuracy was achieved for at least one variety with each sample processing method. Results were even more remarkable with dried powdered leaves, for which 100% accuracy was attained in the test set validation for all three varieties. The lowest

**Table 4**  
PLS-DA model results of the spectra pre-treatment and study of the types of pre-processed samples.

Real class	Data set	Assigned class		
		SNV + Mean center	SNV + 1st derivative + Mean center	SNV + De-trending + Mean center
Dried-powdered leaves				
Avijor	Cross-validation	87.4%	99.2%	86.6%
	Test set validation	97.5%	100%	97.5%
Guara	Cross-validation	89.9%	99.2%	89.1%
	Test set validation	96.6%	100%	96.6%
Pentacebas	Cross-validation	97.5%	100%	97.5%
	Test set validation	99.2%	100%	99.2%
Dried leaves				
Avijor	Cross-validation	97.5%	99.2%	95.0%
	Test set validation	95.0%	98.3%	93.3%
Guara	Cross-validation	95.0%	100%	93.3%
	Test set validation	92.5%	97.5%	91.7%
Pentacebas	Cross-validation	97.5%	99.2%	93.3%
	Test set validation	97.5%	100%	98.3%
Fresh leaves				
Avijor	Cross-validation	100%	97.5%	100%
	Test set validation	99.2%	100%	99.2%
Guara	Cross-validation	99.2%	97.5%	99.2%
	Test set validation	98.3%	99.2%	98.3%
Pentacebas	Cross-validation	99.2%	100%	99.2%
	Test set validation	99.2%	99.2%	99.2%

accuracy obtained with this spectral pre-treatment was 97.5% at both the cross-validation and test set validation levels. No relevant differences between the other two spectral pre-treatments were observed, so de-trending did not appear to have a significant effect. It is important to note that in the case of fresh leaves, similar results were obtained with the three different spectral pre-treatments.

### 3.3.2. Sample processing study

Each sample processing method had its advantages and disadvantages. Fresh leaves did not require any processing, so measurement was faster and easier than it was with the other types of samples. However, the water content of the leaves was a disadvantage, because it generated wide bands in the NIR spectra. This could make discrimination between varieties more difficult. Samples can be dehydrated to circumvent the effects of water, but this process is time-consuming (48 h), so it is not the best option if rapid identification is required.

To evaluate which of the processed samples was the most suitable for varietal classification, the advantages and disadvantages of each were considered together with the PLS-DA classification results obtained with SNV spectral pre-treatment and application of the SG first derivative and mean centering (Table 4).

The results obtained with the three types of sample processing at the calibration level could be considered quite good, although those obtained with fresh leaves were less stellar. The dried powdered leaves provided a higher percentage of correct classifications. For the test set validation, high percentages of correct classifications were obtained with all varieties and processed sample types. The results provided by the dried leaves were not as good as those obtained with the other two processed sample types, although the *Pentacebas* variety was correctly classified in 100% of the test set validations. Fresh leaves provided almost perfect classification, and nearly 100% correct classification was attained with dried powdered leaves. Taking only the PLS-DA results into account, the best sample processing method was drying and powdering the leaves. Considering the methodological aspects, using fresh leaves was the fastest and easiest option. The biggest drawback of fresh leaves was their water content, but this did not seem to hinder discrimination between the varieties studied.

In the ASCA-ANOVA model performed in assay two (Table 3), the strongest effect on dried leaves was contributed by the leaf age factor. The age factor accounted for more variability than even the tree variety

factor, which could be problematic. Fresh leaves exhibited more favorable behavior in the ASCA-ANOVA model. Results of the ASCA-ANOVA model with dried powdered leaves were similar to those obtained with fresh leaves, but the residual was higher.

## 4. Conclusions and perspectives

In this study, we defined a methodology for construction of a classification model that could discriminate between *Prunus dulcis* varieties using NIRS. We also identified the most important sampling and analysis aspects. In assay one, differences were seen in the PCA whether or not the midvein was included. The central leaf region provided more useful information for discriminating between almond tree varieties, because it contained both the primary vein and the laminar tissues. We also attempted to determine which surface of the leaves, adaxial or abaxial, was the most suitable for analysis. Despite the spectral differences observed, the comparison made using the PLS-DA model indicated this was not an important aspect.

In assay two, no notable differences were detected between trees of the same variety, which indicated that trees within each variety were quite homogeneous. Differences were observed at the PCA level between young and adult leaves, which indicated age was important to consider during the sampling process.

The best results from the PLS-DA models in assay three were obtained with dried powdered leaves when SNV was used for spectral pre-treatment with application of the SG first derivative (15-point window, second order) and mean centering. However, fresh leaves appeared to be the easiest and most suitable samples for laboratory or industrial analysis. These results indicated that both fresh leaves and dried powdered leaves could be useful for discriminating between *Prunus dulcis* varieties using NIR spectroscopy.

All the information gathered in the present study will be used to build a classification model that includes more *Prunus dulcis* varieties. The potential of NIR spectroscopy for the classification of almond tree varieties and its implementation as a quality control tool in the nursery plant industry will be studied.

## Funding

This work was supported by Generalitat de Catalunya through a grant from Program of Industrial Doctorates (DI-COF 2017) and by The



Spanish Ministry of Economy and Competitiveness (project AGL 2015-70106-R).

## Acknowledgment

The authors thank Thermo Scientific for temporary assignment of the NIR equipment.

## References

- [1] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, H. Erlich, Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction, *Cold Spring Harbor Symp. Quant. Biol.* 51 (1986) 263–273, <https://doi.org/10.1101/SQB.1986.051.01.032>.
- [2] P.K. Gupta, J.K. Roy, M. Prasad, Single nucleotide polymorphism a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plant.pdf, *Curr. Sci.* 80 (2001) 524–535, <https://doi.org/10.2307/24104242>.
- [3] M. Makky, P. Soni, In situ quality assessment of intact oil palm fresh fruit bunches using rapid portable non-contact and non-destructive approach, *J. Food Eng.* 120 (2014) 248–259, <https://doi.org/10.1016/J.JFOODENG.2013.08.011>.
- [4] M.V. Reboucas, J.B. dos Santos, D. Domingos, A.R.C.G. Massa, Near-infrared spectroscopic prediction of chemical composition of a series of petrochemical process streams for aromatics production, *Vib. Spectrosc.* 52 (2010) 97–102, <https://doi.org/10.1016/J.VIBSPEC.2009.09.006>.
- [5] M. Verstraeten, D. Van Hauwermeiren, M. Hellings, E. Hermans, J. Geens, C. Vervaeke, I. Nopens, T. De Beer, Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process, *Int. J. Pharm.* 546 (2018) 247–254, <https://doi.org/10.1016/j.ijpharm.2018.05.043>.
- [6] Z. Seregély, T. Deák, G.D. Bisztray, Distinguishing melon genotypes using NIR spectroscopy, *Chemometr. Intell. Lab. Syst.* 72 (2004) 195–203, <https://doi.org/10.1016/j.chemolab.2004.01.013>.
- [7] Q. Fan, Y. Wang, P. Sun, S. Liu, Y. Li, Discrimination of Ephedra plants with diffuse reflectance FT-NIRS and multivariate analysis, *Talanta* (2010), <https://doi.org/10.1016/j.talanta.2009.09.018>.
- [8] F.M. Durgante, N. Higuchi, A. Almeida, A. Vicentini, Species spectral signature: Discriminating closely related plant species in the amazon with near-infrared leaf spectroscopy, *For. Ecol. Manag.* 291 (2013) 240–248, <https://doi.org/10.1016/j.foreco.2012.10.045>.
- [9] Y.N. Shao, C.Q. Xie, L.J. Jiang, J.H. Shi, J.J. Zhu, Y. He, Discrimination of tomatoes bred by spaceflight mutagenesis using visible/near infrared spectroscopy and chemometrics, *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* (2015), <https://doi.org/10.1016/j.saa.2015.01.018>.
- [10] L. Zhang, S.S. Wang, Y.F. Wang, J.R. Pan, C. Zhu, Discrimination of transgenic rice based on near infrared reflectance spectroscopy and partial least squares regression discriminant analysis, *Rice Sci.* (2015), <https://doi.org/10.1016/j.rsci.2015.09.004>.
- [11] L. de Oliveira Moura, D. de Carvalho Lopes, A.J. Steidle Neto, L. de Castro Louback Ferraz, L. de Almeida Carlos, L.M. Martins, Evaluation of techniques for automatic classification of lettuce based on spectral reflectance, *Food Anal. Methods* 9 (2016) 1799–1806, <https://doi.org/10.1007/s12161-015-0366-5>.
- [12] G.D. Batten, Plant analysis using near infrared reflectance spectroscopy: The potential and the limitations, *Aust. J. Exp. Agric.* 38 (1998) 697–706, <https://doi.org/10.1071/EA97146>.
- [13] P. Skolik, M.R. McAinsh, F.L. Martin, *Biospectroscopy for plant and crop science*, in: C.S. João Lopes (Ed.), *Compr. Anal. Chem.* Elsevier, 2018, pp. 15–49, <https://doi.org/10.1016/BS.COAC.2018.03.001>.
- [14] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra, *J. Pharm. Biomed. Anal.* 21 (1999) 115–132, <https://doi.org/10.1080/13554790490495140>.
- [15] L. Xu, Y.P. Zhou, L.J. Tang, H.L. Wu, J.H. Jiang, G.L. Shen, R.Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, *Anal. Chim. Acta* 616 (2008) 138–143, <https://doi.org/10.1016/j.aca.2008.04.031>.
- [16] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, *TrAC Trends Anal. Chem. (Reference Ed.)* 28 (2009) 1201–1222, <https://doi.org/10.1016/j.trac.2009.07.007>.
- [17] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Appl. Spectrosc.* 43 (1989) 772–777, <https://doi.org/10.1366/00037028940220201>.
- [18] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Anal. Chem.* 36 (1964) 1627–1639, <https://doi.org/10.1021/ac60214a047>.
- [19] R.S.K. Barry, M. Wise, Neal B. Gallager, Rasmus Bro, Jeremy M. Shaver, Willem Windig, *Chemometrics Tutorial for PLS Toolbox and Solo*, (2006), <https://doi.org/10.1016/j.cplett.2004.08.130>.
- [20] P. Wang, Z. Yu, Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review, *J. Pharm. Anal.* 5 (2015) 277–284, <https://doi.org/10.1016/j.jpaha.2015.04.001>.
- [21] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: Taking the magic away, *J. Chemom.* 28 (2014) 213–225, <https://doi.org/10.1002/cem.2609>.
- [22] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models, PLS-DA, *Anal. Methods* 5 (2013) 3790, <https://doi.org/10.1039/c3ay40582f>.
- [23] L. Sjöhle, S. Wold, Analysis of variance (ANOVA), *Chemometr. Intell. Lab. Syst.* 6 (1989) 259–272, [https://doi.org/10.1016/0169-7439\(89\)80095-4](https://doi.org/10.1016/0169-7439(89)80095-4).
- [24] L. Sjöhle, S. Wold, Multivariate analysis of variance (MANOVA), *Chemometr. Intell. Lab. Syst.* 9 (1990) 127–141, [https://doi.org/10.1016/0169-7439\(90\)80094-M](https://doi.org/10.1016/0169-7439(90)80094-M).
- [25] G. Zwanenburg, H.C.J. Hoefsloot, J.A. Westerhuis, J.J. Jansen, A.K. Smilde, ANOVA–principal component analysis and ANOVA–simultaneous component analysis: A comparison, *J. Chemom.* 25 (2011) 561–567, <https://doi.org/10.1002/cem.1400>.
- [26] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA–simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data, *Bioinformatics* 21 (2005) 3043–3048, <https://doi.org/10.1093/bioinformatics/bti476>.
- [27] J.J. Jansen, H.C.J. Hoefsloot, J. Van Der Greef, M.E. Timmerman, J.A. Westerhuis, A.K. Smilde, ASCA: Analysis of multivariate data obtained from an experimental design, *J. Chemom.* 19 (2005) 469–481, <https://doi.org/10.1002/cem.952>.
- [28] O.E. de Noord, E.H. Theobald, Multilevel component analysis and multilevel PLS of chemical process data, *J. Chemom.* 19 (2005) 301–307, <https://doi.org/10.1002/cem.933>.
- [29] N. Suzuki, R.M. Rivero, V. Shulaev, E. Blumwald, R. Mittler, Abiotic and biotic stress combinations, *New Phytol.* 203 (2014) 32–43, <https://doi.org/10.1111/nph.12797>.
- [30] J.C.M.S. Moura, C.A.V. Bonine, J. de Oliveira Fernandes Viana, M.C. Dornelas, P. Mazzafera, Abiotic and biotic stresses and changes in the lignin content and composition in plants, *J. Integr. Plant Biol.* 52 (2010) 360–376, <https://doi.org/10.1111/j.1744-7909.2010.00892.x>.



**Capítulo 2. Multivariate  
Classification of *Prunus dulcis*  
Varieties Using Leaves of  
Nursery Plants and Near-  
Infrared Spectroscopy**



OPEN

# Multivariate Classification of *Prunus dulcis* Varieties using Leaves of Nursery Plants and Near-Infrared Spectroscopy

Sergio Borraz-Martínez <sup>1,2\*</sup>, Joan Simó <sup>4</sup>, Anna Gras <sup>1</sup>, Mariàngela Mestre<sup>2</sup> & Ricard Boqué <sup>3</sup>

The emergence of new almond tree (*Prunus dulcis*) varieties with agricultural interest is forcing the nursery plant industry to establish quality systems to keep varietal purity in the production stage. The aim of this study is to assess the capability of near-infrared spectroscopy (NIRS) to classify different *Prunus dulcis* varieties as an alternative to more expensive methods. Fresh and dried-powdered leaves of six different varieties of almond trees of commercial interest (*Avijor*, *Guara*, *Isabelona*, *Marta*, *Pentacebas* and *Soleta*) were used. The most important variables to discriminate between these varieties were studied through of three scientifically accepted indicators (Variable importance in projection, selectivity ratio and vector of the regression coefficients). The results showed that the 7000 to 4000  $\text{cm}^{-1}$  range contains the most useful variables, which allowed to decrease the complexity of the data set. Concerning to the classification models, a high percentage of correct classifications (90–100%) was obtained, where dried-powdered leaves showed better results than fresh leaves. However, the classification rate of both kinds of leaves evidences the capacity of the near-infrared spectroscopy to discriminate *Prunus dulcis* varieties. We demonstrate with these results the capability of the NIRS technology as a quality control tool in nursery plant industry.

The almond market is in expansion. The world production in the 2017/2018 season reached the record of 1.3 million tons, 6% above the registered in the previous season, according to the latest estimates of the United States Department of Agriculture (USDA)<sup>1</sup>. The Food and Agriculture Organization Corporate Statistical Database (FAOSTAT)<sup>2</sup> shows that, in Spain, in the period 2015–2017, the almond tree harvested area increased by 15%, which was reflected in an increase in production of 21%. These data demonstrate the importance of almond tree in world agriculture. For this reason, new varieties of almond trees have been increasing<sup>3,4</sup>. Varietal control in the production stage is a great challenge. The appearance of varietal mixtures within a batch, which should be homogeneous, is an important trouble, not only because the customer receives unwanted vegetal material, but also because nursery plant companies may face expensive fines and the deterioration of their corporative image.

The current method used for the identification of plants consists of performing molecular studies of the vegetal material in order to obtain molecular profiles and identifying the variety by using microsatellites and single nucleotide polymorphisms (SNPs)<sup>5</sup>. These biomolecular techniques are reliable. However, they are expensive and time consuming, and therefore cannot be applied routinely and with a high sample throughput. For this reason, the most common varietal control system consists of ensuring a correct and detailed traceability from the origin of the plant to its delivery to the customer, which is not always reliable. In this context, near-infrared spectroscopy (NIRS), combined with chemometrics techniques, is potentially a rapid, accurate, and non-destructive alternative.

<sup>1</sup>Dept. of Agri-Food Engineering and Biotechnology, Universitat Politècnica de Catalunya, Campus Baix Llobregat, Esteve Terrades 8, 08860, Castelldefels, Spain. <sup>2</sup>Center of Initial Materials, Agromillora Iberia S.L.U, Ctra. BV-2247 km. 3, 08770, Sant Sadurn d'Anoia, Spain. <sup>3</sup>Universitat Rovira i Virgili, Dept. of Analytical Chemistry and Organic Chemistry, Campus Sescelades, 43007, Tarragona, Spain. <sup>4</sup>Fundació Miquel Agustí, Campus Baix Llobregat, Esteve Terrades 8, 08860, Castelldefels, Spain. \*email: [sergio.borraz@upc.edu](mailto:sergio.borraz@upc.edu)

Varieties	Parents	Breeder	No. of fresh leaves	No. of dried-ground leaves	Harvesting time
<i>Avijor</i>	Ferragnès x Tuono	INRA	50	50	October 2018
<i>Guara</i>	Unknown	CITA	50	50	October 2018
<i>Isabelona</i>	Blanquerna x Bella d'Aurons	CITA	50	50	October 2018
<i>Marta</i>	Ferragnès x Tuono	CEBAS-CSIC	50	50	October 2018
<i>Pentacebas</i>	S5133 x Lauran	CEBAS-CSIC	50	50	October 2018
<i>Soleta</i>	Blanquerna x Bella d'Aurons	CITA	50	50	October 2018
Total			300	300	

**Table 1.** Description of the different varieties of *Prunus dulcis* studied. (INRA = Institut National de la Recherche Agronomique (France); CITA = Centro de Investigación y Tecnología Agroalimentaria de Aragón (Spain); CEBAS-CSIC = Centro de Edafología y Biología Aplicada del Segura (Spain)).

In the last years, NIRS has gained importance in the agriculture sector as an interesting tool for monitoring and quality assessment of agricultural products<sup>6–8</sup>. Various studies have been published concerning the geographic origin<sup>9–11</sup> and species discrimination in grapevine<sup>12</sup>, tea<sup>13</sup>, tomato<sup>14</sup> and coffee<sup>15</sup>.

Moreover, several studies have shown the existence of an empirical relationship between leaf spectral properties and leaf physiological conditions<sup>16,17</sup>.

In the present study, six varieties of *Prunus dulcis* of agricultural interest were chosen: *Avijor*, *Guara*, *Isabelona*, *Marta*, *Pentacebas* and *Soleta*. All of them are genetically close because they come from crosses of traditional varieties, and some of them even share parents. In fact, this is the case of *Avijor* and *Marta*, both coming from the cross between “*Ferragnès x Tuono*”, and *Isabelona* and *Soleta* come from the cross between “*Blanquerna x Bella d'Aurons*”. Due to their genetic proximity, the morphology of these six varieties is difficult to distinguish.

Despite the present study focuses on specific varieties of *Prunus dulcis*, as the leaves' composition is similar in all the angiosperm species<sup>18</sup>, the knowledge achieved in this work could be applied to discriminate other plant species.

The present study is the continuation of a previous research, where a sampling methodology was developed and optimized to analyze leaves from *Prunus dulcis* varieties using near infrared spectroscopy<sup>19</sup>. In the cited study, several sources of variability affecting the measurements were investigated, such as the regions of the leaves analyzed and the age of the leaves, and how these factors affect to the spectral signature of the varieties. Moreover, the most suitable preprocessing of the leaves and the best spectral pre-treatment were determined. In the present study, the information gathered from the previous research was used to assess the potential of NIRS to classify several varieties of almond trees (*Prunus dulcis*) that are genetically close and morphologically not distinguishable. The specific objectives are to: 1) develop a classification model capable of discriminating between six varieties of *Prunus dulcis*; 2) investigate the most important variables for the discrimination of the varieties; and 3) compare the classification results obtained with dried-ground leaves and fresh leaves.

## Material and Methods

**Description of the sampling field.** The studied almond trees belong to the mother-plant field of the Centre of Initial Materials of Agromillora Iberia S.L.U., which is located in Sant Sadurní d'Anoia (Catalonia, Spain). The sanitary quality of the sampled trees was verified by means of molecular analysis, which Agromillora S.L.U. perform periodically as a quality control. Moreover, the trees were certified by the company. The same mother-plant field was used in a previously study<sup>19</sup>.

**Sampling protocol.** Six *Prunus dulcis* varieties of agricultural interest were used. One hundred leaves were collected from ten trees for each variety. In total, six hundred leaves were collected from which three hundred fresh leaves were analyzed without any prior treatment, and the rest was dried and ground. The number of samples per variety was identical. Table 1 details the samples used in this study. The leaf samples collected were introduced in plastic bags with an identification code and stored at 4 °C until they were analyzed.

**Sample pre-processing.** To obtain dried-powdered leaves, fresh leaves were heated in an oven at 65 °C for 48 h. Once dried, a grinder was used to obtain a homogeneous powder. Further, the powder was stored in a desiccator with silica gel to avoid moisture<sup>19</sup>.

**Acquisition of NIR spectra.** The NIR spectra acquisition method shown below was developed in a previous study<sup>19</sup>. An Antaris II FT-NIR analyzer (Thermo Scientific, USA), equipped with an integrating sphere module was used to scan the samples, which were measured in the spectral range of 12000–3800 cm<sup>-1</sup> (833–2630 nm). The instrument configuration used consisted in the average of 32 scans by spectrum with a resolution of 4 cm<sup>-1</sup>. Each sample was analyzed in triplicate and the average of the replicates was used in the subsequent discrimination models. Every 20 minutes a background spectrum was collected. The powdered leaf samples were measured in a standard sample cup available with the instrument and fresh leaves were placed directly over the sphere. In both cases, samples were covered to prevent interference from environmental light. The reflectance spectra were mathematically transformed to absorbance by means of the log(1/R), where R is the reflectance. The room temperature was maintained at ~25 °C, and the humidity remained constant throughout the spectral acquisition process.

**Data preparation.** When applying supervised multivariate classification methods, it is important to validate mathematical models by using an independent test data set. We applied two difference strategies to split the original dataset into calibration and test set: random split and the Kennard–Stone algorithm<sup>20</sup>. Similar results were obtained, so finally we decided to use the Kennard–Stone algorithm, which was applied to each class separately to split the data set into calibration set (70% of the leaves) and a test set (30% of the leaves).

**Spectral data pre-treatment.** To enhance the spectral features and reduce systematic noise, such as baseline variation, light scattering, and path length differences, a mathematical pre-processing of the original spectra was necessary. The spectral pretreatment was optimized in a previous study<sup>19</sup> and consisted of a combination of standard normal variate (SNV) with *Savitzky–Golay* (SG) first (1<sup>st</sup>) derivative filter and mean centering. SNV is a normalization procedure for spectral light scattering correction. It is used to correct additive and multiplicative effects in spectra caused by particle size variation. SNV calculates the standard deviation of all the variables in a given sample spectrum. The entire data set is then normalized by this value, which yields a unit standard deviation ( $s = 1$ ) for the sample spectrum<sup>19,21</sup>. SG first derivative was applied to remove the baseline drift and enhance small spectral differences. The SG derivative method includes a smoothing step, the SG algorithm, which corrects the additional noise caused by the application of the derivative. The SG algorithm requires the selection of the order of the polynomial, order of the derivative, and filter width, which corresponds to the size of the window<sup>19,22</sup>. Herein, a 15-point window and second order polynomial were selected. Finally, mean centering was applied, which consists of calculating the mean value of each column and subtracting it from each individual value in the column. After mean centering, the mean value of each column equals zero, and each row of mean-centered data reflects only how it differs from the average sample in the original data matrix<sup>19,23</sup>.

The PLS\_Toolbox (Eigenvector Research Incorporated, Manson, USA) with MATLAB R2017b (MathWorks, Natick, USA) were used to perform the spectral pre-treatments.

**Partial Least-Squares discriminant analysis (PLS-DA).** PLS-DA<sup>24</sup> is based on the PLS regression algorithm and seek to find the linear combinations of the original variables (latent variables (LVs)) that have maximum covariance with the Y-variables (classes)<sup>24</sup>. Unlike PLS regression, in PLS-DA the Y-block is coded with dummy variables. Each sample is codified with one (1) if the sample belongs to a given class, and zero (0) if the sample does not belong to a given class. The result of a PLS-DA model is a series of prediction values for the different classes, that is, values around one for the class of interest and values around zero for the rest of classes. Finally, a threshold is calculated and optimized that optimally discriminates the different classes. With this PLS-DA model, validated with the test set of samples, it is then possible to predict the class of an unknown sample. In our case, six classes of leaves, one for each variety, were used, and therefore, the Y-block contained six columns.

A Venetian blinds cross-validation, with a data split of 10 and one sample per blind (thickness), was used to find the optimal number of factors (LVs) for the PLS-DA model<sup>19</sup>. The number of factors that showed the lowest classification error was selected as optimal. After the internal validation (optimization), the prediction ability of the model was assessed using samples that were not included in the calibration (30% of the total data set). The results are shown in Tables 2 and 3 (test-set validation).

To study the model results three different statistical parameters were calculated: sensitivity, specificity and accuracy. The sensitivity measures the proportion of actual positives that are correctly identified as such, the specificity measures the proportion of actual negatives that are correctly identified as such, and the accuracy measures the proportion of correctly classified samples, that is, the sum of true positives and true negatives divided by the total number of samples.

**Study of the important variables.** The prediction accuracy of the full-spectrum PLS-DA model is negatively affected by the water absorption and other unrelated or collinear spectral variables. By selecting the optimal wavelengths, it is possible to reduce the complexity of the multivariate calibration model, by reducing the computational requirements and avoiding the inclusion of noisy variables. Spectral variable selection is then essential to generate models with better prediction accuracies<sup>25</sup>. However, it is crucial to use different variable selection methods in order to ensure that no important variables are eliminated.

The variable importance in projection (VIP) score, selectivity ratio and vector of regression coefficients were used to study the most important variables required to discriminate between the classes.

The VIP score is a parameter used for calculating the cumulative values of the influence of individual X-variables in the model<sup>26</sup>. In other words, VIP scores estimate the importance of each variable in the projection used in a PLS model<sup>27</sup>. Scores close to or greater than 1 are considered relevant.

The selectivity ratio provides a simple numerical assessment of the importance of each variable in a regression model<sup>28</sup>. By using the y-vector as a target, it is possible to transform the PLS components to obtain a single predictive target-projected component analogously to the predictive component in the orthogonal partial least squares discriminant analysis (OPLS-DA). By calculating the ratio between the explained and residual variance of the spectral variables on the target-projected component, a selectivity ratio plot is obtained and can be used for variable selection<sup>29</sup>. The larger the selectivity ratio, the more useful a variable for prediction. Variables with lower selectivity ratios may be excluded without degrading the performance of the model<sup>28</sup>.

The vector of the regression coefficients is an informative tool to select variables in multivariate calibration. The regression vector obtained from a multivariate regression can be defined as the expected change in the response per unit change in the variable if all the variables and responses are linearly related<sup>30,31</sup>. Variables with low regression coefficients do not contribute significantly to the prediction and can be removed.

## Results and Discussion

**Spectra investigation.** Figure 1 shows the mean raw spectra of the fresh and dried-powdered leaves. It can be observed that the shape of the spectra from both types of samples differs. This is due to the water content, which generates broad bands in the NIR spectra. However, both types of spectra present a low information zone that corresponds to the region between 7500 and 10000  $\text{cm}^{-1}$ . In contrast, the remaining region, from 4000 to 7500  $\text{cm}^{-1}$ , seems to be more informative, and the importance of this region will be evaluated below. It is noted that both regions were also observed in others studies performed on tea and bamboo leaves<sup>13,32</sup>.

By using NIR tables, it was possible to identify the main chemical groups behind the different bands. These bands in the NIR region correspond to the overtones and combination bands.

Regarding fresh leaves, two main broad bands were observed at 5000 and 7000  $\text{cm}^{-1}$ , corresponding to O-H vibrations and more specifically the combination O-H stretching and first overtone O-H stretching, respectively. Water is the main responsible of these bands as observed in a study developed on grapevine leaves<sup>33</sup>, although the O-H bond is also found in carbohydrates. Another important peak is observed between 4000 and 4500  $\text{cm}^{-1}$ . This absorption is attributed to the combination C-H stretching. The C-H bond is present in monosaccharides and, together with oxygen, conforms the aldehyde group, which is also present in carbohydrates.

Concerning the dried-powdered leaves, the range between 4000 to 7500  $\text{cm}^{-1}$  includes the main bands, which correspond to the combination of C-H (7692–7042  $\text{cm}^{-1}$ ), N-H (5000–4545  $\text{cm}^{-1}$ ), and O-H stretching (5000–4545  $\text{cm}^{-1}$ ) and the first overtone of C-H (6061–5556  $\text{cm}^{-1}$ ), N-H and O-H (7143–6250  $\text{cm}^{-1}$ ) stretching. All the bonds mentioned before are commonly found in carbohydrates and proteins. In addition, a small band can be observed between 8000 and 9000  $\text{cm}^{-1}$  approximately. This band correspond to the second overtone of the C-H stretching (9091–8163  $\text{cm}^{-1}$ ).

**Classification model using the whole spectra.** Table 2 presents the results of the classification models for fresh and dried-powdered leaves. Concerning the dried-powdered leaf results, the classification accuracy in the cross-validation and test set validation was greater than 95% for all the varieties. In most cases, the test set validation showed a higher classification accuracy than that of cross-validation, except for *Avijor* class, due to the erroneous assignation of two *Guara* samples. This is not a common situation, as in general, better accuracy is reached for cross-validation than for test-set validation. However, sometimes this can happen when the test set does not contain a high number of samples. *Avijor* was presented the highest results for the cross-validation test, with a classification accuracy of 99.0%. In case of the test set validation, the best classification rate was obtained for the *Marta* class, with a classification accuracy of 100%. The lowest classification rate was obtained for *Guara*, which was the only variety showing an accuracy below 97%. Three *Guara* samples were misclassified in the test set validation, two samples were classified in the *Avijor* class, and one sample was assigned to the *Isabelona* class. Errors in the assignation of *Guara* samples were also reported in the cross-validation. Consequently, the *Guara* class showed a low sensitivity in both data sets. *Isabelona* and *Marta* showed the same sensitivity in both, cross-validation and test set validation. In addition, a correct assignation of all the samples of the validation data set was achieved for both varieties. *Marta* showed a higher specificity than that of *Pentacebas* because one *Guara* sample and one *Pentacebas* sample were misclassified in the *Isabelona* class. Thus, *Marta* presented better classification accuracy than that of *Isabelona* (100% and 97.8% accuracy, respectively). *Soleta* and *Pentacebas* showed the same sensitivity in the test set validation. The specificity was higher for *Soleta* than that of *Pentacebas* because one *Soleta* sample was erroneously assigned to *Pentacebas*, which was also reflected in a higher classification accuracy for *Soleta*.

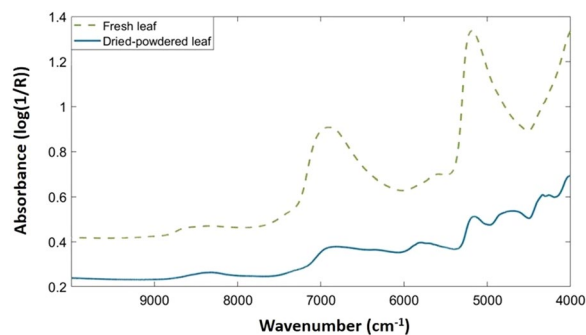
Concerning the fresh leaf results, except for *Isabelona*, the rest of the varieties showed a classification accuracy higher than 90% in both cross-validation and test set validation. Specifically, an accuracy between 89.5–98.6% and between 92.2–100% was obtained for the cross-validation and test set validation, respectively. The best classification results were obtained for *Marta* variety, with 98.6% and 100% of accuracy for the cross-validation and test set validation, respectively. *Pentacebas* and *Avijor* reached an accuracy of 100% in the test set validation. *Marta*, *Pentacebas* and *Avijor* presented a relatively low sensitivity in the cross-validation, but the samples were correctly predicted in test set validation. The lowest classification accuracy was obtained for *Isabelona* and *Guara*. *Isabelona* had a 89.5% accuracy in the cross-validation and 92.2% in the test set validation. *Guara* was the only variety showing a lower accuracy in the test set validation than that in the cross-validation. Four samples of *Guara* were assigned to *Isabelona*, and one sample of *Isabelona* was predicted as *Guara*. Similarly, in the dried-powdered leaf model, the *Guara* class was not correctly modelled. *Soleta* showed a high sensitivity in test set validation as only one sample of *Guara* was misclassified. However, due to the erroneous assignment of two *Guara* samples and one *Isabelona* sample, the classification accuracy of *Soleta* class was reduced. The highest specificity was obtained for *Marta*, *Pentacebas*, and *Avijor*.

For both models the prediction accuracy was high despite the genetic proximity of the varieties. However, the dried-powdered leaf model presented a higher accuracy in both the cross-validation and test set validation than that of the fresh leaf model. The homogeneity, and especially, the absence of water in the dried-powdered leaves seem to explain these results. The vibration of the water molecules generates broad bands in the near-infrared range that can hide other features, as it can be seen in the Fig. 1 and in other studies<sup>34</sup>. However, in of fresh leaf model, three varieties reached a 100% classification accuracy in the test set validation, in contrast, this was only achieved for one variety in the dried-powdered model. *Guara* was the worst classified variety in both models. In contrast, *Marta* showed a high classification accuracy in both models.

The full-spectrum PLS-DA model results were used as a reference to check that the variable selection process was correctly performed.

**Study of the most important variables.** The VIP score, selectivity ratio, and vector of the regression coefficients were used to study the most important variables to discriminate between the classes. The results of these indicators are shown in Fig. 2.





**Figure 1.** NIR mean raw spectra of fresh (green dashed line) and dried-powdered (blue solid line) leaves.

Real class	Data set	Sensitivity	Specificity	Accuracy
<b>Dried-powdered leaves</b>				
Avijor	Cross-validation	0.971	0.994	99.0%
	<b>Test set Validation</b>	<b>1.000</b>	<b>0.973</b>	<b>97.8%</b>
Guara	Cross-validation	0.829	0.983	95.7%
	<b>Test set Validation</b>	<b>0.800</b>	<b>1.000</b>	<b>96.7%</b>
Isabelona	Cross-validation	0.971	0.960	96.2%
	<b>Test set Validation</b>	<b>1.000</b>	<b>0.973</b>	<b>97.8%</b>
Marta	Cross-validation	0.971	0.983	98.1%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
Pentacebas	Cross-validation	0.886	0.989	97.1%
	<b>Test set Validation</b>	<b>0.933</b>	<b>0.987</b>	<b>97.8%</b>
Soleta	Cross-validation	0.886	0.994	97.6%
	<b>Test set Validation</b>	<b>0.933</b>	<b>1.000</b>	<b>98.9%</b>
<b>Fresh leaves</b>				
Avijor	Cross-validation	0.857	0.977	95.7%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
Guara	Cross-validation	0.857	0.943	92.9%
	<b>Test set Validation</b>	<b>0.600</b>	<b>0.987</b>	<b>92.2%</b>
Isabelona	Cross-validation	0.657	0.943	89.5%
	<b>Test set Validation</b>	<b>0.867</b>	<b>0.933</b>	<b>92.2%</b>
Marta	Cross-validation	0.943	0.994	98.6%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
Pentacebas	Cross-validation	0.914	0.983	97.1%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
Soleta	Cross-validation	0.743	0.954	91.9%
	<b>Test set Validation</b>	<b>0.933</b>	<b>0.960</b>	<b>95.6%</b>

**Table 2.** PLS-DA results of the classification of six varieties of *Prunus dulcis* using the entire spectra.

Regarding the dried-powdered leaf model, the VIP score (Fig. 2a) showed important variables located in the range from 4000 to 6000  $\text{cm}^{-1}$ . Similar results were obtained with the regression vector (Fig. 2b). However, around 6600  $\text{cm}^{-1}$ , other important variables appeared in the selectivity ratio plot (Fig. 2c). In the range from 10000 to 7000  $\text{cm}^{-1}$ , no important variable was detected. Therefore, it was decided to cut the spectra from 6700  $\text{cm}^{-1}$  in order to include only relevant variables.

Concerning the fresh leaf model, the VIP score and regression vector (Fig. 2d,e) showed important variables in the range between 6000 and 4000  $\text{cm}^{-1}$ . A small gap was detected between 7000 and 6000  $\text{cm}^{-1}$ , and the variables inside this range did not help the discriminant study. However, around 7000  $\text{cm}^{-1}$ , important variables can be found. The same results were obtained for the selectivity ratio indicator (Fig. 2f), except for the appearance of a band between 7000 and 6000  $\text{cm}^{-1}$ . The rest of variables were not useful to discriminate the almond tree varieties. Considering this, the fresh leaf spectra were cut from 7500  $\text{cm}^{-1}$ .

In both cases, fresh and dried-powdered leaves, the variables selected contain the bands related to C-H, O-H and N-H bonds, which can mainly be associated to proteins and carbohydrates, and water too in the fresh leaf. The amount of these bonds and their configuration inside the molecules affect the shape of the spectrum and, therefore, have an influence in the spectral fingerprint of the varieties. However, the region removed corresponded to wavenumbers with no important bands in the raw spectra (Fig. 1).

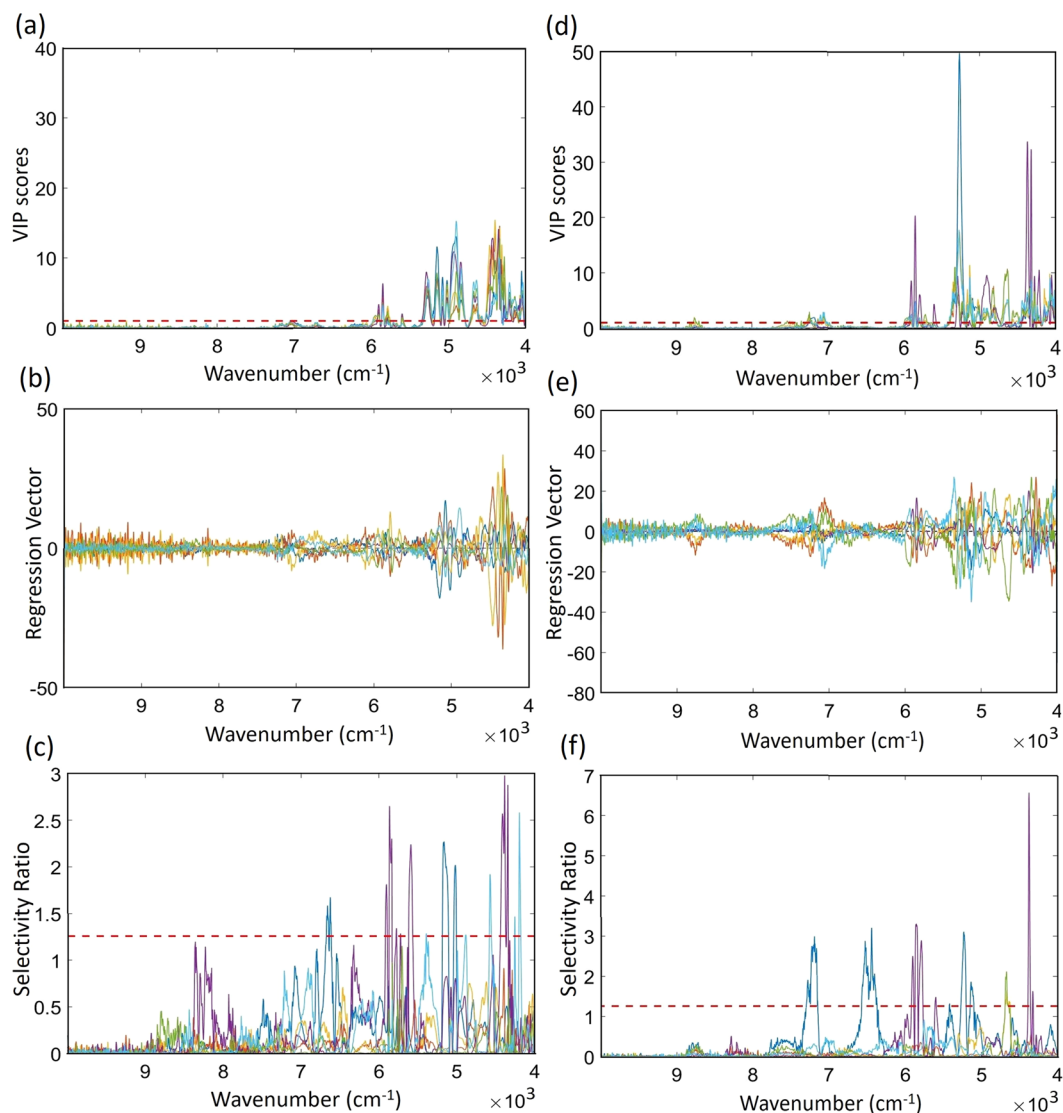
Real class	Data set	Sensitivity	Specificity	Accuracy
<b>Dried-powdered leaves</b>				
<i>Avijor</i>	Cross-validation	0.971	0.994	99.0%
	<b>Test set Validation</b>	<b>1.000</b>	<b>0.973</b>	<b>97.8%</b>
<i>Guara</i>	Cross-validation	0.829	0.977	95.2%
	<b>Test set Validation</b>	<b>0.867</b>	<b>0.987</b>	<b>96.7%</b>
<i>Isabelona</i>	Cross-validation	0.914	0.943	93.8%
	<b>Test set Validation</b>	<b>0.933</b>	<b>0.987</b>	<b>97.8%</b>
<i>Marta</i>	Cross-validation	1.000	0.989	99.0%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
<i>Pentacebas</i>	Cross-validation	0.886	0.983	96.7%
	<b>Test set Validation</b>	<b>0.933</b>	<b>1.000</b>	<b>98.9%</b>
<i>Soleta</i>	Cross-validation	0.800	0.994	96.2%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
<b>Fresh leaves</b>				
<i>Avijor</i>	Cross-validation	0.857	0.977	95.7%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
<i>Guara</i>	Cross-validation	0.829	0.943	92.4%
	<b>Test set Validation</b>	<b>0.600</b>	<b>0.987</b>	<b>92.2%</b>
<i>Isabelona</i>	Cross-validation	0.686	0.949	90.5%
	<b>Test set Validation</b>	<b>0.867</b>	<b>0.907</b>	<b>90.0%</b>
<i>Marta</i>	Cross-validation	0.914	0.983	97.1%
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100%</b>
<i>Pentacebas</i>	Cross-validation	0.886	0.971	95.7%
	<b>Test set Validation</b>	<b>0.867</b>	<b>0.133</b>	<b>97.8%</b>
<i>Soleta</i>	Cross-validation	0.743	0.960	92.4%
	<b>Test set Validation</b>	<b>0.933</b>	<b>0.960</b>	<b>95.6%</b>

**Table 3.** PLS-DA results of the classification of six varieties of *Prunus dulcis* after variable selection.

**Model using the most important wavenumbers.** A new PLS-DA model was created using only the most important variables selected in the previous section. Table 3 shows the discrimination model results for dried-powdered and fresh leaves.

Concerning the dried-powdered leaf results, *Avijor* and *Guara* showed the same results as those found by using the entire spectra. Therefore, the removed wavenumbers did not contain important information for the discrimination study. *Isabelona* showed a small decrease of the classification accuracy in the cross-validation with respect to the results obtained using the entire spectrum (93.8% and 96.2%, respectively). This difference was not reflected in the classification accuracy of the test set. *Marta* class showed 100% of accuracy in the test set classification, and a more accurate result in the cross-validation compared to that of the model using all the variables. *Pentacebas* and *Soleta* showed a lower classification accuracy in the cross-validation compared to the ones obtained in the previous model. However, more accurate results were obtained for the test set classification for both varieties. Specifically, the classification accuracy of *Pentacebas* increased from 97.8% to 98.9%. *Soleta* achieved 100% of classification accuracy. In general, the PLS-DA results improved slightly when performing variable selection. The variables removed not only did not provide useful information for varietal differentiation, but their presence increased the complexity of the multivariate calibration models.

Concerning the fresh leaf results, *Avijor* and *Guara* classes showed the same results as when using the entire spectra. The classification accuracy of *Isabelona* increased in the cross-validation compared to that of the previous model, but the result in test set validation was lower, 90.0% of classification accuracy in contrast to 92.2% obtained by using the entire spectra. *Soleta* also showed a greater classification accuracy in the cross-validation, but the same classification rate was obtained in both models. In the cross-validation, *Marta* and *Pentacebas* classes showed lower classification accuracy compared to the results obtained using the entire spectra. However, the classification accuracy of *Marta* in the test set validation was similar to that in the previous model. In contrast, a lower accuracy was found in the test set validation, 95.6% of classification accuracy by using the selected variables compared to 100% obtained by using all the variables. In general, the PLS-DA results with variable selection were slightly worse than the ones obtained using the whole spectra, which is in contrast with the results achieved in the model with dried-powdered leaves. This can be due to the water content in the fresh leaves, which produces broad spectral bands that can hide other spectral features and decrease the effectivity of the variable selection. The usefulness of variable selection depends then of the type of sample, and the shape of the spectra seems to have an especial influence. In the case of dried-powdered leaves, the raw spectra show more peaks than fresh leaves due to the absence of water, and this seems to favor the discrimination of the varieties.



**Figure 2.** Variables selected for the dried-powdered and fresh leaves models. (a) VIP score of the dried-powdered leaves, (b) regression vector of the dried-powdered leaves, (c) selectivity ratio of the dried-powdered leaves (d) VIP score of the fresh leaves, (e) Regression vector of the fresh leaves, and (f) Selectivity ratio of the fresh leaves.

## Conclusions

The results of this study demonstrated that NIR spectroscopy has the capability to discriminate *Prunus dulcis* varieties with a notable accuracy, up to 100%. The correct classification rate was higher by using dried-powdered leaves than that by using fresh leaves, which was mainly due to the presence of water in fresh leaves. In addition, the sample homogeneity in the dried-powdered leaves may improve the classification rates. However, three varieties reached 100% of classification accuracy in the fresh leaf model, which demonstrated that fresh leaves could also be used in classification studies. Moreover, selecting the most relevant variables allowed reducing the complexity of the data set and increased the accuracy of the model for dried-powdered leaves. In contrast, the classification rate in the model for fresh leaves was better using the whole spectra. This information could be used in other studies of vegetal species discrimination. The main limitations were the genetic proximity of the varieties and the water influence in the fresh samples.

This is the first study that represents an advance in the research and implementation of the NIRS technology in the nursery plant industry as a varietal discrimination tool. Future work will focus on new classification models in order to improve the results obtained in the present study and to develop a deep comparison with the biomolecular techniques.

Received: 12 September 2019; Accepted: 9 December 2019;

Published: 24 December 2019

## References

1. USDA. Available at, <https://www.usda.gov/> (Accessed: 10th May 2019).
2. United Nations. FAOSTAT. Available at, <http://www.fao.org/faostat/en/#home> (Accessed: 10th May 2019).
3. Socias i Company, R. & Felipe, A. J. 'Belona' and 'Soleta' Almonds. *HortScience* **42**, 704–706 (2007).
4. Romero, A., Vargas, F. J., Tous, J., Ninot, A. & Miarnau, X. New almond varieties from IRTA's breeding programme: (1) chemical composition. *in Acta Horticulturae* **912**, 477–484 (2011).
5. Gupta, P. K., Roy, J. K. & Prasad, M. Single nucleotide polymorphism: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plant. *Curr. Sci.* **80**, 524–535 (2001).
6. Jamshidi, B., Mohajerani, E. & Jamshidi, J. Developing a Vis/NIR spectroscopic system for fast and non-destructive pesticide residue monitoring in agricultural product. *Meas. J. Int. Meas. Confed.* **89**, 1–6 (2016).
7. Afandi, S. D. *et al.* Nitrogen Content Estimation of Rice Crop Based on Near Infrared (NIR) Reflectance Using Artificial Neural Network (ANN). *Procedia Environ. Sci.* **33**, 63–69 (2016).
8. Barbin, D. F. *et al.* Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. *Food Res. Int.* **61**, 23–32 (2014).
9. Giraud, A. *et al.* Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis. *Food Control* **99**, 137–145 (2019).
10. Vitale, R. *et al.* A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. *Chemom. Intell. Lab. Syst.* **121**, 90–99 (2013).
11. Wang, P. & Yu, Z. Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review. *J. Pharm. Anal.* **5**, 277–284 (2015).
12. Maimaitiyiming, M., Miller, A. J. & Ghulam, A. Discriminating Spectral Signatures Among and Within Two Closely Related Grapevine Species. *Photogramm. Eng. Remote Sens.* **82**, 51–62 (2016).
13. Chen, Y. *et al.* Study on discrimination of white tea and albino tea based on near-infrared spectroscopy and chemometrics. *J. Sci. Food Agric.* **94**, 1026–1033 (2014).
14. Shao, Y. *et al.* Discrimination of tomatoes bred by spaceflight mutagenesis using visible/near infrared spectroscopy and chemometrics. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **140**, 431–436 (2015).
15. Mees, C. *et al.* Identification of coffee leaves using FT-NIR spectroscopy and SIMCA. *Talanta* **177**, 4–11 (2018).
16. Curran, P. J., Dungan, J. L. & Peterson, D. L. Estimating the foliar biochemical concentration of leaves with reflectance spectrometry: Testing the Kokaly and Clark methodologies. *Remote Sens. Environ.* **76**, 349–359 (2001).
17. Pacumbaba, R. O. & Beyl, C. A. Changes in hyperspectral reflectance signatures of lettuce leaves in response to macronutrient deficiencies. *Adv. Sp. Res.* **48**, 32–42 (2011).
18. Pallardy, S. *Physiology of Woody Plants. Physiology of Woody Plants* (Elsevier, 2008). doi:10.1071/pc980272.
19. Borraz-Martínez, S., Boqué, R., Simó, J., Mestre, M. & Gras, A. Development of a methodology to analyze leaves from *Prunus dulcis* varieties using near infrared spectroscopy. *Talanta* **204**, 320–328 (2019).
20. Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**, 137–148 (1969).
21. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **43**, 772–777 (1989).
22. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
23. Rinnan, Å., Berg, Fvanden & Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends Anal. Chem.* **28**, 1201–1222 (2009).
24. Ballabio, D. & Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods* **5**, 3790–3798 (2013).
25. Wu, Z. *et al.* Rapid Determination of Process Variables of Chinese Rice Wine Using FT-NIR Spectroscopy and Efficient Wavelengths Selection Methods. *Food Anal. Methods* **8**, 1456–1467 (2015).
26. Galindo-Prieto, B., Eriksson, L. & Trygg, J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J. Chemom.* **28**, 623–632 (2014).
27. Vip - Eigenvector Research Documentation Wiki. Available at, <https://www.wiki.eigenvector.com/index.php?title=Vip> (Accessed: 15th November 2019)
28. Sratio - Eigenvector Research Documentation Wiki. Available at, <https://www.wiki.eigenvector.com/index.php?title=Sratio> (Accessed: 15th November 2019)
29. Rajalahti, T. *et al.* Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom. Intell. Lab. Syst.* **95**, 35–48 (2009).
30. Mosteller, F. *Data Analysis and Regression: A Second Course in Statistics (Addison-Wesley Series in Behavioral Science)*. Addison-Wesley Series in Behavioral Science: Quantitative Methods (Addison-Wesley Series in Behavioral Science, 1977).
31. Teófilo, R. F., Martins, J. P. A. & Ferreira, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemom.* **23**, 32–48 (2009).
32. Wang, Y. Z., Dong, W. Y. & Kouba, A. J. Fast Discrimination of Bamboo Species Using VIS/NIR Spectroscopy. *J. Appl. Spectrosc.* **83**, 826–831 (2016).
33. Páscoa, R. N. M. J., Lopo, M., Teixeira dos Santos, C. A., Graça, A. R. & Lopes, J. A. Exploratory study on vineyards soil mapping by visible/near-infrared spectroscopy of grapevine leaves. *Comput. Electron. Agric.* **127**, 15–25 (2016).
34. Stuart, B. H. *Infrared Spectroscopy: Fundamentals and Applications. Infrared Spectroscopy: Fundamentals and Applications*, <https://doi.org/10.1002/0470011149> (John Wiley & Sons, Ltd, 2005).

## Acknowledgements

The authors thank to Thermo Scientific for temporary assignment of the NIR equipment. In addition, the authors thank the financial support of the Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia (project DI-COF 2017) and the Spanish Ministry of Economy and Competitiveness (project AGL2015-70106-R, AEI/FEDER, UE).

## Author contributions

Sergio Borraz-Martínez carried out the entire experimental process and performed the statistical models together with Ricard Boqué. Moreover, he wrote the main manuscript text including the references and all the figures and tables. Joan Simó and Anna Gras helped in the experimental design and in the writing of the article. Also, they helped in the revision of the manuscript before its submission. Mariàngela Mestre selected the almond tree varieties and helped in their study. In addition, she helped in the sampling process and contributed an agronomic point of view. Ricard Boqué contributed in the development of the models shown in the present article as an expert in analytical chemistry.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.B.-M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019



**Capítulo 3. Diferenciación  
varietal y control del estado  
fisiológico mediante cámara  
hiperespectral**





# Diferenciación varietal y control del estado fisiológico mediante cámara hiperespectral

En este capítulo se muestran los resultados obtenidos a partir del trabajo realizado con una cámara hiperespectral, en colaboración con los compañeros de Agromillora Sur (Chile). La cámara hiperespectral empleada fue adquirida por Agromillora Sur con el objetivo de desarrollar una serie de proyectos. El interés mutuo por esta técnica dio lugar a una estrecha colaboración que culminó con una estancia en Chile, durante la cual se realizaron la mayor parte de los ensayos que se muestran en este capítulo.

## 1. Introducción

La necesidad de aumentar la eficiencia y la productividad de los cultivos está generando que surja un gran interés, en el sector agrícola, por la aplicación de nuevas tecnologías que permitan monitorizar y extraer información que ayude en la gestión de los cultivos. En este sentido, una de las nuevas corrientes que más está creciendo es la agricultura de precisión, dentro de la cual se emplean herramientas tan diversas como vehículos aéreos no tripulados [1,2], satélites [3,4], diversos tipos de sensores [5,6] y cámaras espectrales [7,8], entre otros. Además de la utilización de la inteligencia artificial para el análisis de los datos [9,10].

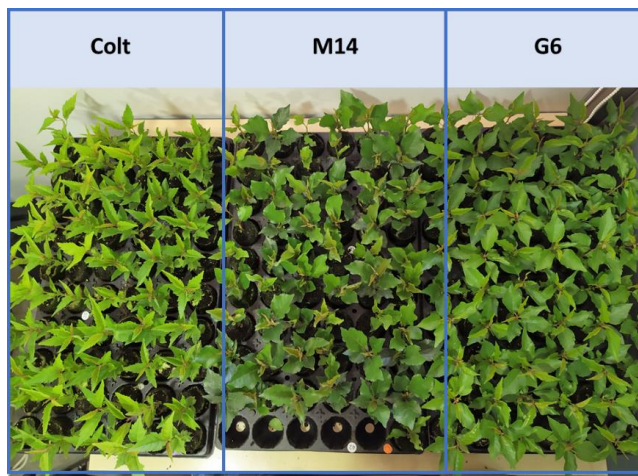
En este contexto, se planteó la utilización de una cámara espectral para afrontar los objetivos del presente estudio. Con dicha finalidad, se diseñaron y se llevaron a cabo una serie de ensayos dirigidos al estudio de la discriminación varietal y el estado fisiológico de plantas de vivero. Al desarrollarse esta investigación en las instalaciones de Agromillora Sur, no se pudo disponer de las variedades de almendro de interés. Por ello, estas fueron reemplazadas por tres variedades de portainjerto de cerezo, como son *Colt*, *Gisela 6 (G6)* y *Maxma 14 (M14)*. Los tres portainjertos escogidos son suficientemente similares como para ser confundidos por ojos poco entrenados,

aunque no presentan la misma proximidad genética encontrada en las variedades de almendro. No obstante, fueron consideradas como buenos especímenes de ensayo dada su importancia para Agromillora Sur.

## 2. Material y Métodos

### 2.1 Ensayos realizados

1. **Estudio de diferenciación varietal:** Para este estudio se emplearon 48 portainjertos de cerezo por cada una de las tres variedades estudiadas, *Colt*, *Gisela 6* y *Maxma 14*. En total se analizaron 144 plantas. Todas ellas presentaban un tamaño comprendido entre los 6 y los 10 cm de altura (**Figura 1**).



**Figura 1.** Imagen de portainjertos de cerezo de las variedades Colt, Maxma 14 (M14) y Gisela 6 (G6).

2. **Estudio sobre la detección de deficiencias nutricionales:** Para este estudio se separaron de la línea de producción 26 plantas de la variedad Colt, las cuales fueron privadas de la fertiirrigación habitual durante un período de 4 semanas. Durante este período, únicamente se fueron regando con agua, siguiendo el ciclo normal de riego de la línea de producción. De esta forma, se indujo la aparición de deficiencias nutricionales en estas 26 plantas. A la hora del análisis, los

espectros obtenidos de las plantas con deficiencias nutricionales fueron comparados con los espectros de 26 plantas de Colt que sí habían seguido el ciclo normal de fertirrigación, las cuales se consideraron como plantas control. Todas ellas presentaban un tamaño comprendido entre los 10 y los 15 cm de altura. La **Figura 2**, muestra una imagen de las plantas empleadas en el ensayo de detección de deficiencias nutricionales.



**Figura 2.** Portainjertos de la variedad Colt empleados en el ensayo de detección de deficiencias nutricionales.

- 3. Detección de la presencia de *Pseudomonas syringae*:** Para estudiar la detección precoz de *Pseudomonas*, se procedió a la inoculación de *Pseudomonas syringae* se realizó a través de un método denominado infiltración. Este método consistió en inocular la bacteria mediante el uso de una jeringuilla sin aguja. Para ello la jeringuilla se llenó previamente con 1 ml de cultivo bacteriano líquido con una concentración de  $10^6$  Unidades Formadoras de Colonias (UFC). A continuación, se colocó la punta de la jeringuilla en contacto con la cara adaxial del limbo de la hoja y se ejerció presión del líquido sobre la misma mediante el embolo de la jeringuilla. Este proceso se repitió sobre varios puntos de una misma hoja en tres hojas distintas por planta. La infiltración se realizó con el máximo cuidado posible para evitar generar daños en el tejido foliar. Una vez realizada la inoculación, las

plantas fueron analizadas cada 2 días, realizándose un total de 11 mediciones. Para este estudio se emplearon 50 plantas de la variedad de portainjerto de cerezo Colt, 25 plantas fueron inoculadas y las otras 25 plantas restantes fueron empleadas como control. Las plantas control fueron sometidas al mismo proceso de infiltración, con la diferencia de que se empleó como solución de inoculación agua esterilizada en autoclave. Por último, resaltar que el presente ensayo de detección de *Pseudomonas syringae* se enfocó como un estudio preliminar. Dicho estudio fue realizado por mis compañeros de Agromillora Sur, junto con mi participación en el diseño experimental, asesoramiento en la adquisición de datos y en el análisis de los mismos.

## 2.2 Cámara Hiperespectral y adquisición de datos

Para la realización del estudio se empleó una cámara hiperespectral de tipo *push-broom* OCI Hyperspectral Camera (BaySpec, San José, CA, EUA). Las muestras fueron analizadas en un rango espectral comprendido entre los 400 y los 1000 nm (25000-10000  $\text{cm}^{-1}$ ), el cual incluye toda la región del visible (400-780 nm) y una parte del infrarrojo cercano (780-1000 nm). La cámara hiperespectral se situó en un soporte regulable en altura y las muestras se desplazaron manualmente por debajo de la cámara, de esta forma, se obtuvieron imágenes apicales de los portainjertos. Al soporte se le añadieron tres luces halógenas para estandarizar la fuente de luz. Se escogieron las luces halógenas porque tienen un espectro de emisión estable a lo largo de todo el espectro de estudio. La cámara hiperespectral fue operada a través de un ordenador al cual estaba conectada mediante dos puertos USB 3.0. La adquisición de datos se realizó en una sala bajo condiciones de luz controladas.

## 2.3 Tratamiento de los datos

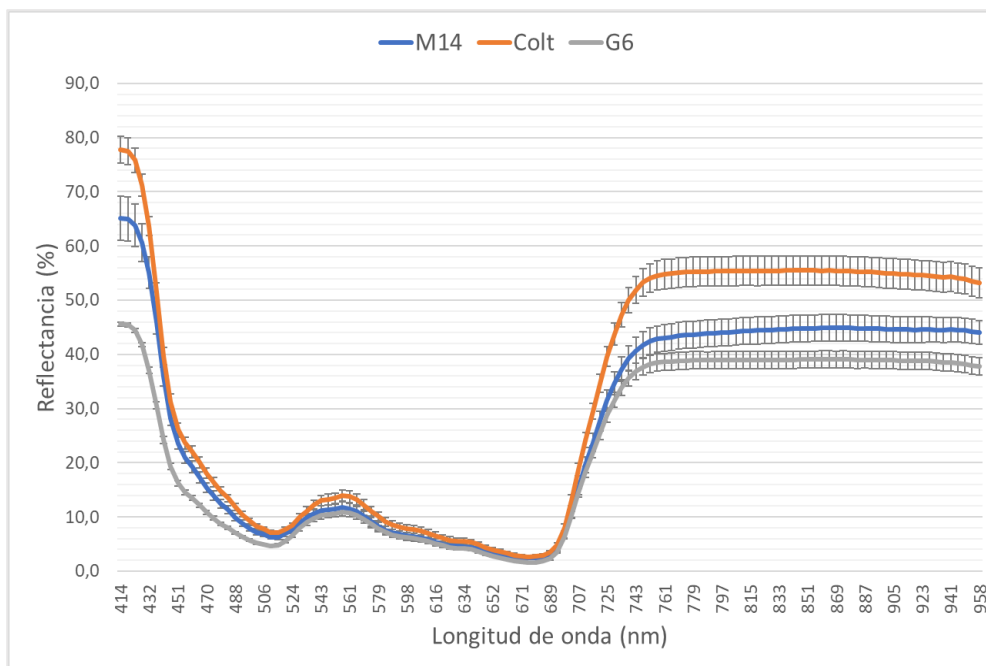
Previamente a la construcción de los modelos, los datos espectrales fueron pretratados matemáticamente aplicando un autoescalado. A continuación, los datos

se separaron aleatoriamente en dos subconjuntos. El conjunto de entrenamiento, compuesto por el 70 % de las muestras, fue utilizado para la calibración del modelo. Por otro lado, el conjunto de test, constituido por el 30 % de las muestras, se empleó para validar el modelo. Ambos grupos presentaban una distribución proporcional de las clases para evitar una representación no balanceada.

### 3. Resultados y Discusión

#### 3.1 Resultados del estudio sobre la diferenciación varietal

Los resultados del estudio de las huellas espectrales de las tres variedades de portainjerto de cerezo, *Maxma 14 (M14)*, *Gisela 6 (G6)* y *Colt*, se muestran en el gráfico de la **Figura 3**.



**Figura 3.** Espectros de reflectancia promedio de las variedades de portainjerto de cerezo Colt (naranja), Gisela 6 (G6) (gris) y Maxma 14 (M14) (azul). Se incluye un intervalo de confianza del 95% de la media (Estudio de la distribución normal realizado).

Se observaron diferencias significativas en dos regiones del espectro. Estas corresponden a las regiones comprendidas entre 414 a 430 nm (primera parte de la región del azul), y la región de 797 nm a 958 nm (infrarrojo cercano). También hay una amplia extensión donde no hay diferencias significativas entre los espectros promedio de las tres variedades, la cual corresponde a la región entre los 506 y 725 nm. Esta región se amplía si consideramos los solapamientos entre las variedades Colt y M14 en la región del azul, y entre M14 y G6 en la región del borde rojo (*red edge*, 680-730 nm). Toda esta región de solapamiento coincide con las longitudes de onda en las que se produce absorción de radiación electromagnética por parte de los fotosistemas de las plantas, como son la banda del azul (440-500 nm) y, especialmente, en la banda del rojo (600-700 nm). Esto es debido a que la absorción de luz por parte de los fotosistemas es un proceso esencial en la fisiología vegetal, lo que favorece a que sea un carácter muy conservado [11].

Por lo que respecta al infrarrojo cercano, se trata de una longitud de onda que no aporta suficiente energía para la fotosíntesis, pero al mismo tiempo, causa un perjudicial incremento de la temperatura en los fotosistemas debido a la gran absorbancia de esta radiación por parte de las moléculas de agua [12] (incremento del estado energético vibracional). Por esta razón, los organismos vegetales presentan una alta capacidad de reflexión de la radiación infrarroja como mecanismo de protección, lo que explica la gran reflectancia espectral observada en esta región. Las diferencias entre los espectros promedio de las tres variedades se asocian a la diferente capacidad de cada una de ellas para reflejar la radiación del infrarrojo cercano.

Como se puede observar en la [Figura 3](#), la región comprendida entre los 414 y los 430 nm se localiza en un extremo del rango de detección de la cámara, además está constituida por tan solo cuatro variables (mediciones). Ambos aspectos podrían afectar a su eficacia como región de interés para la diferenciación de las variedades de portainjerto.

Se construyó un modelo de clasificación *Partial least square discriminant analysis* (PLS-DA) con la finalidad de evaluar y validar las diferencias observadas entre los espectros promedio de las tres variedades. Los resultados del modelo PLS-DA se muestran en la **Tabla 1**.

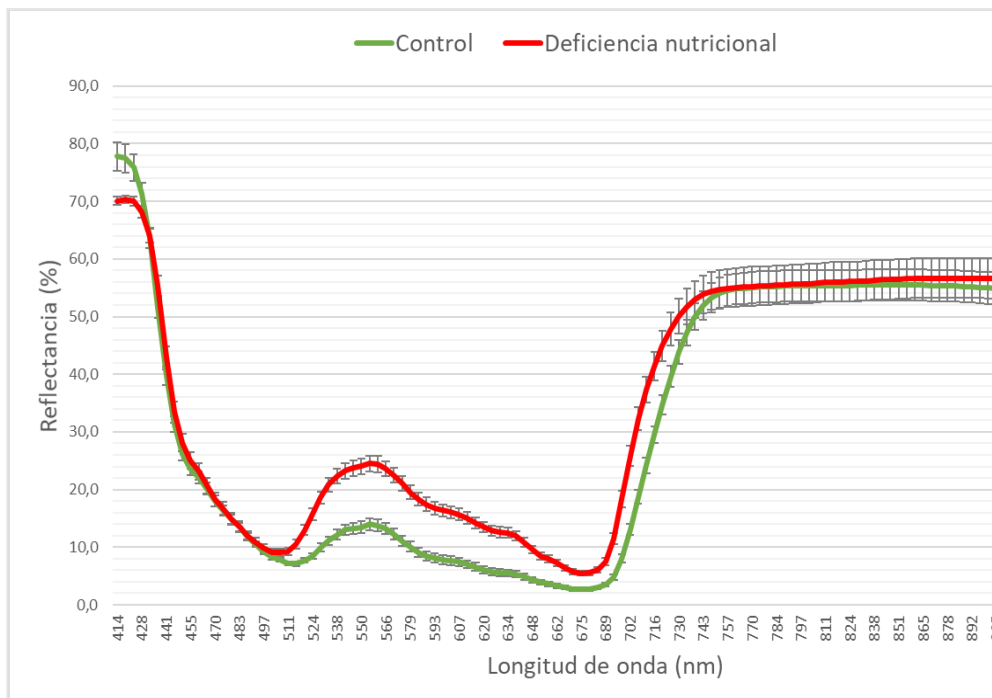
Se obtuvo una exactitud en la clasificación entre 97.7 y 100 % dependiendo de la clase. La variedad mejor identificada fue *Gisela 6*, siendo la única que alcanzó un 100 % de exactitud en la validación del test. Dicho resultado se vio favorecido por una menor variabilidad de las muestras que componían la clase *Gisela 6* respecto a las otras dos clases. Este aspecto es observable a través de los intervalos de confianza mostrados en la **Figura 3**. Por otro lado, se puede extraer del análisis de la sensibilidad y la especificidad, que la tasa de error en las clases *Colt* y *Maxma 14* se debe al error en la asignación de alguna muestra de la clase *Colt* en a clase *Maxma 14*. De esta forma, los resultados de modelo PLS-DA corroboran lo observado gráficamente en la **Figura 3**.

**Tabla 1.** Resultados del modelo PLS-DA para la diferenciación de tres variedades de portainjertos de cerezo.

<i>Variedades</i>	<i>Conjunto de datos</i>	<i>Sensibilidad</i>	<i>Especificidad</i>	<i>Exactitud</i>
<i>Colt</i>	Cross-validation	0.941	0.955	95.0 %
	<b>Test set Validation</b>	<b>0.929</b>	<b>1.000</b>	<b>97.7 %</b>
<i>Gisela 6</i>	Cross-validation	1.000	0.971	98.0 %
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100 %</b>
<i>Maxma 14</i>	Cross-validation	0.853	0.970	93.1 %
	<b>Test set Validation</b>	<b>1.000</b>	<b>0.966</b>	<b>97.7 %</b>

### 3.2 Resultados del estudio de deficiencia nutricional

La **Figura 4** muestra los espectros promedio de dos conjuntos de portainjertos de cerezo de la variedad Colt, uno de ellos compuesto por plantas con deficiencia nutricional y el otro compuesto por plantas sin deficiencia nutricional, al cual se le denominó control.



**Figura 4.** Espectros de reflectancia promedio de plantas de Colt control (verde) y plantas de Colt con deficiencia nutricional (roja). Se incluye un intervalo de confianza del 95% de la media (Estudio de la distribución normal realizado).

Se observó una región espectral, comprendida entre los 500 y los 720 nm, en la cual existen diferencias significativas entre los espectros promedio de las plantas control y las plantas con deficiencia nutricional. Esta región incluye las bandas del azul, el verde y el rojo, además de la zona de transición entre esta última y el NIR, llamada borde rojo (*red edge*). En la región espectral descrita son claves los pigmentos que intervienen en la fotosíntesis, especialmente la clorofila. Por lo tanto,



los resultados indican que la deficiencia nutricional afectó a la producción de clorofila, reduciéndose la concentración de esta en las hojas.

Por otro lado, se construyó un modelo PLS-DA con el fin de evaluar mejor la posibilidad de diferenciar ambos grupos. Los resultados de dicho modelo se muestran en la **Tabla 2**.

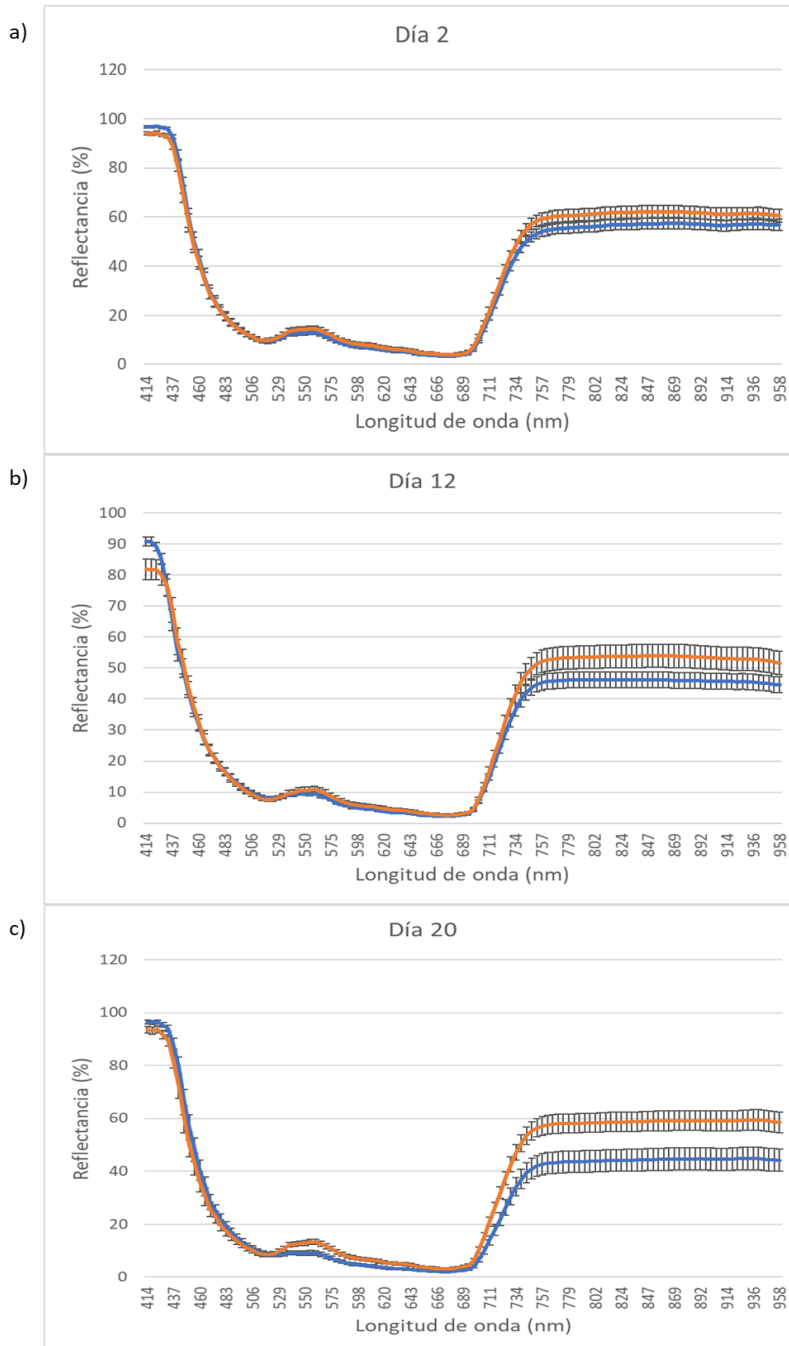
**Tabla 2.** Resultado del modelo PLS-DA para el estudio de deficiencias nutricionales

<i>Clases</i>	<i>Conjunto de datos</i>	<i>Sensibilidad</i>	<i>especificidad</i>	<i>Exactitud</i>
<i>Control</i>	Cross-validation	1.000	1.000	100 %
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100 %</b>
<i>Deficiencia</i>	Cross-validation	1.000	1.000	100 %
	<b>Test set Validation</b>	<b>1.000</b>	<b>1.000</b>	<b>100 %</b>

Tal como se observa en la **Tabla 2**, se logró diferenciar perfectamente ambos grupos de plantas, alcanzándose una exactitud en la clasificación del 100%. Esto indica, que es posible identificar la presencia de deficiencias nutricionales a partir de la huella espectral, pudiendo ser clasificadas las plantas según la presencia o ausencia de dichas deficiencias.

### 3.3 Resultados del estudio de detección de *Pseudomonas syringae*

Las plantas infectadas con *Pseudomonas syringae* y las del grupo control fueron monitorizadas durante 22 días. La **Figura 5** muestra los resultados de tres de las once mediciones realizadas a lo largo del período de monitorización. Se puede observar como el espectro promedio del grupo control se mantuvo sin grandes cambios en las sucesivas mediciones. En cambio, la huella espectral promedio del conjunto de plantas inoculadas con *Pseudomonas syringae*, sí que experimentó una variación.



**Figura 5.** Espectros de reflectancia promedios del conjunto de plantas infectadas con *Pseudomonas syringae* y el conjunto de plantas control (naranja) en tres días diferentes: a) 2 días después de la inoculación b) 12 días después de la inoculación; c) 20 días después de la inoculación. Se incluye un intervalo de confianza del 95% de la media (Estudio de la distribución normal realizado).

En la primera medición, correspondiente a dos días después de la inoculación (**Figura 5, a**), el espectro promedio de ambos grupos es muy similar a lo largo de todo el espectro analizado. En cambio, al cabo de 12 días de la inoculación (**Figura 5, b**), ya se pudieron apreciar ciertos cambios, especialmente en la región correspondiente al infrarrojo cercano. En dicha región se observó una ligera caída de la reflectancia en el espectro promedio del conjunto de plantas inoculadas con *Pseudomonas*, mientras que el espectro del grupo control se mantuvo estable entre el 50-60% de reflectancia. Esta tendencia de reducción de la reflectancia en el infrarrojo cercano en el grupo de plantas con *Pseudomonas* se hizo aún más notable a los 20 días de la inoculación (**Figura 5, c**), donde el porcentaje de radiación reflejada se situó cerca del 40%. En este punto, se observaron diferencias significativas entre los espectros promedio de los dos grupos. Además, en los espectros promedios adquiridos el día 20, se pudo percibir una reducción de la reflectancia en la banda del verde (entre 529 y 575 nm) en el grupo inoculado con bacteria, junto con un cambio en la región del borde rojo. A pesar de estos cambios espectrales, es importante remarcar que no se constató la aparición de síntomas observables a simple vista en las plantas inoculadas con *Pseudomonas syringae* a lo largo del ensayo.

Con la finalidad de evaluar la posibilidad de detección de portainjertos infectados con *Pseudomonas syringae* se realizó un modelo de clasificación PLS-DA empleando los datos del día 20. Los resultados del mismo se muestran en la **Tabla 3**.

Se logró una exactitud del 93,3% en la clasificación del conjunto de test. Este resultado aun fue algo superior en la validación cruzada, en la cual se alcanzó un 94.3% de exactitud. Centrándonos en los parámetros de sensibilidad y especificidad en la clasificación del conjunto de test, estos nos indican que el error se debe únicamente a que una muestra del grupo control fue erróneamente asignada a la clase *Inoculadas con ssp*. Debido al limitado número de muestras del ensayo, 25 portainjertos por clase, el error en la clasificación de tan sólo una de ellas redujo en más de un 6% la exactitud del modelo.

**Tabla 3.** Resultados modelo PLS-DA para la detección de portainjertos inoculados con *Pseudomonas syringae* (ssp).

<i>Clases</i>	<i>Conjunto de datos</i>	<i>Sensibilidad</i>	<i>especificidad</i>	<i>Exactitud</i>
<i>Inoculadas con Pseudomonas syringae</i>	Cross-validation	0.944	0.941	94.3 %
	<b>Test set Validation</b>	<b>1.000</b>	<b>0.875</b>	<b>93.3%</b>
<i>Control</i>	Cross-validation	0.941	0.944	94.3 %
	<b>Test set Validation</b>	<b>0.875</b>	<b>1.000</b>	<b>93.3 %</b>

Los resultados del modelo PLS-DA, junto con los gráficos de los espectros promedios, hacen pensar en la viabilidad de la detección de *Pseudomonas syringae* mediante cámaras espectrales. A pesar de ello, será necesario realizar más estudios, incrementando el tamaño muestral, con el fin de lograr una validación robusta de la detección prematura y asintomática de *Pseudomonas syringae* en portainjertos de cerezo.

#### 4. Conclusiones

Se realizaron tres estudios mediante el uso de una cámara hiperespectral, empleando como material de estudio tres variedades de portainjerto de cerezo.

Mediante el primer ensayo, se lograron diferenciar las variedades *Colt*, *Gisela 6* y *Maxma 14* con una alta exactitud en la clasificación comprendida entre el 97.7 y el 100%. Demostrándose de esta forma, la eficacia de la cámara hiperespectral para la diferenciación de variedades de plantas de vivero.

Por otro lado, se observó que los cambios en la huella espectral de las plantas derivados de una deficiencia nutricional, pueden ser detectados mediante el uso de una cámara hiperespectral. Los datos capturados por la cámara, fueron empleados para la construcción de un modelo matemático capaz de diferenciar con una exactitud del 100% plantas sanas de plantas con deficiencias nutricionales.

Por último, se logró detectar la presencia de *Pseudomonas syringae* en plantas inoculadas previamente con la bacteria, pero sin presencia de síntomas. Este resultado, suponen un avance en la detección prematura de *Pseudomonas syringae* en plantas de vivero.

Los resultados alcanzados en este estudio demuestran el potencial de la cámara hiperespectral como herramienta de control y monitorización de las plantas de vivero.

## Referencias

- [1] V. Puri, A. Nayyar, L. Raja, Agriculture drones: A modern breakthrough in precision agriculture, *J. Stat. Manag. Syst.* 20 (2017) 507–518. <https://doi.org/10.1080/09720510.2017.1395171>.
- [2] D.C. Tsouros, S. Bibi, P.G. Sarigiannidis, A review on UAV-based applications for precision agriculture, *Inf.* 10 (2019) 349. <https://doi.org/10.3390/info10110349>.
- [3] D. Murugan, A. Garg, T. Ahmed, D. Singh, Fusion of drone and satellite data for precision agriculture monitoring, in: 11th Int. Conf. Ind. Inf. Syst. ICIIS 2016 - Conf. Proc., Institute of Electrical and Electronics Engineers Inc., 2016: pp. 910–914. <https://doi.org/10.1109/ICIINFS.2016.8263068>.
- [4] D. Murugan, A. Garg, D. Singh, Development of an Adaptive Approach for Precision Agriculture Monitoring with Drone and Satellite Data, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 10 (2017) 5322–5328. <https://doi.org/10.1109/JSTARS.2017.2746185>.
- [5] S.A. Kumar, P. Ilango, The Impact of Wireless Sensor Network in the Field of Precision Agriculture: A Review, *Wirel. Pers. Commun.* 98 (2018) 685–698. <https://doi.org/10.1007/s11277-017-4890-z>.

- [6] D. Thakur, Y. Kumar, A. Kumar, P.K. Singh, Applicability of Wireless Sensor Networks in Precision Agriculture: A Review, *Wirel. Pers. Commun.* 107 (2019) 471–512. <https://doi.org/10.1007/s11277-019-06285-2>.
- [7] B. Lu, P.D. Dao, J. Liu, Y. He, J. Shang, Recent advances of hyperspectral imaging technology and applications in agriculture, *Remote Sens.* 12 (2020) 2659. <https://doi.org/10.3390/RS12162659>.
- [8] L. Deng, Z. Mao, X. Li, Z. Hu, F. Duan, Y. Yan, UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras, *ISPRS J. Photogramm. Remote Sens.* 146 (2018) 124–136. <https://doi.org/10.1016/j.isprsjprs.2018.09.008>.
- [9] T. van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: A systematic literature review, *Comput. Electron. Agric.* 177 (2020) 105709. <https://doi.org/10.1016/j.compag.2020.105709>.
- [10] A. Chlingaryan, S. Sukkarieh, B. Whelan, Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, *Comput. Electron. Agric.* 151 (2018) 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>.
- [11] R.S. Gupta, Origin and Spread of Photosynthesis Based upon Conserved Sequence Features in Key Bacteriochlorophyll Biosynthesis Proteins, *Mol. Biol. Evol.* 29 (2012) 3397–3412. <https://doi.org/10.1093/molbev/mss145>.
- [12] R. Zwigelaar, A review of spectral properties of plants and their potential use for crop/weed discrimination in row-crops, *Crop Prot.* 17 (1998) 189–206. [https://doi.org/10.1016/S0261-2194\(98\)00009-X](https://doi.org/10.1016/S0261-2194(98)00009-X).

**Capítulo 4. Varietal Quality  
Control in the Nursery Plant  
Industry Using Computer Vision  
and Deep Learning Techniques**





# Varietal quality control in the nursery plant industry using computer vision and deep learning techniques

Sergio Borraz-Martínez<sup>1,2</sup>  | Francesc Tarrés<sup>5</sup>  | Ricard Boqué<sup>3</sup>  |  
Mariàngela Mestre<sup>2</sup> | Joan Simó<sup>1,4</sup>  | Anna Gras<sup>1</sup> 

<sup>1</sup>Department of Agri-Food Engineering and Biotechnology, Universitat Politècnica de Catalunya, Campus Baix Llobregat, Esteve Terrades 8, Castelldefels, 08860, Spain

<sup>2</sup>Center of Initial Materials, Agromillora Iberia S.L.U, Ctra. BV-2247 km. 3, Sant Sadurní d'Anoia, 08770, Spain

<sup>3</sup>Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, Campus Sescelades, Tarragona, 43007, Spain

<sup>4</sup>Fundació Miquel Agustí, Campus Baix Llobregat, Esteve Terrades 8, Castelldefels, 08860, Spain

<sup>5</sup>Signal Theory and Communications Department, UPC Barcelona Tech, Barcelona, Spain

## Correspondence

Sergio Borraz-Martínez, Department of Agri-Food Engineering and Biotechnology, Universitat Politècnica de Catalunya, Campus Baix Llobregat, Esteve Terrades 8, Castelldefels 08860, Spain.  
Email: sborrazma@gmail.com;  
sergio.borraz@upc.edu

## Funding information

Spanish Ministry of Economy and Competitiveness, Grant/Award Numbers: AGL2015-70106-R, AEI/FEDER, UE, AEI/FEDER, UE, AGL2015-70106-R; Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia, Grant/Award Number: DI-COF 2017

## Abstract

Computer vision coupled to deep learning is a promising technique with multiple applications in the industry. In this work, the potential of this technique has been assessed in the classification of two varieties of almond trees (*Prunus dulcis*), *Soleta* and *Pentacebas*. For that, a convolutional neural network named VGG16 was used. The most appropriate configuration for model training was studied, which included the comparison between two different filling modes (*reflect* and *nearest*) in the data augmentation step, the evaluation of the batch size and the analysis of the image sizes. The robustness of the model was also checked, and information was obtained about how the model extracts the information from the images. The results showed that the *reflect* fill mode was more effective than the *nearest* one. The best results were obtained using batches with 30 and 40 images, with an image size of (224 × 224) pixels. The verification of the robustness proved the capability of the technique as a promising tool for plant varietal identification.

## KEYWORDS

computer vision, convolutional neural network, deep learning, nursery plant, varietal mixture

## 1 | INTRODUCTION

Computer vision coupled to deep learning (CV-DL) techniques are getting more popularity in different areas such as pedestrian detection and tracking,<sup>1</sup> human health,<sup>2</sup> and structural health monitoring in civil structures,<sup>3</sup> especially for their great capability of object recognition. This capability has been shown in the successive Large Scale Visual Recognition Challenge (LSVRC) competitions.<sup>4</sup> Many studies have been developed to solve different problems and overcome challenges in agro-food quality control,<sup>5,6</sup> precision agriculture,<sup>7,8</sup> control of fruits and vegetables,<sup>9,10</sup> plant disease detection,<sup>11,12</sup> and plant phenotyping.<sup>13,14</sup>

We have applied CV-DL to solve the problem of varietal discrimination in the nursery plant industry. During the production process of the plants, varietal mixtures can appear in the batches due to traceability errors. When this happens and the product is commercialized, customers complain, the company must pay a fine, and its corporative image is damaged. Nowadays, the varietal identification is usually performed by means molecular methods based on DNA analysis, specifically DNA amplification by the polymerase chain reaction (PCR)<sup>15</sup> together with analysis of genetic variations, such as single nucleotide polymorphism (SNP).<sup>16</sup> Despite these molecular techniques have an excellent accuracy, they are very expensive for routine analysis of a large number of samples. Other studies have focused on the implementation of near-infrared (NIR) spectroscopy as a tool for varietal control.<sup>17,18</sup>

In the literature, there are studies related to plant identification using CV.<sup>19–22</sup> In our case, the challenge is to discriminate between two almond trees (*Prunus dulcis*) varieties, *Soleta* and *Pentacebas*, which are morphologically very similar, genetically close, and almost indistinguishable for the human eye. Figure 1 shows images of both varieties.

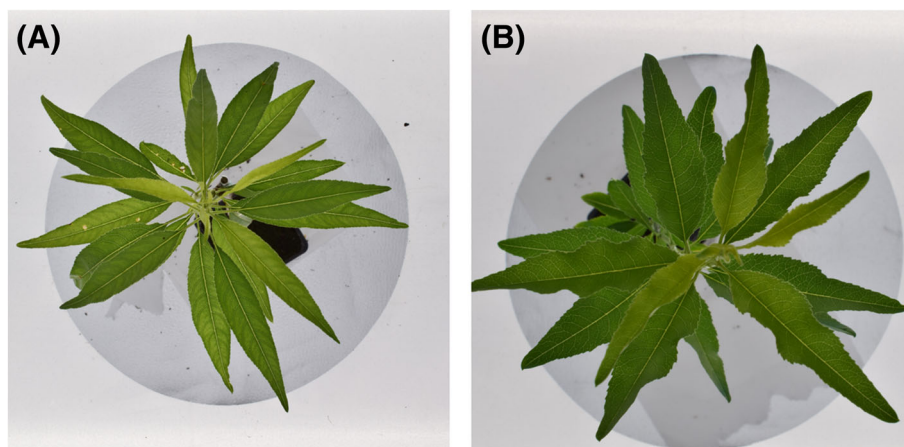
In other CV systems, the final classification relies on the features designed by the engineer. The programmer has to choose the features of the image that will be more representative to classify the leaves of the plant. The final quality of the system depends on this previous knowledge of the best features. But the designer does not know if the features applied are the best or another selection, or combination would produce better results. In this work, it was wanted to obtain a classification system without using any prior knowledge about the best-suited features. With DL, the programmer inputs data as well as the outputs expected from the data, and the system automatically searches the best features.<sup>23</sup> Therefore, CV-DL is more suitable to face complex problems, as the case of study.

For that, a database with images from the two varieties of interest was created, and a convolutional neural network (CNN) was used to build a classification model. The specific objectives of the study are (1) to study the training configuration of the neural network; (2) to explore what features of the plant is the model using in the modeling/training step; and (3) to check the robustness of the technique.

## 2 | MATERIAL AND METHODS

### 2.1 | Image acquisition

The red–green–blue (RGB) images used were acquired using a photo studio box and a Nikon D5300 camera. Two LED lamps (Color temperature 5500 K, color rendering index of 93%, and 13 000 lumens) were placed at the bottom of the box to achieve the correct illumination of the plant. The position of the light source was optimized in order to highlight



**FIGURE 1** Images of almond tree plants. (A) *Pentacebas* variety; (B) *Soleta* variety

the features of the leaves: the venation and the borders. To avoid pictures with too much light intensity or reflectance effects, a textile diffuser was placed between the light source and the canopy of the plant. The images were collected from the top of the plants, thus obtaining an apical view. Before image acquisition, the setup of the camera was adjusted (diaphragm opening =  $f/29$ , shutter speed =  $1/30$ , and International Organization for Standardization [ISO] sensitivity = 200). This parameter set was selected to produce a well-balanced histogram. In total, 824 images (one image per plant) were taken from two *P. dulcis* varieties, 412 images of *Soleta*, and 412 of *Pentacebas*. Table 1 shows more details about the number of images used in the study.

## 2.2 | Image preparation

The rectangular raw images obtained from the camera (image size of  $6000 \times 4000$  pixels) were cropped to an image size range between  $4000 \times 4000$  and  $2500 \times 2500$  pixels to obtain squared images, which constituted the dataset used in this work. The model was adjusted to be fed with images of  $224 \times 224$  pixels, so just before introducing them into the model, their resolution was reduced automatically to  $224 \times 224$  pixels. Moreover, the pixel component levels were normalized ( $1.0/255$ ) in order to adjust their values between 0 and 1. These two last modifications are indispensable to decrease the computation time.

The whole dataset was randomly split into a training set (75% of images), a validation set (15% of images), and a test set (10% of images). Table 1 shows more details about the number of samples of the three datasets.

## 2.3 | Convolutional neural network

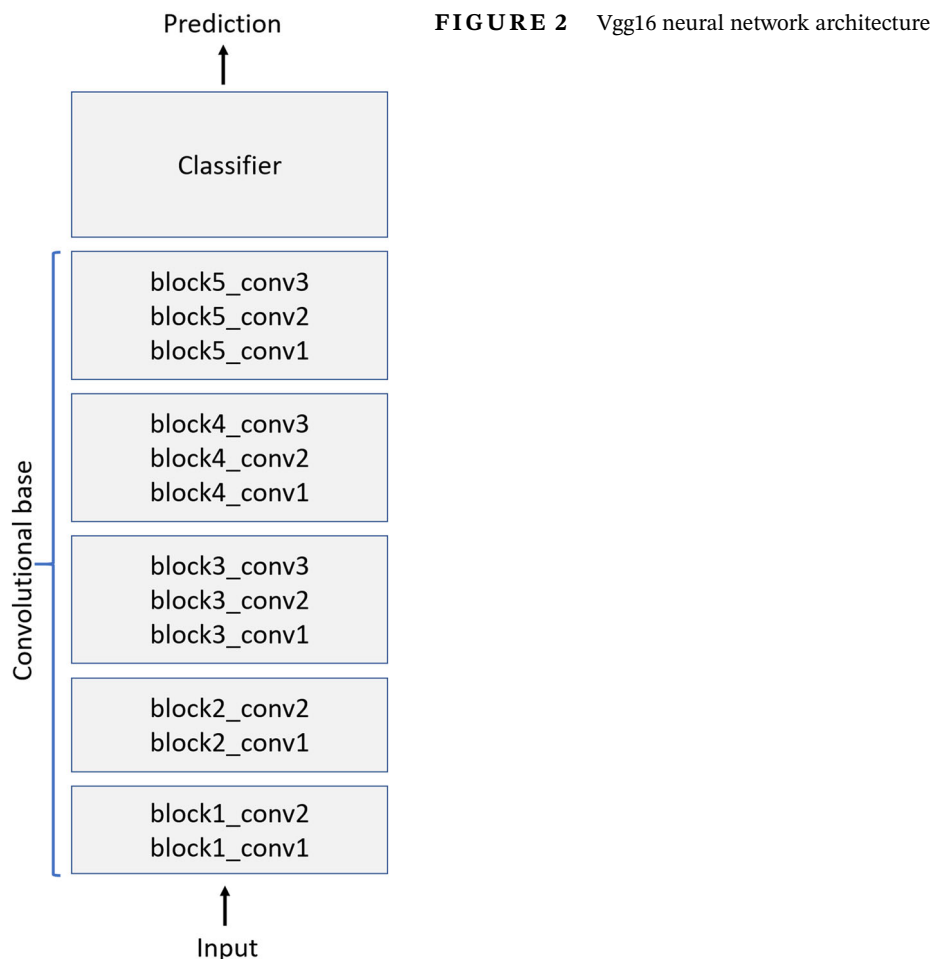
A CNN is a type of network architecture commonly used in image analysis. It is characterized by having a reduced number of tunable parameters, because each neuron of the convolutional layer represents a small kernel that is applied to the previous layer using a convolution operator. Therefore, not all neurons in one layer are connected to all neurons in the previous layer. The neurons of the CNN form three dimensional layers, with height, width, and depth, also named channels.<sup>24</sup> Next to the convolutional layers, there are pooling layers that reduce the resolution of the output features in higher-level layers. In this way, higher layers have a global view of the image contents while lower layers are specialized in detecting local features.

The Vgg16 CNN<sup>25</sup> was used to perform the study. Vgg16 was developed by the Geometry Group at the University of Oxford to obtain state of the art results in the ILSVRC-2014 competition.<sup>4,26</sup> The convolutional base of the CNN is constituted by 16 layers, which are ordered in five blocks. Moreover, after the convolutional base, the Vgg16 contains a classifier that is composed by three fully connected layers. Blocks 1 and 2 have two convolutional layers each one and 64 and 128 filters, respectively. Blocks 3–5 contain three convolutional layers and 256, 512, and 512 filters, respectively. Every convolutional layer use filters with a receptive field of  $3 \times 3$ . The convolution stride is  $1 \times 1$  pixel, and the spatial padding of the convolutional layer input is such that the spatial resolution is preserved after convolution.<sup>25</sup> In addition, one max-pooling layer performed over a  $2 \times 2$  pixel window is located at the end of each block. The classifier is composed of three fully connected layers. Figure 2 shows a scheme of the Vgg16 architecture.

The convolutional base of the Vgg16 neural network was downloaded pretrained (with the parameters adjusted) with ImageNet dataset<sup>27</sup> from TensorFlow.<sup>28</sup> ImageNet is formed by 1.4 million labeled images and 1000 different classes, and it is commonly used to pretrain CNNs in CV. Then, a new classifier was added, constituted by one flatten layer followed by two dense layers. The CNNs learn local features of the input image. For this reason, the purpose of these fully connected layers is to collect the local information and combine it through nonlinear filtering, which is learned

**TABLE 1** Details of the number of samples in the datasets used

	Percentage %	<i>Pentacebas</i>	<i>Soleta</i>	Overall dataset
Training set	75	309	309	618
Validation set	15	62	62	124
Test set	10	41	41	82
Total		412	412	824



from the training set. Usually, a minimum of two fully connected layers is used. The flatten layer allows transforming the 3-D tensor of the CNN to a 1-D tensor, which is required by the dense layer. The first dense layer included 256 neurons with rectified linear unit (ReLU) activation. This layer is necessary because it integrates all the local features learned by the last CNN layer, taking into account nonlinear combinations. Otherwise, without this inner layer, outputs would be a form of regression of the convolutional layers. Finally, the second dense layer contained one neuron with sigmoid activation. The output of the sigmoid will be an estimation of the confidence of having an input image of one or the other class.

During the training process, only the classifier and the convolutional layers of Block 5 were trained. The rest of parameters retained the values acquired from the training with the ImageNet dataset. We decided to train the Block 5 because this way, it is possible to improve the accuracy of the classification model. This method, which involves training the last CNN layer, is known as *Fine-tuning*.

## 2.4 | Data augmentation study

To avoid overfitting, a data augmentation technique was used. The overfitting occurs when the model fits very specific features of the training images that are not representative of the whole class. Thus, the accuracy in the classification of new images is reduced.

Data augmentation consists of obtaining multiple images from the original one. This is achieved by modifying the original image. The common modifications are rotation, shear, horizontal flip, zoom, and shift of the image. The goal is to increase the diversity of the training set, which helps to improve the robustness of the model and its prediction accuracy.

When using data augmentation, it is important to check the modified images. Misuse of data augmentation can lead to a reduction in the effectiveness of the model. An important aspect of data augmentation is what to do with pixels that

are empty due to image rotation or shifting. This can be dealt with using the argument *fill\_mode*. This argument defines the method for filling empty pixels. There are several strategies, the most common being *nearest* and *reflect*. The command *nearest* fills the pixels with the values of the contiguous pixels. In contrast, *reflect* generates a specular image. Both methods were compared to investigate which one generates less differences with the original images and thus leads to better results in the test.

Data augmentation was only applied to the training set because in the validation and test sets, it is preferable to use original images. In addition to the fill mode, other image modifications were applied: shear in a range of 0.2, zoom in a range of 0.4, horizontal flip, random rotation in a range of  $180^\circ$ , and width and height shift in a range of 0.2.

## 2.5 | Batch size and optimizer

Batch size is an important aspect to consider before starting to train a DL model. A batch is a set of input images that are passed through the network with fixed weights. The loss function is computed, and gradient backpropagation is applied to modify the weights for each batch. The steps per epoch (weight updates) come from the division of the number of images of the training/validation set by the batch size assigned. The validation step consists of the division of the number of images contained in the validation set by the batch size. Both concepts show the number of iterations by epoch. If the batch size is small, the number of iterations will be high, but each learning step will be performed with few images, which means little information. In contrast, if the batch size is large, the number of iterations will be low, but each learning step of the model will be robust because it will be based on many images.

We studied different batch sizes to evaluate what ratio between number of images per batch and number of iterations per epoch was the most suitable. The batch sizes used were 20 (30 steps per epoch/6 validation steps), 30 (20 steps per epoch/4 validation steps), and 40 (15 steps per epoch/3 validation steps) samples per batch. The number of epochs was defined based on the convergence between training and validation accuracies. Moreover, the progression of the classification error (loss value) along of the training process was taken into account. The models were trained in a range between 80 and 90 epochs, because overfitting was observed above 100 epochs.

## 2.6 | Comparison between image sizes

The Vgg16 CNN can be fed with two image sizes,  $150 \times 150$  or  $224 \times 224$  pixels. The advantage of using  $224 \times 224$  pixel images is their higher resolution. This allows keeping more fine features, which can provide useful information for the identification and the discrimination of the classes. However, this strategy has two main disadvantages; first, the number of operations to adjust the parameters of the convolutional base of the Vgg16 neural network increases; second, the number of parameters of the classifier also increases, which translates in longer computing times.

In this study,  $224 \times 224$  pixel images were used because we decided to prioritize resolution instead of time. However, we compared the performance when using images of  $224 \times 224$  and  $150 \times 150$  pixels. The dataset images, with a size range between  $4000 \times 4000$  and  $2500 \times 2500$  pixels, were automatically reduced to images with a resolution of  $150 \times 150$  pixels through adjusting the setup of the model. To perform the comparison, we applied the training setup resulting from the study of the previous sections.

## 2.7 | Model robustness assessment

The robustness of the model trained with the configuration mentioned above was studied. To perform this analysis, five models were built using different random splits of the images into training, validation, and test sets. The variability in the test accuracy obtained helped us to assess the robustness of the DL technique to classify the *Prunus* varieties.

## 2.8 | Visualization of the intermediate activations

It is interesting to know how the model performs the analysis of the images in order to understand how it works. To try to answer this question, an input (image) was introduced, and the output was observed in different convolutional

layers. The output is the activation of the layer, which provides information about how the input is decomposed into the different filters learned by the network.<sup>23</sup> By doing this, feature maps were created, which are composed of channels or filters. Each channel performs a specific analysis of the input, although all the filters in a layer perform a similar decomposition of the input image. This characteristic allows us to have an overview of the features extracted by each layer of the neural network. Finally, to plot the feature maps, a 2-D image was created with an independent plot of the contents of every channel of a layer of the CNN.

### 3 | RESULTS AND DISCUSSION

#### 3.1 | Data augmentation study

The images were analyzed by using the *nearest* and *reflect* fill modes. Several modified images were obtained using both methods. Figure 3 shows three examples from each method. It is important to remark that the images shown in Figure 3 are an example of the images that are feeding the neural network.

Regarding the *nearest* fill mode, despite the modified images showed an appearance very similar to the original ones, several anomalies or deformations were detected, which are highlighted with a circle in Figure 3. These distortions arose because some of the empty pixels were filled using the value of the leaf pixels, which resulted in an elongation of the original leaves. The presence of this deformed leaves in the training images represents a significant alteration of the original images.

Concerning the *reflect* fill mode, the modified images showed an alteration too, which is highlighted with a square in Figure 3. The *reflect* fill mode generated a deformation of the original images too but with the difference that created a specular image of the original one.

Table 2 shows the results of the DL models. Using the *nearest* fill mode, an accuracy average of 95.1% in the test set was obtained, whereas the *reflect* fill mode provided an average accuracy of 96.7%. When the results obtained using the different batch sizes, in all cases, better results were obtained with the *reflect* fill mode. This result confirms that image deformations caused by the *nearest* fill mode negatively affected the accuracy of the model. To understand better this result, it is important to remark that data augmentation is only used during the training of the model. Therefore, the

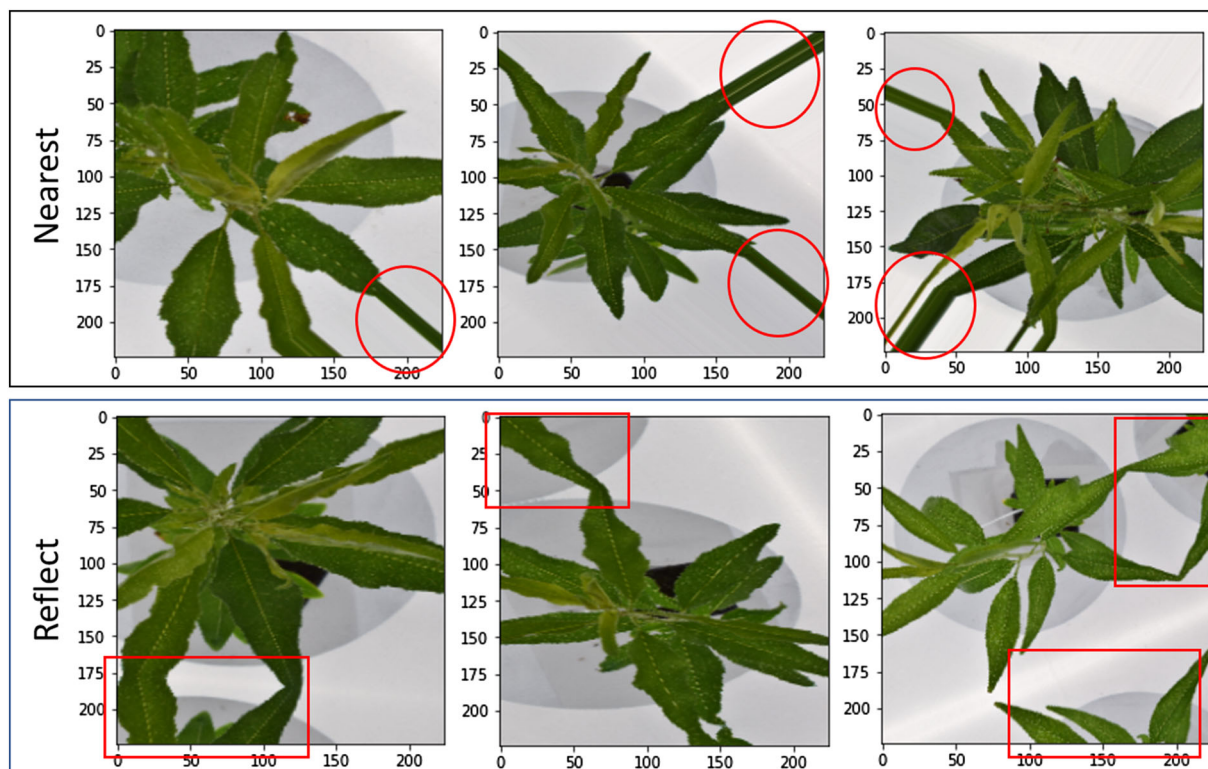


FIGURE 3 Study of the image modifications performed by *nearest* and *reflect* fill modes

**TABLE 2** Result obtained in data augmentation and batch size studies

Data augmentation	Batch size	Test accuracy
Nearest	20	93.9
	30	96.4
	40	95.1
Reflect	20	95.1
	30	<b>97.6</b>
	40	<b>97.6</b>

**TABLE 3** Comparison between images with  $150 \times 150$  and  $224 \times 224$  pixels

Size of the images	Test accuracy	Training time (s)
$150 \times 150$	95.1	<b>230</b>
$224 \times 224$	<b>97.6</b>	275

model will learn any alteration generated by data augmentation, even if the features are false. In addition, these deformations will not be present in the validation and test sets because data augmentation was not applied there. Thus, the *nearest* fill mode increased the gap between the training set and the test set.

### 3.2 | Batch size and optimizer study

Three different batch sizes, 20, 30, and 40, were studied. As it can be seen in Table 2, the best result of this comparison was achieved using the *reflect* fill mode together with a batch size of 30 and 40, reaching a 97.5% accuracy in the test set. In contrast, the worst result was obtained by applying the *nearest* fill mode together with a batch size of 20. Although a batch size of 20 allows a greater number of iterations, each learning step is performed based on a smaller number of images, and this affects the test accuracy. It seems more adequate to use a batch size of 30 or 40 images. Both provided the same result using the *reflect* fill mode, but the batch size of 30 achieved a better accuracy using the *nearest* fill mode.

### 3.3 | Comparison of image sizes

Table 3 shows the results of the comparison between the different image sizes. With  $150 \times 150$  pixel images, an accuracy of 95.1% was obtained in the test set, which is the same accuracy obtained using  $224 \times 224$  pixel images, but with a batch size of 20. This indicates that using images of  $150 \times 150$  pixels, it is possible to obtain results comparable with those obtained using images of  $224 \times 224$  pixels. However, a better comparison is performed with the results obtained using a batch size of 30. In this case, the model trained with  $224 \times 224$  pixel images provided a better result but took an average of 16.4% more computing time than the model trained with  $150 \times 150$  pixel images. When using  $224 \times 224$  pixel images, the Vgg16 had 13.5 million trainable parameters, while using  $150 \times 150$  pixel images the network had 9.1 million of trainable parameters, that is, a 32.6% less parameters. The increase in parameters is higher than the increase in training time. The differences in accuracy are not so great, but the time could be important depending on the requirements of the application.

### 3.4 | Assessment of model robustness

To assess the robustness of the model and discard the possibility that a single choice of the test set was favorable to obtain good results, six models were built with different training, validation, and test sets, which were randomly split. The results are shown in Table 4.

TABLE 4 Study of the robustness of the model

Models	Test accuracy
1	96.3
2	96.3
3	97.6
4	96.3
5	97.6
<b>Average</b>	<b>96.8</b>

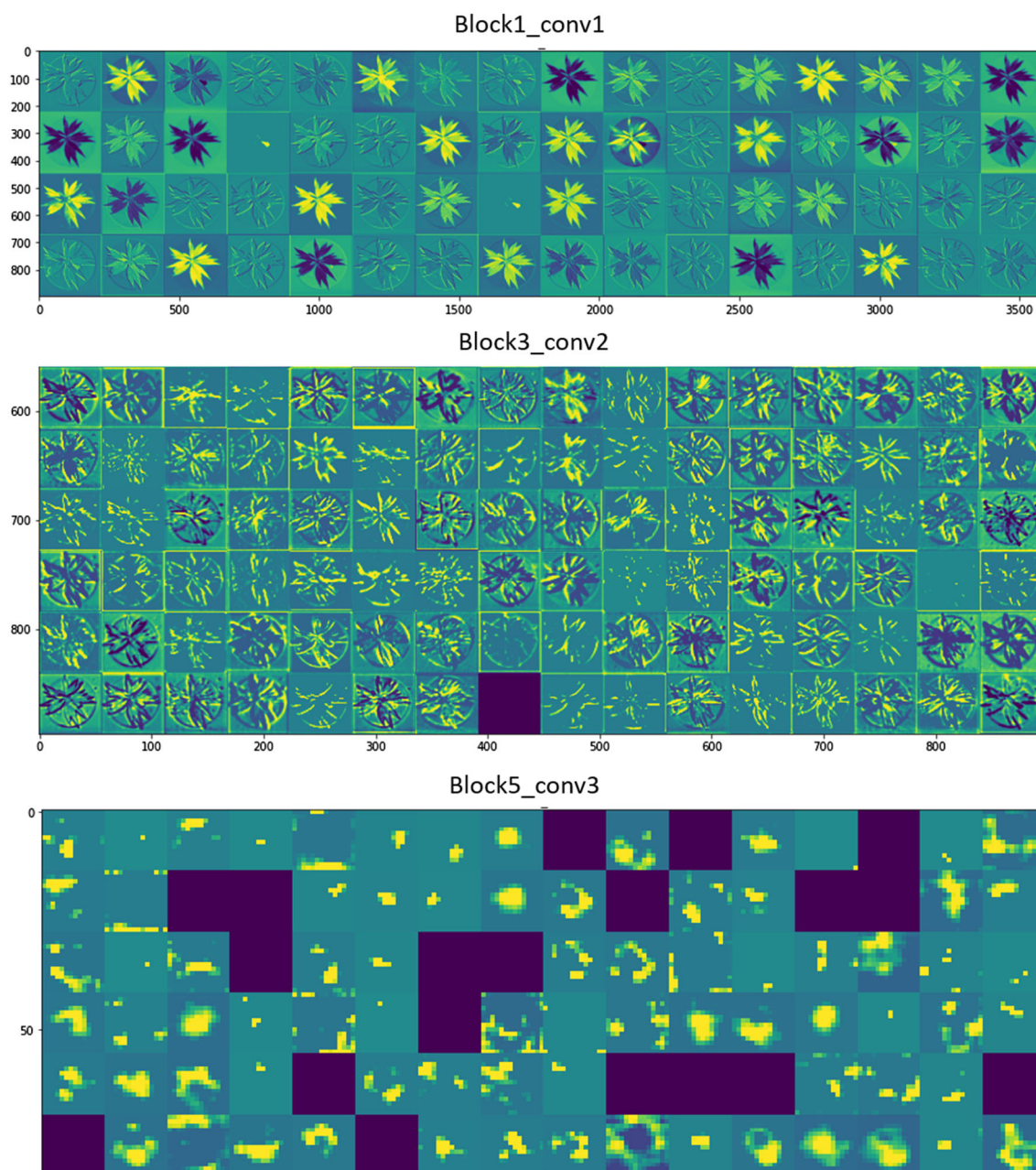


FIGURE 4 Intermediate activations of the neural network



The average accuracy obtained was 96.8%, which was a little lower than the best result achieved, 97.6% accuracy, which was reached twice. It is important to note that the difference between 96.3% and 97.6% corresponds to just a mistake in one image. Therefore, all the models showed similar behavior.

It is important to highlight that this robustness study was developed under specific illumination conditions, which are described in the Section 2. Therefore, a change in these conditions could yield different results. For this reason, it will be important to keep the same illumination conditions in future implementations.

### 3.5 | Feature maps

The feature maps of the 18 layers of the convolutional base of the Vgg16 (convolutional and max pooling layers included) were studied. From these 18 layers, we chose three representative layers of the CNN, because of the large size of the images. Figure 4 shows the progress of the image along the neural network by visualizing the intermediate activations. It is important to note that due to the high number of filters in each layer, Figure 4 only shows a subset of them.

In the first layers, which were represented by the block1\_conv1 layer, it is easy to identify the plant in the image. This indicates that this layer works with the whole image and retains almost all the information present in the original image. For this reason, the activation of the filters showed images like the input image. The channels of this first layer process low-level information such as contours, edges, corners, and therefore extract information about the leaf shape and margins. As we move into the neural network, the plant is increasingly more difficult to recognize. In the block3\_conv2, which is located in the middle of the network, it was only possible to identify small parts of the plants because the feature map was less visually interpretable. This layer encodes higher-level features, such as parts of the leaves, which implies less information about the visual contents, but more information related to the class of the input image. When observing the activation of the last layer of the convolutional base, block5\_conv3, the abstraction level is very high, and no part resembling the plant can be identified. As the block5\_conv3 layer encodes higher-level concepts than block3\_conv1 layer, the visual interpretation of the feature maps is even more abstract.

When comparing the resulting image in each filter of the same layer, we were able to observe the activation similarities between them, which reaffirms the idea that each layer works as an analysis block, where each channel performs an own extraction of information but all of them in the same analysis level.

It is also remarkable that, as we go into the CNN layers, the number of filters in blank increases. In the first layer, all filters are activated by the input image, but in the block3\_conv2, one filter in blank was found, and in the last layer, the number of filters in blank was 19. This means the pattern encoded by the filter is not found in the input image.<sup>23</sup>

## 4 | CONCLUSIONS

In this study, we investigated the most suitable training configuration of a CNN to obtain the best results in the classification of two almond plant varieties of *P. dulcis*, *Soleta* and *Pentacebas*, and the potential of CV together with DL models was evaluated.

The importance of the argument *fill\_mode* in the configuration of the data augmentation has been demonstrated. The results indicate that the *reflect* mode is more suitable than the *nearest* mode, because the latter generated some distortions of the images, which affected the accuracy of the model. In addition, the study of the batch size showed that batches with 30 or 40 images are the best option. The advantages and disadvantages of working with  $224 \times 224$  or  $150 \times 150$  pixel images were also analyzed. We found that the test accuracy was better for  $224 \times 224$  pixel images but with an increase of the computing time. The robustness of the technique has also been proven. Finally, the study of how the neural network works allowed us to have an idea of the image analysis done by the model and what features of the plant are used to discriminate the varieties.

The results of this study show the capacity of CV-DL to discriminate *P. dulcis* varieties. Future works will try to expand the dataset and the number of varieties and to assess the use of specific images from different structures of the plant, such as leaves and their venation.

## FUNDING INFORMATION

In addition, the authors thank the financial support of the Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia (project DI-COF 2017) and of the Spanish Ministry of Economy and Competitiveness (project AGL2015-70106-R, AEI/FEDER, UE).

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/CEM.3320>.

## ORCID

Sergio Borraz-Martínez  <https://orcid.org/0000-0002-5607-9462>

Francesc Tarrés  <https://orcid.org/0000-0003-0920-4782>

Ricard Boqué  <https://orcid.org/0000-0001-7311-4824>

Joan Simó  <https://orcid.org/0000-0002-2853-3381>

Anna Gras  <https://orcid.org/0000-0003-0111-7246>

## REFERENCES

1. Brunetti A, Buongiorno D, Trotta GF, Bevilacqua V. Computer vision and deep learning techniques for pedestrian detection and tracking: a survey. *Neurocomputing*. 2018;300:17-33. <https://doi.org/10.1016/j.neucom.2018.01.092>
2. Yeung S, Rinaldo F, Jopling J, et al. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *Npj Digit Med*. 2019;2(1):1-5. <https://doi.org/10.1038/s41746-019-0087-z>
3. Bao Y, Tang Z, Li H, Zhang Y. Computer vision and deep learning-based data anomaly detection method for structural health monitoring. *Struct Heal Monit*. 2019;18(2):401-421. <https://doi.org/10.1177/1475921718757405>
4. ImageNet Large Scale Visual Recognition Competition 2014 (ILSVRC2014) <http://www.image-net.org/challenges/LSVRC/2014/> (accessed May 7, 2020).
5. Ma J, Sun DW, Qu JH, et al. Applications of computer vision for assessing quality of Agri-food products: a review of recent research advances. *Crit Rev Food Sci Nutr*. 2016;56(1):113-127. <https://doi.org/10.1080/10408398.2013.873885>
6. Nyalala I, Okinda C, Nyalala L, et al. Tomato volume and mass estimation using computer vision and machine learning algorithms: cherry tomato model. *J Food Eng*. 2019;263:288-298. <https://doi.org/10.1016/j.jfoodeng.2019.07.012>
7. Patrício DI, Rieder R. Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. *Computers and Electronics in Agriculture*. Elsevier B.V. October 1, 2018, pp 69–81. <https://doi.org/10.1016/j.compag.2018.08.001>
8. Zheng Y-Y, Kong J-L, Jin X-B, Wang X-Y, Su T-L, Zuo M. CropDeep: the crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors*. 2019;19(5):1058-1079. <https://doi.org/10.3390/s19051058>
9. Arakeri MP, Lakshmana. Computer vision based fruit grading system for quality evaluation of tomato in agriculture industry. In *Procedia Computer Science*; Elsevier BV, 2016; Vol. 79, pp. 426–433. <https://doi.org/10.1016/j.procs.2016.03.055>
10. Bhargava A, Bansal A. Fruits and vegetables quality evaluation using computer vision: a review. *J King Saud Univ - Comput Inform Sci*. 2018;5:1-15. <https://doi.org/10.1016/j.jksuci.2018.06.002>
11. Wang G, Sun Y, Wang J. Automatic image-based plant disease severity estimation using deep learning. *Comput Intell Neurosci*. 2017; 2017:1-8. <https://doi.org/10.1155/2017/2917536>
12. Chouhan SS, Singh UP, Jain S. Applications of computer vision in plant pathology: a survey. *Arch Comput Meth Eng*. 2020;27(2):611-632. <https://doi.org/10.1007/s11831-019-09324-0>
13. Jiang Yu, Li Changying. Convolutional Neural Networks for Image-Based High-Throughput Plant Phenotyping: A Review. *Plant Phenomics*. 2020;2020:1-22. <https://doi.org/10.34133/2020/4152816>
14. Bauer A, Bostrom AG, Ball J, et al. Combining computer vision and deep learning to enable ultra-Scale aerial Phenotyping and precision agriculture: a case study of lettuce production. *Hortic Res*. 2019;6(1):1-12. <https://doi.org/10.1038/s41438-019-0151-5>
15. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol*. 1986;51(Pt 1):263-273. <https://doi.org/10.1101/SQB.1986.051.01.032>
16. Gupta PK, Roy JK, Prasad M. Single nucleotide polymorphism: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plant. *Curr Sci*. 2001;80(4):524-535. <https://doi.org/10.2307/24104242>
17. Borraz-Martínez S, Simó J, Gras A, Mestre M, Boqué R. Multivariate classification of *Prunus Dulcis* varieties using leaves of nursery plants and near infrared spectroscopy. *Sci Rep*. 2019;9(1):1-9. <https://doi.org/10.1038/s41598-019-56274-5>
18. Borraz-Martínez S, Boqué R, Simó J, Mestre M, Gras A. Development of a methodology to analyze leaves from *Prunus Dulcis* varieties using near infrared spectroscopy. *Talanta*. 2019;204:320-328. <https://doi.org/10.1016/j.talanta.2019.05.105>
19. Barré P, Stöver BC, Müller KF, Steinhage V. LeafNet: a computer vision system for automatic plant species identification. *Eco Inform*. 2017;40:50-56. <https://doi.org/10.1016/j.ecoinf.2017.05.005>
20. Hemming J, Rath T. Computer-vision-based weed identification under field conditions using controlled lighting. *J Agric Eng Res*. 2001; 78(3):233-243. <https://doi.org/10.1006/jaer.2000.0639>

21. Luckow A, Cook M, Ashcraft N, Weill E, Djerekarov E, Vorster B. Deep Learning in the Automotive Industry: Applications and Tools. In *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*; Institute of Electrical and Electronics Engineers Inc, 2016; pp. 3759–3768. <https://doi.org/10.1109/BigData.2016.7841045>
22. Sun Y, Liu Y, Wang G, Zhang H. Deep learning for plant identification in natural. *Environ.* 2017;2017:1-6. <https://doi.org/10.1155/2017/7361042>
23. Chollet F. Deep learning with Python. Manning Publications, 2017.
24. Buduma N, Locascio N. Fundamentals of deep learning; 2019; Vol. 29.
25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition; 2015.
26. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis.* 2015;115(3):211-252. <https://doi.org/10.1007/s11263-015-0816-y>
27. ImageNet. <http://www.image-net.org/> (accessed May 7, 2020).
28. TensorFlow. <https://www.tensorflow.org/> (accessed May 7, 2020).





**How to cite this article:** Borraz-Martínez S, Tarrés F, Boqué R, Mestre M, Simó J, Gras A. Varietal quality control in the nursery plant industry using computer vision and deep learning techniques. *Journal of Chemometrics.* 2021;e3320. <https://doi.org/10.1002/cem.3320>



**Capítulo 5. Combining  
Computer Vision and Deep  
Learning to Classify Varieties of  
*Prunus dulcis* for the Nursery  
Plant Industry**



# Combining computer vision and deep learning to classify varieties of *Prunus dulcis* for the nursery plant industry

Sergio Borraz-Martínez<sup>1,2</sup>  | Joan Simó<sup>1,3</sup>  | Anna Gras<sup>1</sup>  |  
Mariàngela Mestre<sup>2</sup> | Ricard Boqué<sup>4</sup>  | Francesc Tarrés<sup>5</sup> 

<sup>1</sup>Department of Agri-Food Engineering and Biotechnology, Universitat Politècnica de Catalunya, Castelldefels, Spain

<sup>2</sup>Center of Initial Materials, Agromillora Iberia SLU, Sant Sadurní d'Anoia, Spain

<sup>3</sup>Fundació Miquel Agustí, Castelldefels, Spain

<sup>4</sup>Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, Tarragona, Spain

<sup>5</sup>Signal Theory and Communications Department, UPC Barcelona Tech, Barcelona, Spain

## Correspondence

Sergio Borraz-Martínez, Department of Agri-Food Engineering and Biotechnology, Universitat Politècnica de Catalunya, Campus Baix Llobregat, Esteve Terrades 8, Castelldefels 08860, Spain.  
Email: sergio.borraz@upc.edu

## Funding information

Spanish Ministry of Economy and Competitiveness, Grant/Award Number: AGL2015-70106-R, AEI/FEDER, UE; Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia, Grant/Award Number: DI-COF 2017

## Abstract

Varietal control to avoid unwanted varietal mixtures is an important objective for the nursery plant industry. In this study, we have developed and analyzed the capabilities of a computer vision system based on deep learning for the control of plant varieties in the nursery plant industry and for evaluating its capabilities. For this purpose, three datasets of nursery plant images were compared. The datasets came from two varieties of almond trees (*Prunus dulcis*) named *Soleta* and *Pentacebas*. Each dataset contained images with three different scales: whole plant, leaf, and venation. The Gradient-weighted Class Activation Mapping (Grad-CAM) technique was used to unveil the most important features to discriminate between both varieties. The three datasets provided classification accuracies above 97% in the test set, being the leaf dataset, with a 98.8% accuracy, the one providing the best results. Concerning the most important features of the plants, the Grad-CAM showed that they are located in the center of the leaf, that is, the venation. In conclusion, we have shown that computer vision is a promising technique for the control of plant varietal mixtures.

## KEYWORDS

computer vision, convolutional neural network, deep learning, nursery plant, varietal mixture

## 1 | INTRODUCTION

Rapid discrimination between vegetal varieties is a key requirement in nursery plant production.<sup>1</sup> The appearance of varietal mixtures within a batch, which should be homogeneous, is an important trouble, not only because the customer receives unwanted vegetal material but also because nursery plant companies may face expensive legal suits and

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Journal of Chemometrics* published by John Wiley & Sons Ltd.

the deterioration of their corporative image.<sup>2</sup> In this context, we propose using computer vision based on deep learning techniques (CV-DL) as a possible solution to sort out this problem.

Computer vision is an old discipline that has recently undergone a qualitative improvement thanks to the use of deep learning algorithms. The goals of computer vision are to achieve a higher level of operational efficiency and productivity, as well as a higher level of automatization.<sup>3</sup> Moreover, it is one of the new technologies that drives the so-called Industry 4.0.

Deep learning is a specific subfield of machine learning, which in turn is included in the field of artificial intelligence. Deep learning could be defined as a modality of learning representations from data that emphasize learning successive layers of increasingly meaningful representations.<sup>4</sup> Each layer may be viewed as a nonlinear transform of the outputs of the precedent layer. The transformation parameters (weights) are learned using a supervised strategy. In other words, it consists of training a model from examples of the classes that we want to characterize, in our case the two varieties of almond trees. During the training process, the model calculates the classification error, which is represented as a mathematical function named *loss*, and with a small set of instructions, it autonomously modifies parameters (weights) of the model to reduce the loss, which allows improving the accuracy of the classification.<sup>5</sup> This process is known as model *learning*. The *deep* in *deep learning* is a reference to the large number of interconnected successive layers, stacked on top of each other and composed of computational neurons, which is called *neural network*. Nowadays, neural networks involve tens or even hundreds of layers. The structure of deep learning neural networks is divided into three sections. The first is the *input layer*, which receives the input images that feed the model. The middle section contains the *hidden layers*, a group of successive layers where the representation of the data takes place. This part constitutes the body of the neural network. Finally, the last section receives the name of *output layer*. This layer predicts the class to which a sample belongs based on the representation made by the hidden layers.

It was decided to use CV-DL because, in other computer vision systems, the final classification is based on the features designed by the engineer. Then, the final quality of the system depends on this previous knowledge of the best features. However, in the case of study, the most important features are very difficult to determine. Therefore, it was considered more appropriate to use CV-DL.

In the agronomic sector, the use of computer vision along with deep learning algorithms is recent but is becoming a promising technique with growing popularity.<sup>6</sup> This technology is fast and does not modify or destroy the samples. Moreover, it can be easily implemented or adapted to the production line, enabling the collection of useful information that can be used as a quality control tool or to monitor a process. These advantages make computer vision an attractive tool for the agricultural industry. Despite being an emerging technology, there are several studies where CV-DL has been applied, for example, for crop type classification<sup>7,8</sup>; crop yield estimation,<sup>9,10</sup> plant phenological estimation,<sup>11,12</sup> or even plant recognition.<sup>13,14</sup> In short, CV-DL techniques have great potential in the automation of the agronomic industry, and their capacities and efficiency have been improved in the last decade.<sup>15,16</sup>

The objective of this study was to develop a methodology to implement computer vision system for the control of plant varieties in the nursery plant industry and evaluate its capabilities. The study was developed working with two varieties of *Prunus dulcis*, *Pentacebas* and *Soleta*, which are genetically close and almost morphologically indistinguishable. Three datasets of different image scales were used, namely, whole plant, leaf, and venation. It is important to note that this study focuses on the automatic control of nursery plants in the greenhouse, which means we have avoided the manipulation of the plants. This way, we tried to simulate a possible automatic line of image acquisition without sampling. The specific objectives were (1) to compare between different image scales: whole plant, leaf, and venation, to determine which is most suitable for varietal discrimination; (2) to determine which morphological characteristics are most relevant for the classification of varieties; and (3) to evaluate the neural network complexity required for the case of study.

## 2 | MATERIAL AND METHODS

### 2.1 | Image acquisition

The RGB images used were acquired using a Nikon D5300 camera. To keep homogeneous light conditions, plants were introduced inside a professional photo studio box with a dimension of 80 × 80 × 80 cm. The photo studio had integrated two LED lamps (color temperature 5,500 K, color rendering index of 93%, and 13,000 lumens), which were placed at the bottom of the box. The idea was to illuminate the leaves from the bottom up to highlight their venous architecture. A textile diffuser was positioned between the canopy plant and the light source to avoid the reflectance of



the LED lights on the leaves. The camera was placed over the box and subjected by a holder; this way, apical images were collected. This method was developed in a previous study.<sup>17</sup> A balanced sampling was performed and 412 images per variety were used, which means 824 images for each dataset. In total, 2,472 images were employed in the study. Table 1 shows more details about the number of samples used in the study.

## 2.2 | Image preparation

The whole plant image dataset was cropped in  $4,000 \times 4,000$ ,  $3,000 \times 3,000$ , and  $2,500 \times 2,500$  pixels; the leaf dataset was composed of images with a range between  $2,000 \times 2,000$  and  $1,125 \times 1,125$  pixels; finally, the images of the venation dataset were cropped in a range of  $600 \times 600$  and  $300 \times 300$  pixels. Different images sizes were used to increase the robustness of the model against scale variations. The leaf and venation datasets were obtained from the same set of original images by different cropping factors, so that leaf and venation datasets were constituted by cropping the images contained in the whole plant dataset. This strategy was chosen to reduce the acquisition time of the photos and to increase the robustness of the comparison. In addition, the resolution of all images in all sets was reduced to  $224 \times 224$  pixels before introducing them into the model. Moreover, the pixels were normalized (1.0/255) with numeric values between zero and one. The reduction of resolution was performed to reduce the computational cost and the normalization to accelerate learning in the deep learning models, thereby decreasing computational time. Figure 1 shows examples of the images of the three data sets used.

The set of images used in this study was randomly split into a training set (75% of the images), a validation set (15% of the images), and a test set (10% of the images). Table 1 shows more details about the datasets. The validation set was used to verify the correct convergence of the model and to optimize the selection of hyper-parameters. The test set was used for the final assessment of the network performance.

## 2.3 | Data augmentation

Overfitting tends to occur when neural networks work with small datasets (<1,000 images), and the model fits very specific features of the training images that are not representative of the whole class. Therefore, it reduces the capability of the model to correctly classify images that do not belong to the training dataset. Hence, overfitting means that the system achieves excellent accuracy in modeling the training set at the cost of reducing accuracy in the prediction of validation and test datasets. To reduce overfitting, data augmentation was employed.

Data augmentation allows increasing the training data set from the available data. This method consists of generating additional images by modifying the original ones using pre-established random transformations.<sup>18</sup> In this way, it is possible to increase the diversity and the number of images during the training process. Data augmentation reduces overfitting and helps the model to better predict validation/test data and to improve its classification accuracy and robustness.<sup>19</sup> Data augmentation was only applied to the training set, and it consisted of shear in a range of 0.2, zoom in a range of 0.4, horizontal flip, random rotation in a range of  $180^\circ$ , width and height shift in a range of 0.2.

## 2.4 | Convolutional neural network

A convolutional neural network (CNN) is a type of network architecture that is widely used in image analysis; for this reason, CNN is commonly used in computer vision for image recognition.<sup>20</sup> Its structural design made it possible to

TABLE 1 Description of the study

	Percentage %	Pentacebas	Soleta	Overall dataset	Overall study
Training set	75	309	309	618	1,854
Validation set	15	62	62	124	372
Test set	10	41	41	82	246
Total		412	412	824	2,472

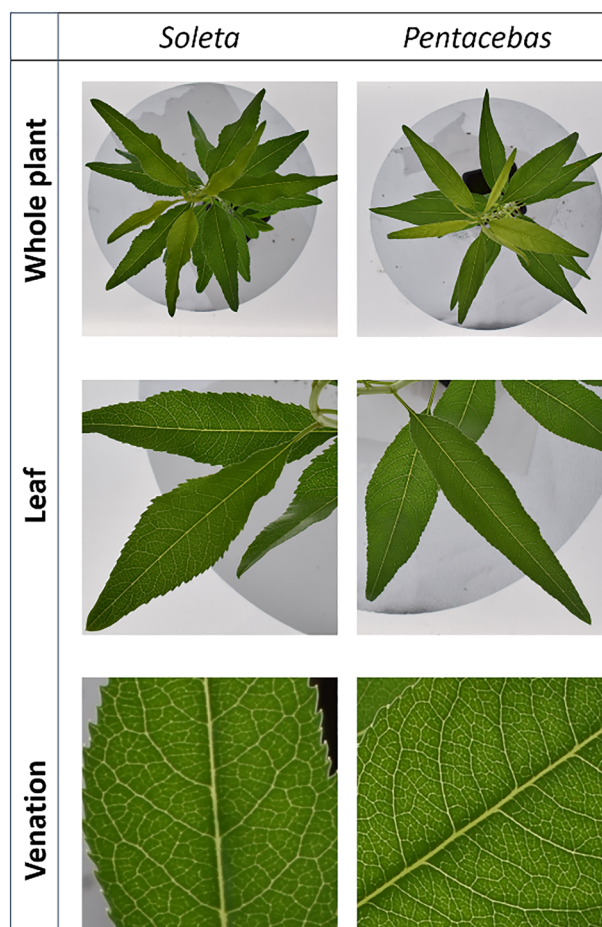


FIGURE 1 Examples of images used in the study

reduce the number of parameters of our model, which means a significant reduction in cost and computation time. The layers of a convolutional network have neurons arranged in three dimensions, so the layers have width, height, and depth. Neurons in a convolutional layer are applied using convolution to small local regions of the previous layer, so we avoid wasting completely connected neurons.<sup>5</sup> The convolution is translationally invariable, and kernels may detect different types of features at different scales in every layer. Furthermore, CNN also contains pooling layers that reduce the amount of information obtained from the convolutional layers, creating a condensed version of the information they contain.

In this study, two convolutional neural networks have been used: the VGG16 convolutional neural network<sup>21</sup> and the Shallow network. Both networks have the same types of layers, but with different structural complexity. Figure 2 shows the scheme of both neural networks.

The Shallow CNN was created specifically for this study. The goal was to build a CNN with less structural complexity than the VGG16 network, that is, with a lower number of parameters. The convolutional base of the Shallow CNN is organized in five blocks. Blocks 1 and 2 contain two convolutional layers, and blocks 3–5 contain one convolutional layer. Concerning the number of filters, blocks 1 and 2 have 32 and 64 filters, respectively. Blocks 3–5 have 128 filters. Every convolutional layer uses a filter with a receptive field of  $3 \times 3$ . The convolution stride is one pixel, and the spatial padding of the convolutional layer input is such that the spatial resolution is preserved after convolution. In addition, one max-pooling layer over a  $2 \times 2$  pixel window is located at the end of each block. The classifier is composed of two fully connected layers. The first one has 512 neurons with ReLu activation. The last one has one neuron with sigmoid activation.

VGG16 was developed by the Geometry Group at the University of Oxford to obtain the state-of-the-art results in the Large Scale Visual Recognition Challenge (LSVRC-2014) competition.<sup>21</sup> This architecture has become very popular and widely used in many image recognition problems. It consists of 16 layers divided into five blocks, which are named convolutional base, together with a fully connected layer named classifier. Every block in the VGG16 is formed by a concatenation of two or three convolutional layers followed by a pooling layer with a stride of  $2 \times 2$ . The classifier is

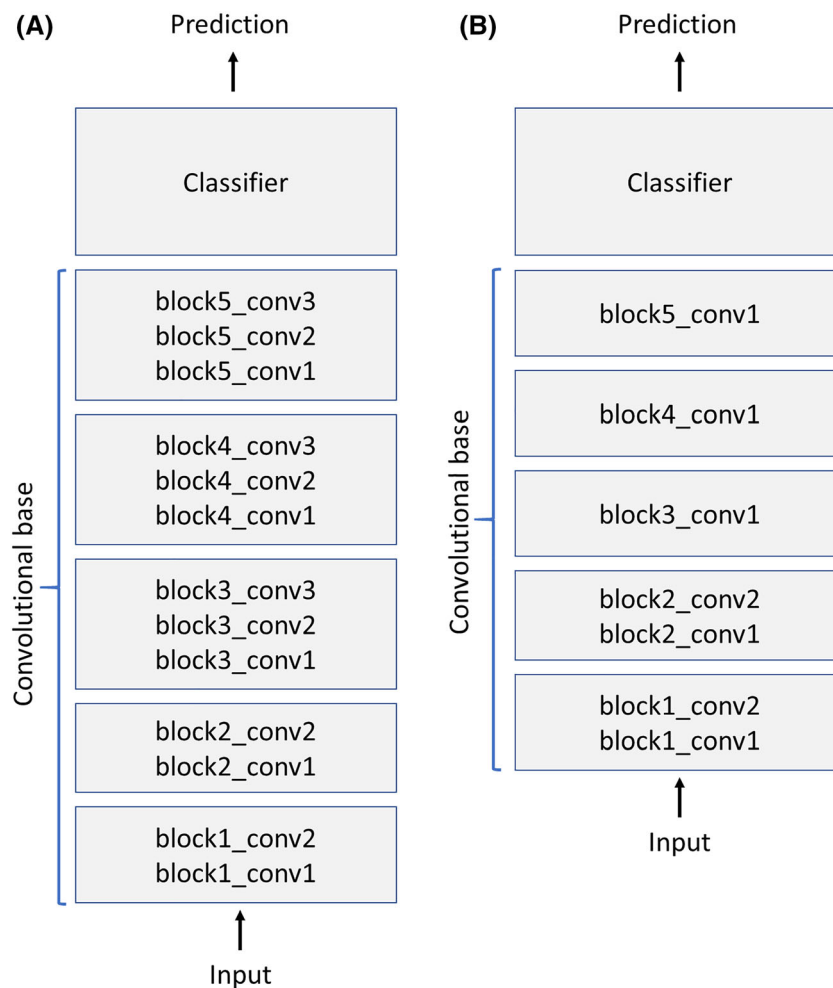


FIGURE 2 Schema of the two convolutional neural networks used. (A) VGG16; (B) Shallow

composed of three fully connected layers. The convolutional base of the VGG16 neural network was downloaded from Keras (<https://keras.io/>). Then, it was inserted into a new classifier, which consisted of one flattened layer followed by two dense layers. The convolutional neural networks learn local features of the input image. For this reason, the purpose of these fully connected layers is to collect the local information. The flatten layer allows transforming the 3D tensor of the CNN to a 1D tensor, which is required by the dense layer. The first dense layer contained 256 neurons with ReLU activation. This layer integrates all the local features learned by the last CNN layer, taking into account nonlinear combinations. Otherwise, without this inner layer, outputs would be a form of regression of the convolutional layers. The second dense layer had one neuron with sigmoid activation. The sigmoid activation means that the output of the neural network will be in a range between zero and one. As in this case, there are two classes; each image was encoded with a value above 0.5 if the image belonged to the *Soleta* class and below 0.5 if the image belonged to the *Pentacebas* class.

The complexity of the structure increases the time and the difficulties during the training process. Therefore, it was considered interesting to compare both CNN to assess the complexity required to develop the plant varietal control tool.

## 2.5 | The training steps

### 2.5.1 | Training of the VGG16 CNN

To facilitate the training process and try to get better results, the transfer learning technique was used. Transfer learning<sup>22</sup> is one of the most important techniques in deep learning as it allows us to use a pre-trained neural network.<sup>23,24</sup>

This is an advantage because training a neural network from scratch requires a huge dataset and a lot of computation time due to the large number of parameters that need to be adjusted. Through transfer learning, it is possible to download a pre-trained network and use its parameters as an initial good weight guess, and the fine tune the model by training it with our dataset. In this way, the representations previously learned by the network can be used to extract interesting features from new samples. In this study, a VGG16 convolutional network model, pre-trained with the dataset of ImageNet, was downloaded. This database includes 1.4 million labeled images from 1,000 different classes. ImageNet is one of the most common datasets used to pre-train CNNs in computer vision.

There are two modalities of transfer learning, *feature extraction* and *fine-tuning*. Based on the results obtained in previous tests,<sup>17</sup> fine-tuning was chosen. *Fine-tuning*<sup>4</sup> consists of freezing the parameters from the initial layers and training the last layers of the convolutional base, specifically the layers from block 5, together with the classifier. In this way, it is possible to slightly adjust the more abstract representations of the model being reused, in order to make them more relevant for the problem at hand.<sup>4</sup> Through this strategy, the trainable parameters were reduced by 45.4%, from more than 16 million to 9 million.

The three datasets were trained using the same strategy, except for the number of epochs, which was chosen based on the convergence between training and validation, shown by the accuracy and loss. The model of the whole plant dataset was trained during 90 epochs because it tended to suffer from overfitting above 100 epochs. In contrast, the leaf and venation models were trained during 110 epochs. The batch size was 30 for the three datasets, and a learning rate of  $10^{-5}$  was used. Regarding the optimizer, an Adam with a beta 1 = 0.9, beta 2 = 0.999, and epsilon =  $1e-0.7$  was used. Besides, a dropout of 0.5 was applied to avoid the overfitting and to increase the robustness of the models.

## 2.5.2 | Training of the shallow CNN

Since the Shallow network has fewer parameters than VGG16, it was preferred to train it from scratch, allowing us to simplify the process. The strategy followed consisted of training the model using an RMSprop optimizer with a learning rate of  $10^{-5}$  during 40 epochs. Then, the CNN was trained again with 30 epochs but changing the RMSprop optimizer by the Adam optimizer (beta 1 = 0.9, beta 2 = 0.999, and epsilon =  $1e-0.7$ ). Throughout the training process, a batch size of 30 epochs was used.

## 2.6 | Gradient-weighted class activation mapping (grad-CAM)

It is not just important to know what type of dataset works best, but it is also interesting to assess why this happens. In the same way, visualization of the activation maps may help to understand misclassifications. The interpretation of how CNN works to assign a label is usually difficult.<sup>25</sup> Grad-CAM<sup>26</sup> is a technique that consists of producing heat maps of class activation over input images. A class activation heat map is a 2D grid of scores associated to a specific output class, computed for every location in any input image, indicating how important each location is for the class under consideration.<sup>4</sup> This way, false color images are used to highlight the most significant regions of the image used in the final classification. Therefore, we relied on the information provided by the Grad-CAM technique to know what features the model used to classify the two classes of each dataset.

# 3 | RESULTS AND DISCUSSION

## 3.1 | Comparison between datasets

The images were collected simulating a typical production line that could be found in any nursery plant company. The idea was to mimic an automatic image acquisition process. Table 2 shows the results obtained in the deep learning models.

The model for the leaf dataset provided the best result in the test. The plant and venation datasets got the same accuracy, slightly below the leaf dataset.

Before starting the discussion, it is important to remark that there is a resolution reduction step just before the images are introduced in the CNN. So that, in the images of the whole plant dataset, the features related to the

TABLE 2 Results obtained in the comparison of datasets

Dataset	Test accuracy (%)
Whole plant	97.6
Leaf	98.8
Venation	97.6

edge of the leaves and the venation are negligible. This means that the whole plant only provides information about the color, the shape, the relation between length/width of the leaf, and the disposition of the leaves. Concerning the images of the leaf dataset, they contain information about the shape, edge features, and principal veins of the leaves. These characteristics seem to be the most critical to discriminate between the two classes, *Soleta* and *Pentacebas*. Using the leaf dataset, a 98.8% of accuracy was achieved in the test, which means the model only misclassified one image. This misclassification corresponds to a *Pentacebas* image that was erroneously assigned to the *Soleta* class with a value of 0.61. It is important to remark that the rest of the images, which were correctly classified in both *Soleta* and *Pentacebas* classes, had confidence values very close to one and zero, respectively. This result is surprisingly good considering that they are two genetically very close varieties of *P. dulcis*. This agrees with the botanical science that uses leaf feature taxonomic keys for plant identification.<sup>27,28</sup> Regarding venation, some studies highlight the usefulness of the venation pattern for the identification of vegetal species.<sup>29</sup> In our study, the venation model provided good results but worse than the ones of the leaf dataset. We believe that the methodology chosen to collect the venation images might not be the most suitable. Although the source of light was positioned under the plant to emphasize the venation characters, the illumination was not homogeneous in all the leaves. This lack of homogeneity may have increased the variability within a class, thus increasing the difficulty of varietal identification. Perhaps this approach is underestimating the real potential of venation for varietal discrimination, but this study focused on the automatic acquisition of plant images. Thus, we did not consider sampling individual leaves to obtain better images of the venation.

### 3.2 | Important features

The class activation heat map for each class is shown in Figure 3. For the whole plant, high activation regions of the CNN can be appreciated. These areas were located mainly in the center of the leaves. Leaves with greater activation than others can also be seen, which is because the grad-CAM image represents the features analyzed by one neuron. Other CNN neurons generate complementary activation, so this heat map is extrapolated to the other leaves.

Concerning the leaf dataset, a high activation was observed in the center and edges of the leaves. This indicates that the shapes of the edge of the leaves, along with the venation, are interesting features for identification and discrimination between the two classes. By comparing the heat map activation of leaves and whole plants, it is easy to find similarities. This implies the model searches and uses the same features in both cases. The main difference between leaf and whole plant datasets is the scale. In the whole plant dataset, more leaves can be used, each with its own features, but the clarity of the edge shape and the venation is less than in the leaf dataset due to the reduced resolution. In contrast, in the leaf dataset, there is less information due to the limited vision field that is focused on fewer leaves, but the edges and venation are observed with a better resolution than when the image is of the entire plant. Plant-scale images provide a global image compared to leaf-scale images. Despite this, the leaf scale ones offer better information to classify the images correctly because they show better the characteristics that differentiate both classes.

The venation dataset did not show any high activation zone, in contrast to what was observed in the other two cases. This indicates that there are no specific features useful for classification other than the entire image. Thus, all the information contained in the image is important for discrimination. By looking at Figure 3 from top to bottom, it can be observed that the model is focused on the center of the leaves in the three datasets, which means the most important characteristics are located in the center of the leaf, that is, the venation. Despite the result obtained with the venation dataset being good, it is not the best due to the variability caused by the position of the leaves in respect to the source of light, which affected the visualization of the veins of the leaves. The way the beam of light pass through the leaves makes the veins are clearly observed in some images but are difficult to appreciate in others. This issue could have made classification difficult as there was no clear pattern. The leaf dataset was the best in the comparison, which can

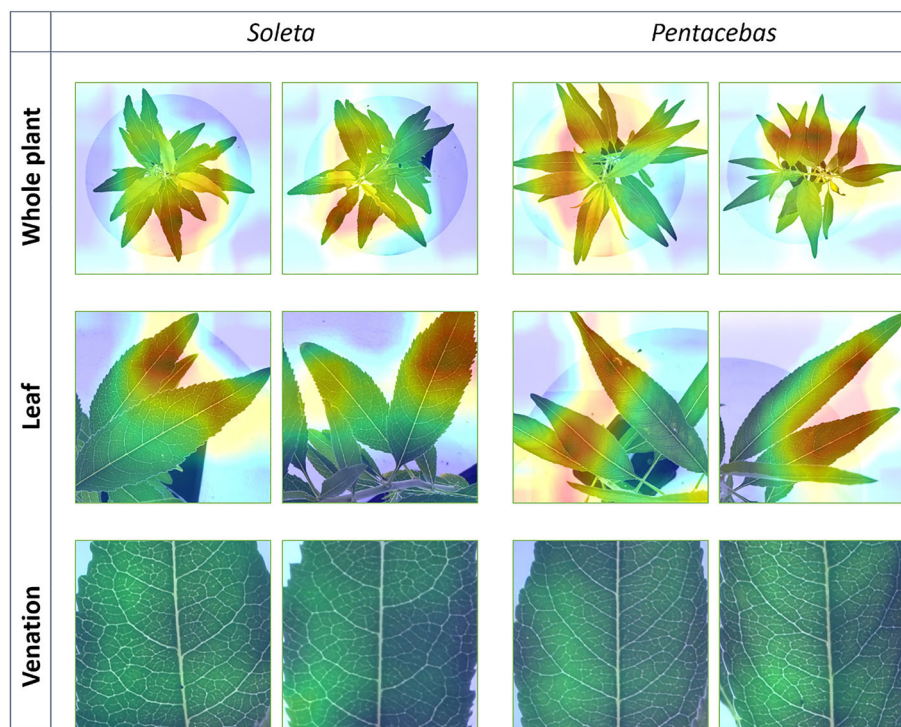


FIGURE 3 Heat map activation from the three datasets

TABLE 3 Comparison between Shallow and VGG16 networks

	Accuracy	Number of images misclassified	Number of parameters ( $\times 10^6$ )
VGG16	98.8	1	13.5
Shallow	93.9	5	1.5

be explained because leaves contain information extracted from the venation, together with other important features, such as the shape of the edge. Due to the method chosen to collect the images, edge shape was a more robust characteristic than venation, and this can help to improve the discrimination between *Soleta* and *Pentacebas* varieties.

### 3.3 | Comparative between CNNs

The leaf dataset was used to compare the VGG16 and Shallow neural networks, because it was the dataset producing the best results in the previous section. Table 3 shows the results of the comparison between both CNNs.

Concerning the classification rate, the Shallow network reached a 93.9% of accuracy, which is a good result given the difficulties to discriminate both *P. dulcis* varieties. Despite the successful classification, the Shallow network provided a classification rate of almost 5% lower than the VGG16 network. This means five images misclassified by the Shallow network against only one image erroneously classified by the VGG16 network. To evaluate these results, it is important to remark that during the training process, 13.5 million parameters were adjusted in the VGG16 network, against only 1.5 million parameters in the Shallow network. Training the VGG16 network requires more time and has a higher computational cost, and definitely, we always have a trade-off between training time and accuracy of the model.

The Shallow network proved to be effective, but in the industry, accuracy is very important, even more than the time required to develop an innovation. In addition, once the VGG16 network has been trained, there is not a significant difference in the time required to do a prediction between both networks. For this reason, the VGG16 network seems to be more appropriate in this case.

## 4 | CONCLUSIONS

In this study, computer vision and deep learning models have been applied for plant varietal discrimination. Three datasets (whole plant, leaf, and venation) have been compared, each of them containing images with a different scale from two varieties of *P. dulcis*, *Soleta* and *Pentacebas*, in nursery plant state. The best results were obtained for the leaf data set, with a classification accuracy of 98.8%, while whole plant and venation datasets obtained 97.6% of accuracy in the test. It was shown that the edge shapes of the leaves, and especially the venation, are the most important characteristic for the identification and discrimination of the classes. Although the venation features had great importance, the image acquisition method used in the study may have limited their potential. Finally, the performances of two CNNs were compared. The VGG16 network performed better than the Shallow network, indicating that the enormous difficulty of discriminating both varieties of *P. dulcis* requires a complex neural network architecture.

These results show that computer vision together with deep learning is a promising technique for automatic control of varietal mixtures. Moreover, it will be interesting to develop a method that allows us to obtain good images of the venation of the leaves without having to take individual leaf samples.

Finally, although these models were built using images of nursery almond trees, the idea is to extend the models to other plant species.

### FUNDING INFORMATION

The authors thank the financial support of the Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia (project DI-COF 2017) and by the Spanish Ministry of Economy and Competitiveness (project AGL2015-70106-R, AEI/FEDER, UE).

### PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/cem.3388>.

### DATA AVAILABILITY STATEMENT

Data available on request from the authors.

### ORCID

Sergio Borraz-Martínez  <https://orcid.org/0000-0002-5607-9462>

Joan Simó  <https://orcid.org/0000-0002-2853-3381>

Anna Gras  <https://orcid.org/0000-0003-0111-7246>

Ricard Boqué  <https://orcid.org/0000-0001-7311-4824>

Francesc Tarrés  <https://orcid.org/0000-0003-0920-4782>

### REFERENCES

- Borraz-Martínez S, Boqué R, Simó J, Mestre M, Gras A. Development of a methodology to analyze leaves from prunus dulcis varieties using near infrared spectroscopy. *Talanta*. 2019;204:320-328. doi:10.1016/j.talanta.2019.05.105
- Borraz-Martínez S, Simó J, Gras A, Mestre M, Boqué R. Multivariate classification of prunus dulcis varieties using leaves of nursery plants and near infrared spectroscopy. *Sci Rep*. 2019;9(1):1-9. doi:10.1038/s41598-019-56274-5
- Lu Y. Industry 4.0: a survey on technologies, applications and open research issues. *J Ind Inf Integr*. 2017;6:1-10. doi:10.1016/j.jii.2017.04.005
- Chollet F. *Deep Learning With Python*. Manning Publications; 2018.
- Buduma N, Locascio N. *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*. Vol. 29. O'Reilly Media, Inc; 2017.
- Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. *Comput Electron Agric*. 2018;147(July 2017):70-90. doi:10.1016/j.compag.2018.02.016
- Kussul N, Lavreniuk M, Skakun S, Shelestov A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci Remote Sens Lett*. 2017;14(5):778-782. doi:10.1109/LGRS.2017.2681128
- Debats SR, Luo D, Estes LD, Fuchs TJ, Caylor KK. A generalized computer vision approach to mapping crop fields in heterogeneous agricultural landscapes. *Remote Sens Environ*. 2016;179:210-221. doi:10.1016/j.rse.2016.03.010
- Wang Q, Nuske S, Bergerman M, Singh S. *Automated Crop Yield Estimation for Apple Orchards*. Springer, Heidelberg; 2013:745-758. doi:10.1007/978-3-319-00065-7\_50.
- Dorj UO, Lee M, Yun S. An yield estimation in citrus orchards via fruit detection and counting using image processing. *Comput Electron Agric*. 2017;140:103-112. doi:10.1016/j.compag.2017.05.019

11. Yalcin, H. Plant phenology recognition using deep learning: deep-pheno. In 2017 6th International Conference on Agro-Geoinformatics. IEEE; 2017:1-5. doi:10.1109/Agro-Geoinformatics.2017.8046996.
12. Mochida K, Koda S, Inoue K, et al. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience*. 2018;8(1):1-12. doi:10.1093/gigascience/giy153
13. Zhu X, Zhu M, Ren H. Method of plant leaf recognition based on improved deep convolutional neural network. *Cogn Syst Res*. 2018;52: 223-233. doi:10.1016/j.cogsys.2018.06.008
14. Barré P, Stöver BC, Müller KF, Steinhage V. LeafNet: a computer vision system for automatic plant species identification. *Ecol Inform*. 2017;40:50-56. doi:10.1016/j.ecoinf.2017.05.005
15. Dutta S. An overview on the evolution and adoption of deep learning applications used in the industry. *Wiley Interdiscip Rev Data min Knowl Discov*. 2018;8(4):1-12. doi:10.1002/widm.1257
16. Cheng Y-P, Li C-W, Chen Y-C. Apply computer vision in GUI automation for industrial applications. *Math Biosci Eng*. 2019;7526-7545. doi:10.3934/mbe.2019378
17. Borraz-Martínez S, Tarrés F, Boqué R, Mestre M, Simó J, Gras A. Varietal quality control in the nursery plant industry using computer vision and deep learning techniques. *J Chemometr*. 2020:1-11. doi:10.1002/cem.3320
18. Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018*. Institute of Electrical and Electronics Engineers Inc.; 2018:117-122. doi: 10.1109/IIPhDW.2018.8388338.
19. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):1-48. doi:10.1186/s40537-019-0197-0
20. Qin Z, Yu F, Liu C, Chen X. How convolutional neural networks see the world—a survey of convolutional neural network visualization methods. *Math Found Comput*. 2018;1(2):149-180. doi:10.3934/mfc.2018008
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc*. 2014.
22. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191
23. Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. *Front Plant Sci*. 2016;7(September): 1-10. doi:10.3389/fpls.2016.01419
24. Mehdipour Ghazi M, Yanikoglu B, Aptoula E. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*. 2017;235(January):228-235. doi:10.1016/j.neucom.2017.01.018
25. Lipton ZC. The myths of model interpretability. *Commun ACM*. 2018;61(10):36-43. doi:10.1145/3233231
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput vis*. 2020;128(2):336-359. doi:10.1007/s11263-019-01228-7
27. Cope JS, Corney D, Clark JY, Remagnino P, Wilkin P. Plant species identification using digital morphometrics: a review. *Expert Systems with Applications Pergamon*. 2012;39(8):7562-7573. doi:10.1016/j.eswa.2012.01.073
28. Hearn DJ. Shape analysis for the automated identification of plants from images of leaves. *Taxon*. 2009;58(3):934-954. doi: 10.1002/tax.583021
29. Lee SH, Chan CS, Mayo SJ, Remagnino P. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognit*. 2017;71:1-13. doi:10.1016/j.patcog.2017.05.015

**How to cite this article:** Borraz-Martínez S, Simó J, Gras A, Mestre M, Boqué R, Tarrés F. Combining computer vision and deep learning to classify varieties of *Prunus dulcis* for the nursery plant industry. *Journal of Chemometrics*. 2022;e3388. doi:10.1002/cem.3388



## Discusión



## Discusión

---

A lo largo del proyecto se han empleado diferentes métodos y herramientas analíticas con la finalidad de lograr los objetivos planteados en su inicio. En este caso, los objetivos son un reflejo de las necesidades de la empresa, las cuales deben ser satisfechas para lograr una mejora en el sistema productivo de la misma. No todas las tecnologías estudiadas surgieron del planteamiento inicial del proyecto, sino que en su mayoría se fueron incorporando a medida que el proyecto se fue desarrollando, como posibles soluciones a los desafíos e inconvenientes que fueron apareciendo.

### 4.1 Espectroscopia de infrarrojo cercano

La espectroscopia de infrarrojo cercano fue la primera tecnología que se empleó en el presente estudio debido a sus numerosas ventajas, como su rapidez de análisis, sencillez de manejo, bajo coste y análisis no destructivo [39–41]. Pese a que el espectrofotómetro FT-NIR Antaris II (Thermo Scientific, USA) no se adecuaba plenamente a los requisitos de implementación en la línea de producción de Agromillora, sus buenas prestaciones hicieron que fuese considerado como un buen punto de partida para evaluar el potencial de la espectroscopia NIR para discriminar las variedades de interés. Dichas cualidades consistían en un amplio rango espectral, capaz de cubrir toda la región del infrarrojo cercano, su alta resolución y la posibilidad de múltiples métodos de registro de datos. Prestaciones difíciles de encontrar en equipos más portátiles.

En primer lugar, todos los esfuerzos se centraron en el desarrollo de una metodología de trabajo. A pesar de los numerosos estudios que se pueden encontrar en la bibliografía en los que se emplea la espectroscopía NIR como herramienta de análisis [136,137], existía una falta de conocimiento y de definición de protocolos por lo que respecta a su uso en muestra vegetales. Por ello fue necesario investigar aspectos como la región de las hojas a analizar, información sobre la heterogeneidad

## Discusión

de las muestras, fuentes de variabilidad, comparativas entre diferentes métodos de preprocesamiento de las muestras foliares y de pretratamientos espectrales. Toda esta información (Capítulo 1), resultó de gran utilidad a la hora de diseñar un muestreo foliar y de definir el mejor tratamiento de las muestras y de los espectros.

Como se ha comentado, la firma espectral deriva de la composición molecular y estructural de las muestras analizadas [138–140]. Para idear una herramienta que permita diferenciar variedades de almendro (*Prunus dulcis*), es necesario identificar la firma espectral de cada una de las variedades de interés. Al tratarse de organismos vivos, y más específicamente, de plantas de hoja caduca, la composición molecular de las hojas a lo largo del año está expuesta a importantes variaciones [141–143]. La mayoría de ellas derivan del estado fisiológico de la planta, dentro del cual se incluyen los cambios a lo largo del ciclo vegetativo y aparición de factores de estrés, tanto de origen biótico (enfermedades e insectos) como abiótico. Por lo que respecta a este último, se deben incluir también aquellos ocasionados por el propio manejo de las plantas, como son los ciclos de riego, aportación de nutrientes y la aplicación de tratamientos fitosanitarios. Un ejemplo de estas fuentes de variabilidad, fácilmente observable en la **Figura 2 del Capítulo 1**, es la influencia de las moléculas de agua en la huella espectral del infrarrojo cercano. La molécula de agua tiene una gran absorbancia en la región del NIR, causando la aparición de grandes y amplias bandas en el espectro. De manera que, el propio estado hídrico de la planta, el cual depende del aporte de agua a través de los ciclos de riego, puede influir en la huella espectral.

Lo descrito anteriormente, viene reforzado por los resultados obtenidos en el modelo *Analysis of variance – simultaneous component analysis (ASCA)* [144,145] (**Tabla 3 del Capítulo 1**), donde se analiza la aportación de los factores, variedad, árbol y edad de las hojas, a la variabilidad total detectada en los espectros. Los resultados, además de mostrar que el factor variedad es importante, indican que hay una serie de factores no contemplados en el modelo (Residuo) que aportan una parte muy importante de la variabilidad detectada. Dentro de estos factores no

contemplados encontraríamos los mencionados anteriormente. También hay que mencionar que las muestras empleadas en el modelo ASCA provenían de árboles de uno de los campos de planta madre de Agromillora. De esta forma, se puede realizar la suposición que, bajo condiciones controladas, como es el caso de un invernadero [146], todas las fuentes de variabilidad, anteriormente descritas, pueden verse amortiguadas, de manera que su efecto podría ser menor al mostrado en el modelo.

Todos estos factores causan un incremento de la variabilidad de los perfiles espectrales de cada una de las variedades de interés. Por lo tanto, suponen un reto a la hora de poder definir un perfil espectral para cada una de las clases.

Teniendo presente toda la información obtenida del estudio expuesto en el Capítulo 1 y siguiendo la metodología definida en ese mismo capítulo, se construyeron una serie de modelos de clasificación multivariantes (PLS-DA) a partir de las firmas espectrales. Para ello se estudiaron seis variedades de almendro, *Avijor*, *Guara*, *Isabelona*, *Marta*, *Pentacebas* y *Soleta*. Todas ellas presentan una gran proximidad genética debido a que provienen de cruces de variedades tradicionales. De hecho, algunas de ellas provienen del mismo cruzamiento, como es el caso de *Avijor* y *Marta*, las cuales se originaron del cruce entre “*Ferragnès x Tuono*”, e *Isabelona* y *Soleta*, provenientes del cruce entre “*Blanquerna x Bella d’Aurons*”. Dicha cercanía genética hace que sean fenotípicamente muy similares, incrementando notablemente la dificultad de identificación y diferenciación. Por esta razón, supuso un gran logro el hecho de haber alcanzado una tasa de acierto en la clasificación de entre el 96.7% y el 100% (**Tablas 2, Capítulo 2**) en hoja deshidratada y triturada. Además de una exactitud en la predicción de entre el 92.2% y el 100% en hoja fresca (**Tablas 2, Capítulo 2**). Estos resultados demuestran el potencial de la espectroscopia NIR para discriminar las variedades de almendro de interés.

Si se comparan los resultados obtenidos con la bibliografía. La primera gran diferencia es que la mayoría de estudios realizados hasta la fecha se centran en la

## Discusión

diferenciación e identificación de individuos a nivel de especie (nivel taxonómico) [137,147,148]. Como es el caso del estudio realizado por Qui Fan, *et al.*, 2010, sobre tres especies de *Ephedra*, en el cual se alcanza una tasa de acierto en la clasificación de entre el 84.2% y el 95.0%. También hay publicaciones enfocadas en la diferenciación de variedades, como la realizada por Lorena de Oliveira Moura, *et al.*, 2015. En el citado estudio, el objetivo era caracterizar tres variedades de lechuga con el fin de facilitar su clasificación. Los resultados mostraron una eficacia del 100%. No obstante, las variedades de lechuga empleadas eran morfológicamente muy diferentes entre ellas. Características a nivel de coloración y estructura foliar influyen en el espectro de reflectancia. En el caso de las variedades de almendro estudiadas en el presente trabajo, presentan una gran proximidad fenotípica que las hace prácticamente indistinguibles. Lo cual incrementa la dificultad de la diferenciación.

Por otro lado, siguiendo métodos quimiométricos, se identificaron las regiones espectrales de mayor interés, pudiendo reducir el rango espectral de trabajo casi a la mitad. La selección de las longitudes de onda óptimas para la clasificación permite reducir la complejidad del modelo de calibración multivariante, por medio de la reducción de los requisitos computacionales y evitar la incorporación al modelo de variables que únicamente aportan ruido [149].

Pensando en la posible implementación de la espectroscopia de infrarrojo cercano en la línea de producción de Agromillora, los mayores inconvenientes podrían ser la limitación en el tamaño de la región de análisis y dificultades en su automatización. Cuando se trabaja con un espectrofotómetro, la región o superficie de análisis se limita a poco más de un centímetro de diámetro. Este hecho implicó realizar varias mediciones de cada hoja, como método para incrementar la superficie de hoja analizada. Esto es debido a que la espectroscopia es una tecnología de proximidad que se utiliza principalmente para analizar pequeñas áreas con el objetivo de obtener muestras espectrales sin definición espacial [76]. Esta limitación impide el poder analizar plantas enteras, ya sea individualmente o en grupos. A la hora de

implementar la técnica, y teniendo presente las fuentes de variabilidad, sería conveniente el poder extraer información de la planta en su conjunto. No obstante, los resultados de los modelos PLS-DA presentes en el Capítulo 2, muestran que no es estrictamente necesario analizar la planta entera para lograr una correcta discriminación de las variedades de interés.

Estos primeros trabajos permiten tener una idea bastante clara de las cualidades de la espectroscopía de infrarrojo cercano. Llegados a este punto, se quisieron buscar otras alternativas que pudieran corregir en algún aspecto las debilidades detectadas en esta técnica, como puede ser el muestreo masivo o la influencia de las condiciones ambientales sobre la adquisición de las muestras. Por ello, surgió el interés de investigar las cámaras espectrales.

## 4.2 Cámara hiperespectral

La investigación de las cámaras espectrales surgió dentro del marco de búsqueda de técnicas que permitieran salvar los inconvenientes de la espectroscopia NIR a la hora de lograr los objetivos planteados.

Al igual que la espectroscopia NIR, la cámara espectral presenta una serie de características que la convierten en una técnica interesante. Sus principales ventajas consisten en que no requiere de muestreo manual e individualizado, ya que permite analizar conjuntos de plantas simultáneamente y presenta mayor robustez frente ambientes más adversos, como los de un invernadero. Además, permite tener un control de la región de análisis gracias a la adquisición de imágenes.

En referencia al estudio de diferenciación varietal mostrado en el Capítulo 3, se lograron observar diferencias significativas a nivel de espectro promedio (**Figura 3, Capítulo 3**) entre las tres variedades de portainjertos estudiadas, *Colt*, *Gisela 6* y *Maxma 14*. Dichas diferencias fueron analizadas matemáticamente mediante la construcción de un modelo de discriminación (PLS-DA). A través de este modelo

## Discusión

quimiométrico se logró clasificar las tres variedades con una tasa de acierto de entre el 97.7 y el 100%, dependiendo de la clase (**Tabla 1, Capítulo 3**). De esta forma, se demuestra la capacidad de poder diferenciar variedades de portainjerto de cerezo a través de sus imágenes espectrales. Este resultado va en línea con los obtenidos por otros grupos de investigación que emplearon esta tecnología para diferenciar especies arbóreas en bosques tropicales [150], especies de plantas herbáceas [151] y en zonas de pasto [152].

La primera región espectral en la que se aprecian diferencias entre los espectros promedios se localiza en la banda del azul, entre 414 y 430 nm. Esta se encuentra en el extremo del rango espectral en el que opera la cámara. Por otro lado, también se observaron diferencias en el porcentaje de reflectancia promedio en la región del infrarrojo cercano (780-1000 nm). El espectro de reflectancia en dicha región se relaciona con la configuración de los espacios aéreos que forman la estructura interna de las hojas [153,154]. El resto del espectro de reflectancia de las tres variedades, correspondiente a la región en la que intervienen los pigmentos fotosintéticos, se encuentra altamente correlacionado debido a que presentan una composición química muy similar. Por lo tanto, la componente o característica principal que permite diferenciar las variedades *Colt*, *Gisela 6* y *Maxma 14* a través de sus respectivos espectros de reflectancia es la estructura foliar.

Por lo que respecta a las carencias nutricionales, se observó una reducción de la absorbancia en las bandas del azul y del rojo, además de un incremento en la reflectancia en la región del verde. Dichos cambios espectrales, al hallarse en la región del visible, se relacionan con afectaciones en los pigmentos de las hojas, especialmente de la clorofila [153]. Al verse afectada la clorofila, la capacidad fotosintética de las plantas también se vio reducida, lo cual se observa explícitamente en la reducción de la absorbancia de radiación en las regiones del azul y el rojo, correspondientes a las regiones de máxima absorbancia de la clorofila. Por otro lado, también se observó un desplazamiento hacia la izquierda del borde rojo (*red edge*),



el cual se relaciona con la carencia de clorofila [155,156] y el estado fisiológico de la planta [157]. Tal es así, que muchos de los índices vegetales empleados en la agricultura de precisión y en investigación emplean el borde rojo como un indicador del estado de la clorofila [158]. El borde rojo es el resultante de dos propiedades ópticas de las plantas: La absorbancia de la clorofila, que genera una reflectancia baja, y la dispersión de la luz derivada de la estructura interna de las hojas, que causa una reflectancia alta [159,160].

Los citados cambios espectrales pueden ser empleados para la construcción de modelos matemáticos que permitan identificar automáticamente plantas con deficiencias nutricionales. Así se demuestra con los resultados obtenidos en el modelo PLS-DA, en el cual se logró un 100% de exactitud en la clasificación de plantas sanas (control) y plantas con deficiencias nutricionales.

En relación con el control del estado fitosanitario, se logró detectar diferencias espectrales entre un grupo de plantas de la variedad *Colt*, inoculadas con *Pseudomonas syringae* y el grupo control. No solo se observaron diferencias a nivel espectral, sino que se logró identificar el grupo de plantas inoculadas con una exactitud del 93.3% mediante un modelo PLS-DA. La detección e identificación de fitopatologías es un aspecto fundamental para una aplicación adecuada y específica de las medidas fitosanitarias en la producción de cultivos [88], de ahí la importancia de estos resultados. El empleo de cámaras espectrales supone una ventaja operativa respecto a los métodos tradicionales de control visual de patologías, y también económica en comparación a las pruebas moleculares (PCR, ELISA, etc.). Mediante las cámaras espectrales es posible monitorizar los cultivos sin realizar muestreos ni ningún análisis destructivo, además de poder detectar el estrés vegetal de forma precoz, incluso antes de la aparición de síntomas visible [161,162]. En este sentido, es posible encontrar en la bibliografía trabajos en los que se han empleado imágenes espectrales para la detección de hongos en el trigo [163], enfermedades en plantas de tabaco [164] y la evaluación de la afectación en la cebada [165], entre otros.

## Discusión

Aunque se trata de un estudio preliminar, estos resultados animan a pensar en la capacidad de las cámaras espectrales para la detección temprana de fitopatologías, y en especial *Pseudomonas syringae*. Por ello, estos resultados se consideran un buen avance para mejorar la monitorización y el manejo de las plantas, especialmente porque brinda la oportunidad de poder actuar antes de que se produzcan daños importantes.

Respecto a su implementación, la cámara hiperespectral empleada mostro ser muy versátil y apta para trabajar en ambiente agrícola, gracias a que se empleó un modelo diseñado para trabajar en el exterior. Únicamente sería necesaria la construcción de un soporte adecuado para la cámara en función de la tarea a desarrollar. Un factor que sí debe tenerse en cuenta es la luz ambiental, ya que el espectro de reflectancia se ve afectado por variaciones en la intensidad de luz incidente. En este sentido existen dos opciones. Por un lado, se puede estandarizar la luz incidente mediante la incorporación de luz artificial, tal como se hizo en el presente estudio. Por otro lado, es posible trabajar utilizando la luz natural, siempre y cuando se realicen frecuentemente calibraciones y se intente evitar la presencia de sombras.

Comparativamente, la adquisición de imágenes espectrales supone una ventaja operativa respecto a la obtención de espectros sin definición espacial. Esta característica la convierte en un instrumento más adecuado para ser implementado en la línea de producción de Agromillora, respecto a un espectrofotómetro de infrarrojo cercano. Por otro lado, se debe tener en cuenta que los cambios espectrales relacionados con el estado fisiológico podrían afectar a la robustez de la técnica, al igual que sucedía con la espectroscopia NIR. Por ejemplo, la región del infrarrojo cercano, en la cual se han detectado diferencias espectrales entre las tres variedades de portainjerto, es susceptible de experimentar variaciones en función del estado fisiológico de la planta [159], tal como se ha observado en los resultados obtenidos en el ensayo de detección de *Pseudomonas syringae*, **Figura 5 del Capítulo**

3. En este sentido, parece difícil el poder utilizar la misma información espectral para la diferenciación varietal y el control fitosanitario de las plantas.

### 4.3 Visión artificial junto *Deep learning* (CV-DL)

Se combina las cualidades de la visión artificial, como método de adquirir información, y la inteligencia artificial, como herramienta de análisis. Se trata de una tecnología que parte con desventaja en relación a las dos anteriores, debido a que CV-DL no es capaz de captar tanta información de las muestras como la cámara espectral, al trabajar únicamente con imágenes digitales (RGB). Pero esto a su vez, aporta una serie de beneficios que se van a ir desglosando a continuación.

Debido a las particularidades y a la complejidad de los datos empleados, primero de todo, fue necesario definir la configuración más idónea del modelo (**Capítulo 4**). Esto supone siempre un paso clave, ya que ayuda a encontrar la configuración óptima de los parámetros de entrenamiento del sistema, lo cual repercute en la eficacia final del modelo. Por otro lado, es conveniente resaltar, que la configuración del modelo depende de la estructura de la red neuronal convolucional y de los datos de estudio, de ahí la importancia de realizar esta etapa. En la **Tabla 2** y en la **Figura 3 del Capítulo 4**, se demuestra la importancia de ser cuidadoso a la hora de emplear técnicas como el *data augmentation*, ya que a pesar de reducir la aparición del indeseado *overfitting*, también puede originar artefactos que afecten negativamente a la clasificación. Además, se evaluó la robustez del modelo de inteligencia artificial, quedando demostrada la solidez del sistema (**Tabla 4, Capítulo 4**). Una vez definida la configuración óptima de entrenamiento del modelo, se pasó a evaluar las capacidades de la visión artificial junto al potencial de análisis de datos de la inteligencia artificial, en este caso *deep learning*.

Al encontrarse la información integrada en imágenes digitales (RGB), la identificación se basa principalmente en el análisis de rasgos fenotípicos de carácter morfológico. Con relación a los datos mostrados en la **Figura 3 del Capítulo 5**, las

## Discusión

características más importantes para la identificación de las variedades estudiadas son el patrón de la nervadura y los bordes de las hojas. Estos rasgos morfológicos son más estables que otras características como la intensidad del color de las hojas, el cual puede verse modificado por la aparición de deficiencias nutricionales o la fertirrigación [166–168]. A pesar de ello, se trata de una estrategia arriesgada teniendo en cuenta que las variedades son morfológicamente muy similares entre ellas. No obstante, los resultados de la **Tabla 2 del Capítulo 5** muestran una exactitud en la clasificación en el conjunto de test del 98.8%, en el caso del conjunto de datos de hoja, y del 97.6 % en los conjuntos de datos de planta y nervadura. Este resultado equivale a errar en la clasificación de tan solo una imagen de las ochenta y dos que formaban el grupo de test en los datos de hoja, y de dos imágenes en el caso de los conjuntos de imágenes de planta y venación. Estos datos son sorprendentes si se tiene en cuenta la gran similitud morfológica entre ambas variedades (**Figura 1, Capítulo 4**), lo que demuestra la eficacia de la visión artificial junto con *deep learning* para la identificación de las variedades de almendro.

De estos resultados, es conveniente resaltar que el patrón de nerviación de la hoja es sumamente importante. Tal y como se observa en la **Figura 3 del Capítulo 5**, la red convolucional emplea principalmente la información presente en el centro de las hojas para realizar la identificación y clasificación de las imágenes. No obstante, el margen de la hoja también juega un papel importante, ya que en las imágenes en las que ambas características son más visibles, *data set* de hoja, es donde se ha alcanzado una mejor tasa de acierto en la clasificación. Este resultado va en concordancia con los obtenidos por otros grupos de investigación [169–171], con la diferencia de que en los citados estudios se trabajó con el patrón de inervación de especies vegetales como las leguminosas, y en el presente trabajo se han utilizado variedades de almendro, lo que dificulta aún más esta diferenciación. El patrón de nervadura es la característica principal debido a que la cercanía genética de las variedades hace que no existan diferencias evidentes en otras características como

la forma, el tamaño, el color o la textura de las hojas, ya que comparten las mismas propiedades visuales [172]. Esto se observa claramente si se comparan las dos imágenes que componen la **Figura 1 del Capítulo 4**.

También es conveniente comentar que para lograr la máxima eficacia del sistema fue necesario la utilización de una red neuronal compleja, como es el caso de la red neuronal convolucional Vgg16. Los datos mostrados en la **Tabla 3 del Capítulo 5** así lo evidencian. La red neuronal Shallow (más sencilla que la Vgg16), pese a mostrar una exactitud en la clasificación del 93.3%, se encuentra lejos del 98.6% alcanzado por la Vgg16. La implementación de una red neuronal computacional más compleja supone un mayor trabajo a la hora de entrenar el modelo, ya que la red Vgg16 necesita ajustar 12 millones de parámetros más que la red neuronal Shallow. A pesar de ello, la diferencia de complejidad entre ambas redes neuronales únicamente se limita a la fase de entrenamiento, ya que una vez entrenadas ambas presentan un comportamiento similar por lo que respecta a complejidad y velocidad de respuesta.

La CV-DL no es sensible a los cambios que se producen a nivel molecular, siempre y cuando estos cambios no generen efectos visibles muy agudizados. De esta forma, esta técnica no se ve tan afectada por las fuentes de variabilidad relacionadas con el estado fisiológico de la planta. Esto supone una ventaja a la hora de crear modelos para la diferenciación varietal, ya que permite reducir la variabilidad de los perfiles fenotípicos de las variedades de interés, lo que a su vez favorece la robustez de los modelos. Esto se ve reforzado por los datos obtenidos en el ensayo de robustez de la técnica mostrados en la **Tabla 4 del Capítulo 4**.

Los aspectos morfológicos también pueden verse sometidos a variación por la aparición de graves daños de origen diverso, como enfermedades, daños mecánicos, etc., los cuales pueden dificultar la identificación de las variedades por parte del sistema. Este inconveniente se ve minimizado al tratarse de plantas cultivadas en

invernadero y por la aplicación de protocolos de estandarización del material vegetal por parte de la empresa. De todos modos, es un aspecto que debe tenerse en cuenta.

Por lo que respecta al control fitosanitario, el CV-DL únicamente podría indicar la presencia de alguna enfermedad cuando la planta presentara síntomas visibles [173]. En cualquier caso, podría ayudar como herramienta automatizada de detección de fitopatologías y estrés vegetal [174–176]. Para una detección prematura sería necesario la obtención de mayor información espectral.

Respecto a las cualidades para ser implementada en la línea de producción de Agromillora, destacar que se trata de una técnica sencilla, la cual consiste en una cámara digital colocada en un soporte que envía las imágenes que va adquiriendo a un ordenador, en el cual se procesa la información. Este proceso es similar al de la cámara hiperespectral, con la diferencia de que el paso de las imágenes a través de la red espectral es bastante más rápido que el tiempo necesario para construir los hiper cubos de las imágenes espectrales. Además, los buenos resultados obtenidos con los tres conjuntos de datos, planta, hoja y venación, mostrados en la [Tabla 2 del Capítulo 5](#), evidencian la flexibilidad en la adquisición de imágenes.

## 4.4 Aplicación de los métodos estudiados

El cometido de este apartado es aportar información respecto a los precios de adquisición aproximados de los equipos e instrumentos empleados en el presente proyecto, con el fin de incorporar el coste económico como otro factor comparativo. Los precios han sido calculados a partir de los costes reales de adquisición de los instrumentos juntamente con la información aportada por diferentes fabricantes. No se han considerado los posibles costes de implementación asociados a cada uno de ellos, ya que no se disponen de prototipos desarrollados. Además, en base a la experiencia y los datos, se ha realizado una estimación del tiempo necesario para la construcción de los modelos, *Tiempo de modelización*, y el tiempo necesario para analizar un conjunto de plantas una vez obtenido el modelo, *Tiempo operacional*. Las

estimaciones de tiempo deben ser entendida a nivel orientativo, no en sus valores absolutos. Toda esta información se muestra en la [Tabla 2](#).

**Tabla 2:** Resumen del precio de adquisición de los instrumentos utilizados en el presente estudio

<b>Dispositivo</b>	<b>Rango de precios en euros (10<sup>3</sup>)</b>	<b>Tiempo de modelización (semanas)</b>	<b>Tiempo operacional (horas)</b>
Espectrofotómetro NIR	30 - 40	4-5*	2-72**
Cámara hiperespectral	20 - 30	3	3
Cámara multiespectral	3 - 10	3	2
Visión artificial	1 - 3	4	1

\*4 si se trata de hoja fresca; 5 si se trata de hoja deshidratada y triturada.

\*\*2 si se trata de hoja fresca; 72 si se trata de hoja deshidratada y triturada.

Como se observa en la [Tabla 2](#), el instrumento que presenta un mayor coste de adquisición es el espectrofotómetro de infrarrojo cercano. En el lado opuesto, el más económico sería la visión artificial. Su bajo coste de adquisición favorecería su implementación, debido a que requiere de una menor inversión inicial. Por lo que respecta a las cámaras espectrales, el coste de adquisición de una cámara hiperespectral es alto, aproximándose al de un espectrofotómetro de infrarrojo cercano. Por el contrario, las cámaras multiespectrales presentan un coste mucho más contenido. Una vez se conocen las regiones de interés del espectro VNIR, es posible sustituir la cámara hiperespectral, cuya finalidad es más de investigación, por una cámara multiespectral, más barata y más fácil de usar.

Respecto al tiempo de modelización, la espectroscopia de infrarrojo cercano presenta un tiempo muy variable, dependiendo del tipo de preprocesamiento de las muestras. En este caso, la mayor inversión de tiempo radica en el muestreo de las hojas y en la preparación de las mismas, en particular las del grupo de hoja deshidratada y triturada. En cambio, la construcción de los modelos PLS-DA a partir de los espectros infrarrojos es relativamente rápida. Las técnicas basadas en la

## Discusión

adquisición de datos mediante cámaras, no requieren de la fase de muestreo ni de un procesamiento de las muestras, lo que permite ahorrar bastante tiempo en la fase de adquisición de datos. No obstante, las diferencias se encuentran a la hora de analizar los datos adquiridos. Un modelo PLS-DA se construye mucho más rápido que un modelo en el que se emplea una red convolucional, debido al tiempo necesario para entrenar la red.

Por lo que respecta a tiempo operacional, la tecnología más rápida es la visión artificial ya que el análisis de la imagen es automático e inmediato. Las cámaras espectrales requieren de más tiempo para construir los hipercubos, además en este caso, las imágenes espectrales deberían de ser visualizadas por personal cualificado para extraer el máximo de información. La diferencia de tiempo entre una cámara hiperespectral y una multiespectral es la complejidad de los datos a analizar, siendo más sencilla la multiespectral. Por último, en la espectroscopia de infrarrojo cercano volveríamos a encontrar una gran diferencia dependiendo del procesamiento de las muestras. En el caso de trabajar con hoja fresca, es en la fase de muestreo donde se pierde más tiempo, ya que el análisis de las muestras es inmediato. Si hablamos de hoja deshidratada y triturada, el tiempo se dispara, ya que es necesario introducir las muestras en una estufa para, posteriormente, triturarlas hasta obtener un polvo homogéneo.

## 4.5 Nivel de maduración tecnológica

La **Figura 23** muestra los diferentes niveles de la escala *Technological Readiness Level (TRL)*. La citada escala sirve de guía para determinar el grado de desarrollo de una tecnología, siendo el nivel 1 el inicial y el nivel 9 el de máxima madurez.

En lo que concierne a la espectroscopia de infrarrojo cercano, se logró validar la técnica a nivel de laboratorio. En el caso de querer implantarse en este ambiente, el método habría alcanzado el nivel 7, siendo necesario validar la técnica dentro del flujo de trabajo de Agromillora (nivel 8). Sin embargo, si la implementación de la



técnica se realiza a nivel de invernadero, esta se encontraría en un nivel 4 de madurez tecnológica.

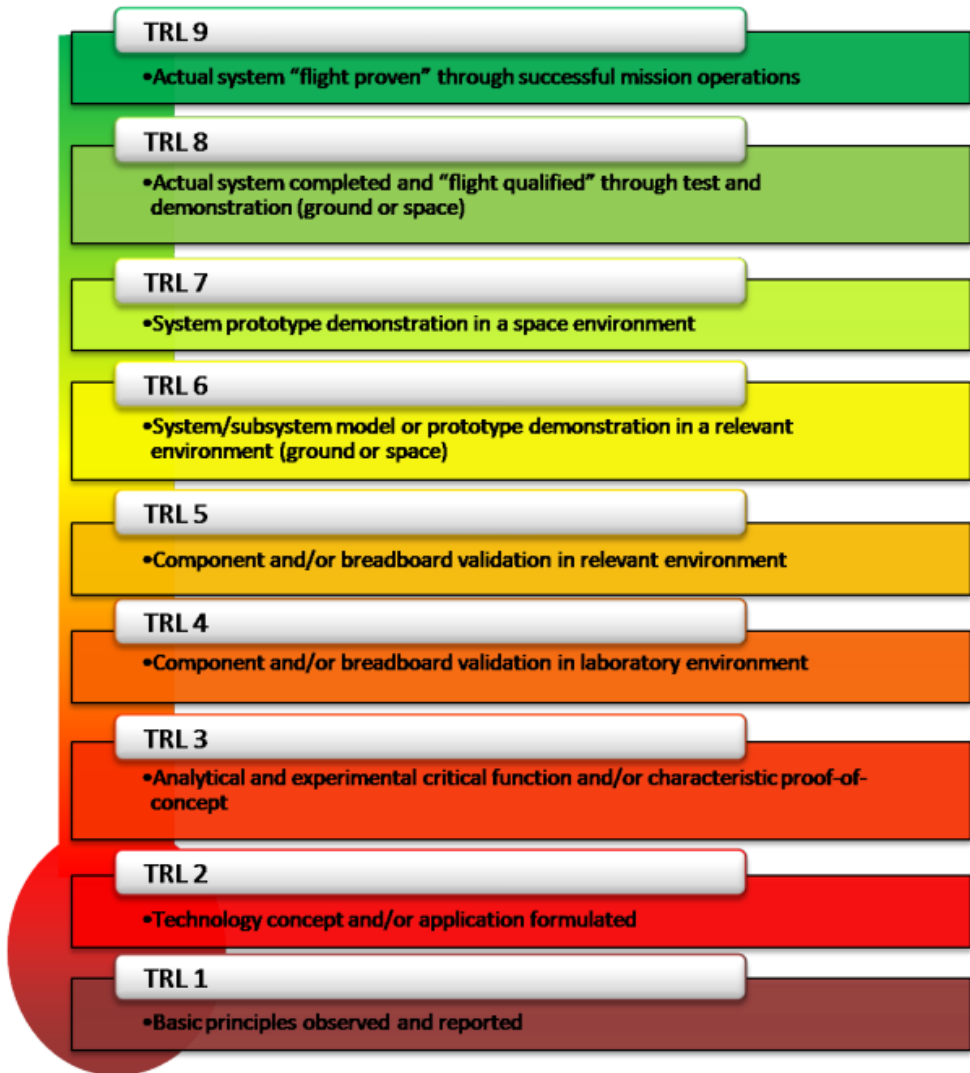


Figura 23. Escala *Technological Rediness Level (TRL)* [178].

Por lo que respecta a la cámara espectral y la visión artificial. En ambos casos se realizaron pruebas con el material vegetal simulando las condiciones del proceso productivo. Por ello, se considera que el nivel de madurez tecnológica alcanzado en

## Discusión

ambos casos es del 5 en la escala TRL, el cual consiste en alcanzar una validación del nuevo método o técnica en un ambiente que simule las condiciones reales.

## **Conclusiones**



## Conclusiones

---

1. Se logró desarrollar protocolos de adquisición de muestras, obtención de datos y análisis de los mismos para las tres técnicas utilizadas en el estudio, espectroscopia de infrarrojo cercano, cámara hiperespectral y visión artificial acoplada a *deep learning*, adecuándolos a las características de cada una y al tipo de muestra utilizado en el estudio.
2. Se logró construir y optimizar modelos quimiométricos, en el caso de la espectroscopía de infrarrojo cercano y cámara espectral, y modelos de redes neuronales computacionales, en el caso de la visión artificial, obteniendo de ellos altas tasas de acierto en la clasificación de las variedades de estudio.
3. La espectroscopia NIR demostró ser una técnica capaz de diferenciar, con una exactitud entre el 96.7% y el 100%, las variedades de almendro de interés bajo las condiciones de cultivo, a pesar de las diversas fuentes de variabilidad que afectan a la técnica, como la presencia de agua. En este sentido se desarrolló un método de deshidratación de las muestras previo análisis con el fin de evitar la influencia del agua en los espectros. Además, en la bibliografía la mayoría de estudios realizados de clasificación mediante espectroscopia de infrarrojo cercano se realizaron a nivel de especie, no a nivel de variedad, como es el caso del presente estudio.
4. La cámara hiperespectral ha mostrado ser una herramienta eficaz para la diferenciación de tres variedades de portainjerto de cerezo y, especialmente, en la gestión de los cuidados de los cultivos. En este sentido, mediante la cámara espectral se ha logrado identificar plantas con deficiencias nutricionales y afectadas por *Pseudomonas syringae*. De esta forma se demuestra la gran versatilidad de las cámaras espectrales.
5. La visión artificial acoplada a *deep learning* ha permitido diferenciar, con gran exactitud, dos variedades de almendro, *Soleta* y *Pentacebas*. Además, ha

## Conclusiones

demostrado ser una herramienta robusta, ya que se basa en características fenotípicas estables. En este sentido, se ha demostrado que el patrón de nerviación foliar es la característica más importante, lo cual coincide con la conclusión de otros grupos de investigación. No obstante, no es una herramienta adecuada para la detección prematura de fitopatologías, aunque si lo es para la identificación y control de las mismas una vez aparecen síntomas visibles.

6. De los resultados obtenidos se desprende que la mejor estrategia sería la utilización de varias herramientas o tecnologías de manera independiente o acopladas. Se podría implementar la cámara espectral para el control del estado fisiológico de las plantas y la visión artificial acoplada a *deep learning* para la identificación varietal. También sería interesante fusionar ambas técnicas y, de esta forma, aprovechar el potencial de las cámaras espectrales a la hora de adquirir información de las muestras junto con la gran capacidad de análisis de datos aportado por la inteligencia artificial.

## **Bibliografía**





## Bibliografía

---

- [1] Generalitat de Catalunya, Doctorats Industrials, (2016).  
<http://doctoratsindustrials.gencat.cat/ca/pages/home> (accessed November 17, 2020).
- [2] UPC - UPC Universitat Politècnica de Catalunya, (n.d.).  
<https://www.upc.edu/es/> (accessed November 17, 2020).
- [3] Universitat Rovira i Virgili | La universidad pública de Tarragona, (n.d.).  
<https://www.urv.cat/es/> (accessed November 17, 2020).
- [4] Agromillora, Innovación y liderazgo en el sector Viverístico, (n.d.).  
<https://www.agromillora.com/> (accessed November 17, 2020).
- [5] IRTA - Instituto de Investigación y Tecnología Agroalimentarias, (n.d.).  
<http://www.irta.cat/es/> (accessed November 23, 2020).
- [6] P.K. Gupta, J.K. Roy, M. Prasad, Single nucleotide polymorphism: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plant, *Curr. Sci.* 80 (2001) 524–535.  
<https://doi.org/10.2307/24104242>.
- [7] A. Ariani, J.C. Berny Mier y Teran, P. Gepts, Genome-wide identification of SNPs and copy number variation in common bean (*Phaseolus vulgaris* L.) using genotyping-by-sequencing (GBS), *Mol. Breed.* 36 (2016) 1–11.  
<https://doi.org/10.1007/s11032-016-0512-9>.
- [8] M.F. Clark, Immunosorbent Assays in Plant Pathology, *Annu. Rev. Phytopathol.* 19 (1981) 83–106.  
<https://doi.org/10.1146/annurev.py.19.090181.000503>.

## Bibliografía

- [9] Z. Wang, D. Yu, X. Li, M. Zeng, Z. Chen, L. Bi, J. Liu, L. Jin, D. Hu, S. Yang, B. Song, The Development and Application of a Dot-ELISA Assay for Diagnosis of Southern Rice Black-Streaked Dwarf Disease in the Field, *Viruses*. 4 (2012) 167–183. <https://doi.org/10.3390/v4010167>.
- [10] R. Balodi, S. Bisht, A. Ghatak, K.H. Rao, Plant disease diagnosis: Technological advancements and challenges, *Indian Phytopathol.* 70 (2017) 275–281. <https://doi.org/10.24838/ip.2017.v70.i3.72487>.
- [11] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, H. Erlich, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction., *Cold Spring Harb. Symp. Quant. Biol.* 51 Pt 1 (1986) 263–73. <https://doi.org/10.1101/SQB.1986.051.01.032>.
- [12] L. Schena, F. Nigro, A. Ippolito, D. Gallitelli, Real-time quantitative PCR: A new technology to detect and study phytopathogenic and antagonistic fungi, *Eur. J. Plant Pathol.* 110 (2004) 893–908. <https://doi.org/10.1007/s10658-004-4842-9>.
- [13] S.B. Visnovsky, P. Panda, K.R. Everett, A. Lu, R.C. Butler, R.K. Taylor, A.R. Pitman, A PCR diagnostic assay for rapid detection of plant pathogenic pseudomonads, *Plant Pathol.* 69 (2020) 1311–1330. <https://doi.org/10.1111/ppa.13204>.
- [14] Home | Food and Agriculture Organization of the United Nations, (n.d.). <https://www.fao.org/home/en> (accessed November 17, 2020).
- [15] Desarrollo Sostenible – United Nations Sustainable Development Sites, (n.d.). <https://www.un.org/sustainabledevelopment/es/> (accessed November 17, 2020).
- [16] P. Frankelius, C. Norrman, K. Johansen, Agricultural Innovation and the Role

- of Institutions: Lessons from the Game of Drones, *J. Agric. Environ. Ethics.* 32 (2017) 681–707. <https://doi.org/10.1007/s10806-017-9703-6>.
- [17] S. Wolfert, L. Ge, C. Verdouw, M.J. Bogaardt, Big Data in Smart Farming – A review, *Agric. Syst.* 153 (2017) 69–80. <https://doi.org/10.1016/j.agsy.2017.01.023>.
- [18] D.C. Rose, J. Chilvers, Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming, *Front. Sustain. Food Syst.* 2 (2018) 87. <https://doi.org/10.3389/fsufs.2018.00087>.
- [19] A. Kaloxylou, R. Eigenmann, F. Teye, Z. Politopoulou, S. Wolfert, C. Shrank, M. Dillinger, I. Lampropoulou, E. Antoniou, L. Pesonen, H. Nicole, F. Thomas, N. Alonistioti, G. Kormentzas, Farm management systems and the Future Internet era, *Comput. Electron. Agric.* 89 (2012) 130–144. <https://doi.org/10.1016/j.compag.2012.09.002>.
- [20] M.S. Mekala, P. Viswanathan, A Survey: Smart agriculture IoT with cloud computing, in: 2017 Int. Conf. Microelectron. Devices, Circuits Syst. ICMDCS 2017, Institute of Electrical and Electronics Engineers Inc., 2017: pp. 1–7. <https://doi.org/10.1109/ICMDCS.2017.8211551>.
- [21] C.-J. Chen, Y.-Y. Huang, Y.-S. Li, C.-Y. Chang, Y.-M. Huang, An AIoT Based Smart Agricultural System for Pests Detection, (n.d.). <https://doi.org/10.1109/ACCESS.2020.3024891>.
- [22] M. Carolan, Publicising Food: Big Data, Precision Agriculture, and Co-Experimental Techniques of Addition, *Sociol. Ruralis.* 57 (2017) 135–154. <https://doi.org/10.1111/soru.12120>.
- [23] I.D. López, J.C. Corrales, A smart farming approach in automatic detection of favorable conditions for planting and crop production in the upper basin of

- Cauca River, in: *Adv. Intell. Syst. Comput.*, Springer Verlag, 2018: pp. 223–233. [https://doi.org/10.1007/978-3-319-70187-5\\_17](https://doi.org/10.1007/978-3-319-70187-5_17).
- [24] D.I. Patrício, R. Rieder, Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review, *Comput. Electron. Agric.* 153 (2018) 69–81. <https://doi.org/10.1016/j.compag.2018.08.001>.
- [25] U.R. Mogili, B.B.V.L. Deepak, Review on Application of Drone Systems in Precision Agriculture, *Procedia Comput. Sci.* 133 (2018) 502–509. <https://doi.org/10.1016/j.procs.2018.07.063>.
- [26] E.R. Hunt, C.S.T. Daughtry, What good are unmanned aircraft systems for agricultural remote sensing and precision agriculture?, *Int. J. Remote Sens.* 39 (2018) 5345–5376. <https://doi.org/10.1080/01431161.2017.1410300>.
- [27] A. Kamilaris, A. Kartakoullis, F.X. Prenafeta-Boldú, A review on the practice of big data analysis in agriculture, *Comput. Electron. Agric.* 143 (2017) 23–37. <https://doi.org/10.1016/j.compag.2017.09.037>.
- [28] A. Weersink, E. Fraser, D. Pannell, E. Duncan, S. Rotz, Opportunities and Challenges for Big Data in Agricultural and Environmental Analysis, *Annu. Rev. Resour. Econ.* 10 (2018) 19–37. <https://doi.org/10.1146/annurev-resource-100516-053654>.
- [29] S. Mao, Y. Li, Y. Ma, B. Zhang, J. Zhou, Kai Wang, Automatic cucumber recognition algorithm for harvesting robots in the natural environment using deep learning and multi-feature fusion, *Comput. Electron. Agric.* 170 (2020). <https://doi.org/10.1016/j.compag.2020.105254>.
- [30] H.A.M. Williams, M.H. Jones, M. Nejati, M.J. Seabright, J. Bell, N.D. Penhall, J.J. Barnett, M.D. Duke, A.J. Scarfe, H.S. Ahn, J.Y. Lim, B.A. MacDonald, Robotic kiwifruit harvesting using machine vision, convolutional neural

- networks, and robotic arms, *Biosyst. Eng.* 181 (2019) 140–156.  
<https://doi.org/10.1016/j.biosystemseng.2019.03.007>.
- [31] L. Klerkx, E. Jakku, P. Labarthe, A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda, *NJAS - Wageningen J. Life Sci.* 90–91 (2019) 100315.  
<https://doi.org/10.1016/j.njas.2019.100315>.
- [32] J. Stilgoe, R. Owen, P. Macnaghten, Developing a framework for responsible innovation, *Res. Policy.* 42 (2013) 1568–1580.  
<https://doi.org/10.1016/j.respol.2013.05.008>.
- [33] SDG Progress Report, (n.d.). <http://www.fao.org/sdg-progress-report/es/> (accessed November 17, 2020).
- [34] J. Gabàs Masip, Maxwell: La teoría electromagnética de la luz, *Arbor.* 191 (2015) a265–a265. <https://doi.org/10.3989/arbor.2015.775n5004>.
- [35] T.L. Dimitrova, A. Weis, The wave-particle duality of light: A demonstration experiment, *Am. J. Phys.* 76 (2008) 137–142.  
<https://doi.org/10.1119/1.2815364>.
- [36] J. Hendry, The development of attitudes to the wave-particle duality of light and quantum theory, 1900-1920, *Ann. Sci.* 37 (1980) 59–79.  
<https://doi.org/10.1080/00033798000200121>.
- [37] A. Manickavasagan, H. Jayasuriya, Imaging with electromagnetic spectrum: Applications in food and agriculture, Springer-Verlag Berlin Heidelberg, 2014.  
<https://doi.org/10.1007/978-3-642-54888-8>.
- [38] S.N. Jha, Nondestructive evaluation of food quality: Theory and practice, Springer Berlin Heidelberg, 2010. <https://doi.org/10.1007/978-3-642-15796-7>.

## Bibliografía

- [39] R. Vitale, M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics, *Chemom. Intell. Lab. Syst.* 121 (2013) 90–99. <https://doi.org/10.1016/j.chemolab.2012.11.019>.
- [40] V. Cortés, J. Blasco, N. Aleixos, S. Cubero, P. Talens, Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review, *Trends Food Sci. Technol.* 85 (2019) 138–148. <https://doi.org/10.1016/j.tifs.2019.01.015>.
- [41] D.F. Barbin, A.L. de S.M. Felicio, D.W. Sun, S.L. Nixdorf, E.Y. Hirooka, Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview, *Food Res. Int.* 61 (2014) 23–32. <https://doi.org/10.1016/j.foodres.2014.01.005>.
- [42] E.S. Barr, The infrared pioneers-I. Sir William Herschel, *Infrared Phys.* 1 (1961) 1-IN6. [https://doi.org/10.1016/0020-0891\(61\)90037-9](https://doi.org/10.1016/0020-0891(61)90037-9).
- [43] T. Davies, The history of near infrared spectroscopic analysis: Past, present and future - From sleeping technique to the morning star of spectroscopy, *Analisis.* 26 (1998) 17–19. <https://doi.org/10.1051/analisis:199826040017>.
- [44] P. Williams, Application of Near Infrared Reflectance Spectroscopy to Analysis of Cereal Grains and Oilseeds, *Cereal Chem.* 52 (1975) 561–576.
- [45] W.F. McClure, K.H. Norris, W.W. Weeks, Rapid Spectrophotometric Analysis of the Chemical Composition of Tobacco, *Beitrag Zur Tab. Int. Contrib. to Tob. Res.* 9 (1977) 13–18. <https://doi.org/10.2478/cttr-2013-0421>.
- [46] K.H. Norris, R.F. Barnes, J.E. Moore, J.S. Shenk, Predicting Forage Quality by Infrared Reflectance Spectroscopy, *J. Anim. Sci.* 43 (1976) 889–897. <https://doi.org/10.2527/jas1976.434889x>.

- [47] M. Plans Pujolràs, *Utilització de l'espectroscòpia d'infraroig com a eina per a la selecció i millora de la mongeta seca (Phaseolus vulgaris L.)*, Universitat Politècnica de Catalunya, 2014.  
<https://upcommons.upc.edu/handle/2117/95491> (accessed January 14, 2021).
- [48] P. Prajapati, R. Solanki, V. Modi, T. Basuri, A Brief Review on NIR Spectroscopy and its Pharmaceutical Applications, *Int. J. Pharm. Chem. Anal.* 3 (2016) 117. <https://doi.org/10.5958/2394-2797.2016.00018.6>.
- [49] M. Verstraeten, D. Van Hauwermeiren, M. Hellings, E. Hermans, J. Geens, C. Vervaet, I. Nopens, T. De Beer, Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process, *Int. J. Pharm.* 546 (2018) 247–254.  
<https://doi.org/10.1016/j.ijpharm.2018.05.043>.
- [50] L.M. Kandpal, J. Tewari, N. Gopinathan, J. Stolee, R. Strong, P. Boulas, B.K. Cho, Quality assessment of pharmaceutical tablet samples using Fourier transform near infrared spectroscopy and multivariate analysis, *Infrared Phys. Technol.* 85 (2017) 300–306.  
<https://doi.org/10.1016/j.infrared.2017.07.016>.
- [51] S. Macho, M.S. Larrechi, Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry, *TrAC - Trends Anal. Chem.* 21 (2002) 799–806.  
[https://doi.org/10.1016/S0165-9936\(02\)01202-5](https://doi.org/10.1016/S0165-9936(02)01202-5).
- [52] M. V. Reboucas, J.B. dos Santos, D. Domingos, A.R.C.G. Massa, Near-infrared spectroscopic prediction of chemical composition of a series of petrochemical process streams for aromatics production, *Vib. Spectrosc.* 52 (2010) 97–102. <https://doi.org/10.1016/J.VIBSPEC.2009.09.006>.

## Bibliografía

- [53] A. Alishahi, H. Farahmand, N. Prieto, D. Cozzolino, Identification of transgenic foods using NIR spectroscopy: A review, *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* 75 (2010) 1–7. <https://doi.org/10.1016/j.saa.2009.10.001>.
- [54] B. Jamshidi, E. Mohajerani, J. Jamshidi, Developing a Vis/NIR spectroscopic system for fast and non-destructive pesticide residue monitoring in agricultural product, *Meas. J. Int. Meas. Confed.* 89 (2016) 1–6. <https://doi.org/10.1016/j.measurement.2016.03.069>.
- [55] G. Mishra, S. Srivastava, B.K. Panda, H.N. Mishra, Rapid Assessment of Quality Change and Insect Infestation in Stored Wheat Grain Using FT-NIR Spectroscopy and Chemometrics, *Food Anal. Methods.* 11 (2018) 1189–1198. <https://doi.org/10.1007/s12161-017-1094-9>.
- [56] P. Jain, R. Maddila, A.M. Joshi, A precise non-invasive blood glucose measurement system using NIR spectroscopy and Huber’s regression model, *Opt. Quantum Electron.* 51 (2019) 1–15. <https://doi.org/10.1007/s11082-019-1766-3>.
- [57] K.B. Beć, J. Grabska, C.W. Huck, Near-Infrared Spectroscopy in Bio-Applications, *Molecules.* 25 (2020) 2948. <https://doi.org/10.3390/molecules25122948>.
- [58] N. Shanas, W. Querido, A. Dumont, E. Yonko, E. Carter, J. Ok, J.P. Karchner, M.F. Barbe, S. Ali, C. Patil, C. Raggio, N. Pleshko, Clinical application of near infrared fiber optic spectroscopy for noninvasive bone assessment, *J. Biophotonics.* 13 (2020). <https://doi.org/10.1002/jbio.201960172>.
- [59] J.J. Roberts, A. Power, J. Chapman, S. Chandra, D. Cozzolino, *Vibrational Spectroscopy Methods for Agro-Food Product Analysis*, in: *Compr. Anal. Chem.*, Elsevier B.V., 2018: pp. 51–68. <https://doi.org/10.1016/bs.coac.2018.03.002>.



- [60] B.H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, John Wiley & Sons, Ltd, Chichester, UK, 2004.  
<https://doi.org/10.1002/0470011149>.
- [61] A. Peguero Gutiérrez, *La espectroscopia NIR en la determinación de propiedades físicas y composición química de intermedios de producción y productos acabados*, (2010). <https://doi.org/10.1109/PDCAT.2008.72>.
- [62] C. Pasquini, *Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications*, *J. Braz. Chem. Soc.* 14 (2003) 198–219.  
<https://doi.org/10.1590/S0103-50532003000200006>.
- [63] H.W. Siesler, *Near-Infrared Spectra, Interpretation*, 3rd ed., Elsevier Ltd., 2017. <https://doi.org/10.1016/B978-0-12-409547-2.12173-0>.
- [64] M. Blanco, I. Villarroya, *NIR spectroscopy: A rapid-response analytical tool*, *TrAC - Trends Anal. Chem.* 21 (2002) 240–250.  
[https://doi.org/10.1016/S0165-9936\(02\)00404-1](https://doi.org/10.1016/S0165-9936(02)00404-1).
- [65] W.F. McClure, *204 Years of near Infrared Technology: 1800–2003*, *J. Near Infrared Spectrosc.* Vol. 11, Issue 6, Pp. 487-518. 11 (2003) 487–518.  
<https://www.osapublishing.org/abstract.cfm?uri=jnirs-11-6-487> (accessed November 25, 2020).
- [66] Y. Ozaki, T. Genkawa, Y. Futami, *Near-Infrared Spectroscopy*, 3rd ed., Elsevier Ltd., 2017. <https://doi.org/10.1016/B978-0-12-409547-2.12164-X>.
- [67] Q. Zhao, X. Lv, Y. Jia, Y. Chen, G. Xu, L. Qu, *Rapid determination of the fat, moisture, and protein contents in homogenized chicken eggs based on near-infrared reflectance spectroscopy*, *Poult. Sci.* 97 (2018) 2239–2245.  
<https://doi.org/10.3382/ps/pey070>.
- [68] Z. Zhu, S. Chen, X. Wu, C. Xing, J. Yuan, *Determination of soybean routine*

- quality parameters using near-infrared spectroscopy, *Food Sci. Nutr.* 6 (2018) 1109–1118. <https://doi.org/10.1002/fsn3.652>.
- [69] N. Prieto, O. Pawluczyk, M.E.R. Dugan, J.L. Aalhus, A Review of the Principles and Applications of Near-Infrared Spectroscopy to Characterize Meat, Fat, and Meat Products, *Appl. Spectrosc.* 71 (2017) 1403–1426. <https://doi.org/10.1177/0003702817709299>.
- [70] D.P. Perez, M.T. Sanchez, G. Cano, A. Garrido, Authentication of green asparagus varieties by near-infrared reflectance spectroscopy, *J. Food Sci.* 66 (2001) 323–327. <https://doi.org/10.1111/j.1365-2621.2001.tb11340.x>.
- [71] C. Yang, W.S. Lee, J.G. Williamson, Classification of blueberry fruit and leaves based on spectral signatures, *Biosyst. Eng.* 113 (2012) 351–362. <https://doi.org/10.1016/j.biosystemseng.2012.09.009>.
- [72] Z. Long, W. Shan-shan, D. Yan-fei, P. Jia-rong, Z. Cheng, Discrimination of Transgenic Rice Based on Near Infrared Reflectance Spectroscopy and Partial Least Squares Regression Discriminant Analysis, *Rice Sci.* 22 (2015) 245–249. <https://doi.org/10.1016/j.rsci.2015.09.004>.
- [73] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, *TrAC - Trends Anal. Chem.* 28 (2009) 1201–1222. <https://doi.org/10.1016/j.trac.2009.07.007>.
- [74] Y. Garini, I.T. Young, G. McNamara, Spectral imaging: Principles and applications, *Cytom. Part A.* 69A (2006) 735–747. <https://doi.org/10.1002/cyto.a.20311>.
- [75] A.F.H. Goetz, Three decades of hyperspectral remote sensing of the Earth: A personal view, *Remote Sens. Environ.* 113 (2009) S5–S16. <https://doi.org/10.1016/j.rse.2007.12.014>.

- [76] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, J.J. Sousa, Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry, *Remote Sens.* 9 (2017) 1110. <https://doi.org/10.3390/rs9111110>.
- [77] A.F.H. Goetz, G. Vane, J.E. Solomon, B.N. Rock, Imaging spectrometry for earth remote sensing, *Science* (80-. ). 228 (1985) 1147–1153. <https://doi.org/10.1126/science.228.4704.1147>.
- [78] A. Barreto, J.P. Cruz-Tirado, R. Siche, R. Quevedo, Determination of starch content in adulterated fresh cheese using hyperspectral imaging, *Food Biosci.* 21 (2018) 14–19. <https://doi.org/10.1016/j.fbio.2017.10.009>.
- [79] M. Kamruzzaman, Y. Makino, S. Oshita, S. Liu, Assessment of Visible Near-Infrared Hyperspectral Imaging as a Tool for Detection of Horsemeat Adulteration in Minced Beef, *Food Bioprocess Technol.* 8 (2015) 1054–1062. <https://doi.org/10.1007/s11947-015-1470-7>.
- [80] Y. Liu, H. Pu, D.W. Sun, Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications, *Trends Food Sci. Technol.* 69 (2017) 25–35. <https://doi.org/10.1016/j.tifs.2017.08.013>.
- [81] J. Qin, M. Kim, K. Chao, D. Chan, S. Delwiche, B.-K. Cho, Line-Scan Hyperspectral Imaging Techniques for Food Safety and Quality Applications, *Appl. Sci.* 7 (2017) 125. <https://doi.org/10.3390/app7020125>.
- [82] M. Halicek, H. Fabelo, S. Ortega, G.M. Callico, B. Fei, In-Vivo and Ex-Vivo Tissue Analysis through Hyperspectral Imaging Techniques: Revealing the Invisible Features of Cancer, *Cancers (Basel)*. 11 (2019) 756. <https://doi.org/10.3390/cancers11060756>.

## Bibliografía

- [83] B. Fei, Hyperspectral imaging in medical applications, in: *Data Handl. Sci. Technol.*, Elsevier Ltd, 2020: pp. 523–565. <https://doi.org/10.1016/B978-0-444-63977-6.00021-3>.
- [84] Stuart, McGonigle, Willmott, Hyperspectral Imaging in Environmental Monitoring: A Review of Recent Developments and Technological Advances in Compact Field Deployable Systems, *Sensors*. 19 (2019) 3071. <https://doi.org/10.3390/s19143071>.
- [85] J. Nieke, M. Rast, Towards the copernicus hyperspectral imaging mission for the environment (CHIME), in: *Int. Geosci. Remote Sens. Symp.*, Institute of Electrical and Electronics Engineers Inc., 2018: pp. 157–159. <https://doi.org/10.1109/IGARSS.2018.8518384>.
- [86] D. Caballero, R. Calvini, J.M. Amigo, Hyperspectral imaging in crop fields: precision agriculture, in: *Data Handl. Sci. Technol.*, Elsevier Ltd, 2020: pp. 453–473. <https://doi.org/10.1016/B978-0-444-63977-6.00018-3>.
- [87] W.H. Maes, K. Steppe, Perspectives for Remote Sensing with Unmanned Aerial Vehicles in Precision Agriculture, *Trends Plant Sci*. 24 (2019) 152–164. <https://doi.org/10.1016/j.tplants.2018.11.007>.
- [88] S. Thomas, M.T. Kuska, D. Bohnenkamp, A. Brugger, E. Alisaac, M. Wahabzada, J. Behmann, A.K. Mahlein, Benefits of hyperspectral imaging for plant disease detection and plant protection: a technical perspective, *J. Plant Dis. Prot.* 125 (2018) 5–20. <https://doi.org/10.1007/s41348-017-0124-6>.
- [89] P. Moghadam, D. Ward, E. Goan, S. Jayawardena, P. Sikka, E. Hernandez, Plant disease detection using hyperspectral imaging, in: *DICTA 2017 - 2017 Int. Conf. Digit. Image Comput. Tech. Appl.*, Institute of Electrical and Electronics Engineers Inc., 2017: pp. 1–8. <https://doi.org/10.1109/DICTA.2017.8227476>.

- [90] F.M. Lacar, M.M. Lewis, I.T. Grierson, Use of hyperspectral imagery for mapping grape varieties in the Barossa Valley, South Australia, IGARSS 2001. Scanning Present Resolv. Futur. Proceedings. IEEE 2001 Int. Geosci. Remote Sens. Symp. (Cat. No.01CH37217). 6 (2001) 2875–2877. <https://doi.org/10.1109/IGARSS.2001.978191>.
- [91] O. Rudjord, O.D. Trier, Tree species classification with hyperspectral imaging and lidar, in: Work. Hyperspectral Image Signal Process. Evol. Remote Sens., IEEE Computer Society, 2016. <https://doi.org/10.1109/WHISPERS.2016.8071665>.
- [92] S. Thomas, M.T. Kuska, D. Bohnenkamp, A. Brugger, E. Alisaac, M. Wahabzada, J. Behmann, A.K. Mahlein, Benefits of hyperspectral imaging for plant disease detection and plant protection: a technical perspective, J. Plant Dis. Prot. 125 (2018) 5–20. <https://doi.org/10.1007/s41348-017-0124-6>.
- [93] A. Robles-Kelly, C.P. Huynh, Imaging Spectroscopy for Scene Analysis, Springer London, London, 2013. <https://doi.org/10.1007/978-1-4471-4652-0>.
- [94] D. Wu, D.W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review - Part I: Fundamentals, Innov. Food Sci. Emerg. Technol. 19 (2013) 1–14. <https://doi.org/10.1016/j.ifset.2013.04.014>.
- [95] F. Chollet, Deep Learning with Python, Manning Publications, 2018.
- [96] J. Torres, Deep Learning. Introducción práctica con Keras (Primera parte), (2018) 1–207. <https://torres.ai/deep-learning-inteligencia-artificial-keras/> (accessed February 3, 2021).
- [97] N. Buduma, N. Locascio, Fundamentals of Deep Learning, 2017.
- [98] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444.

## Bibliografía

<https://doi.org/10.1038/nature14539>.

- [99] J. Torres, Deep Learning. Introducción práctica con Keras (Segunda parte), (2019) 1–175. <https://torres.ai/deep-learning-inteligencia-artificial-keras-2a-parte/#AccesoLibro2aParte> (accessed February 3, 2021).
- [100] Francois Chollet, Keras: the Python deep learning API, Dropout Layer. (2015). [https://keras.io/api/layers/regularization\\_layers/dropout/](https://keras.io/api/layers/regularization_layers/dropout/) (accessed February 3, 2021).
- [101] S.J. Pan, Q. Yang, A survey on transfer learning, *IEEE Trans. Knowl. Data Eng.* 22 (2010) 1345–1359. <https://doi.org/10.1109/TKDE.2009.191>.
- [102] M. Mehdipour Ghazi, B. Yanikoglu, E. Aptoula, Plant identification using deep neural networks via optimization of transfer learning parameters, *Neurocomputing*. 235 (2017) 228–235. <https://doi.org/10.1016/j.neucom.2017.01.018>.
- [103] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 Int. Interdiscip. PhD Work. IIPhDW 2018, Institute of Electrical and Electronics Engineers Inc., 2018: pp. 117–122. <https://doi.org/10.1109/IIPHDW.2018.8388338>.
- [104] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2014). <http://www.robots.ox.ac.uk/> (accessed March 19, 2020).
- [105] ImageNet Large Scale Visual Recognition Competition 2014 (ILSVRC2014), (n.d.). <http://www.image-net.org/challenges/LSVRC/2014/> (accessed May 7, 2020).
- [106] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale

- Visual Recognition Challenge, *Int. J. Comput. Vis.* 115 (2015) 211–252.  
<https://doi.org/10.1007/s11263-015-0816-y>.
- [107] J.N. Miller, J.C. Miller, *Statistics and Chemometrics for Analytical Chemistry* Sixth edition, 6th ed., Prentice Hall, 2010. [www.pearsoned.co.uk/Miller](http://www.pearsoned.co.uk/Miller) (accessed December 2, 2020).
- [108] D. Massart, B. Vandeginste, L. Buydens, P. Lewi, *Handbook of chemometrics and qualimetrics*, (1998). <https://dl.acm.org/doi/abs/10.5555/552536> (accessed November 26, 2020).
- [109] B.R. Kowalski, P.C. Jurs, T.L. Isenhour, C.N. Reilley, T.L. Isenhour, P.C. Jurs, *Computerized Learning Machines Applied to Chemical Problems Interpretation of Infrared Spectrometry Data*, *Anal. Chem.* 41 (1969) 1945–1949. <https://doi.org/10.1021/ac50159a026>.
- [110] P.C. Jurs, B.R. Kowalski, T.L. Isenhour, C.N. Reilley, *Computerized Learning Machines Applied to Chemical Problems Molecular Structure Parameters from Low Resolution Mass Spectrometry*, *Anal. Chem.* 42 (1970) 1387–1394. <https://doi.org/10.1021/ac60294a015>.
- [111] P.C. Jurs, B.R. Kowalski, T.L. Isenhour, C.N. Reilley, *Computerized Learning Machines Applied to Chemical Problems. Investigation of Convergence Rate and Predictive Ability of Adaptive Binary Pattern Classifiers*, *Anal. Chem.* 41 (1969) 690–695. <https://doi.org/10.1021/ac60275a025>.
- [112] P.C. Jurs, B.R. Kowalski, T.L. Isenhour, *Computerized Learning Machines Applied to Chemical Problems Molecular Formula Determination from Low Resolution Mass Spectrometry*, *Anal. Chem.* 41 (1969) 21–27. <https://doi.org/10.1021/ac60270a002>.
- [113] B.R. Kowalski, P.C. Jurs, T.L. Isenhour, C.N. Reilley, *Computerized Learning*

- Machines Applied to Chemical Problems. Multicategory Pattern Classification by Least Squares, *Anal. Chem.* 41 (1969) 695–700.  
<https://doi.org/10.1021/ac60275a026>.
- [114] S. Wold, Spline Functions in Data Analysis, *Technometrics*. 16 (1974) 1–11.  
<https://doi.org/10.1080/00401706.1974.10489142>.
- [115] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra, *J. Pharm. Biomed. Anal.* 21 (1999) 115–132.  
<https://doi.org/10.1080/13554790490495140>.
- [116] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, *J. Pharm. Biomed. Anal.* 44 (2007) 683–700.  
<https://doi.org/10.1016/j.jpba.2007.03.023>.
- [117] M.Z. Zhu, B. Wen, H. Wu, J. Li, H. Lin, Q. Li, Y. Li, J. Huang, Z. Liu, The Quality Control of Tea by Near-Infrared Reflectance (NIR) Spectroscopy and Chemometrics, *J. Spectrosc.* 2019 (2019).  
<https://doi.org/10.1155/2019/8129648>.
- [118] E. Teye, E. Anyidoho, R. Agbemaflé, L.K. Sam-Amoah, C. Elliott, Cocoa bean and cocoa bean products quality evaluation by NIR spectroscopy and chemometrics: A review, *Infrared Phys. Technol.* 104 (2020) 103127.  
<https://doi.org/10.1016/j.infrared.2019.103127>.
- [119] T. Næs, T. Isaksson, Different Strategies for Handling Non-Linearity Problems in NIR Calibration, *NIR News*. 5 (1994) 4–11.  
<https://doi.org/10.1255/nirn.227>.
- [120] M. Esteki, J. Simal-Gandara, Z. Shahsavari, S. Zandbaaf, E. Dashtaki, Y. Vander



- Heyden, A review on the application of chromatographic methods, coupled to chemometrics, for food authentication, *Food Control*. 93 (2018) 165–182. <https://doi.org/10.1016/j.foodcont.2018.06.015>.
- [121] N.P. Kalogiouri, V.F. Samanidou, Liquid chromatographic methods coupled to chemometrics: a short review to present the key workflow for the investigation of wine phenolic composition as it is affected by environmental factors, *Environ. Sci. Pollut. Res.* (2020) 1–15. <https://doi.org/10.1007/s11356-020-09681-5>.
- [122] A. Kanginejad, A. Mani-Varnosfaderani, Chemometrics advances on the challenges of the gas chromatography–mass spectrometry metabolomics data: a review, *J. Iran. Chem. Soc.* 15 (2018) 2733–2745. <https://doi.org/10.1007/s13738-018-1461-5>.
- [123] Z. Liu, Y. Yuan, Y. Zhang, Y. Shi, G. Hu, J. Zhu, K.M. Rogers, Geographical traceability of Chinese green tea using stable isotope and multi-element chemometrics, *Rapid Commun. Mass Spectrom.* 33 (2019) 778–788. <https://doi.org/10.1002/rcm.8405>.
- [124] L. Du, W. Lu, Z. (Julia) Cai, L. Bao, C. Hartmann, B. Gao, L. (Lucy) Yu, Rapid detection of milk adulteration using intact protein flow injection mass spectrometric fingerprints combined with chemometrics, *Food Chem.* 240 (2018) 573–578. <https://doi.org/10.1016/j.foodchem.2017.07.107>.
- [125] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Appl. Spectrosc.* 43 (1989) 772–777. <https://doi.org/10.1366/0003702894202201>.
- [126] D. MacDougall, H. Martens, P. Geladi, Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat, *Appl. Spectrosc.* Vol. 39, Issue 3, Pp. 491-500. 39 (1985) 491–500.

- <https://www.osapublishing.org/as/abstract.cfm?uri=as-39-3-491> (accessed December 29, 2018).
- [127] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, *Chemometrics Tutorial for PLS \_ Toolbox and Solo*, 2006.
- [128] K. Norris, P. Williams, Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size., *Cereal Chem.* 61 (1984) 158–165.  
[http://online.cerealsgrains.org/publications/cc/backissues/1984/Documents/Chem61\\_158.pdf](http://online.cerealsgrains.org/publications/cc/backissues/1984/Documents/Chem61_158.pdf) (accessed October 2, 2020).
- [129] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Anal. Chem.* 36 (1964) 1627–1639.  
<https://doi.org/10.1021/ac60214a047>.
- [130] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9).
- [131] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, *Anal. Methods.* 5 (2013) 3790–3798.  
<https://doi.org/10.1039/c3ay40582f>.
- [132] K.H. Esbensen, D. Guyot, F. Westad, L.P. Houmoller, *Multivariate Data Analysis - in Practice: An Introduction to Multivariate Data Analysis and Experimental Design*, 5th ed., AS, Camo Process, Oslo. Norway, 2002.
- [133] S. Wold, Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models, *Technometrics.* 20 (1978) 397–405.  
<https://doi.org/10.1080/00401706.1978.10489693>.
- [134] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: Taking

- the magic away, *J. Chemom.* 28 (2014) 213–225.  
<https://doi.org/10.1002/cem.2609>.
- [135] LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S, *Philos. Trans. R. Soc. London.* 53 (1763) 370–418.  
<https://doi.org/10.1098/rstl.1763.0053>.
- [136] Q. Fan, Y. Wang, P. Sun, S. Liu, Y. Li, Discrimination of Ephedra plants with diffuse reflectance FT-NIRS and multivariate analysis, *Talanta.* 80 (2010) 1245–1250. <https://doi.org/10.1016/j.talanta.2009.09.018>.
- [137] F.M. Durgante, N. Higuchi, A. Almeida, A. Vicentini, Species Spectral Signature: Discriminating closely related plant species in the Amazon with Near-Infrared Leaf-Spectroscopy, *For. Ecol. Manage.* 291 (2013) 240–248.  
<https://doi.org/10.1016/j.foreco.2012.10.045>.
- [138] D.H. Card, D.L. Peterson, P.A. Matson, J.D. Aber, Prediction of leaf chemistry by the use of visible and near infrared reflectance spectroscopy, *Remote Sens. Environ.* 26 (1988) 123–147. [https://doi.org/10.1016/0034-4257\(88\)90092-2](https://doi.org/10.1016/0034-4257(88)90092-2).
- [139] M. Petit Bon, H. Böhner, S. Kaino, T. Moe, K. Anne Bråthen, C. Kari Anne Bråthen, One leaf for all: Chemical traits of single leaves measured at the leaf surface using near-infrared reflectance spectroscopy, *Methods Ecol Evol.* 11 (2020) 1061–1071. <https://doi.org/10.1111/2041-210X.13432>.
- [140] I.E. Renner, V.A. Fritz, Using Near-infrared reflectance spectroscopy (NIRS) to predict glucobrassicin concentrations in cabbage and brussels sprout leaf tissue, *Plant Methods.* 16 (2020) 136. <https://doi.org/10.1186/s13007-020-00681-7>.

## Bibliografía

- [141] S. Muhammad, B.L. Sanden, B.D. Lampinen, D.R. Smart, S. Saa, K.A. Shackel, P.H. Brown, Nutrient Storage in the Perennial Organs of Deciduous Trees and Remobilization in Spring – A Study in Almond (*Prunus dulcis*) (Mill.) D. A. Webb, *Front. Plant Sci.* 11 (2020). <https://doi.org/10.3389/fpls.2020.00658>.
- [142] L.M. Da Ros, S.D. Mansfield, Biotechnological mechanism for improving plant remobilization of phosphorus during leaf senescence, *Plant Biotechnol. J.* 18 (2020) 470–478. <https://doi.org/10.1111/pbi.13212>.
- [143] F.S. Bonora, H.F. Nahrung, R.A. Hayes, T. Scharaschkin, G. Pegg, D.J. Lee, Changes in leaf chemistry and anatomy of *Corymbia citriodora* subsp. *variegata* (Myrtaceae) in response to native and exotic pathogens, *Australas. Plant Pathol.* 49 (2020) 641–653. <https://doi.org/10.1007/s13313-020-00739-z>.
- [144] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data, *Bioinformatics.* 21 (2005) 3043–3048. <https://doi.org/10.1093/bioinformatics/bti476>.
- [145] J.J. Jansen, H.C.J. Hoefsloot, J. Van Der Greef, M.E. Timmerman, J.A. Westerhuis, A.K. Smilde, ASCA: Analysis of multivariate data obtained from an experimental design, *J. Chemom.* 19 (2005) 469–481. <https://doi.org/10.1002/cem.952>.
- [146] F. Rodríguez, M. Berenguel, J.L. Guzmán, A. Ramírez-arias, *Modeling and Control of Greenhouse Crop Growth.*, 1st ed., Springer International Publishing, Cham, 2015. <https://doi.org/10.1007/978-3-319-11134-6>.
- [147] C. Mees, F. Souard, C. Delporte, E. Deconinck, P. Stoffelen, C. Stévigny, J.M. Kauffmann, K. De Braekeleer, Identification of coffee leaves using FT-NIR spectroscopy and SIMCA, *Talanta.* 177 (2018) 4–11.

<https://doi.org/10.1016/j.talanta.2017.09.056>.

- [148] M. Maimaitiyiming, A.J. Miller, A. Ghulam, Discriminating Spectral Signatures Among and Within Two Closely Related Grapevine Species, *Photogramm. Eng. Remote Sens.* 82 (2016) 51–62. <https://doi.org/10.14358/pers.82.2.51>.
- [149] Z. Wu, E. Xu, F. Wang, J. Long, X.X.A. Jiao, Z. Jin, Rapid Determination of Process Variables of Chinese Rice Wine Using FT-NIR Spectroscopy and Efficient Wavelengths Selection Methods, *Food Anal. Methods.* 8 (2015) 1456–1467. <https://doi.org/10.1007/s12161-014-0021-6>.
- [150] J.-B. Feret, G.P. Asner, Tree Species Discrimination in Tropical Forests Using Airborne Imaging Spectroscopy, *IEEE Trans. Geosci. Remote Sens.* 51 (2013) 73–84. <https://doi.org/10.1109/TGRS.2012.2199323>.
- [151] S.K. Meerdink, D.A. Roberts, K.L. Roth, J.Y. King, P.D. Gader, A. Koltunov, Classifying California plant species temporally using airborne hyperspectral imagery, *Remote Sens. Environ.* 232 (2019) 111308. <https://doi.org/10.1016/j.rse.2019.111308>.
- [152] J. Lopatin, F.E. Fassnacht, T. Kattenborn, S. Schmidtlein, Mapping plant species in mixed grassland communities using close range imaging spectroscopy, *Remote Sens. Environ.* 201 (2017) 12–23. <https://doi.org/10.1016/j.rse.2017.08.031>.
- [153] M.A. Cochrane, Using vegetation reflectance variability for species level classification of hyperspectral data, *Int. J. Remote Sens.* 21 (2000) 2075–2087. <https://doi.org/10.1080/01431160050021303>.
- [154] I. Sobhan, Species discrimination from a hyperspectral perspective, 2007. [http://www.itc.nl/library/papers\\_2007/phd/sobhan.pdf](http://www.itc.nl/library/papers_2007/phd/sobhan.pdf) (accessed May 1, 2021).

## Bibliografía

- [155] I. Filella, J. Peñuelas, The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status, *Int. J. Remote Sens.* 15 (1994) 1459–1470. <https://doi.org/10.1080/01431169408954177>.
- [156] D. Li, T. Cheng, K. Zhou, H. Zheng, X. Yao, Y. Tian, Y. Zhu, W. Cao, WREP: A wavelet-based technique for extracting the red edge position from reflectance spectra for estimating leaf and canopy chlorophyll contents of cereal crops, *ISPRS J. Photogramm. Remote Sens.* 129 (2017) 103–117. <https://doi.org/10.1016/j.isprsjprs.2017.04.024>.
- [157] F. Boochs, G. Kupfer, K. Dockter, W. Kuhbaüch, Shape of the red edge as vitality indicator for plants, *Int. J. Remote Sens.* 11 (1990) 1741–1753. <https://doi.org/10.1080/01431169008955127>.
- [158] T. Zheng, N. Liu, L. Wu, M. Li, H. Sun, Q. Zhang, J. Wu, Estimation of Chlorophyll Content in Potato Leaves Based on Spectral Red Edge Position, *IFAC-PapersOnLine.* 51 (2018) 602–606. <https://doi.org/10.1016/j.ifacol.2018.08.131>.
- [159] W. Collins, REMOTE SENSING OF CROP TYPE AND MATURITY., *Photogramm. Eng. Remote Sensing.* 44 (1978) 42–55.
- [160] D.N.H. Horler, M. Dockray, J. Barber, The red edge of plant leaf reflectance, *Int. J. Remote Sens.* 4 (1983) 273–288. <https://doi.org/10.1080/01431168308948546>.
- [161] P.J. Zarco-Tejada, C. Camino, P.S.A. Beck, R. Calderon, A. Hornero, R. Hernández-Clemente, T. Kattenborn, M. Montes-Borrego, L. Susca, M. Morelli, V. Gonzalez-Dugo, P.R.J. North, B.B. Landa, D. Boscia, M. Saponari, J.A. Navas-Cortes, Previsual symptoms of *Xylella fastidiosa* infection revealed in spectral plant-trait alterations, *Nat. Plants.* 4 (2018) 432–439. <https://doi.org/10.1038/s41477-018-0189-7>.

- [162] S. Mahesh, D.S. Jayas, J. Paliwal, N.D.G. White, Hyperspectral imaging to classify and monitor quality of agricultural materials, *J. Stored Prod. Res.* 61 (2015) 17–26. <https://doi.org/10.1016/j.jspr.2015.01.006>.
- [163] H. Zhang, J. Paliwal, D. S. Jayas, N. D. G. White, Classification of Fungal Infected Wheat Kernels Using Near-Infrared Reflectance Hyperspectral Imaging and Support Vector Machine, *Trans. ASABE.* 50 (2007) 1779–1785. <https://doi.org/10.13031/2013.23935>.
- [164] H. Zhu, B. Chu, C. Zhang, F. Liu, L. Jiang, Y. He, Hyperspectral Imaging for Presymptomatic Detection of Tobacco Disease with Successive Projections Algorithm and Machine-learning Classifiers, *Sci. Rep.* 7 (2017) 1–12. <https://doi.org/10.1038/s41598-017-04501-2>.
- [165] S. Thomas, J. Behmann, A. Steier, T. Kraska, O. Muller, U. Rascher, A.K. Mahlein, Quantitative assessment of disease severity and rating of barley cultivars based on hyperspectral imaging in a non-invasive, automated phenotyping platform, *Plant Methods.* 14 (2018) 1–12. <https://doi.org/10.1186/s13007-018-0313-8>.
- [166] W.H. Yang, S. Peng, J. Huang, A.L. Sanico, R.J. Buresh, C. Witt, Using leaf color charts to estimate leaf nitrogen status of rice, in: *Agron. J.*, 2003: pp. 212–217. <https://doi.org/10.2134/agronj2003.2120>.
- [167] G. Xu, F. Zhang, S.G. Shah, Y. Ye, H. Mao, Use of leaf color images to identify nitrogen and potassium deficient tomatoes, *Pattern Recognit. Lett.* 32 (2011) 1584–1590. <https://doi.org/10.1016/j.patrec.2011.04.020>.
- [168] Y. Sun, C. Tong, S. He, K. Wang, L. Chen, Identification of Nitrogen, Phosphorus, and Potassium Deficiencies Based on Temporal Dynamics of Leaf Morphology and Color, *Sustainability.* 10 (2018) 762. <https://doi.org/10.3390/su10030762>.

## Bibliografía

- [169] G.L. Grinblat, L.C. Uzal, M.G. Larese, P.M. Granitto, Deep learning for plant identification using vein morphological patterns, *Comput. Electron. Agric.* 127 (2016) 418–424. <https://doi.org/10.1016/j.compag.2016.07.003>.
- [170] M.G. Larese, R. Namías, R.M. Craviotto, M.R. Arango, C. Gallo, P.M. Granitto, Automatic classification of legumes using leaf vein image features, *Pattern Recognit.* 47 (2014) 158–168. <https://doi.org/10.1016/j.patcog.2013.06.012>.
- [171] J.W. Tan, S.W. Chang, S. Abdul-Kareem, H.J. Yap, K.T. Yong, Deep Learning for Plant Species Classification Using Leaf Vein Morphometric, *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17 (2020) 82–90. <https://doi.org/10.1109/TCBB.2018.2848653>.
- [172] M.G. Larese, R.M. Craviotto, M.R. Arango, C. Gallo, P.M. Granitto, Legume identification by leaf vein images classification, in: *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, Springer, Berlin, Heidelberg, 2012: pp. 447–454. [https://doi.org/10.1007/978-3-642-33275-3\\_55](https://doi.org/10.1007/978-3-642-33275-3_55).
- [173] S.P. Mohanty, D.P. Hughes, M. Salathé, Using deep learning for image-based plant disease detection, *Front. Plant Sci.* 7 (2016) 1–10. <https://doi.org/10.3389/fpls.2016.01419>.
- [174] S.S. Chouhan, U.P. Singh, S. Jain, Applications of Computer Vision in Plant Pathology: A Survey, *Arch. Comput. Methods Eng.* 27 (2020) 611–632. <https://doi.org/10.1007/s11831-019-09324-0>.
- [175] G. Wang, Y. Sun, J. Wang, Automatic Image-Based Plant Disease Severity Estimation Using Deep Learning, *Comput. Intell. Neurosci.* 2017 (2017) 8. <https://doi.org/10.1155/2017/2917536>.
- [176] Y.A. Nanekaran, D. Zhang, J. Chen, Y. Tian, N. Al-Nabhan, Recognition of



- plant leaf diseases based on computer vision, *J. Ambient Intell. Humaniz. Comput.* 1 (2020) 1–18. <https://doi.org/10.1007/s12652-020-02505-x>.
- [177] G. Polder, E. Pekkeriet, M. Snickers, A Spectral Imaging System for Detection of Botrytis in Greenhouses, *Proc. EFITA-WCCA-CIGR Conf. Sustain. Agric. through ICT Innov.* 23-27 June, 2013, Turin, Italy. (2013) 24–27. [www.gezondekas.eu](http://www.gezondekas.eu) (accessed January 15, 2022).
- [178] Technology Readiness Level | NASA, (n.d.). [https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt\\_accordion1.html](https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html) (accessed January 18, 2021).