# Early diagnosis of bacterial plant diseases based on proximal sensing from a precision agriculture perspective

Mafalda Alexandra Reis Pereira
Doctoral Program in Agrarian Sciences
Department of Geosciences, Environment and Spatial Plannings
2023

**Supervisor**
Mário Manuel de Miranda Furtado Campos Cunha, Associate Professor, Faculty of Science of the University of Porto

**Co-supervisors**
Fernando Manuel dos Santos Tavares, Associate Professor, Faculty of Science of the University of Porto

Filipe Baptista Neves dos Santos, Coordinator of TEC4AGRO-FOOD, Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)

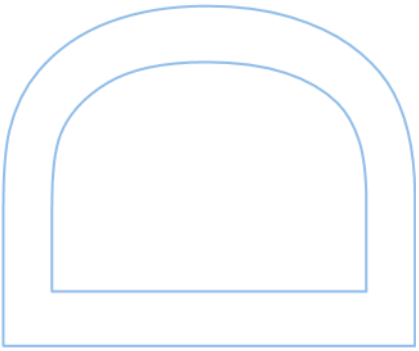

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

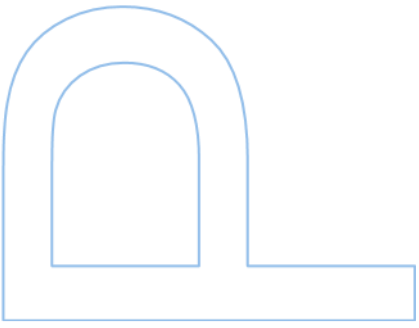
**Note:** This page should only be included in case the jury the thesis with recommendation of correction of errors or inaccuracies identified and expressly mentioned during the defence.

All the corrections determined by the jury, and only those, were made.

The Supervisor,

Porto, _____/_____/_____





# Preliminary note

During the thesis preparation, and according to the applicable terms, under n.º 2, Article 4.º of the General Regulations for the Third Cycles of Studies of the University of Porto (*Regulamento Geral dos Terceiros Ciclos de Estudos da Universidade do Porto*) and of Article 31.º of Law Decree (L.D. / *D.L. - Decreto Lei*) n.º 74/2006, of 24th March, with the new wording introduced by L.D. n.º 65/2018, of 16th August, a consistent set of research works, that are already published or submitted for publication in indexed international scientific journals and subject to peer review, was used and make part of the Chapter II and III of this thesis. It is important to note that these studies were carried out in collaboration with other authors, but the candidate clarifies that she played a main role in conceiving, obtaining, analyzing, and discussing the results, as well as in preparing the final published version of each one of them.

The present thesis was carried out under the supervision of Dr. Mário Cunha, Associate Professor at the Department of Geosciences, Environment and Spatial Planning of the Faculty of Sciences of the University of Porto and Researcher at the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência), under the co-supervision of Dr. Fernando Tavares, Associate Professor in the Biology Department of the Faculty of Sciences of the University of Porto and Researcher at the Biodiversity and Genetic Resources Research Center (CIBIO-InBio) and the co-supervision of Filipe Neves dos Santos, Researcher at the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC). Most of the work was carried out at the University of Porto and INESC TEC, which acted as host institutions, and in commercial orchards located in Guimarães (Portugal).

This work was supported by the Foundation for Science and Technology (FCT - *Fundação para a Ciência e Tecnologia*) through the doctoral grant SFRH/BD/146564/2019.

*“The mind adapts and converts to its own purposes the obstacle to our acting. The impediment to action advances action. What stands in the way becomes the way.”*

*— Marcus Aurelius, Meditations*

# Acknowledgments

O meu sincero e profundo agradecimento à minha família por todo o suporte e incentivo que me deram nos últimos quatro anos. Aos meus pais, Alberto e Maria Luísa, por todo o carinho, amor, ajuda e conselhos que me deram. Obrigado por serem os meus exemplos de força, amor e suporte. À minha irmã Erica por me ensinar sempre a perseguir os meus sonhos, mesmo que nem sempre seja simples. À minha irmã gémea Jéssica por demonstrar sempre otimismo, coragem e força mesmo nos momentos mais delicados e difíceis. És o exemplo de resiliência e bravura que quero seguir! Às minhas sobrinhas Deolinda, Amanda e Maria Luísa por me ensinarem a ver o mundo com mais curiosidade e espírito de aventura. Ao Bernardo por me ter ajudado a redescobrir a esperança, por me ensinar a ter confiança e a esquecer o medo, por compreender as minhas decisões e revelar a minha melhor 'versão'.

Ao meu orientador professor Mário Cunha pelos ensinamentos passados nestes últimos 10 anos; pela oportunidade e confiança oferecidas para me acompanhar neste regresso ao mundo académico; por me auxiliar na realização deste projeto. Aos meus coorientadores: professor Fernando Tavares por me revelar o maravilhoso mundo da microbiologia; por me ensinar a ser (ainda mais) exigente e rigorosa com o meu trabalho; por demonstrar o seu lado humano que tanto me inspirou a ser uma melhor investigadora. Ao Filipe Neves dos Santos pela pronta disponibilidade e alegria; pelos ensinamentos e pela paciência; por me demonstrar que é possível ser um excelente líder; criar e unir uma grande equipa.

À minha equipa TRIBE do INESCTEC (Alexandre Magno, André Aguiar, André Baltazar, Daniel Silva, Domingos Bento, Francisco Terra, Francisco Oliveira, Germano Moreira, Humberto Rocha, Héber Sobreira, Igor Portis, Isabel Pinheiro, José Sarmento, João Castro, Leandro Rodrigues, Luís Santos, Miguel Marques, Pedro Moura, Renan Tosin, Ricardo Neves, Rui Coutinho, Sandro Magalhães, Tatiana Pinho e Vítor Tinoco) obrigada pelos conhecimentos partilhados, pelo companheirismo e ajuda.

À minha equipa do Microbial Diversity Evolution (MDE, Kayla Silva, Leonor Martins, Sofia Martins) por todos os bons momentos, ensinamentos e apoio. À Cristiana Correia, Élia Fogeiro, Jessy Silva, João Prada, Mafalda Pinto, Maria João, Nuno Ponte, Paulo Pinto, Pedro Aguiar, Rafael Mendes, Rafaela Santos, Rita Fernandes, Rose Sousa, Sara Pinto, Sara Sario, e Telmo Vieira cada um de vós enriqueceu o meu percurso académico.

# Resumo

As doenças bacterianas nas plantas têm um impacto agronómico, ambiental, económico e humanitário significativo, o que justifica a procura e desenvolvimento de ferramentas para diagnósticos rigorosos e precoces (ou seja, antes do aparecimento de sintomas visuais ou aquando dos primeiros sinais de infeção). As doenças da pinta e mancha bacteriana do tomateiro (causadas, respetivamente, pela bactéria gram-negativa *Pseudomonas syringae* pv. *tomato*, Pst, e *Xanthomonas euvesicatoria*, Xeu), assim como o cancro bacteriano do kiwi (desencadeado por *Pseudomonas syringae* pv. *actinidiae*, Psa) são exemplos de distúrbios biológicos cuja gestão sustentável e proactiva é atualmente desafiadora. Estes agentes patogénicos são responsáveis por causar modificações nas características biofísicas, moleculares e estruturais das plantas hospedeiras, levando a alterações no seu comportamento espectral.

O presente trabalho tem como objetivo explorar a possibilidade de aplicar sensores de deteção foliar proximal, in vivo, para uma avaliação rápida e não destrutiva das características espectrais das plantas. Para além disso, também é investigada a possibilidade de detetar perfis espectrais não conformes, derivados de uma infeção bacteriana, que permitam estabelecer um diagnóstico indireto da doença. A possibilidade de discriminação do agente etiológico é também estudada.

Foram testados diferentes sensores hiperespectrais proximais, no âmbito deste trabalho, para providenciar um diagnóstico rápido, precoce e em tempo real de doenças nas culturas do tomate (herbáceas) e kiwi (lenhosas). Os ensaios de caso foram realizados em condições controladas (Porto, Portugal) e de campo (Guimarães, Portugal).

No **Capítulo I**, é apresentado o principal tópico desta tese, juntamente com a estrutura da mesma. O **Capítulo II** fornece o enquadramento do trabalho, abordando os principais conceitos, contexto e pertinência do tema. O **Caso de Estudo 1** é introduzido neste capítulo, tratando-se de uma revisão crítica que apresenta o estado da arte relacionado às técnicas atuais e inovadoras de deteção proximal realizadas para diagnosticar doenças nas plantas. Os resultados da pesquisa identificaram as principais metodologias atualmente utilizadas globalmente, bem como suas vantagens e principais limitações. Destaca-se a necessidade de métodos diagnósticos complementares. As abordagens de deteção proximal emergiram dos resultados como ferramentas adequadas para preencher esta lacuna de investigação. A sua descrição é feita, juntamente com a explicação da sua aplicação no diagnóstico de doenças nas plantas.

Adicionalmente, é fornecida uma breve contextualização da análise de dados para extração de informação, identificando diferentes soluções de *Machine Learning* (ML) e Quimiometria aplicadas na modelação de dados. São ainda identificadas métricas para avaliação e comparação dos diferentes modelos.

No **Capítulo III**, são apresentados os casos de estudo realizados no âmbito deste trabalho. O **Caso de Estudo 2** investigou a combinação de espectroscopia hiperespectral (de transmitância e de reflectância) com duas abordagens de modelação preditiva para a discriminação de doenças bacterianas em folhas de tomateiro e kiwi, adquiridas, respetivamente, em condições controladas e de campo. A primeira abordagem explorou a combinação de Índices de Vegetação (IVs), uma abordagem paramétrica amplamente utilizada para reduzir a dimensionalidade dos dados espectrais, com a *Flexible Discriminant Analysis* (FDA), um algoritmo de ML com um método de seleção de variáveis integrado. A segunda abordagem, por sua vez, investigou a ferramenta *Gaussian Process Classification Band Analysis Tool* (GPC-BAT), um algoritmo de ML supervisionado integrado no software ARTMO. Ambas as abordagens permitiram a identificação das diferentes classes em estudo (envolvendo classificação binária e multi-classe), utilizando os dois conjuntos de dados, no entanto, a ferramenta GPC-BAT apresentou melhores resultados. Para além disso, os comprimentos de onda nas regiões do azul (450 nm), verde (550 nm), *red-edge* (680 a 754 nm) e infravermelho próximo (NIR, 795 a 1000 nm) foram identificados como relevantes em ambos os casos de estudo deste artigo. Eles apresentam um significado biológico interessante, uma vez que coincidem com as regiões de absorção espectral de vários pigmentos fotossintéticos, da água e componentes estruturais das folhas. Todos esses compostos são afetados pela ação das bactérias Psa, Pst e Xeu em folhas de kiwi e tomateiro, respetivamente.

O **Caso de Estudo 3** investiga uma técnica *in situ* e não destrutiva para a discriminação do cancro bacteriano em folhas de kiwi usando reflectância hiperspectral e abordagens preditivas aplicadas. Diversas metodologias de Seleção de Variáveis (SV) e abordagens supervisionadas de ML foram avaliadas. Os resultados mostraram que as folhas não sintomáticas apresentaram o comportamento espectral característico da vegetação verde e fotossinteticamente ativa, enquanto as amostras sintomáticas revelaram desvios nas suas assinaturas espectrais nas regiões visíveis (VIS) e NIR. Diversas características espectrais localizadas nas regiões do azul (350-500 nm), verde (500-600 nm), vermelho (600-750 nm) e NIR (superior a 750 nm) foram destacadas pelas diferentes técnicas de SV. Todas as abordagens de classificação desenvolvidas puderam discriminar amostras não sintomáticas e sintomáticas, apoiando a

implementação de medições espectrais pontuais para a discriminação de doenças em culturas de campo.

O **Caso de Estudo 4** propõe uma metodologia para investigar o potencial de modelos preditivos baseados em pontos de medição hiperspectral (POM, transmitância) e ML para diagnóstico *in situ* e discriminação precoces da pinta e da mancha bacteriana do tomateiro. Um modelo de classificação multi-temporal (18 dias) foi desenvolvido, consistindo numa estratégia de pré-processamento de normalização vinculada a uma *Linear Discriminant Analysis* (LDA) visando a redução da dimensionalidade dos dados, e um algoritmo de ML supervisionado, *Support Vector Machines* (SVM) para modelação preditiva. Os resultados revelaram a aptidão do modelo para classificar corretamente tecidos saudáveis e doentes (inoculados com *Pseudomonas* spp. ou *Xanthomonas* spp.), mesmo em estados não sintomáticos (ou seja, quando apenas amostras sem sintomas visíveis foram consideradas). O modelo demonstrou, além disso, a capacidade de discriminar folíolos afetados por diferentes agentes patogénicos, tanto em estados não sintomáticos como em estados sintomáticos. Quarenta e quatro comprimentos de onda foram, para além disso, identificados como os mais relevantes, localizados principalmente nas regiões do azul-verde e vermelha do espectro eletromagnético (coincidentes com bandas de absorção de clorofila, carotenoides, compostos fenólicos e feofitinas).

Por fim, o **Caso de Estudo 5** explorou o uso de dados hiperespectrais de transmissão POM em combinação com o algoritmo *Data Driven Soft Independent Modeling of Class Analogy* (DD-SIMCA) para a avaliação precoce da pinta e mancha bacterianas do tomateiro em condições controladas, utilizando amostras saudáveis como classe alvo, e a *Multivariate Curve Resolution – Alternating Least Squares* (MCR-ALS) como meio para recuperar os perfis espectrais puros de tecidos saudáveis e de tecidos doentes. Os resultados da pesquisa demonstraram que o DD-SIMCA pode classificar amostras espectrais medidas em tecidos de folhas saudáveis (alvo) como amostras regulares. Além disso, esta abordagem de classificação baseada numa classe consegue classificar amostras espectrais medidas em folhas inoculadas com Pst e Xeu como não regulares, mesmo antes dos sintomas macroscópicos característicos destas doenças se tornem visíveis a olho nu. Os perfis espectrais das folhas saudáveis, inoculadas com Pst e inoculadas com Xeu também foram estimados e validados por meio do cálculo do MCR-ALS.

Os cinco casos de estudo apoiam uma abordagem integrada para a avaliação espectroscópica proximal in situ, não destrutiva, de doenças bacterianas em culturas

herbáceas e lenhosas, tanto em condições controladas quanto de campo. Os resultados do presente trabalho demonstram que os modelos baseados em dados de reflectância e transmitância hiperespectrais podem ser usados para distinguir tecidos saudáveis de tecidos doentes (mesmo em estados de infeção não sintomáticos) e para avaliar as variações fisiológicas que permitem a discriminação de diferentes agentes patogénicos.

Apesar destes resultados encorajadores, o presente trabalho reconhece que o *Technology Readiness Level* (TRL) destas abordagens ainda é baixo e deve ser melhorado. Da mesma forma, a falta de protocolos padronizados para a aquisição e modelação de dados hiperespectrais é abordada para uniformizar os processos de diagnóstico e reduzir o ruído e interferências espectrais indesejadas. São ainda recomendados mais estudos focados em diferentes interações hospedeiro-patógeno.

**Palavras-chave:** Deteção Proximal, Doença das Plantas, Modelação Classificativa, Cancro Bacteriano do Kiwi, Mancha Bacteriana do Tomateiro, Pinta Bacteriana do Tomateiro

# Abstract

Bacterial plant diseases have an important agronomic, environmental, economic, and humanitarian impact that justifies researching tools for rigorous and early diagnosis (i.e., prior to the appearance of visual symptoms or at the first signs of infection). Bacterial tomato speck and spot (caused, respectively, by the gram-negative bacteria *Pseudomonas syringae* pv. *tomato,* Pst, and *Xanthomonas euvesicatoria*, Xeu) and kiwi bacterial canker (triggered by *Pseudomonas syringae* pv. *actinidiae*, Psa) are examples of biological disorders whose sustainable and proactive management is currently challenging. These pathogens are responsible for causing modifications in hosts' biophysical, molecular, and structural characteristics, leading to modifications in plants' spectral behavior.

The present work aims to explore the suitability of in-vivo foliar proximal sensing for a quick and non-destructive assessment of plants' spectral traits, along with the identification of non-conform spectral profiles derived from bacterial infection, leading to an indirect disease diagnosis. Furthermore, the possibility of pathogen discrimination was also explored.

The research project tested different non-contact hyperspectral proximal sensors for a rapid, early, real-time disease diagnosis in tomato (herbaceous) and kiwi (woody) crops. The case studies were developed in controlled (Porto, Portugal), and field (Guimarães, Portugal) conditions.

In **Chapter I** the main topic of this thesis is introduced along with its outline, **Chapter II** provides the conceptual background of the research by addressing the main concepts, context, and pertinence of the theme, and introduces **Case Study 1**, a critical review presenting the state of the art related to current and innovative proximal sensing techniques performed to diagnose plant diseases. The research outcomes identified the main methodologies globally currently used, as well as their advantages and principal constraints. The necessity of complementary diagnostic methods is highlighted. Proximal sensing approaches emerged from the results as suitable tools for filling this gap. Their description is made, along with their application in plant disease diagnosis. In addition, a brief contextualization of data analysis for information extraction is provided, identifying different Machine Learning (ML) and Chemometric solutions applied in data modeling. Furthermore, metrics for model evaluation and comparison are identified.

In **Chapter III,** were introduced the case studies performed in the aim of this work. **Case Study 2** studied the combination of hyperspectral spectroscopy (transmittance and

reflectance) with two predictive modeling approaches for bacterial disease discrimination in tomato and kiwi leaves, acquired respectively in controlled and field conditions. The first approach explored the combination of Vegetation Indices (VIs), a widely used standard parametric approach suitable for reducing spectral data dimensionality, with a Flexible Discriminant Analysis (FDA), an ML algorithm with a built-in feature selection method. The second one, in turn, investigated the suitability of the Gaussian Process Classification with a Band Analysis Tool (GPC-BAT), an ARTMO software-supervised ML algorithm. Both approaches allowed the identification of the different classes in the study (the binary and multi-class), using the two datasets, but GPC-BAT showed better results. Furthermore, wavelengths in the blue (450 nm), green (550 nm), red-edge (680 to 754 nm), and near-infrared (NIR, 795 to 1000 nm) were identified as relevant in both case studies. They present an interesting biological significance since they coincide with the spectral absorption regions of several photosynthetic pigments, water content, and structural components of leaves. All these compounds are affected by the action of Psa, Pst, and Xeu bacteria in kiwi and tomato leaves, respectively.

**Case Study 3** investigates an in-situ, non-destructive technique for discrimination of bacterial canker on kiwi leaves using hyperspectral reflectance and applied predictive modeling approaches. Several Feature Selection (FS) methodologies and supervised ML approaches were evaluated. Outcomes showed that non-symptomatic leaves presented the characteristic spectral behavior of green and photosynthetically active vegetation, while symptomatic samples revealed deviations in their spectral signatures in the visible (VIS) and NIR regions. Several spectral features located in the blue (350–500 nm), green (500–600 nm), red (600–750 nm), and NIR (higher than 750 nm) regions were highlighted by the different FS techniques. All the developed classification approaches could discriminate non-symptomatic and symptomatic samples, supporting the implementation of spectral point measurements for in-field crop disease discrimination.

**Case Study 4** a methodology is proposed to investigate the potential of hyperspectral point-of-measurement (POM, transmittance) and ML-based predictive models for early in-situ diagnosis and discrimination of bacterial speck and spot of tomato. A multi-temporal (18 days) classification model was developed, consisting of a normalizing pre-processing strategy linked with a Linear Discriminant Analysis (LDA) aiming for data dimensionality reduction, and a supervised ML algorithm Support Vector Machines (SVM) for prediction. The outcomes revealed the model's fitness for correctly classifying healthy and diseased tissues (inoculated with *Pseudomonas* spp. and *Xanthomonas* spp.), even at non-symptomatic stages (i.e., when only samples without

visual symptoms were considered). The model showed, furthermore, the capacity of discriminating leaflets affected by different pathogens in both non-symptomatic and symptomatic stages. Forty-four spectral features were, moreover, identified as the most relevant being mainly located in the blue-green and red regions of the electromagnetic spectrum (coinciding with chlorophyll, carotenoids, phenolic compounds, and pheophytins absorption bands).

Lastly, **Case Study 5** explores the suitability of hyperspectral POM transmittance in combination with Data-Driven Soft Independent Modeling of Class Analogy (DD-SIMCA) for the early assessment of tomato bacterial speck and spot in controlled conditions, using healthy samples as a target class, and Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS) as a means to retrieve pure profiles of healthy and diseased tissues. The research outcomes demonstrated that DD-SIMCA can classify spectral samples collected in healthy leaflets' tissues (target) as regular samples. Furthermore, this one-class-classifier approach can categorize spectral assessments collected in diseased leaflets inoculated with both Pst and Xeu as non-regular, even before macroscopic symptoms and characteristics of these diseases become visible to the human eye. The spectral profiles of healthy, Pst and Xeu-diseased leaflets were also retrieved and validated through the computation of MCR-ALS.

The five articles support an integrated approach for the in-situ, non-destructive, proximal spectroscopy assessment of bacterial diseases in herbaceous and woody crops, both in controlled and field conditions. Our results demonstrate that both reflectance and transmittance hyperspectral-based models can be used for distinguishing healthy from diseased tissues (even at non-symptomatic stages) and for assessing physiological variations that allow pathogen discrimination. Despite these encouraging results, the present work recognizes that the Technology Readiness Level (TRL) of these approaches is still low and must be improved. Likewise, the lack of standardized protocols for hyperspectral data acquisition and modeling is addressed to uniformize the diagnosis processes and reduce noise and undesired spectral interferences. More research focused on different host-pathogen interactions is also suggested.

**Keywords:** Proximal Sensing, Plant Disease Diagnosis, Predictive Modelling, Kiwi Bacterial Canker, Tomato Bacterial Spot, Tomato Bacterial Speck

# Publications and communications

The following outputs were reached in the context of the present PhD thesis:

## Publications in international peer-reviewed journals

### Published

*Scientific research articles*

**Reis Pereira, M.**; Santos, F.N.d.; Tavares, F.; Cunha, M. Enhancing host-pathogen phenotyping dynamics: early detection of tomato bacterial diseases using hyperspectral point measurement and predictive modeling. Frontiers in Plant Science. 2023. 14:1242201. https://10.3389/fpls.2023.1242201. Research Topic: Advanced AI Methods for Plant Disease and Pest Recognition. Q1: Plant Sciences 27/238. JIF: 5.6, percentile 88.9.

**Reis-Pereira, M.**; Tosin, R.; Martins, R.; Neves dos Santos, F.; Tavares, F.; Cunha, M. Kiwi Plant Canker Diagnosis Using Hyperspectral Signal Processing and Machine Learning: Detecting Symptoms Caused by Pseudomonas syringae pv. actinidiae. Plants. 2022, 11, 2154. https://doi.org/10.3390/plants11162154. Special Issue: Detection and Diagnostics of Bacterial Plant Pathogens. Q1: Plant Sciences 43/238. JIF: 4.5, percentile 82.1.

*Conference proceedings (indexed in Scopus/WoS)*

**Reis Pereira, M.**; Tosin, R.; Martins, R.; Santos, F.N.d.; Tavares, F.; Cunha, M. Enhancing Kiwi Bacterial Canker Leaf Assessment: Integrating Hyperspectral-based Vegetation Indexes in Predictive Modeling. Engineering proceedings. 2023, 48, 22. https://doi.org/10.3390/CSAC2023-14920

**Reis-Pereira, M.**; Martins, R.C.; Silva, A.F.; Tavares, F.; Santos, F.; Cunha, M. Unravelling Plant-Pathogen Interactions: Proximal Optical Sensing as an Effective Tool for Early Detect Plant Diseases. Chemistry Proceedings. 2021, 5, 18. https://doi.org/10.3390/CSAC2021-10560

### Submitted

**Reis-Pereira, M.**; Tosin, R.; Verrelst, J.; Caicedo, J.; Tavares, F.; Santos, F. N. d.; Cunha, M. Plant disease diagnosis based on hyperspectral sensing: comparative analysis of parametric spectral indices and machine learning Gaussian process classification approaches.

**Reis-Pereira, M.**; Santos, F. N. d.; Tavares, F.; Cunha, M. Digital assessment of plant diseases: a critical review and analysis of optical sensing technologies for early plant disease diagnosis.

**Reis-Pereira, M.**; Mazivila, S. J.; Tavares, F.; Santos, F. N. d.; Cunha, M. Early plant disease diagnosis through hyperspectral point-of-measurement data coupled to DD-SIMCA as one-class classification and multivariate curve resolution

## Publications in national technical magazines

**Reis-Pereira, M.**; Tavares, F.; Neves dos Santos, F.; Cunha, M. Sensores óticos de proximidade para diagnóstico avançado das doenças das plantas, Agrotec, Suplemento "Inovação na FCUP em Saúde das Plantas - do Laboratório à Indústria", 2020.

## Oral communications in scientific conferences

**Reis-Pereira, M.**; Martins, R.; Silva, F.; Neves dos Santos, F.; Tavares, F.; Cunha, M. "Tracking changes on host physiological traits promoted by *Xanthomonas euvesicatoria*: proximal optical sensing as an innovative tool for plant disease detection". 4th Annual Conference of the EuroXanth COST Action, 2021.

**Reis-Pereira, M.**; Tavares, F.; Neves dos Santos, F.; Cunha, M. Diagnostics of bacterial plant diseases: proximal optical sensors as new tools for early detection. 4º Encontro Biologia Funcional e Biotecnologia de Plantas, Porto, 2020.

## Poster communications at scientific conferences

**Reis-Pereira, M.**; Tavares, F.; Neves dos Santos, F.; Cunha, M. Early assessment of tomato bacterial spot through proximal hyperspectral sensing. 14th European Conference on Precision Agriculture Unleashing the potential of Precision Agriculture (ECPA2023), 133, Bologna (Italy), 2-6 July 2023.

**Reis-Pereira, M.**; Tavares, F.; Neves dos Santos, F.; Cunha, M. A review on the main challenges in early diagnostics of plant diseases based on proximal sensing. II Plant Pests and Diseases Forum - Redefining Concepts, Mechanisms & Management Tools, 2021.

## Scientific datasets published

**Reis-Pereira, M.**; Santos, F.; Tavares, F.; Cunha, M. Hyperspectral spectroscopic transmittance data collected in-vivo healthy and diseased tomato leaflets in controlled conditions - dataset I. Zenodo. 2024. https://doi.org/10.5281/zenodo.10498387

**Reis-Pereira, M.**; Tavares, F.; Santos, F.; Cunha, M. Hyperspectral spectroscopic transmittance data collected in-vivo healthy and diseased tomato leaflets in controlled conditions - dataset II. Zenodo. 2024. https://doi.org/10.5281/zenodo.10498473

**Reis-Pereira, M.**; Tavares, F.; Santos, F.; Cunha, M. Hyperspectral spectroscopic reflectance data collected in-vivo non-symptomatic and symptomatic kiwi leaves in field conditions. Zenodo. 2024. https://doi.org/10.5281/zenodo.10498541

**Reis-Pereira, M.**; Martins, L.; Moura, P.; Tavares, F.; Santos, F.; Cunha, M. RGB images of healthy and bacterial-inoculated Tobacco plants captured under different LED light sources. Zenodo. 2024. https://doi.org/10.5281/zenodo.10515075

# Table of contents

# List of Tables

**Case Study 4**



**APPENDIX | SUPPLEMENTARY MATERIALS**

**Appendix B | Paper II**



**Appendix E | Poster presentation - 14th European Conference on Precision Agriculture Unleashing the potential of Precision Agriculture (ECPA2023)**

# List of Figures

**CHAPTER I | GENERAL INTRODUCTION**



**CHAPTER II | CONCEPTUAL FRAMEWORK**



**Case Study 1**

**Case Study 5**

**CHAPTER V | REMARQUES AND PERSPECTIVES**



**APPENDIX | SUPPLEMENTARY MATERIALS**

**Appendix A | Paper I**



**Appendix C | Oral communication - 4th Annual Conference of the EuroXanth COST Action**

# Abbreviations

| | |
|---|---|
| AVI | Ashburn Vegetation Index |
| BCK | Bacterial Canker of Kiwi |
| BRI | Browing Reflectance Index |
| BT | Briteiros |
| Chlgreen | Chlorophyll Green |
| CI | Coloration Index |
| CM | Confusion Matrix |
| CT | Caldas das Taipas |
| DD-SIMCA | Data-Driven Soft Independent Modeling of Class Analogy |
| FDA | Flexible Discriminant Analysis |
| FS | Feature Selection |
| FN | False Negative |
| FP | False Positive |
| GI | Simple Ratio Greenness Index |
| GLM | General Linear Model |
| LASSO | Lasso Regularized Generalized Linear Models |
| LDA | Linear Discriminant Analysis |
| MCR-ALS | Multivariate Curve Resolution – Alternating Least Squares |
| mSR | Modified Simple Ratio |
| NIR | Near-Infrared |
| PCA | Principal Component Analysis |
| PLS | Partial Least Squares |
| POM | Point of Measurement |
| PS | Proximal Sensing |
| PSA | *Pseudomonas syringae* pv. *actinidae* |
| PST | *Pseudomonas syringae* pv. *tomato* |
| PVIhyp | Hyperspectral Perpendicular Vegetation Index |
| Rre | Reflectance at the Inflexion Point |
| RTM | Radiative Transference Model |
| SFFS+JM | Sequential Forward Floating Selection Search Strategy and the Jeffries–Matusita Distance |
| SFVS | Stepwise Forward Variable Selection Method Using Wilk's Lambda Criterion |
| SVM | Support Vector Machines |
| SVM-L | Support Vector Machines with Linear kernel |

| | |
|---|---|
| SVM-LW | Support Vector Machines with Linear kernel with class weights |
| SVM-R | Support Vector Machines with Radial basis function kernel |
| SVM-RW | Support Vector Machines with Radial basis function kernel with class weights |
| TN | True Negative |
| TP | True Positive |
| TRL | Technology Readiness Level |
| UV | Ultraviolet |
| $VI_S$ | Vegetation Indexes |
| VIS | Visible |
| XEU | *Xanthomonas euvesicatoria* |

# Chapter I |
# General Introduction

# Introduction

On a global scale, pests and diseases contribute to yield losses in different crops that range from 20% to 40%, causing important economic impacts affecting diverse countries, and regions (Oerke, Dehne et al. 2012). Moreover, these biotic agents led to reduced income for producers, and restricted product availability, along with higher prices for consumers (Savary, Ficke et al. 2012, Savary, Bregaglio et al. 2017, Nelson 2020). The abundance and quality of fruits and vegetables, important sources of nutrients, are also affected because they are particularly vulnerable to diseases (due to current global breeding and agronomic practices) (Chakraborty and Newton 2011, Savary, Bregaglio et al. 2017).

For these reasons, plant pests and pathogens can affect a country's ability to import or export crops and their derived products around the world or even to move them within its borders. In fact, these transactions may provide pathways for the entry and spread of pathogenic microorganisms, also providing ideal conditions for pathogen adaptation and change (Macleod, Pautasso et al. 2010, Savary, Bregaglio et al. 2017).

Therefore, many agronomic, environmental, economic, and humanitarian reasons justify the non-destructive, early diagnosis of plant diseases, i.e. performed before or at the appearance of the first macroscopic characteristic lesions (symptoms) (Mahlein, Oerke et al. 2012, Martinelli, Scalenghe et al. 2014, Mahlein 2015). This proactive approach provides an opportunity for timely intervention, enabling effective control measures, and preventing infection's spread. Moreover, it allows for adjustments in crop management practices before the entire production site succumbs to an infection or incurs damage.

The non-destructive nature of the diagnosis process, furthermore, allows the identification of production areas affected by diseases, making possible the target application of phytosanitary products or other plant protective methods. This results in a more precise and efficient management approach and leads to the reduction of phytosanitary product usage, with a beneficial impact on the environment, ecosystem services, producer's income, and agricultural product quality (Lowe, Harrison et al. 2017). Hence, a Precision Agriculture perspective may be followed, i.e., a cropping strategy that gathers, processes, and analyzes temporal, individual, and spatial data to support management decisions according to estimated variability. Its main goal lies in improving resource use efficiency, productivity, quality, profitability, and sustainability of agricultural production (ISPA 2024).

Addressing the future challenges in Precision Agriculture involves overcoming key hurdles such as i) developing targeted and precise practices based on plant physiology, and ii) integrating them seamlessly with cutting-edge high-throughput (HTP) technologies based on the assessment of biophysical indicators of plant diseases, enabling early diagnosis. This integration contributes to the real-time application of laboratory information, enabling a spatio-temporal-functional approach to enhance Precision Agriculture practices. While this biophysical-integrative methodology is readily applicable in controlled laboratory environments, its translation to in-field agricultural applications remains a challenge. Globally, there is a growing application of HTP solutions in Precision Agriculture, with the establishment of sophisticated and costly HTP plant screening platforms that enable non-invasive measurements. New efforts are currently in progress to narrow the gap between laboratory feasibility and on-field implementation, aiming to fully unleash the potential of Precision Agriculture (Cunha, Martins et al. 2022).

The detection and identification (i.e., diagnosis) of plant pests and diseases constitutes, currently, a global key challenge in agriculture. The standard existing methods often are focused on direct crop scouting, aiming at the assessment of etiological agent's indicator signs, that are visible to the human eye. Despite its inherent importance, this technique may be time-consuming and demanding, depending on the crop type and production area size (which, in many commercial sites, is very large). Moreover, to be performed this approach mainly relies on the presence of visible symptoms caused by the pathogen in the host plant, which usually only occur in the middle to late stages of the infection process, when damages in the plants are mostly irreversible (Lowe, Harrison et al. 2017). Also, it is important to be aware that different pathogens can cause identical visual symptoms, as well as other abiotic stress (e.g., nutritional deficiencies, water deficit, among others), thus symptoms alone may not be enough for an accurate diagnosis, or for determining the disease-causing agents (Agrios 2012, Mora-Romero, Félix-Gastélum et al. 2022).

Other common direct diagnostic techniques may surpass this drawback, such as molecular and serological laboratory methods, which have revolutionized the detection of plant diseases since they allow large sample processing and rigorous pathogen identification. However, these biotechnologies may not be effective in early diagnosis tasks when the sample analyzed does not show any lesion characteristic of the disease, i.e., when the sample is non-symptomatic (also known as asymptomatic, presymptomatic, or symptomless). In fact, since pathogens often do not spread uniformly inside plants, destructive molecular methods can be non-diagnostic at this early stage

(Bock, Poole et al. 2010, Veys, Chatziavgerinos et al. 2019). Furthermore, these methods have other limitations, as the sample preparation is destructive (not allowing disease field mapping), the intensive labor needed, and producing specific antibodies can be inefficient with the presence of inhibitors reducing the sensitivity of nucleic acid-based methods (Martinelli, Scalenghe et al. 2014, Veys, Chatziavgerinos et al. 2019). Thus, they do not provide immediate results, requiring usually at least two days to be completed (Martinelli, Scalenghe et al. 2014, Moghadam, Ward et al. 2017).

Therefore, providing alternatives for visual-based and biotechnological methods (wet labs) processes of crop disease diagnosis, mainly used in the agriculture and horticulture sectors, with more automated, objective, and sensitive approaches, is crucial in sustainable crop production.

In this regard, indirect diagnostic methods (proxy) have been under development in recent years. They are based on the evidence that when plant-pathogen interactions occur, specific compounds are produced by both the host and pathogen, resulting in changes in plants' biophysical and structural composition, even before visual symptom development. Given that these biophysical modifications affect the optical properties of the plant's tissues (e.g., reflectance, transmittance, and emittance) it opens up prospects for the use of optical-photonic techniques to detect these changes in early disease stages and indirectly provides indicators of the plant's health status condition (Mahlein 2015, Moghadam, Ward et al. 2017).

Several researchers applied optical spectral sensors for disease diagnosis on laboratory (Barthel, Dordevic et al. 2021), greenhouse (Cen, Huang et al. 2022), and field (Nguyen, Sagan et al. 2021) scales. Nonetheless, there is not always a clear definition between the disease/infection and the classification of the causal agent, which limits the interpretation of results. Additionally, several studies are performed at the onset of visual symptom development, not being suitable for early diagnosis. Other constraints identified are related to i) the pathogen being analyzed, since usually Fungi are used in plant disease studies, ii) in contrast with Bacteria, which are still scarcely studied; iii) and, the conditions for data acquisition, as different sensors are not standardized, and spectral data processing techniques have not yet been sufficiently explored (e.g. Machine Learning, ML, and Chemometrics). These insufficiencies may lead to limited portability of sensor-based techniques for field conditions.

Thus, although it is considered that there is knowledge regarding the early diagnosis of diseases through the usage of these optical sensors, many of the explored protocols and techniques are yet at very initial stages, and the spectral devices usually

present low ‘Technology readiness levels' (TRL) (Mankins 1995). Therefore, the transference of this knowledge to agricultural producers has not yet occurred. Integrating this expertise is not yet trivial and requires research in line with the objectives of the present thesis.

This thesis's primary objective is to explore, test, and validate the application of proximal optical sensed data for the early assessment and diagnosis of bacterial plant diseases. In this regard, the study i) investigates the suitability of different measuring systems, namely Hyperspectral Spectroscopy sensors (measuring reflectance and transmittance data), RGB cameras, and Thermography for early diagnostic tasks; ii) tests different data handling (e.g., feature selection and dimensionality reduction), and modeling (predictive classification using Machine Learning or Chemometrics algorithms); and iii) explores the approaches portability between crops (i.e, tomato – herbaceous, annual plant - and kiwi – woody, perineal crop), experimental environmental conditions (laboratory and field), and bacterial species (belonging to the genus *Pseudomonas* spp. and *Xanthomonas* spp).

The pathosystems (i.e., host-pathogen ecosystems) analyzed in this work, namely i) *Pseudomonas syringae* pv. *tomato* in tomato, ii) *Xanthomonas euvesicatoria* in tomato, and iii) *Pseudomonas syringae* pv. *actinidiae* in kiwi were chosen due to their agronomic and economic relevance, mostly related to aesthetic and yield losses (intimately linked to important monetary losses) they cause in several plants with agronomic interest (such as tomato, pepper, tomato, eggplant, kiwi, fruit trees – pear, cherry, peach, citrus, walnut, among others).

This thesis aims to respond to a specific group of research questions:

1. On spectral signatures measured in healthy and diseased crop samples (pathosystems).
    1.1. Do healthy and bacterial-diseased biological samples exhibit distinct spectral signatures? If yes, i) what specific spectral characteristics differentiate them? ii) In which spectral regions they are located? iii) Can these differences be detected before the development of visual symptoms?
    1.2. Do diseased biological samples affected by different pathogens differ in spectral signatures? If yes, i) what are the specific spectral characteristics distinguishing different pathogens? ii) In which spectral regions are these pathogen-specific differences located? iii) Is it possible to detect them before visual symptom development?

2. On diagnostic sensing approaches for early bacterial diseases in plants (sensing technologies).
   2.1. Which strategies were tested for early bacterial diagnosis in plants? i) How does each diagnostic approach perform regarding sensitivity and specificity? ii) What is the biological/agronomic significance of the information provided by each sensing approach? iii) Are there specific advantages or limitations associated with each tested approach?
3. On predictive modeling strategies (agronomic application).
   3.1. Which predictive classification modeling strategies were developed for bacterial diagnosis? i) How does each modeling strategy perform regarding accuracy and reliability? ii) What is the biophysical meaning associated with the predictions made by each model? Iii) Is the performance consistent across different stages of disease development?

The responses to these questions provide supporting evidence for using proximal optical sensors for early bacterial disease approach in both laboratory and field conditions. Furthermore, it's plausible that the protocols and results obtained here will also serve as 'concept proof' for other diseases with great agro-economic impact.

The objectives and research questions of this thesis support more efficient and sustainable agronomic practices in terms of protection measures, so it is considered that they are aligned with the *European Green Deal*, and its *Farm to Fork strategy,* fostering innovation targeting a healthy and environmentally friendly food system (Fetting 2020). Moreover, they also are in line with the United Nations Sustainable Development Goals (SDGs) (Bernstein 2017, Carlsen and Bruggemann 2022), namely: SDG2 - End hunger, achieve food security and improved nutrition and promote sustainable agriculture; SDG12 - Ensure sustainable consumption and production patterns; SDG13 - Take urgent action to combat climate change and its impacts; SDG15 - Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss.

The present thesis is organized into five Chapters, including five scientific articles (Papers) (**Case Studies**, Figure 1). In brief, after this introduction, **Chapter II** arranges the conceptual foundation, where key concepts are explored, and the theme relevance is contextualized (Figure 1). This Chapter includes a critical review (**Case Study 1**) concerning the state of the art related to current and innovative Proximal Sensing (PS) technologies used to diagnose diseases in crops (Figure 1).

**Chapter III** describes different case studies, four of which consist of peer-reviewed articles and two are preliminary assays. All these case studies explore distinct aspects of bacterial disease assessment and diagnosis through the application of proximal optical sensors (Figure 1). **Case Study 2** evaluates the application of hyperspectral spectroscopy (both transmittance and reflectance data) combined with two predictive classification approaches for bacterial disease discrimination in tomato (herbaceous crop) and kiwi (woody crop) leaves in-situ laboratory, and field conditions (Figure 1). **Case Study 3** investigated in-situ, non-destructive discrimination of bacterial canker on kiwi leaves using hyperspectral reflectance and applied predictive modeling approaches. Several Feature Selection (FS) methodologies and supervised machine learning approaches were evaluated (Figure 1). **Case Study 4** investigates the potential of hyperspectral point measurement (transmittance) and machine-learning-based predictive models for early in-situ diagnosis and discrimination of bacterial speck and spot of tomato plants (Figure 1). **Case Study 5** explores the suitability of hyperspectral point of measurement (POM) transmittance in combination with Data-Driven Soft Independent Modeling of Class Analogy (DD-SIMCA) for the early assessment of tomato bacterial speck and spot in controlled conditions, using healthy samples as a target class, and Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS) as a mean to retrieve pure profiles of healthy and diseased tissues (Figure 1). Motivated by the findings, **Case Study 6** explores the fusion of hyperspectral data and microscopy imaging in a preliminary assay. The first outcomes of this work resulted in the proof-of-concept for direct research applications (Figure 1). Likewise, **Case Study 7** introduces the first assays performed using Thermal and RGB (Red, Blue, and Green bands) imaging for bacterial disease assessment (Figure 1).

The four datasets used in the previous scientific articles and case studies were, furthermore, published at Zenodo (Research and OpenAIRE 2013), an open repository developed under the European OpenAIRE program and operated by the European Organization for Nuclear Research CERN. They serve as three direct data contributions to future plant research.

**Chapter III** contains a general discussion of this work, summarizing the main contributions of this thesis (Figure 1). In **Chapter IV**, the final remarques and perspectives are made (Figure 1).


**Figure 1** Structure of the thesis. Chapters I and II provide the general introduction and the theoretical background of the work, introducing a critical review of the theme of using proximal sensors for performing the early diagnosis of plant diseases. Chapter III introduces the different case studies analyzed, Chapter IV discusses the main findings made in each one of the studied subjects, and Chapter V indicates the conclusions and perspectives of this thesis.

The **Appendix** Chapter presents several scientific outputs that complement the research performed in previous Chapters (Figure 1). **Appendix A | Paper I** introduced the performance of an in-vivo, in-situ diagnosis of bacterial canker in kiwi leaves in field conditions, using hyperspectral reflectance data and Vegetation Indexes (VIs), for 15 weeks. The findings led to the development of **Case Study 3** located in **Chapter III**. In turn, **Appendix B | Paper II** proposes the computation of a Principal Component Analysis (PCA) as a dimensionality reduction approach for discriminating transmittance hyperspectral signatures collected in healthy and diseased tomato leaflets. The outcomes led to the development of **Case Study 4** located in **Chapter II**. Oral (**Appendix D, E**) and poster communications (**Appendix F, G**), along with the best e-poster award won at the II Plant Pests and Diseases Forum, are also mentioned in the **Appendix**.

Furthermore, a technical article published in a Portuguese agriculture specialty magazine is provided (**Appendix H**).

# Chapter II |
# Conceptual Framework

# 1. Dimensions of plant disease sensing

The Food and Agriculture Organization (FAO) statistics estimate an increase in the world population level in the next 40 years when approximately 9.1 billion people will be reached by 2050. Consequently, there is expected growth in the demand for food, and a need for a steady increase of 70% in agricultural production (Godfray, Beddington et al. 2010, FAO 2018, Nations 2019). Ninety percent of this growth in crop production would be achieved by higher yields and increased cropping intensity, with the remainder resulting from land expansion (Bruinsma 2009, Kopittke, Menzies et al. 2019). Thus, improving the security and nutrition of worldwide crops has been considered an important goal for sustainable development by the United Nations (UN) (Nations 2021). Also, the UN General Assembly proclaimed 2020 as the 'International Year of Plant Health' (IYPH) to raise awareness on how protecting crop health can help end hunger, reduce poverty, protect the environment, and boost economic development (FAO 2020). UN elected 2021 as the 'International Year of Fruits and Vegetables' (IYFV) to raise awareness of the important role of fruits and vegetables in human nutrition, food security, and health, as well as, in achieving UN Sustainable Development Goals (Nations 2020, FAO 2021).

During the cultivation process, crops can be affected by different kinds of biotic (Figure 3) and abiotic stresses, affecting their productivity. Therefore, reductions in yields occur, resulting in reduced income for producers, restricted availability, and higher prices for consumers (Savary, Ficke et al. 2012, Savary, Bregaglio et al. 2017, Nelson 2020). The quality of fruits and vegetables, essential for providing important nutrients for humans, can also be affected (Strange and Scott 2005, Chakraborty and Newton 2011, Shahid, Zaidi et al. 2017).

Other problematics derived from the damages caused by these agents include significant aesthetic losses in agricultural products, which are not generally quantified in monetary terms but reduce some of the pleasure consumers derive from their consumption, influencing their behavior (Oerke 2006, Finlayson 2018); loss of plants that have positive effects on human wellbeing and health, contributing to a loss in comfort and beauty (Maller, Townsend et al. 2006, Hall and Knuth 2019, Elsadek and Liu 2021); and, limitations in the ability of a country to import or export crops and plants around the world or even to move them within its borders, due to the possibility of creating pathways for the entry of new organisms and the creation of the ideal conditions for pathogen change (Macleod, Pautasso et al. 2010, Savary, Bregaglio et al. 2017).

Globally, it is estimated that losses in crop yield caused by pathogenic organisms can range between 20% and 40% (Savary, Ficke et al. 2012, Fried, Chauvel et al. 2017). Current agricultural practices promote the spread of plant disease epidemics and rapid pathogen evolution since they favor intensified monoculture in large areas, genetically uniform plant varieties, and the development of global supply chains and logistic activities (Zhan, Thrall et al. 2015).

Phytosanitary product application, through spraying, is currently the most promoted approach for preventing and treating diseases. Therefore, when a plant disease suddenly appears and spreads on a large scale, its treatment can lead to considerable damage to the environment (Zhang, Yang et al. 2020). Phytosanitary products can be a source of air, soil, and water pollution. After their application, they may volatilize into the air, run off or leach into surface water and groundwater, be taken up by plants or soil organisms that are not the target, or stay in the soil, among other problems (van der Werf 1996). Also, food security can be affected by the intensive usage of these phytosanitary substances (Bonner and Alavanja 2017). Despite this evidence, FAO statistics reported that the worldwide consumption of phytosanitary products tended to increase, and the value rose from approximately 3.09 million tons in 2000 to 4.12 million tons in 2018 (FAO 2020).

Diagnosis plays a pivotal role in the plant disease-management process, encompassing the discernment of both the nature and root cause of a particular disorder. An early, operational, and accurate diagnosis is, thus, fundamental for effective protection measurements. This critical step supports informed decisions, such as whether to pursue treatment or not, and facilitates the selection of the most fitting phytosanitary interventions, including the choice of the most appropriate active substances.

In general, plant disease diagnosis is made through the direct detection of symptoms (e.g. characteristic indicator signs of a disease), encompassing visual (scouting) and laboratory-based techniques. Recently, early diagnostic approaches based on indirect disease diagnosis, through the assessment of specific changes in the optical, biophysical, and molecular plant's properties are being explored. These methods allow a disease diagnosis even before the appearance and development of macroscopic symptoms. Moreover, they allow a more preventive and targeted intervention, playing a crucial sustainable role in the mitigation of crop losses, and promoting plant health in agricultural environments.

Therefore, many agronomic, environmental, economic, and humanitarian reasons justify the development of new early diagnostic methods for plant diseases and their field mapping in line with precision agriculture (Mahlein, Oerke et al. 2012, Martinelli, Scalenghe et al. 2015, Mahlein 2016). Innovative technologies like the global positioning system (GPS) and variable rate spray systems contributed to precision disease management development. This approach allows the application of chemical agricultural products in the right location, at the right moment, and right dose, being in line with the Precision Agriculture 3 R's concept (Cunha and Braga 2022). Precise disease management allows a reduction in phytosanitary product usage, resulting in fewer expenses for the producer, fewer residues in crop production, and environmental contamination (Zhang, Yang et al. 2020).

# 2. Principles of infectious bacterial plant diseases

Different organisms can cause diseases in plants, i.e. cause harmful deviations in the normal physiological functioning of plants, affecting their structure, growth, functions, or other parameters (Surico 2013, Nazarov, Baleev et al. 2020). Phytopathogens differ from each other concerning the set of plant species they affect (host range), the location of the infection they cause, and the age of the organ or tissue they affect (Schumann and D'Arcy 2006, Abdulkhair and Alghuthaymi 2016).

For a disease to occur in any plant system, three components are needed, which must be present at the same time: they represent the disease triangle (consisting of the pathosystem) and include a susceptible plant, a pathogen capable of causing disease, and a favorable environment. If any of these three elements are missing, no disease occurs (Figure 1) (Schumann and D'Arcy 2006, Abdulkhair and Alghuthaymi 2016).

Some characteristics of the host and pathogen highly influence disease development and resistance and include the plant family, age, growth stage, and genetic type, along with pathogen virulence race, inoculum size, and dormant state (Schumann and D'Arcy 2006, Abdulkhair and Alghuthaymi 2016). In turn, the environment in which plant disease occurs also highly influences its appearance and development in all stages of the disease cycle. It influences the development of the plant and its ability to mount defenses against invasion, pathogen dispersion, its capacity to penetrate the plant, and, its subsistence in the absence of the host plant (Schumann and D'Arcy 2006, Sharma 2006). It encompasses a wide range of factors, including meteorological phenomena

such as recent temperatures (such as extreme highs and lows), rainfall or irrigation (amounts, timing, sources), and light intensity or shade. Characteristics of the soil, such as drainage, soil type, and pH, are also important (Schumann and D'Arcy 2006).

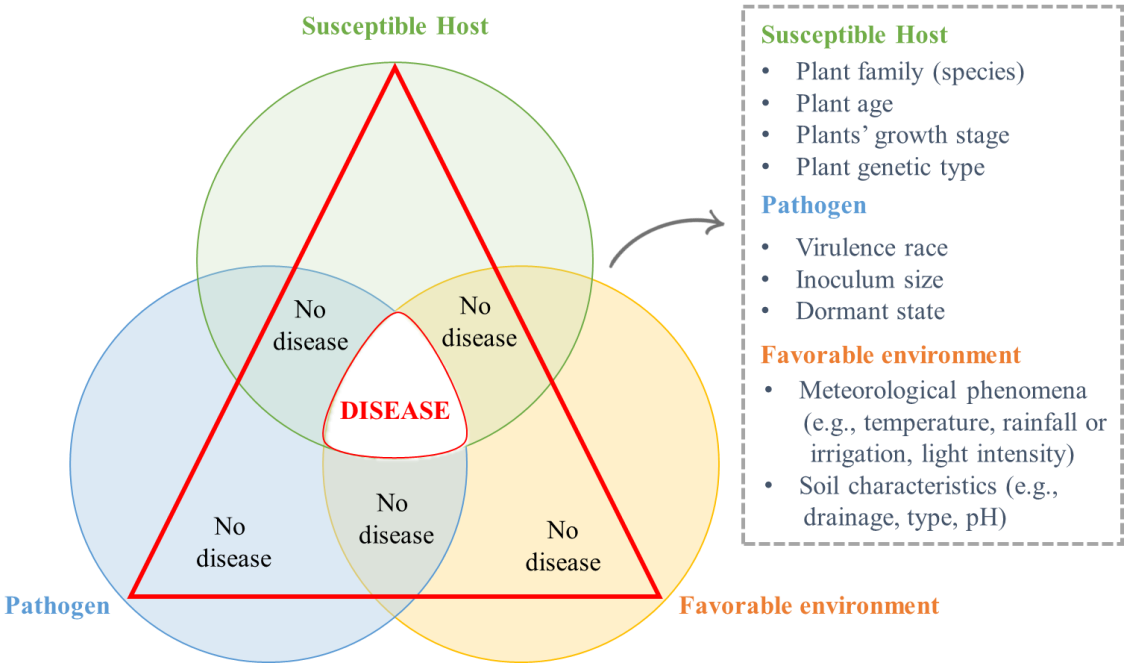



**Figure 1** Illustration of the disease triangle showing the interaction between the susceptible host, pathogen, and surrounding environment as a prerequisite for disease to occur. The triangle may be used as a conceptual model describing the factors that impact the development of an epidemic. Some examples of factors related to the host, pathogen, and environment that may influence disease progression are also provided.

Plant disease development involves a chain of events in a specific order, designated as the disease cycle. It includes the stages of a pathogen's development and the disease's effects on the host plants. These events include inoculation, prepenetration, penetration, infection, colonization (invasion), and growth and reproduction of the pathogen (Figure 2) (Surico 2013, Abdulkhair and Alghuthaymi 2016).

Briefly, the disease on the host plants begins with the arrival and successful penetration by the pathogen. With the development of the invasion and infection process, molecular, physiological, and structural changes start to occur. Ultimately, they result in the appearance of macroscopic signs of infection, i.e., symptoms (Schumann and D'Arcy 2006), which may be localized or systemic (Schumann and D'Arcy 2006). Is important to be aware that different pathogens can cause identical visual symptoms, so these alone are usually not enough for an accurate diagnosis or determining the disease-causing

agents (Schumann and D'Arcy 2006, Agrios 2012). Some examples of the most common visible symptoms regarding bacterial infections may include the appearance of senescent/necrotic lesions, vascular wilt, soft rot, and tumors (Zhang, Yang et al. 2020). An early intervention, preferably before the pathogen infection manifestation and colonization (Figure 3), is of utmost importance to minimize the damage. This can be accomplished through a streamlined, efficient diagnostic process characterized by its early, precise, and operational nature—a paradigm that is yet to be fully developed.

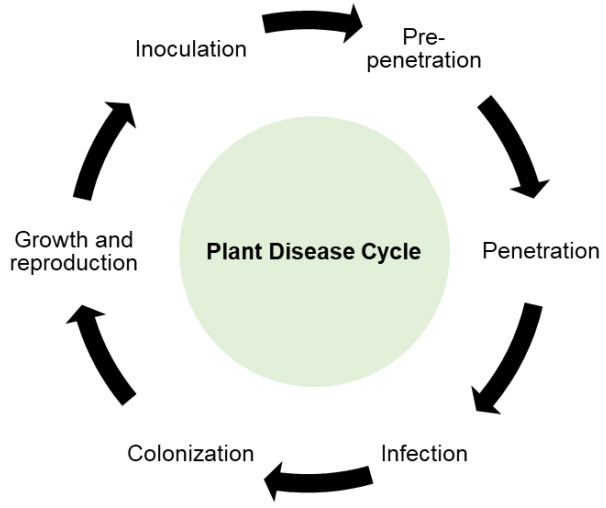



**Figure 2** A generalized diagram displaying infection and disease cycle caused by bacteria.

# 3. Plant-pathogen interactions allow an early disease diagnosis through proximal sensors

Changes in the host plant's physiological, biochemical, and metabolic properties caused by pathogens result in altered optical and metabolic features. Proximal optical sensors can detect these changes, along with the monitorization of the spatiotemporal pattern of disease development (Mahlein, Oerke et al. 2012), which allows the development of several methods of diagnosis. These sensors can eventually be mounted on different platforms (e.g. robots, tractors) to map information about the disease in a precision agriculture approach (Mishra, Polder et al. 2020), as will be discussed in the next subsection.

Plant pigments are one of the first host compounds to be affected and degraded by pathogens, resulting in changes in plant's optical behavior. Chlorophylls (Chl) a and b are the major pigments of plants (accounting for almost 65% of the total pigment content), and their spectral absorption range is mostly concentred in the 410-430 and

(Chl a), 450-470 nm (Chl b), and 600-690 nm (Chl a) bands, located in the blue and red dions, respectively. Green radiation, on the other hand, is less strongly absorbed. In healthy plants, chlorophyll concentration is approximately ten times higher than that of other pigments (e.g. carotenoids, flavonoids, among others), thus masking out the specific absorption features of these compounds (Jacquemoud and Baret 1990). Therefore, plants preferentially absorb red and blue wavelengths, and the green part of the incident light is less absorbed and is consequently mostly reflected, leading to the green appearance of vegetation (Jensen 2009, Jones and Vaughan 2010, Sahoo, Ray et al. 2015, Deshmukh, Janse et al. 2018).

With the disease development and onset, other photosynthetic pigment levels are increasingly more affected, namely carotenoids and polyphenols. The first type of pigment absorbs most effectively between 440 and 480 nm and extends its absorption action into the blue-green region. They include compounds such as yellow lutein pigments, β-carotenes, and xanthophylls (e.g., violaxanthin and zeaxanthin). In turn, polyphenols (e.g. brown pigments) start to appear only when the plant tissues begin to necrose (Jensen 2009, Jones and Vaughan 2010, Sahoo, Ray et al. 2015, Deshmukh, Janse et al. 2018). They include compounds like flavonoids and anthocyanins, which absorb radiation from blue to red spectral ranges with higher intensity in the shorter wavelengths (Jensen 2009, Jones and Vaughan 2010, Sahoo, Ray et al. 2015, Deshmukh, Janse et al. 2018).

Moreover, the optical spectral properties of host plants are also affected in the Near-Infrared (NIR) region (700–1300 nm) and short-wave infrared (SWIR, 1000–2500 nm) when plant leaves structure (e.g., cell layers, cell size, structural components – lignin's, proteins, among others), air spaces, and water content is affected (Jones and Vaughan 2010, Haq and Ijaz 2020). In detail, the major water absorption bands are well documented at 1450, 1940, and 2700 nm, and secondary features at 960, 1120, 1540 1670, and 2200 nm (Ustin, Zomer et al. 1999).

In this regard, is possible to see those changes in plant leaves' biochemistry and cellular composition result in changes in plant's spectral characteristics. Nevertheless, it is important to mention that a leaf's spectral properties are not a static phenomenon over time. Indeed, they continuously change during growth, maturity, senescence, decay, or stress (e.g. plant disease development).

These spectral changes even occur before macroscopic lesions (i.e., symptoms) appear. Therefore, new sensor-based methods considering this evidence can be used to indirectly diagnose different plant pathogen agents (Mahlein, Kuska et al. 2017,

Zhang, Yang et al. 2020, Cheshkova 2022). The measurement of optical plant properties can also be used to assess the spatial-temporal dynamics of the interactions between plants and pathogens. Research using spectral signatures derived from hyperspectral sensors has already been used to study (concept-proof) foliar diseases at the canopy and leaf scales (Mahlein 2011), and this will be discussed in the next subsection.

Considering the previous information, a new interpretation of the disease triangle (Figure 3) can be proposed. Beyond the three existing dimensions, a fourth one related to current and innovative 'Protection measures' should be taken into account, as proposed in Figure 3. This new dimension recognizes that host-pathogen interactions occur in agricultural environments where farmers and producers perform management strategies and may interfere with them. In fact, diverse strategies to mostly prevent, but also treat and irradicate the pathogen action may be performed (e.g., phytosanitary actions), preferably with maximum efficiency, using the most appropriate compounds, doses, and targets (localized application). To achieve this goal, early diagnosis is crucial, and indirect proximal sensors may be an interesting tool (Figure 3).

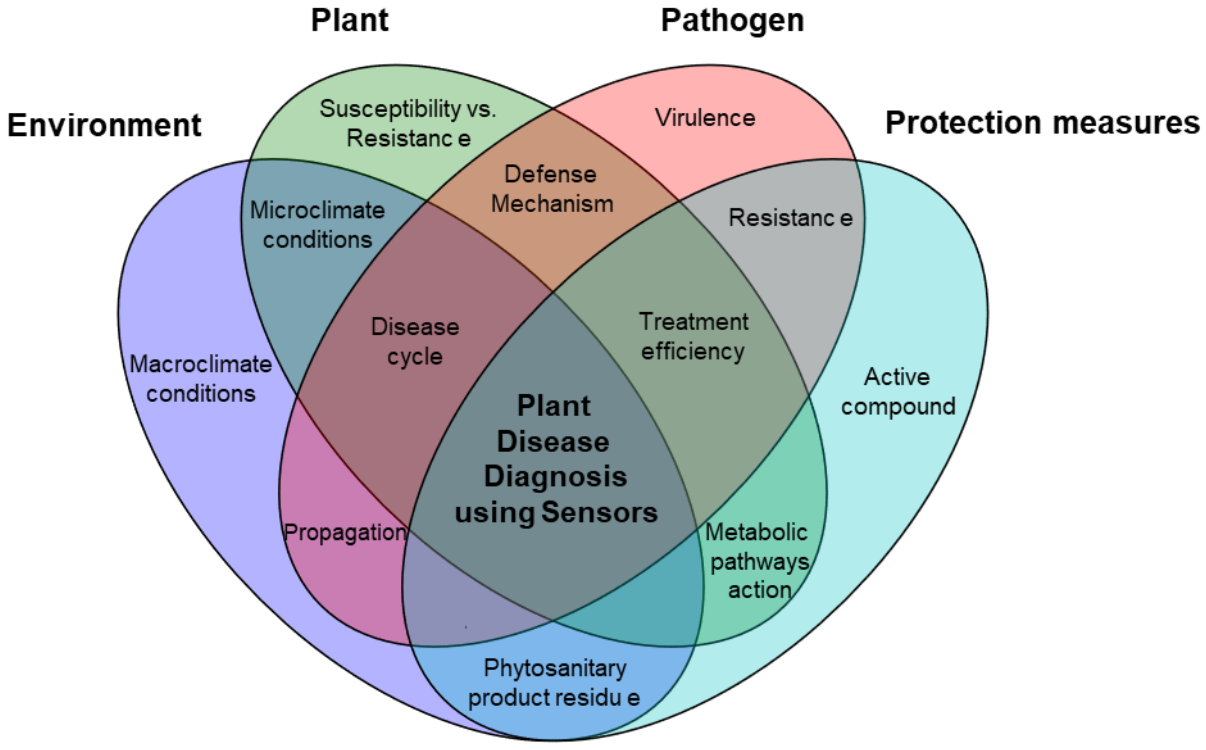



**Figure 3** Proposal of a new interpretation of the standard plant disease triangle, incorporating a new fourth dimension related to 'Protection measures'. Plant disease diagnosis involves a series of complex events, including the interactions between the plant host, the pathogen that affects it, the environment surrounding them, and the plant protection measures applied to mitigate the negative impacts of this interaction.

# Case Study 1

**Reis-Pereira, M.**; Santos, F. N. d.; Tavares, F.; Cunha, M. Digital diagnosis of plant diseases: a critical review and analysis of optical sensing technologies for early plant disease diagnosis.

Paper submitted

Classification according to journal: Review Article

# Digital assessment of plant diseases: A critical review and analysis of optical sensing technologies for early plant disease diagnosis

Mafalda Reis Pereira[1,2], Filipe Neves dos Santos[2], Fernando Tavares[1,3,4], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

*** Correspondence:** Mário Cunha mccunha@fc.up.pt

## Highlights

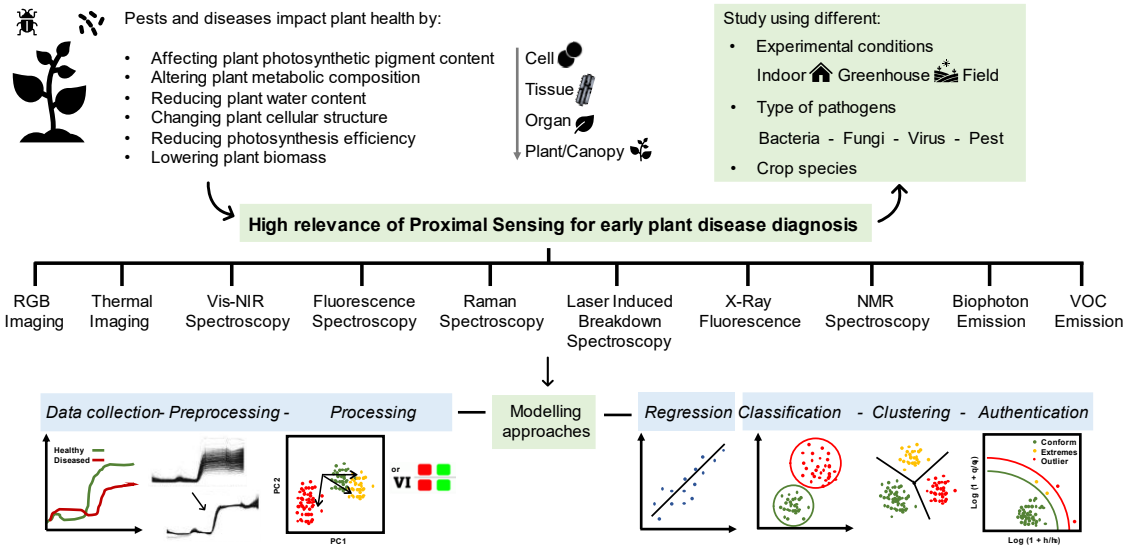
- Hyperspectral spectroscopy was the most applied technology for early plant disease diagnosis.
- Fungi were the most analyzed group of pathogens found in the literature.
- Experimental assays were mainly conducted in laboratory (controlled) conditions.
- Vegetation Indices and Principal Component Analysis were the most used approaches in data processing.
- Classification was the main type of predictive modelling used in the screened scientific articles.

## Abstract

The present critical literature review describes the state-of-the-art innovative proximal (ground-based) solutions for plant disease diagnosis, suitable for promoting more precise and efficient phytosanitary measures. Research and development of new sensors for this purpose are currently a challenge. Present procedures and techniques of diagnosis are dependent on visual characteristics and symptoms to be initiated and

applied, compromising an early intervention. Also, these methods were designed to confirm the presence of pathogens, not having the necessary high throughput and speed required for supporting real-time agronomic decisions in field extensions. Proximal sensor-based systems are a reasonable tool for an efficient and economic disease assessment. This work focused on identifying the application of optical and spectroscopic sensors as a tool for disease diagnosis. Biophoton Emission, Fluorescence Spectroscopy, Laser-Induced Breakdown Spectroscopy, Multi- and Hyperspectral Spectroscopy (HS), Nuclear Magnetic Resonance Spectroscopy, Raman Spectroscopy, RGB Imaging, Thermography, Volatile Organic Compounds assessment, and X-ray Fluorescence were described due to their relevant potential. Nevertheless, some of these techniques revealed a low Technology Readiness Level (TRL). The main conclusions identify HS, single and multi-spatial point observation, as the most applied methods for early plant disease diagnosis studies (82%), combined with distinct feature selection (FeS), dimensionality reduction (DR), and modeling techniques. Vegetation Indices (29%) and Principal Component Analysis (20%) were the most popular FeS and DR approaches used to highlight the most relevant wavelengths contributing to disease diagnosis. In the modeling process, classification was the most applied technique (80%), used mainly for binary and multi-class health status identification. Regression was used in the remaining (20%) scientific works screened. The data was mainly collected in laboratory conditions (69%), and only a smaller number of works were performed in field conditions (20%). Regarding the etiological agent responsible for causing the disease in the study, fungi (53%), and virus (24%) were the most analyzed group of pathogens found in the literature. Overall, proximal sensors are suitable for early plant disease diagnosis both prior to and after symptom appearance, presenting classification accuracies mostly superior to 71% and coefficients of regression superior to 61%. Nevertheless, additional research regarding the study of specific host-pathogen interactions is necessary.

## Graphical abstract




## Keywords

Proximal sensing, Plant disease diagnosis, Predictive Modeling

## 1. Introduction

Food security, intimately linked to crop health, has been a foremost global concern throughout the years. Current threats, such as climate change, the growth of the world population, and the disappearance of several varieties of agricultural plant species, among others, have drawn attention to this problem (Karthikeyan, Chawla et al. 2020). Likewise, biotic stresses (i.e., fungi, bacteria, viruses, and pests) have also become a challenging hurdle to envisage nowadays, since they are responsible for causing crop yield reductions (ranging from 20 to 40%), lower quantity and quality of agricultural products, and lower producers' income and higher product prices, affecting especially the final consumer (Savary, Ficke et al. 2012). They are extremely hard to prevent and treat, and the current phytosanitary solutions available for fighting them are responsible for causing considerable damage to agricultural fields (Savary, Ficke et al. 2012). In this regard, the European Commission announced, in May 2020, two pesticide reduction targets as part of the Farm to Fork Strategy (Commission 2020), aiming for a 50% reduction in the use and risk of chemical pesticides, along with a 50% reduction in the usage of more hazardous pesticides until 2030. Hence, improving precise diagnosis, monitoring, and protection measures is of paramount importance.

Biotic stress diagnosis (i.e., identification of the disease's nature) is typically performed by a variety of direct methods, i.e., approaches that require the observation of characteristic symptoms caused by the pathogen on the host tissues. These usually

occur in the middle to late stages of the infection process, compromising the effectiveness of protection and treatment measures (Lowe, Harrison et al. 2017). Some instances include visual scouting practices, which involve a careful, detailed inspection of crop fields by specialized trained observers, who must be able to detect and identify (diagnose) diseased plants based on the existence of characteristic disease symptoms (Parker, Shaw et al. 1995). This visual recognition of plant stress has been extensively applied in the last decades due to its easy application and usefulness. Nevertheless, it is subjective, error-prone (as symptoms alone are not entirely disease-specific), labor-intensive, time-consuming, and expensive (Khaled, Abd Aziz et al. 2018, Ali, Bachik et al. 2019).

Advancements in biotechnologies have led to the creation and implementation of various serological and molecular laboratory tests renowned for their high objectivity, marked by reliability, precision, and accuracy. These tests have significantly enhanced disease diagnosis capabilities. The most performed are the enzyme-linked immunosorbent assay (ELISA), and the polymerase chain reaction (PCR). Their development has revolutionized plant disease diagnosis due to their capability to process simultaneously a large number of samples and perform precise pathogen identification (Venbrux, Crauwels et al. 2023). These biotech-based methods provide complementary information and are usually used in conjunction. In fact, the European and Mediterranean Plant Protection Organization (EPPO) detailed protocols for plant pathogen diagnosis which integrate these phenotypic, serological, and molecular techniques (Martinelli, Scalenghe et al. 2015). However, these laboratory-based approaches experience limitations in the early phases of the infection process. Pathogens often do not spread uniformly inside plants, leading to inefficacies in these methods when analyzing a diseased sample that exhibits no macroscopic characteristic lesions (symptom) (Martinelli, Scalenghe et al. 2015), occurring notably during non-symptomatic stages. Furthermore, these techniques require detailed sampling procedures, usually involving a destructive sample preparation (i.e., not allowing a follow-up of the disease development) and taking several hours to be concluded (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015). For these reasons, they may not align with the precision agriculture concept, not allowing the full monitoring and optimization of phytosanitary measures in the entire production site. Therefore, the development of new plant disease diagnosis methods should focus on fast, accurate, and selective techniques capable of providing effective and complementary information about the plant’s health status, preferably in vivo and in field conditions.

Implementing these new approaches should ensure early intervention in plant diseases, ideally before the onset of symptoms or at their initial appearance. This proactive measure is crucial in preventing and effectively controlling plant disease progression and spread. Ultimately, it will contribute to adopting more localized and targeted plant protection strategies, including precision applications of phytosanitary products, accurately determining their location, quantity, and specific substances used. This putative shift towards preventive measures over curative ones aligns with a perspective rooted in precision agriculture. As a result, it is anticipated to yield several advantages across agronomic, economic, environmental, and quality aspects.

In the last decades, several approaches have been explored for this early plant disease diagnosis. Among these methods, Proximal sensing (PS), also called Proximal Remote Sensing, Close-Range Remote Sensing, or Ground-level Remote Sensing, stands out, providing the foundation for indirect plant disease diagnosis. These sensors operate from relatively short distances, typically ranging from a few centimeters to a few meters (Martinelli, Scalenghe et al. 2015, Zhang, Yang et al. 2020), to assess molecular, biophysical, and structural modifications promoted by host-pathogen interactions in plants' tissues. Its advantages include offering the benefits of providing a rigorous, sensitive, consistent, standard, high throughput, rapid, and cost-effective diagnostic (Martinelli, Scalenghe et al. 2015, Zhang, Yang et al. 2020). These sensors can be handheld or mounted on terrestrial vehicles (e.g. robots, tractors) or aerial platforms (e.g. drones) to assess and field map plant health (Oerke, Mahlein et al. 2014).

Nevertheless, it is important to address that plants' physiological and phenotypical characteristics are unstable over time, making the disease diagnostic process even more challenging. Similarly, to this dynamic evolution, differences in plant traits may occur within individuals of the same and/or different cultivars and species, demonstrating that plant disease diagnosis can be highly demanding. These difficulties seemed to be surpassed through the application of PS-based solutions.

Despite the multidimensional importance of early plant disease diagnosis and the potential of proximal sensing, its practical implementation remains challenging (Figure 1).

The main objective of the present work is to assemble the information made available in recent years about the different PS techniques showing potential for performing early diagnosis of plant diseases. Although a few literature reviews exist on this research topic, the articles found often lack integration of crucial information concerning the robustness of samples used in each assay, the tested experimental

conditions (e.g., laboratory, greenhouse, field), and/or the model evaluation metrics used (or not used at all) to access the potential of the applied sensing technique for disease prediction. Furthermore, modeling strategies (e.g., data organization, preparation, and modeling) are inconsistently applied across different research articles, leading to ambiguity regarding the relevance of the obtained results. Therefore, this study attempts to provide insights into these topics by evaluating the collected published articles within the context of a functional PS approach for early diagnosing plant diseases. Additionally, it examines the prevailing trends in the literature concerning the most used PS and diagnosis strategies.

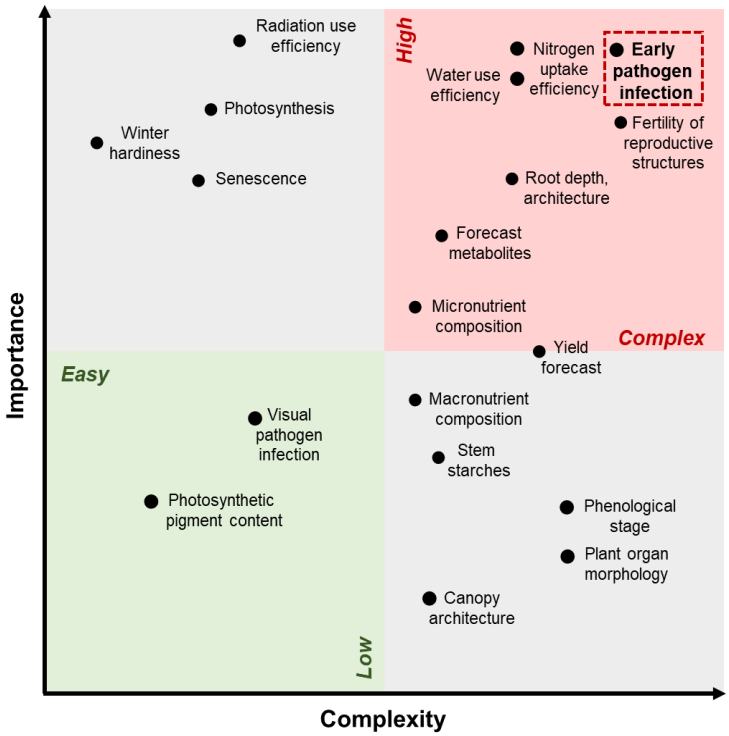



**Figure 1** Analysis of the complexity (x-axis) and importance (y-axis) of the estimation of some relevant agronomic traits by proximal sensing technologies. It is possible to observe that early pathogen infection diagnosis is in the right upper quadrant (in red), revealing the higher relevance and challenge of performing this task. In contrast, disease diagnosis when characteristic symptoms of the disease are visible is less challenging (located in the lower left quadrant, represented in green) but also less important because these lesions only appear in the middle to late stages of the disease infection process, compromising the effectiveness of plant protection measures.

This study, through the analysis of previously published scientific articles, aims to unravel specifically: i) What are the main PS techniques applied to diagnose plant diseases in indoor, greenhouse (or glasshouse), and infield assays? ii) What is the potential capability of these plant sensing devices for early disease diagnosis (prior to

the appearance of the first visual characteristic symptoms), both in terms of capabilities and model metric results? iii) How are computed the processing and modeling strategies most applied in data analysis? Nevertheless, the present review only focuses on research related to ground-level sensors (e.g., handheld, terrestrial platforms, benchtop, microscopes). This article is structured into different sections summarized in Figure 2, which includes (1), after this introduction; (2) the research methods, which briefly describes the applied research strategy; (3) the plant disease diagnosis section, where the main concepts and types of experimental conditions are described; (4) Sensing technologies, where a contextualization of the most applied sensing techniques used in plant disease diagnosis is made, along with the selection of relevant articles aiming an early diagnosis of crop diseases; (5) Methodologies for disease diagnosis, where data handling and modeling approaches found in the screened scientific articles are characterized; (6) Conclusion, which summarizes the main findings.

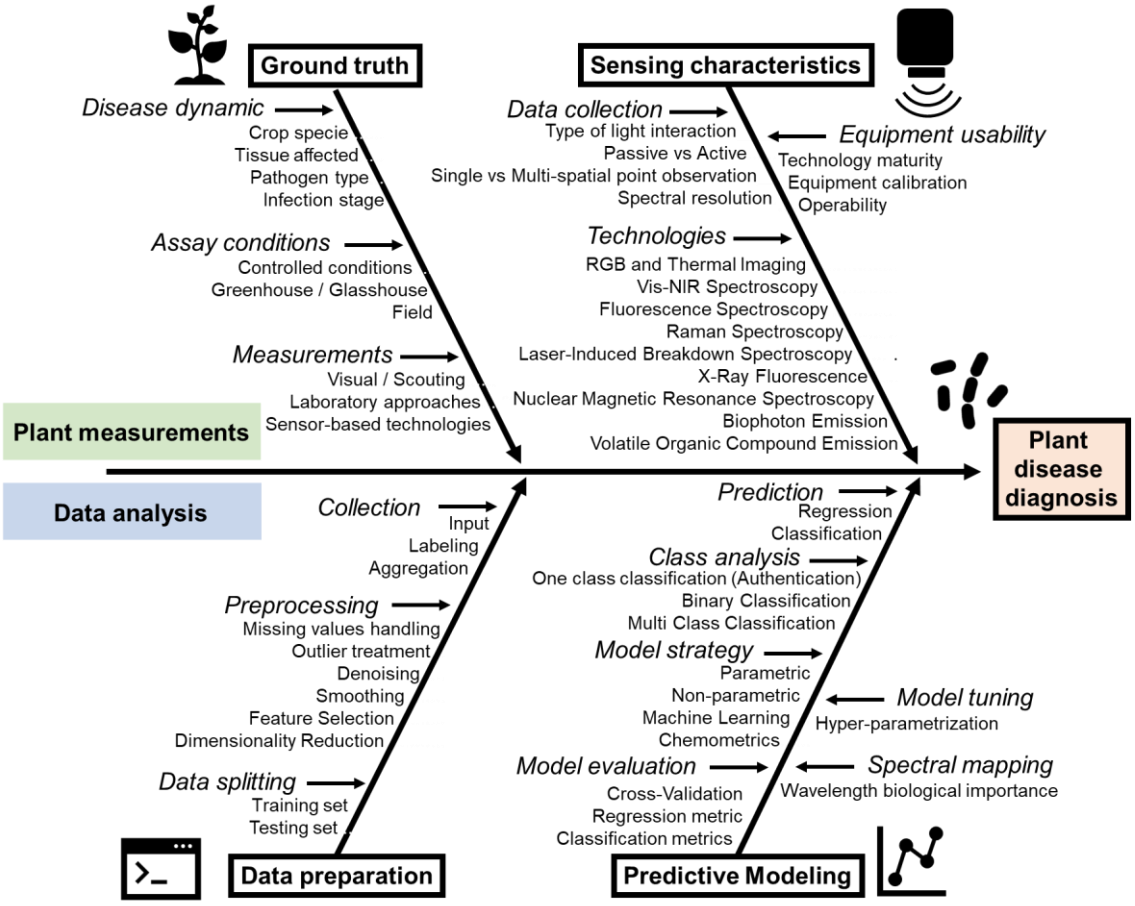



**Figure 2** Diagram showing the review structure, mentioning the main strategies employed for proximal (ground truth) plant disease diagnosis using sensing technologies combined with different predictive modeling approaches. The ‘Plant measurements’ section describes the main factors considered in a plant disease biological assay related

to the plant-pathogen interactions (disease dynamic), the environmental conditions, and the types of diagnosing techniques available. The 'Data analysis' section briefly introduces the main data preparation steps and modeling approaches available.

## 2. Research methods of literature review

This research followed a systematic approach, widely acknowledged for its comprehensive and organized analysis of all articles available on indexed platforms, ensuring their quality (Magalhães, Moreira et al. 2022). Systematic literature reviews frequently apply the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page, Moher et al. 2021). The online tool Parsifal (Freitas and Segatto 2021) was used to hold the ongoing review procedure, allowing an arrangement of the entire research process: procedure design, screening and removal of duplicated articles, quality evaluation, and data extraction (Supplementary Materials I). The present work appraised the primary indexed articles related to the use of optical and spectroscopic sensors in early plant disease diagnosis between 1971 and August of 2023 (Supplementary Materials I). The duplicated articles were removed, and the remaining were evaluated and selected. The chosen studies were fully read and analyzed. Each was assessed according to its quality to confirm if the work fulfills the aims of the current review (Supplementary Materials I). The extraction procedure retrieved information related to the: i) crop(s) studied in the assay; ii) pest(s) or disease(s) in analysis, iii) sensor/technology used, iv) type of sensor (single vs. multi-spatial point observation), v) light source system configuration, vi) sensor parameters (wavelengths studied), vii) environmental conditions (indoor, greenhouse, or infield), viii) modeling approach applied, x) model metric results, and ix) possibility of the current application (evaluated through the Technology Readiness Level – TRL – which is a type of measurement system used to assess the maturity level of a particular technology (Mankins 1995)).

From the total of articles screened (322), 46 research articles were retrieved and considered in the present review. A more detailed section about this subject can be found in Supplementary Materials I.

In the next Chapters, the main findings of this work related to plant disease diagnosis concepts, experimental conditions, plant-light interactions, innovative and emerging sensing technologies, and metrics of applied predictive modeling approaches will be summarized.

## 3. Ground truth plant disease diagnosis – Concepts, pathosystems, and experimental conditions

Different organisms can cause diseases in plants, i.e., promote harmful deviations from the normal physiological functioning of plants, affecting their structure, growth, functions, or other parameters (Surico 2013, Nazarov, Baleev et al. 2020). They can be fungi, bacteria, viruses, protozoa, and insects. Phytopathogens differ concerning the set of plant species they affect (host range), the location of the infection they cause, and the age of the organ or tissue they affect (Schumann and D'Arcy 2006, Abdulkhair and Alghuthaymi 2016).

Plant disease development involves a chain of events occurring in a specific order, called the disease cycle and includes the stages of development of a pathogen and the effects of the disease on the host plants (Nelson 1994). These events include inoculation, prepenetration, penetration, infection, colonization (invasion), and pathogen growth and reproduction (Surico 2013, Abdulkhair and Alghuthaymi 2016). During the first stages of the cycle, the diseased tissues may remain phenotypically similar to healthy ones and are designated as non-symptomatic (also called asymptomatic, pre-symptomatic, and symptomless, among others). In the middle to late stages of the cycle, when diseased tissues start to develop macroscopic lesions characteristics of the infection and consequently become phenotypically distinct from healthy tissues, they are called symptomatic.

Digital sensing techniques have been developed for plant disease diagnosis, in both the early and late stages of the disease cycle. Experimental assays can be performed in three main types of environmental conditions: i) laboratory, ii) greenhouse (glasshouse), and iii) field tests. The firsts are conducted under controlled conditions, where the environmental settings are usually stable (e.g., temperature, light, humidity), and the only variable of interest changing is the etiological agent responsible for causing the disease. It allows the monitoring of infection conditions (severity, authenticity, among others), e.g., (Reis Pereira, Santos et al. 2023), and the controlling of light conditions to better understand the sensor's performance. Thus, this approach provides relevant insights about the interaction between the sensor and the object of interest. The data can be collected on different plant organs, such as leaves, stems, or fruits, to gain insights into disease symptoms and effects on plant physiology. Field tests, in turn, are more challenging to standardize due to uncontrolled environmental conditions and biotic pathogens (e.g., risks of uncontrolled propagation), requiring sensors with higher TRL. Field work, generally preceded by laboratory tests, allows for large-scale tests and

effective operability of digital sensing systems, increasing their TRL. Real-time monitoring is also possible, which is crucial for providing insightful information for supporting immediate decisions about disease management strategies (e.g., early identification of disease outbreaks, and phytosanitary product applications, among others). Greenhouse / Glasshouse experiments constitute an intermediate case between the laboratory (controlled conditions) and field assays since they allow the elaboration of plant pathology studies in larger areas with semi-controlled environmental conditions.

## 4. Sensing technologies for disease diagnosis

### 4.1. Types of plant-light interaction

In recent decades, several sensor-based techniques have been explored for diagnosing plant diseases based on the interaction of biological tissues with radiation as represented in Figure 3. Sensing systems usually measure: 'reflectance', when they assess the quantity of light that is reflected from a surface; 'transmittance' when they measure the ratio of light that falls on a sample and passes through it; and 'emission', when the system captures the light emitted by a sample due to electrons making a transition from a high energy state to a lower energy state (Figure 3). Briefly, sensors that capture reflectance have detector gears that assess an illuminated part of the sample and measure both the specular (wavelengths reflected in a well-defined angle and direction) and diffuse (radiation scattered in many directions). Specular reflection does not convey any information about the internal traits of the sample. Hence, this mode is recommended to assess surface or near-surface modifications of sample traits (Walsh, Blasco et al. 2020).

In turn, some devices assess the samples' emission (i.e., light radiated), resulting from previous light absorption and conversion in distinct wavelengths (Figure 3). They are particularly useful for studying its inherent properties, such as temperature, composition, physiological state, and internal structure. Emission measurements are, therefore, primarily advised for capturing information related to the material's surface and mid-range internal properties (Walsh, Blasco et al. 2020). A particular case of emission is fluorescence, a phenomenon where the biological tissues are excited by absorbing radiation (e.g. ultraviolet – UV –, short-wavelength visible light) emitting longer-wavelength radiation, which is captured by the detector gear (Cerovic, Samson et al. 1999, Belasque, Gasparoto et al. 2008).

In transmittance systems, light directly passes through the sample without any direct light flow from the source to the detector, making this an interesting approach for assessing the internal quality traits of samples (Figure 3). For this reason, it's important

to note that the dimension and shape of the object in the study may influence the measurements since they impact the path traveled by light when crossing the sample (Walsh, Blasco et al. 2020).

Different combinations between the different geometries (e.g., reflectance and transmittance) can thus be beneficial to plant disease studies since they allow the simultaneous retrieval of external (e.g., color, size, and surface appearance) and internal quality characteristics.

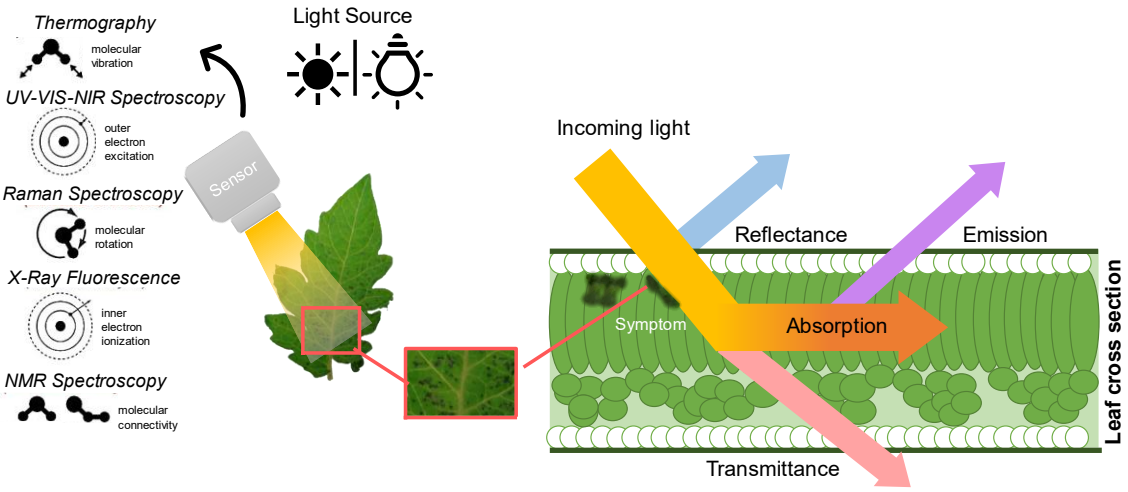



**Figure 3** Diagram showing the possible interactions between the electromagnetic radiation and schematic leaf surface (A). The cross-section illustrates the moment when stimulating light (parting from the sun or other light source) reaches the lesioned area of the tissue and a fraction of it may be promptly reflected and the remaining absorbed. From this, a part can be emitted in longer wavelengths (lower energy) as fluorescence or/and heat or transmitted (B). Different sensing technologies can then be used to measure this radiation, such as Thermography, UV-VIS-NIR Spectroscopy, Raman Spectroscopy, X-Ray Fluorescence, and NMR spectroscopy, among others.

Sensors can also be classified according to the type of light used in the measurement moment. They are designated as passive when no external light source is employed, and assessments are made using the sunlight. In contrast, the sensors are called active when a specific light source is applied in the process. Regarding the type of data measured, sensors used in plant disease studies can be single or multi-spatial point observation devices, SSPO, and MSPO, respectively. The SSPOP typically generates point data (1$^{st}$ order data), which does not provide spatially continuous information about how the data sensed varies within the sensors' field of view. Thus, they gather information from just one specific spot on a plant at a time (e.g., measuring from one spot on a leaf or a stem). In turn, MSPO sensors provide continuous spatial information within the sensor's field of view, sometimes in the form of a digital image,

also including information about the intensity of a given target signal (Manolakis, Lockwood et al. 2016). They collect data from multiple areas of a plant simultaneously or from several plants at once, providing information from different parts of the plant or multiple plants simultaneously. SSPO sensors allow the analysis of biological tissue's chemical composition, molecular structure, and other properties without spatial and color information (Tosin, Monteiro-Silva et al. 2023). In contrast, MSPO sensors usually have this additional information with a certain spatial resolution, allowing an analysis of a plant organ (such as leaf, stem, and root) (Manolakis, Lockwood et al. 2016).

### 4.2. Electromagnetic spectrum and plant physiology

During the past few decades, PS crop studies have predominantly focused on employing reflectance/transmittance techniques (Figure 3) operating within the visible (VIS, 400-700 nm), near-infrared (NIR, 700-1300 nm), short-wave infrared (SWIR, 1300-2500 nm) electromagnetic regions. Also, PS based on emission techniques operates on the thermal infrared range (TIR, 7000-20000 nm) and fluorescence (at 680-740 nm) spectroscopic sensors were used in crop studies (Galieni, D'Ascenzo et al. 2021) (Table 1). This emphasis is related to plant-pathogen interactions, particularly their impact on tissues' spectral behavior within the VIS and red-edge regions (RE, 670-760 nm), due to changes in photosynthetic pigment levels (primarily concerning chlorophyll degradations whose wavelength absorption ranges from 430 to 480 nm, and 640 to 700 nm). Such interactions lead to alterations in photosynthetic pigment levels, notably chlorophyll degradation, which manifests through changes in absorption wavelengths ranging from 430 to 480 nm and 640 to 700 nm (Bhandari, Wang et al. 2015, Buja, Sabella et al. 2021). Furthermore, photosynthesis disturbance can also be observed through the plant's fluorescence emission (450-550 nm, 690-720 nm). As the disease progresses and tissue senescence sets in, the emergence of brown pigments and structural alterations become apparent, leading to discernable spectral variations in the VIS and NIR regions, typically spanning from 680-800 nm (Buja, Sabella et al. 2021). Changes in carotenoid levels are mainly related to β-carotenes, whose primary and secondary absorption peaks are located at 450-480 nm and 600-650 nm, respectively, and to xanthophylls at 520 to 580 nm (Buja, Sabella et al. 2021). In turn, the chlorophyll breakdown can lead to the subsequent formation of pheophytins (brown pigments) characterized by a primary absorption peak at 660-670 nm and a secondary peak around 430-450 nm (Buja, Sabella et al. 2021).

Vibrational spectroscopy (Figure 3) was also used in plant studies, aiming to identify the chemical composition and molecular structure of biological tissues through

the analysis of light scattered from a sample. These approaches do not capture data across specific predefined spectral bands (like multispectral sensors) or numerous contiguous narrow bands (like hyperspectral devices) across the electromagnetic spectrum. An example is NIR spectroscopy which showed the capability of measuring compounds containing the groups -OH, -NH, and -CH (Cozzolino 2014, Türker-Kaya and Huck 2017), which are found in primary and secondary metabolites, important components of plants and plant defenses against pathogens (Conrad, Li et al. 2020). Also, variations in the water content of samples, characteristic of pathogen infection (through the occurrence of desiccation – wilting – and water-soaking lesions), are reflected in NIR spectra (Wang, Zhang et al. 2017, Conrad, Li et al. 2020). In turn, Raman spectroscopy was applied to sense the inelastic scattering of photons by molecular vibrational modes, which carry information about the chemical composition in the focal volume (Payne and Kurouski 2020).

It is important to address that some of these spectral modifications are specific to certain host-pathogen interactions, and their monitoring may lead to the disease diagnosis.

### 4.3. Sensor's characteristics and usability

Table 1 outlines the primary characteristics of the key sensors employed in diagnosing plant diseases, along with their TRL and usage frequency based on the conducted literature review. The main sensing technologies identified during the screening process were Biophoton Emission, Fluorescence Spectroscopy (SSPO and MSPO), Laser-Induced Breakdown Spectroscopy (LIBS), Multi- and Hyperspectral Spectroscopy (including SSPO and MSPO), Nuclear Magnetic Resonance (NMR) Spectroscopy, Raman Spectroscopy, RGB Imaging, Thermography, Volatile Organic Compounds (VOCs) assessment, and X-ray Fluorescence (XRF) MSPO (Table 1). A more detailed description of these techniques can be found in Supplementary Materials II. These sensors may differ from each other in the number of bands they can access (Table 1). As examples, panchromatic sensors capture information across a broad spectrum in a single band; RGB sensors work with red, green, and blue bands; multispectral sensors with broadband, multiple discrete bands across the electromagnetic spectrum; and hyperspectral sensors measure narrow contiguous features (<10 nm) (Liu, Bruning et al. 2020, Galieni, D'Ascenzo et al. 2021). The devices measuring fewer wavelengths are primarily used for qualitative analysis and show limited capability to provide quantitative data. In contrast, higher-precision sensors greatly

improved the ability to study specific atomic or molecular transitions, permitting more precise measurements (Galieni, D'Ascenzo et al. 2021).

The RGB imaging technique was found to have a higher TRL and is frequently used in plant disease studies. However, it is important to note that this technique only allows the diagnosis of plant diseases in symptomatic stages. Consequently, it was not included in the output tables of this study, as elaborated in the subsequent sections. Hyperspectral spectroscopy (including SSPO and MSPO techniques) was the subsequent approach reaching a higher TRL, likely motivated by its frequent application and study in plant disease diagnosis. Notably, it was found to be applied in at least 41 of the screened studies. In contrast, Biophoton emission and VOC assessment have a low TRL and were considered emerging techniques (Table 1).

**Table 1** Main characteristics and Technology Readiness Levels (TRL) of proximal sensors for early plant disease diagnosis.

| Spectral techniques | Energy / Light source | Spectral region | Information | Trait sensed | TRL | Nº Ref. |
|---|---|---|---|---|---|---|
| Biophoton Emission | Passive | VIS-NIR Photomultiplier | Emission spectra | Metabolites | 3 | 4 |
| Fluorescence Spectroscopy | Active (Blue/ Red/ UV LED) | Blue/Red/UV | Emission spectra Image | Metabolites Plant organ* | 4 | 3 |
| LIBS | Active (Laser) | UV-VIS-NIR | Atomic emission | Elemental composition | 4 | 2 |
| Multi- and Hyperspectral Spectroscopy | Passive; Active | UV-VIS-NIR | Reflectance/ Transmittance spectra Hyperspectral data cube (Image) | Metabolites Plant organ* | 5 | 41 |
| NMR | Active | Radiofrequency pulse | NRM spectra | Metabolites | 4 | 5 |
| Raman vibration Spectroscopy | Active (Laser) | IR | Raman spectra | Metabolites | 4 | 5 |
| RGB | Passive | Red, Blue, Green | RGB Image | Plant organ Plant Canopy | 8/9 | 8 |
| Thermography | Passive | NIR, SWIR, MIR | Thermal image | Plant organ* | 4 | 4 |
| VOCs | Electrical power | UV-VIS-NIR | Chromatogram, Mass spectrum, Pattern Recognition | Molecular composition Metabolites | 3 | 4 |
| XRF | Active (X-ray beam) | X-ray | XRF spectra XRF image | Elemental composition Plant organ* | 4 | 4 |

* Organ – Leaf, stem, root, among others; LED – Light-Emitting Diode; Nº Ref. – Number of references (articles); The TRL (Technology Readiness Level) assigned refers not to the technology but to the applications developed for early disease diagnosis. NIR – Near Infrared (700-1300 nm); TIR – Thermal infrared (7000-20000 nm); UV – Ultraviolet (200-400 nm); VIS – Visible (400-700 nm); NMR – Nuclear Magnetic Resonance; VOCs – Volatile Organic Compounds sensing; XRF – X-Ray fluorescence; LIBS – Laser Induced Breakdown Spectroscopy

## 5. Main findings of sensing technologies for plant early disease diagnostic

Literature review focused on the following criteria: (as depicted in the Supplementary Materials Figure S1): i) crop and pathogen studied, ii) the spectral

sensing technique used, iii) its spectral range, iv) the number of samples assessed, v) the methods used to analyze (model) data, vi) and the main statistics used to evaluate the models.

The selected articles were grouped in three tables according to the assay's conditions, namely, laboratory (controlled environment) depicted in Table 2, greenhouse/glasshouse in Table 3, and field in Table 4. Moreover, sensing technology performances were detailed in the next section.

### 5.1. Experimental conditions for sensor application in crop and pathogen setting

In terms of crops analyzed, research mostly referred to the assessment of tomato (16% of the articles) and wheat (13%), followed by sugar beet and soybean (used both in 9%) (Table 2, 3, 4). All these crops are highly economically important due to their widespread cultivation and consumption worldwide, especially wheat, which is considered a staple food source. Additionally, the crops studied were likely chosen for their simplicity in terms of cultivation and maintenance. Some of them (e.g., tomato, sugar beet, soybean, cucumber, pepper, among others) also present short life cycles, which make them suitable for multiple studies in a relatively short period. It was possible to acknowledge that several crops (e.g., avocado, cassava, cotton, among others) were mentioned in a single article.

Concerning the etiological agents studied, it was also possible to observe that Fungi were the more extensively screened pathogen, referred to in 53% of the articles assessed, followed by viruses (24%), bacteria (approximately 18%), and pests (around 9%) (Table 2, 3, 4). Several reasons may be related to these findings, such as economic importance of the impact of the disease/pest, global distribution, visibility and symptomatology, historical emphasis, availability of resources (e.g., pathogen collections), pathogen complexity, ecological significance, and resistance to phytosanitary products. Fungal diseases usually manifest more conspicuously in plants than viral, bacterial, or pest-related diseases (even before symptom appearance). Thus, this may enable the caption and understanding of more changes promoted by plant-pathogen interactions, motivating more research using these organisms.

The experimental conditions observed in the search results showed that most of the screened articles (69%) detailed experiments conducted in laboratory settings. Following this, greenhouse experiments were mentioned in 22% of the articles, while field experiments were documented in 20%, as illustrated in Tables 2, 3, and 4.

## 5.2. Sensing technologies applications

### 5.2.1. Post-symptom plant disease diagnosis

RGB imaging (Supplementary Materials II) was previously identified as suitable for post-symptomatic plant disease studies (involving diagnostic, quantification, and severity studies) (Steddom, McMullen et al. 2004, Turner, Martin et al. 2004). RGB was the most extensively explored, being evaluated in disease diagnosis studies in laboratory (Ferentinos 2018), greenhouse (Mellit, Benghanem et al. 2021), and field conditions (Zhou, Kaneko et al. 2014, Fan, Luo et al. 2022). It was also found that this technique is suitable for disease severity determination (Kang, Huang et al. 2022), pest quantification (Xia, Chon et al. 2015), and recognition/identification of diseases (Fan, Luo et al. 2022). The major drawback of this technique is it requires an extensive sensing area to generate an image, only allowing disease diagnosis after macroscopic visual symptom manifestation, not being suitable for an early diagnosis. For this reason, this technology was not further considered for metric evaluation.

### 5.2.2. Early disease diagnosis (before or at the development of the first symptoms)

Thermography (Supplementary Materials II) was one of the techniques identified as suitable for early diagnosis. It was successfully used to visualize the spatial colonization of apple tissues affected by the scab disease (*Venturia inaequalis*) over and beyond visible symptoms, where hyphae and conidia were only microscopically detectable (Oerke, Fröhling et al. 2011). In this study, leaves inoculated with conidia of *V. inaequalis* show concentric spots of non-standard low leaf temperature even before the appearance of visible scab symptoms (at 8 days after inoculation). The relationship between disease severity and maximum temperature difference, estimated through a regression analysis, achieved a square Pearson's coefficient of determination ($r^2$square) of 0.731 at 9 days after inoculation (after visible symptom appearance). Similarly, Chaerle et al. (1999) studied tobacco plants infected with tobacco mosaic virus (TMV) and found that sites of infection were 0.3–0.4°C warmer than the surrounding tissue approximately 8 hours before the initial appearance of the necrotic lesions (suggesting a presymptomatic diagnosis) (no metrics reported) (Chaerle, Van Caeneghem et al. 1999). Oerke et al. (2006) also analyze cucumber leaves infected by *Pseudoperonospora cubensis* (downy mildew) and the impact of this infection on metabolic processes and transpiration rate (Oerke, Steiner et al. 2006). They concluded that healthy and infected leaves can be discriminated against even before symptoms appear (no metrics shown). Since these two scientific did not present the requested model metrics to integrate this work, they were not further explored in Tables 2, 3, and 4

results. Nevertheless, the authors consider their mention important due to their relevant findings. It is important to address that when Thermography sensors are used, smaller disease lesions might not be distinguishable due to the equipment resolution. Moreover, these minor lesions might not emit enough energy to be detected by sensors with lower resolution.

Hence, 46 (58%) articles were selected to integrate the results in Tables 2, 3, and 4. According to their analysis, our research indicates that hyperspectral spectroscopy was the most used technique, followed by hyperspectral imaging (MSPO, 24%), and Raman spectroscopy (7%). Fluorescence spectroscopy, LIBS (4% of the articles), and XRF (1 article) were also mentioned (Tables 2, 3, 4).

Hyperspectral devices (Supplementary Materials II), both spectroscopy (SSPO) and imaging-based (MSPO), were reported in laboratory (Shuaibu, Lee et al. 2018, Gold, Townsend et al. 2020), greenhouse (Rangarajan, Whetton et al. 2022, Griffel, Delparte et al. 2023), and field (Almoujahed, Rangarajan et al. 2022, Zhang, Jing et al. 2023) disease diagnosis studies (Table 2, 3, and 4, respectively). These sensors have been increasingly used in early disease diagnosis, before symptom appearance, allowing the distinction between healthy, non-symptomatic, and symptomatic tissues (Table 2, 4) (Gold, Townsend et al. 2020, Nguyen, Sagan et al. 2021, Reis Pereira, Santos et al. 2023, Zhu, Su et al. 2023). For example, Rumpf et al. (2010) used hyperspectral spectroscopy to detect and distinguish different pathogens responsible for causing leaf diseases in sugar beet plants, namely *Cercospora beticola*, *Uromyces betae* or *Erysiphe betae* causing Cercospora leaf spot, sugar beet rust, and powdery mildew, respectively (Rumpf, Mahlein et al. 2010). Overall, they could discriminate between healthy and diseased leaves with up to 97% classification accuracy (Rumpf, Mahlein et al. 2010) (Table 2). Multiclass classification between healthy and symptomatic samples of the three diseases achieved an accuracy higher than 86%. The potential for presymptomatic diagnosis demonstrated that, depending on the type and stage of disease infection, the classification accuracy varied from 65 to 90% (Table 2). Likewise, Herrmann et al. (2018) studied the suitability of HS for early predicting Sudden Death Syndrome, caused by *Fusarium virguliforme*, in soybean plants in field conditions (Herrmann, Vosberg et al. 2018). The presymptomatic diagnosis was possible using canopy and leaf spectral data, with a classification accuracy of 82% and 92%, respectively, for validation (Herrmann, Vosberg et al. 2018) (Table 4). More important research on this topic is referred to in Tables 2, 3, and 4.

Fluorescence Spectroscopy (FS, Supplementary Materials II) has also been applied in the assessment of physiological states and stress levels of plants, including disease studies, even in field conditions (Römer, Bürling et al. 2011, Bürling, Hunsche et al. 2012). Römer et al. (2011) applied FS in the study of presymptomatic diagnosis of leaf rust in wheat plants (Römer, Bürling et al. 2011). Only two days after infection, the developed model could diagnose this pathogen with a classification accuracy reaching 93% (Table 2) (Römer, Bürling et al. 2011). Atta et al. (2023) similarly studied the early diagnosis of stripe rust in wheat through the application of light-induced fluorescence spectroscopy (Atta, Saleem et al. 2023). Lower chlorophyll bands were noticeable at 685 and 735 nm in both the non-symptomatic and symptomatic leaf samples (Atta, Saleem et al. 2023). A Partial Least Square Regression (PLSR) model was computed, resulting in a standard error of calibration (SEC) of 0.200, a standard error of prediction (SEP) of 0.140, and a coefficient of determination (R2) of 0.77 (Table 2) (Atta, Saleem et al. 2023).

A commercial fluorescence sensor known in the literature for being suitable for early plant disease diagnosis is Multiplex® (FORCE-A, Orsay, France). It is a hand-held, multi-parametric fluorescence sensor based on light-emitting-diode (LED) excitation and filtered-photodiode detection designed to work in the field under daylight on several crops. Its excitation light sources are mainly located in the UV (365 nm or 340 nm), blue (465 nm), green (520 nm), and red (630 nm), and can measure simultaneously various compounds (e.g. anthocyanins, flavonoids, chlorophyll) (GmbH 2014, ForceA 2019). Bellow et al. (2012) applied the Multiplex diagnosing downy mildew in grapevine leaves (Bellow, Latouche et al. 2012). The authors demonstrated that the stilbene-dependent violet–blue autofluorescence (VBF), an organic compound, had a transitory behavior, increasing to a maximum 6 days post-inoculation (DPI) and then decreasing to a constant lower level when compared to healthy leaves (Bellow, Latouche et al. 2012). On the abaxial side, VBF could discriminate the presence of infection from 1 DPI, and on the adaxial side from 3 DPI (Bellow, Latouche et al. 2012). Yu et al. (2013) also studied Florescence, along with Hyperspectral data to investigate leaf diseases in different barley varieties and estimate leaf chlorophyll concentration (LCC) (Yu, Leufen et al. 2014). The plants of the plot without fungicide treatment showed mild infections with a few punctiform visible symptoms (Yu, Leufen et al. 2014). Detached leaves, allowed in laboratory estimation of LCC with a coefficient of determination ($R^2$) of 0.72 and Root-mean-square error (RMSE) of 1376.3 µg/g, when the validation set was used, and when the blue to far-red fluorescence ratio was used (Yu, Leufen et al. 2014). Support Vector Regression (SVM) was further computed to improve the accuracy in

estimating LCC using fluorescence signals, yielding an $R^2$ of 0.84 and an RMSE of 1021.91 µg/g when the validation set was used (Yu, Leufen et al. 2014).

In plant studies, Raman Spectroscopy (RaS, Supplementary Materials II) methodology has proved to be effective in the detection of the most generally encountered types of phytopathogens, such as fungi (through the impact of mycotoxins), bacteria, viruses, or nematodes (Sylvain and Cecile 2018, Payne and Kurouski 2020). Early assessment success was achieved by Madrile, et al. (2019) in the discrimination of the infection of tomato samples by Tomato yellow leaf curl Sardinia virus (TYLCSV) and Tomato spotted wilt virus (TSWV) (Mandrile, Rotunno et al. 2019). A chemometrics analysis was performed including the computation of a Principal Component Analysis (PCA), and a Partial least squares discriminant analysis (PLS-DA) (Mandrile, Rotunno et al. 2019). Early diagnosis was associated with an accuracy higher than 70% for TYLCSV and 85% for TSWV (Table 2) (Mandrile, Rotunno et al. 2019). Sanchez et al. (2020) employed a hand-held RaS device for the non-invasive and early (non-symptomatic) detection of two haplotypes of Liberibacter disease on tomatoes (Sanchez, Ermolenkov et al. 2020). They detected structural changes in carotenoids, xylan, cellulose, and pectin that were later related to bacterial disease (Sanchez, Ermolenkov et al. 2020). A Partial least squares discriminant analysis (PLS-DA) was performed, allowing 80% accurate diagnostics of Liberibacter disease caused by each of the two different haplotypes (Table 2) (Sanchez, Ermolenkov et al. 2020). Furthermore, Vallejo-Pérez et al. (2021) applied RaS and machine learning to diagnose the bacterial canker in asymptomatic tomato samples (Vallejo-Pérez, Sosa-Herrera et al. 2021). They computed a Principal Component Analysis (PCA) in combination with a multilayer perceptron (PCA+MLP) and, in a different approach, with a linear discriminant analysis (PCA+LDA) (Vallejo-Pérez, Sosa-Herrera et al. 2021). The spectra obtained from diseased leaves showed peaks related to cellular components, and the most outstanding vibrational bands were associated with carbohydrates, carotenoids, chlorophyll, and phenolic compounds (Vallejo-Pérez, Sosa-Herrera et al. 2021). Bands linked with triterpenoids and flavonoid compounds were, furthermore, identified as indicators of the pathogen infection (Vallejo-Pérez, Sosa-Herrera et al. 2021). Classification performance demonstrated an accuracy of 0.99, f1-score of 0.99, sensitivity of 1.0, and specificity of 0.95 were achieved when PCA and MLP were combined (Table 2) (Vallejo-Pérez, Sosa-Herrera et al. 2021).

The other vibrational spectroscopy technique identified, Nuclear Magnetic Resonance (NMR) spectroscopy (Supplementary Materials II), was used in the identification of several metabolites (biomarkers) indicators of several infections (Choi,

Tapias et al. 2004, Ali, Maltese et al. 2012, Freitas, Carlos et al. 2015, Pontes, Ohashi et al. 2016). Pontes et al. (Pontes, Ohashi et al. 2016) studied combining this methodology with chemometrics for diagnosing the citrus greening disease (caused by *Candidatus Liberibacter* spp.) in leaves. They discovered that class discrimination was achievable by the computation of a Principal Component Analysis (PCA), separating healthy from diseased samples (PC1-PC2 score of 76.4%), healthy leaves from the ones infected with insects (PC1-PC2 78.1%), and all the classes from each other (PC1-PC2 81.7%). Studies using NMR similar to XRF, are still insufficient, and further research is advised. Despite these interesting findings, this work was not included in the final result tables since it was performed after insect/symptom visualization.

Laser-Induced Breakdown Spectroscopy (LIBS, Supplementary Materials II) demonstrated its suitability for plant disease studies by enabling the differentiation between healthy and disease plants through the monitoring of changes in their macro and micronutrient composition, even during non-symptomatic stages. In fact, Pereira et al. (2010) proposed a method for the classification of citrus leaves infected by greening disease (caused by the bacteria *Candidatus Liberibacter asiaticus*) through the analysis of LIBS spectra (Pereira, Milori et al. 2010). The authors could develop predictive models for assessing healthy and infected plants, based on relevant differentiations in major, macro-, and microconstituents of samples (both organic and inorganic) of these two health statuses (Pereira, Milori et al. 2010). Classification efficacy of diseased samples ranged from 82 to 97% (Table 3) (Pereira, Milori et al. 2010). Similarly, Ranulfi et al. (2018) studied the green stem and foliar retention (GSFR), caused by the pest *Phelenchoides besseyi*, on soybean plants (Ranulfi, Senesi et al. 2018). The authors concluded that healthy plants had higher concentrations of calcium and magnesium, whereas sick plants had higher concentrations of potassium (Ranulfi, Senesi et al. 2018). Partial Least Square Regression (PLSR) allowed class differentiation with rates higher than 80% (Table 2) (Ranulfi, Senesi et al. 2018).

X-Ray Fluorescence (XRF, Supplementary Materials II) is another technique that was also applied in plant disease investigation (Sharma, Khajuria et al. 2018, Sharma, Khajuria et al. 2020) and diagnosis (Pereira and Milori 2010). Pereira and Milori in 2009 related the usage of X-ray fluorescence and chemometric tools for studying the citrus greening disease (*Candidatus Liberibacter asiaticus*) (Pereira and Milori 2010). They identified the signals for potassium, calcium, iron, copper, and zinc and the region of coherent and incoherent scatterings as significant for distinguishing healthy samples from diseased ones (Table 2) (Pereira and Milori 2010). Rodrigues et al. (2018) employed XRF to acquire chemical images of soybean leaves infected by anthracnose

(*Colletotrichum truncatum*), demonstrating that phosphorus, sulfur, and calcium were concentrated in the diseased regions (Rodrigues, Gomes et al. 2018). Despite these important findings, research is still scarce (Table 2) (Rodrigues, Gomes et al. 2018).

### 5.2.3. Emerging techniques with diagnostic suitability

Biophoton Emission and Volatile Organic Compounds (VOCs) analysis (Supplementary Materials II) were two sensing techniques identified during the screening process and classified as emerging due to their low TRL (3) (Table 1). This indicates that these techniques are mostly in a stage of experimental proof of concept development, and only addressed by a limited number of scientific works.

In terms of emerging techniques for plant disease diagnosis, the detection and quantification of biophotons (Biophoton spectroscopy, Supplementary Materials II) were found to be promising since the emission of these particles increases significantly when plants are infected by pathogens (Kawabata, Miike et al. 2005, Kobayashi, Sasaki et al. 2007). Two types of biophoton emission can be identified according to disease-resistance reactions: relatively weak emissions observed during the early stages of the resistance reaction (Iyozumi, Kato et al. 2005), and strong emissions from cells exhibiting programmed cell death (PCD) during the middle stages of the resistance reaction (Bennett, Mehta et al. 2005). Iyozumi et al. (2002) using sweet potatoes inoculated with *Fusarium oxysporum* showed that photon emissions have their wavelength composition shifted toward a shorter wavelength as compared with that of untreated samples, indicating that this was a luminous phenomenon quantitatively different from the one observed under normal conditions (Iyozumi, Kato et al. 2002). Kawabata et al. (2004) analyzed the impact of spider mites (*Tetranychus kanzawai Kishida*) on kidney bean leaves, demonstrating higher biophoton emission from leaf veins where the pests were crowding (Kawabata, Uefune et al. 2004). Along with this, photon emission intensity augmented with the decrease in chlorophyll content and photosynthesis yield (Kawabata, Uefune et al. 2004).

Moreover, the analysis of Volatile Organic Compounds (VOCs, Supplementary Materials II) emitted by disease plants in contrast with the profile emitted by healthy plants also seems a promising tool for providing non-invasive monitoring of a plant's physiological health status (Table 1) (Cellini, Biondi et al. 2016). Cardoza et al. (2003) performed VOC profiling for discriminating healthy peanut (*Arachis hypogaea*) plants from individuals infected with white mold (fungi, *Sclerotium rolfsii*) (Cardoza, Teal et al. 2003). They detected considerable differences in methyl salicylate (phenolic compound) and 3-octanone, two secondary metabolites (Cardoza, Teal et al. 2003). Likewise, Mauck

et al. (2010) assessed the VOC profile of squash (*Cucurbita pepo* cv. Dixie) to study the Cucumber mosaic virus disease in greenhouse and field conditions (Mauck, De Moraes et al. 2010). The authors showed that diseased individuals presented higher VOC levels when compared to healthy plants, but these compounds were overall qualitatively similar (Mauck, De Moraes et al. 2010). López-Gresav et al. (2010) monitored tomato plants infected with two types of pathogens, namely bacteria Pseudomonas syringae pv. tomato, and citrus exocortis viroid (CEVd), to accompany the VOC profiling evolution (López-Gresa, Maltese et al. 2010). Plants infected with bacteria pathogens presented increased levels of amino acids, rutin, and phenylpropanoids (López-Gresa, Maltese et al. 2010). In turn, individuals affected by plants revealed differences in glucose and malic acid production (López-Gresa, Maltese et al. 2010).

Similarly to what happened in the thermography technique, the scientific articles mentioned in this subsection were not considered in Tables 2, 3, and 4 due to lacking criteria such as model performance analysis and metrics.

**Table 2** Selected findings of the bibliography review of the early diagnosis of several plant pathogens assessment in indoor (controlled) conditions.

| Culture & Pathogen | Spectral sensor | | | Modelling | | Ref. |
|---|---|---|---|---|---|---|
| | *Technique* | *Range* | *Samples* | *Method* | *Statistics* | |
| Citrus, Virus | X-ray fluorescence | Max. 12 keV | 162 samples | SIMCA, KNN, PLS-DA, PCA | Correct classified: 90% (validation) | (Pereira and Milori 2010) |
| Sugar beet, Fungi | Hyperspectral spectroscopy | 450-1050 nm | 15 plants; 630 spectra | SVM | Low severity (<=5%): Acc. 84.3%, Rec. 78.32% | (Rumpf, Römer et al. 2010) |
| Sugar beet, Fungi | Hyperspectral spectroscopy | 350-1100 nm | 15 plants; 630 spectra | Regression | r: 0.85 VIS, ARI 0.54-0.77. REP 0.58-0.75 | (Mahlein, Steiner et al. 2010) |
| Sugarcane, Virus | Hyperspectral spectroscopy | 350-800 nm | 40 leaves | DA | Resubstitution (all dates) 73% | (Grisham, Johnson et al. 2010) |
| Sugar beet, Fungi | Hyperspectral spectroscopy | 400-1050 nm | 30 plants | DT, ANN, SVM | Acc.: 97%, Multi classes: Acc. 86%, Rec. 84-92%. Presymptomatic: 65-90% | (Rumpf, Mahlein et al. 2010) |
| Soybean, Virus | Hyperspectral spectroscopy | 730-1025 nm | 20 plants; 2400 spectra | SIMCA | Sen.: 91.6%, Spe.: 95.8% | (Jinendra, Tamaki et al. 2010) |
| Wheat, Fungi | Fluorescence spectroscopy | 370-800 nm | 72 leaves; 215 spectra | SVM, DT, ANN | Classification: 73.6% 2 DAI, 79.2% 3 DAI, 80.7% 4 DAI | (Römer, Bürling et al. 2011) |
| Oilseed rape, Pest | Hyperspectral imaging | 380-1030 nm | 510 images | G-WNNRA, WNN, GNN, BPNN | G-WNNRA: Calibration R 0.998, $R^2$ 0.996; Validation R 0.953, $R^2$ 0.908 | (Zhao, He et al. 2012) |
| Strawberry, Fungi | Hyperspectral imaging | 400-1000 nm | 5 inoculated, 1 healthy | SAM, SDA, SSM | Acc.: SAM 82.0%, SDA 80.7%, SSM 72.7% | (Yeh, Chung et al. 2013) |
| Eggplant, Bacteria | Hyperspectral spectroscopy | 350-2500 nm | 12 plants | Mean Percent Difference | MDP: 19.51% pre-symptom Day 4, 60.53 maximum Day 8 | (Chew, Hashim et al. 2014) |
| Avocado, Fungi | Hyperspectral spectroscopy | 350-2500 nm | 80 leaves; 800 spectra | SDA, NN - MLP, RBF | Correct classification: MLP 96-99%, RBF 65% | (Abdulridha, Ehsani et al. 2016) |
| Citrus, Virus | Hyperspectral spectroscopy | 400-1100 nm | 150 samples | PCA, N PCA, KNN | Classification: 60-90%. Overall 92% | (Afonso, Guerra et al. 2017) |
| Soybean, Pest | LIBS | 189-966 nm | 70 plants | PCA, CVR+PLSR | Success rate: > 80% | (Ranulfi, Senesi et al. 2018) |
| Apple, Fungi | Hyperspectral imaging | 356-1000 nm | 260 trees, >1000 leaves; 2000 pixels | OSP, DT, EB, Weighted KNN | EB: Acc. 84.3% Overall, 83.2% Healthy, 67.6% Asymptomatic, 89.4-97.3% | (Shuaibu, Lee et al. 2018) |
| Tomato, Pest | Hyperspectral imaging | 400-2500 nm | 42 plants | PLS-SVM | Acc.: 90-100% | (Susič, Žibrat et al. 2018) |
| Wheat, Fungi | Hyperspectral imaging | 375-1017 nm | 184 samples | BPNN | PCA BPNN: $R^2$ 0.92, RMSEP 1.07, RPD 3.36. SPA BPNN: $R^2$ 0.92, RMSEP 1.10, RPD 3.26 | (Yao, Lei et al. 2019) |
| Cucumber, Fungi | Hyperspectral spectroscopy | 450-1100 nm | 152 plants | SIMCA | Acc.: SIMCA > 78% | (Atanassova, Nikolov et al. 2019) |
| Tomato, Virus | Raman spectroscopy | 400-3100 $cm^{-1}$ | 3 plants | PLS-DA | Yellow leaf curl, 14 DAI: Acc. 71%, Sen. 0.80, Spe. 0.67. Spotted wilt, 8 DAI: Acc. 89%, Sen. 0.80, Spe. 1.00 | (Mandrile, Rotunno et al. 2019) |
| Pepper, Virus | Hyperspectral imaging | 400-1000 nm | 20 plants | OR-AC-GAN, MLP | Acc. before symptom: 96.25%, Pixel prediction false positive rate 1.57% | (Wang, Vinson et al. 2019) |
| Potato, Fungi | Hyperspectral spectroscopy | 350-2500 nm | 2039 spectra | PCoA, PLS-DA | Pre-symptomatic: Kap. 0.87 LB, 0.94 EBl. LB vs. EBl: Kap. 0.78, Acc. 91.3% | (Gold, Townsend et al. 2020) |
| Potato, Fungi | Hyperspectral spectroscopy | 350-2500 nm | 1330 spectra | NDSIs, PERMANOVA, PCoA, PLS-DA, RF, PLSR | Acc. 71%, Kap. 0.35; Control vs Pre-symptom Acc. 83%, Kap. 0.37; Pre vs Post-symptom Acc. 71%, Kap. 0.41 | (Gold, Townsend et al. 2020) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Tomato, Bacteria | Raman spectroscopy | 350–2000 $cm^{-1}$ | 36 plants | PLS-DA | Correct classifications: 80% | (Sanchez, Ermolenkov et al. 2020) |
| Grapevine, Fungi | Hyperspectral imaging | 397-1003 nm | 35 plants | LDA, QDA, RDA, SkNN, NB, RPART | NB: Healthy vs. asymptomatic Acc. 76%, Kap. 0.31, EO 0%, EC 75% | (Calamita, Imran et al. 2021) |
| Apple, Bacteria | Hyperspectral spectroscopy | 1100-2498 nm | 862 samples | PLS-DA | Non- vs. symptomatic Acc.: 71%, Sen. 70%, Spe.: 72% | (Barthel, Dordevic et al. 2021) |
| Wheat, Fungi | Hyperspectral imaging | 400-1000 nm | 24 pots | PLS-LDA, PLSR | PLS-LDA: Acc. 85% (3-6 DAI, DS 1-6%); PLSR: $R^2$ 0.818 | (Khan, Liu et al. 2021) |
| Tomato, Bacteria | Raman spectroscopy | 800–1800 $cm^{-1}$ | 297 spectra | PCA+MLP, PCA+LDA | PCA+MLP: Acc. 0.99, Sen. 1.0, Spe. 0.95, PPV 0.98, NPV 1.0, F1 0.99 | (Vallejo-Pérez, Sosa-Herrera et al. 2021) |
| Wheat, Fungi | Fluorescence spectroscopy | 400-900 nm | 66 spectra | PLSR | SEP 0.14, SEC 0.20, $R^2$ 0.77, RMSEP (test) 0.08 | (Atta, Saleem et al. 2023) |
| Tomato, Bacteria | Hyperspectral spectroscopy | 400-800 | 3478 spectra, 9 plants | LDA, SVM | Non-symptomatic (Test set): Pst Acc. 100%, Pre. 0.94, Rec. 1.00, F1 0.97; Xeu Acc. 74%, Pre. 0.77, Rec. 0.74, F1 0.75 | (Reis Pereira, Santos et al. 2023) |
| Barley, Eggplant, Cucumber, Fungi, Bacteria | Hyperspectral imaging | 380-1030 nm | 250 samples; 138 images | PLSR, MLR | $R^2$: Chl-a 0.88, Chl-b 0.88, Car 0.87. RMSE: Chl-a 0.08, Chl-b 0.02, Car 0.01. RPD: Chl-a 2.97, Chl-b 3.17, Car 2.90 | (Zhu, Su et al. 2023) |

Acc. – Accuracy, ANN – Artificial Neural Network, ARI – Anthocyanin Reflectance Index, BPNN – Back Propagation Neural Network, Car – Carotenoids, Chl – Chlorophyll, CR– Cubist Regression, CVR – Classification Via Regression, DA – Discriminant Analysis, DAI – Days After Infection, DGND – Dual– Green Normalized Difference, DGSR – Dual– Green Simple Ratio, DS– Disease Severity, DT – Decision Trees, EB – Ensemble Bagged, EBl – Early Blight, EC – Error of Commission, EO – Error of Omission, ET – Extra Trees, F1 – F1 Score, FDA – Flexible Discriminant Analysis, FiDA – Fisher Discriminant Analysis, GA – Genetic Algorithm, GLM – Generalized Linear Model, GLMVQ – Generalized Matrix Relevance Learning Vector Quantization, GNN – Genetic Neural Network, G– WNNRA – Genetic– Wavelet Neural Network Reconstruction Algorithm, KNN – k– Nearest Neighbor, LB – Late Blight, LDA – Linear Discriminant Analysis, MDP – Mean Percent Difference, MLP – Multilayer Perceptron, MLR – Multiple Linear Regression, NB – Naïve Bayes, NDSI – Normalized Difference Spectral Index, NN – MLP – Neural Networks Multilayer Perceptron, N PCA – N– Way Principal Component Analysis, NPV – Negative Prediction Values, OR– AC– GAN – Outlier Removal Auxiliary Classifier Generative Adversarial Nets, OSP – Unsupervised Feature Selection using Orthogonal Subspace Projection, Kap. – Kappa, PCA – Principal Component Analysis, PCoA – Principal Coordinate Analysis, PERMANOVA – Permutational Multivariate Analysis of Variance, PLS – Partial Least Squares, PLS– DA – Partial Least Squares Discriminant Analysis, PLSR – Partial Least Squares Regression, PPV – Positive Prediction Values, Pre. – Precision, QDA – Quadratic Discriminant Analysis, r / R – Correlation Coefficient, R2– Coefficient of Determination, RBF – Radial Basis Function, RDA – Regularized Discriminant Analysis, Rec. – Recall, REG – Regression, REP – Red Edge Position, RF – Random Forest, RFCNN – Random Forest Convolutional Neural Network, RMSE – Root– Mean– Square Error, RMSEP– Root– Mean– Square Error Prediction, RPART – Recursive Partitioning Regression Tree, RPD – Residual Predictive Deviation, RR – Ridge Regression, SAM – Spectral Angle Mapper, SDA – Stepwise Discriminant Analysis, SEC – Standard Error of Calibration, Sen.– Sensitivity, SEP – Standard Error of Prediction , SIMCA – Soft Independent Modelling of Class Analogies, SkNN – Simple k– Nearest Neighbor, Spe. – Specificity, SSM – Simple Slope Measure, SVM – Support Vector Machine, VQ – Vector Quantification, WNN – Wavelet Neural Network

**Table 3** Selected findings of bibliography review of the early diagnosis of distinct crop diseases in greenhouse/glasshouse conditions. Legend in Table 2 footnote.

| **Culture & Pathogen** | **Spectral sensor** | | | **Modelling** | | **Ref.** |
|---|---|---|---|---|---|---|
| | *Technique* | *Range (nm)* | *Samples* | *Method* | *Statistics* | |
| Citrus, Virus | LIBS | 189-966 | 2560 leaves | SIMCA | Correct predictions: 82–97% | (Pereira, Milori et al. 2010) |
| Apple, Fungi | Thermography | 8000-12000 | > 260 samples | GLM, REG | 9 days: r2 square (standard deviation) 0.731 | (Oerke, Fröhling et al. 2011) |
| Tomato, Fungi | Hyperspectral spectroscopy | 380-1000 | 90 plants; 1350 spectra | LDA | Class. %: 100 (before symptom) | (Marín-Ortiz, Gutierrez-Toro et al. 2020) |
| Cassava, Virus | Hyperspectral spectroscopy | 360-1100 | 27 plants; 450 healthy, 765 disease spectra | KNN, ET, SVM, GMLVQ | Acc.: KNN 0.695-0.735, ET 0.708-0.766, SVM 0.641-0.812, GMLVQ 0.831-0.995 | (Owomugisha, Nuwamanya et al. 2020) |
| Rice, Fungi | Hyperspectral spectroscopy | 350-2500 | 36 pots | SVM, KNN, LDA | Acc.: 65% asymptomatic, 80% early stage, 95% mild stage | (Tian, Xue et al. 2021) |
| Soybean, Fungi | Hyperspectral spectroscopy | 350-2500 | 90 samples | LDA | Acc.: 100% Calibration, 82.51% Validation | (Furlanetto, Nanni et al. 2021) |
| Tomato, Bacteria | Hyperspectral spectroscopy | 400-1000 | Samples: 354 leaves, 179 stem | SVM | Acc.: GA-SVM 90.7% leaves, 92.6% stems. Reliability (2022): Acc. 88.6, F1 0.80 | (Cen, Huang et al. 2022) |
| Cotton, Pest | Hyperspectral spectroscopy | 350-2500 | - | Maximum likelihood (ML) | Acc. > 98% | (Ramamoorthy, Samiappan et al. 2022) |
| Soybean, Bacteria | Hyperspectral imaging | 400-1000 | 106 images | ANOVA | Acc. 57.41 to 62.26% | (Lay, Lee et al. 2023) |

**Table 4** Selected findings of bibliography review of the early diagnosis of different plant diseases including field conditions. Legend in Table 2 footnote.

| Culture & Pathogen | Sensor | | | Modelling | | Ref. |
|---|---|---|---|---|---|---|
| | *Technique* | *Range (nm)* | *Samples* | *Method* | *Statistics* | |
| Sugar beet, Fungi, Greenhouse, Field | Hyperspectral spectroscopy Hyperspectral Imaging | 400-1050 | 1504 samples (630 Indoor, 311 Field) | - | Acc./ Pre.: 89% Healthy, 85-92% Disease. Rec.: 94% Healthy, 74-89% Disease | (Mahlein, Rumpf et al. 2013) |
| Wheat, Fungi, Greenhouse, Field | Hyperspectral spectroscopy | 325-1075 | Spectra: 96 Pot, 51 Field | Regression | $R^2$: 0.845 DGSR (584, 550), 0.845 DGND (584, 550) | (Feng, Shen et al. 2016) |
| Strawberry, Fungi, Indoor, Field | Hyperspectral spectroscopy | 350-2500 | Indoor: 159 spectra. Field: 200 plants, 474 spectra | FiDA, SDA, KNN | Field Acc.: 74%. SDA 71.3%, FDA 70.5%, KNN 73.6% | (Lu, Ehsani et al. 2017) |
| Soybean, Fungi | Hyperspectral spectroscopy | 340-2500 | 500 plants (4844 spectra) | PLSR, PLS-DA | (Validation) Acc.: 82% canopy, 92% leaf. $R^2$: 0.62, RMSE 0.31 | (Herrmann, Vosberg et al. 2018) |
| Grapevine, Virus | Hyperspectral spectroscopy | 350-2500 | 10 plants (120 leaves); 1080 spectra | QDA, NB | Overall Acc. 75-99% QDA. 66-90% NB | (Sinha, Khot et al. 2019) |
| Wheat, Fungi | Hyperspectral spectroscopy | 350-2500 | 432 plots | PLS, RR, RF, CR, PLS-DA, PLS | PLS, RR: $R^2$ 0.64, RMSE 0.063. CR: $R^2$ 0.67, RMSE 0.061. PLSDA: Acc. 0.86 (validation). PLS: $R^2$ 0.71, RMSE 0.068 | (Anderegg, Hund et al. 2019) |
| Grapevine, Virus | Hyperspectral imaging | 400-1000 | 40 images | SVM, RFCNN | Test: VIs SVM Acc. August 7$^{th}$ 67.81%, All 65.70%. PCA + K-PCA Acc. August 7$^{th}$ 77.75%, All 73.62% | (Nguyen, Sagan et al. 2021) |
| Kiwi, Bacteria | Hyperspectral spectroscopy | 400-1010 | 504 spectra, 20 plants | FDA, GLM, PLS, SVM | Acc. 85%, Kap. 0.70, F1 0.84 | (Reis-Pereira, Tosin et al. 2022) |

## 6. Data handling and modeling approaches

### 6.1. Preprocessing approaches

This section introduces the main approaches applied to extract useful information from data collected by PS aimed at an early plant disease diagnosis found in the scientific articles mentioned in Tables 2, 3, and 4. Usually, the first step after data assessment involves a preprocessing approach aiming to identify and handle missing values and outliers (Jinendra, Tamaki et al. 2010, Lu, Ehsani et al. 2017) (Table 2, 4), along with denoising and smoothing methods (Table 2, 4). In this regard, strategies like normalization (used in 7% of the articles), Standard Normal Variate (SNV) (4%), Multiplicative Scatter Correction (MSC) (4%), and Savitzky−Golay filter (22%) are the ones found to be more applied in the articles assessed (Table 2, 3, 4).

In some situations, the data collected may present high dimensionality (e.g., Hyperspectral SSOP and MSPO sensors), resulting from similar or even overlapping spectral information presented in contiguous zones of the spectrum. This redundancy increases the complexity of data analysis and increases the risk of overfitting occurrence when modeling strategies are later computed. Furthermore, biological data can present super-imposed information in the measured spectra at different interference levels (Tosin, Martins et al. 2022). Thus, several Feature Engineering (FE, spectral unmixing) strategies were developed to mitigate the effects of high dimensionality and collinearity, mostly based on identifying and extracting the most relevant and distinctive spectral features (without losing relevant information).

Two main types of FE techniques were identified from the scientific articles studied: Feature Selection (FeS) and Dimensionality Reduction (DR) (Figure 4). FeS involves the identification of a subset of the original spectral features (variables) from the dataset, aiming to retain the most informative and relevant features for the target feature, while not considering redundant or irrelevant ones. As a classic practice, the computation of Vegetation Spectral Indices (VIs) was identified as one of the most used FeS techniques (Figure 4). They consist of the mathematical combination of two or more wavelengths, developed with a biophysical significance, and used to retrieve information related to plants' traits contained in proximal collected data. The VIs most mentioned in the scientific articles screened (Tables 2, 3, and 4) were the Anthocyanin Reflectance Index (ARI) (mentioned in 13% of the articles) (Mahlein, Steiner et al. 2010, Rumpf, Mahlein et al. 2010, Mahlein, Rumpf et al. 2013, Feng, Shen et al. 2016, Calamita, Imran et al. 2021, Khan, Liu et al. 2021), Ashburn Vegetation Index (AVI) (2%) (Calamita, Imran et al. 2021), Cellulose Absorption Index (CAI) (2%) (Chew, Hashim et al. 2014), Chlorophyll Green (Chl) (2%) (Calamita, Imran et al. 2021), Modified Cellulose

Absorption Index (mCAI) (9%) (Mahlein, Steiner et al. 2010, Rumpf, Mahlein et al. 2010, Mahlein, Rumpf et al. 2013, Lu, Ehsani et al. 2017), Modified Chlorophyll Absorption in Reflectance Index (mCARI) (13%) (Chew, Hashim et al. 2014, Feng, Shen et al. 2016, Lu, Ehsani et al. 2017, Khan, Liu et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022), Modified Simple Ratio (mSR) (7%) (Feng, Shen et al. 2016, Lu, Ehsani et al. 2017, Khan, Liu et al. 2021), Normalized Difference Vegetation Index (NDVI) (20%) (Mahlein, Steiner et al. 2010, Rumpf, Mahlein et al. 2010, Mahlein, Rumpf et al. 2013, Chew, Hashim et al. 2014, Lu, Ehsani et al. 2017, Atanassova, Nikolov et al. 2019, Calamita, Imran et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022), Optimized Soil Adjusted Vegetation Index (OSAVI) (7%) (Calamita, Imran et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022), Photochemical Reflectance Index (PRI) (22%) (Mahlein, Rumpf et al. 2013, Chew, Hashim et al. 2014, Feng, Shen et al. 2016, Lu, Ehsani et al. 2017, Anderegg, Hund et al. 2019, Atanassova, Nikolov et al. 2019, Khan, Liu et al. 2021, Nguyen, Sagan et al. 2021, Tian, Xue et al. 2021, Cen, Huang et al. 2022), Simple Ratio 800/680 Pigment Specific Simple Ratio (Cholophyll a) (PSSRa) (7%) (Rumpf, Mahlein et al. 2010, Mahlein, Rumpf et al. 2013, Tian, Xue et al. 2021), Simple Ratio 800/635 Pigment Specific Simple Ratio (Cholophyll b) (PSSRb) (4%) (Mahlein, Rumpf et al. 2013, Lu, Ehsani et al. 2017), Red-Edge Position Linear Interpolation (REP) (4%) (Mahlein, Steiner et al. 2010, Rumpf, Mahlein et al. 2010), Simple Ratio 800/670 Ratio Vegetation Index (RVI) (2%) (Feng, Shen et al. 2016), Structure Intensive Pigment Index (SIPI) (20%) (Mahlein, Steiner et al. 2010, Rumpf, Mahlein et al. 2010, Chew, Hashim et al. 2014, Feng, Shen et al. 2016, Lu, Ehsani et al. 2017, Anderegg, Hund et al. 2019, Khan, Liu et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022), Simple Ratio (SR) (11%) (Rumpf, Mahlein et al. 2010, Lu, Ehsani et al. 2017, Anderegg, Hund et al. 2019, Calamita, Imran et al. 2021, Nguyen, Sagan et al. 2021), Transformed Chlorophyll Absorption in Reflectance Index (TCARI) (9%) (Lu, Ehsani et al. 2017, Khan, Liu et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022), Transformed Vegetation Index (TVI) (13%) (Chew, Hashim et al. 2014, Lu, Ehsani et al. 2017, Atanassova, Nikolov et al. 2019, Khan, Liu et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022), Water Index (WI) (11%) (Feng, Shen et al. 2016, Lu, Ehsani et al. 2017, Nguyen, Sagan et al. 2021, Tian, Xue et al. 2021, Cen, Huang et al. 2022), Simple Ratio 750/710 Zarco-Tejada & Miller (ZM) (4%) (Mahlein, Rumpf et al. 2013, Cen, Huang et al. 2022).

These VIs were specifically relevant when the first broad-band sensors were developed and remain important today. Nonetheless, their application has been decreasing due to the appearance of more narrow-band sensors, capable of collecting high dimensional data (i.e., a high number of spectral features). Since VIs only consider

a limited number of wavelengths, they can lead to information losses when these last types of sensors are used. Moreover, since these VIs are developed for unveiling biophysical information in site-specific conditions, they may present limited transferability between locations.

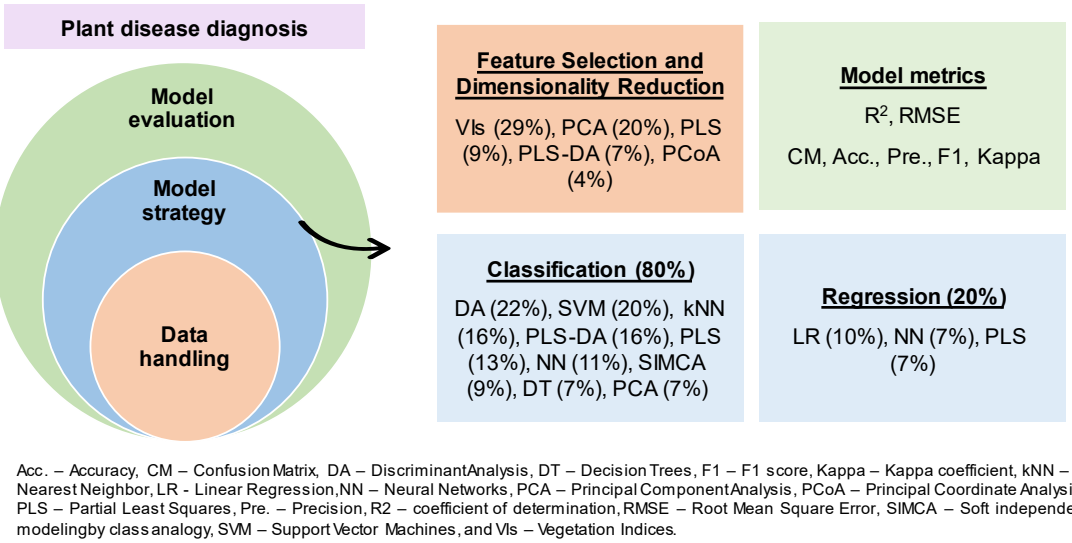
To surpass these difficulties, and with the increased usage of high dimensional sensors (and fewer applications of broad-band devices), several FeS and DR based on non-parametric ML algorithms were developed, and their computation increased to the detriment of the VIs (Figure 4).

Other identified approaches in Tables 2, 3, and 4 involved: spectral data condensation into 10 nm bands (Grisham, Johnson et al. 2010), statistical variance analysis of the wavelengths (Mandrile, Rotunno et al. 2019), Stepwise Discriminant Analysis (SDA) (Abdulridha, Ehsani et al. 2016), Orthogonal Subspace Projection (OSP) (Shuaibu, Lee et al. 2018), Outlier Removal Auxiliary Classifier Generative Adversarial Nets (OR-AC-GAN) (Wang, Vinson et al. 2019), Iterative Random Frog (IRF) (Zhu, Su et al. 2023), Competitive Adaptive Reweighted Sampling (CARS) (Zhu, Su et al. 2023), Successive Projections Algorithm (SPA) (Zhu, Su et al. 2023), and Stepwise Regression (Zhu, Su et al. 2023), wavelength coefficients (Tian, Xue et al. 2021), Sequential Forward Selection (SFS) (Cen, Huang et al. 2022), Simulated Annealing (SA) (Cen, Huang et al. 2022), Genetic Algorithms (GA) (Cen, Huang et al. 2022), RELIEF-F (Mahlein, Rumpf et al. 2013), Variable Importance (Anderegg, Hund et al. 2019), Sequential Forward Floating Selection Search Strategy and the Jeffries–Matusita (SFFS + JM) Distance (Reis-Pereira, Tosin et al. 2022), Stepwise Forward Variable Selection Method using Wilk's Lambda Criterion (SFVS) (Reis-Pereira, Tosin et al. 2022), and a Lasso Regularized Generalized Linear Model (LASSO) (Reis-Pereira, Tosin et al. 2022).

In turn, DR are methods used for transforming the original feature data space into a lower-dimensional representation. Some examples mentioned in the articles screened are included in Tables 2, 3, and 4 such as, the Principal Component Analysis (PCA) (Jinendra, Tamaki et al. 2010, Pereira, Milori et al. 2010, Ranulfi, Senesi et al. 2018, Yao, Lei et al. 2019, Marín-Ortiz, Gutierrez-Toro et al. 2020, Owomugisha, Nuwamanya et al. 2020, Barthel, Dordevic et al. 2021, Nguyen, Sagan et al. 2021, Cen, Huang et al. 2022) (the most applied approach), Partial Least Squares (PLS) (Susič, Žibrat et al. 2018, Sinha, Khot et al. 2019, Gold, Townsend et al. 2020, Ramamoorthy, Samiappan et al. 2022), PLS-Discriminant Analysis (PLS-DA) (Gold, Townsend et al. 2020, Gold, Townsend et al. 2020, Barthel, Dordevic et al. 2021), Principal Coordinate Analysis (PCoA) (Gold, Townsend et al. 2020, Gold, Townsend et al. 2020), Linear Discriminant Analysis (LDA) (Marín-Ortiz, Gutierrez-Toro et al. 2020, Reis Pereira, Santos et al. 2023), Maximum Likelihood (ML) (Ramamoorthy, Samiappan et al. 2022), Successive

Projections Algorithm (SPA) (Yao, Lei et al. 2019), Stepwise Multilinear Regression (SMLR) (Sinha, Khot et al. 2019), Stepwise Wavelengths Selection (STEPWISE) (Furlanetto, Nanni et al. 2021), Vertex Component Analysis (VCA) (Marín-Ortiz, Gutierrez-Toro et al. 2020).

All these FeS and DR approaches simplify the relationships between the spectra and the quality traits of interest and can thus improve data interpretability. Furthermore, they simplify data visualization, decrease the computational cost, help identify or improve useful spectral features, and enhance model performance.




Acc. – Accuracy, CM – Confusion Matrix, DA – Discriminant Analysis, DT – Decision Trees, F1 – F1 score, Kappa – Kappa coefficient, kNN – K Nearest Neighbor, LR - Linear Regression, NN – Neural Networks, PCA – Principal Component Analysis, PCoA – Principal Coordinate Analysis, PLS – Partial Least Squares, Pre. – Precision, R2 – coefficient of determination, RMSE – Root Mean Square Error, SIMCA – Soft independent modeling by class analogy, SVM – Support Vector Machines, and VIs – Vegetation Indices.

**Figure 4** Relationship between plant disease diagnosis analysis (orange), the modeling strategy followed (blue), and the evaluation approach computed for model performance assessing (green). Several Machine Learning (ML) strategies were identified in screening of scientific articles for feature selection and data dimensionality reduction. Furthermore, different chemometric and ML algorithms were also found in the screening process for both classification (qualitative) and regression (quantitative) analysis.

## 6.2. Applied predictive modeling: regression, classification, and authentication

### 6.2.1. Regression and classification approaches

After the preprocessing tasks, the spectral data is used for applied predictive modeling aiming mostly at disease diagnosis (qualification or quantification), being the most common approaches are regression, classification (Figure 4), and authentication.

The regression techniques (quantitative analysis) are used when the variable in analysis (i.e., target variable) is continuous and represented mainly through numeric values, a regression technique is used (quantitative analysis) (Tables 2, 3, 4). These are usually associated with quantification and disease severity studies (Mahlein, Steiner et al. 2010, Oerke, Fröhling et al. 2011, Zhao, He et al. 2012, Feng, Shen et al. 2016,

Anderegg, Hund et al. 2019, Yao, Lei et al. 2019, Khan, Liu et al. 2021, Atta, Saleem et al. 2023, Zhu, Su et al. 2023). On the contrary, when the target variable is based on categorical values (e.g., classes or categories), classification models are computed (qualitative approach) (Table 2, 3, 4). In our research, classification was the most applied strategy, found to be used in 80% of the articles, and regression was used in only 20% (Table 2, 3, 4).

Classification methodologies encompass various approaches, including single-class classification, binary classification (utilized in approximately 38% of the referenced articles), and multi-class classification (observed in around 42% of the studies). These classification methods are often employed through one-against-one or one-against-all analyses, as detailed in Tables 2, 3, and 4.

Multi-class models commonly involve categorizing spectral samples into distinct categories such as 'healthy,' 'asymptomatic' (or variations like 'non-symptomatic' or 'symptomless'), and 'symptomatic.' Additionally, these models may classify samples based on specific diseases, typically encompassing three to four diseases (Rumpf, Mahlein et al. 2010, Mahlein, Rumpf et al. 2013, Abdulridha, Ehsani et al. 2016, Susič, Žibrat et al. 2018, Reis Pereira, Santos et al. 2023), different cultivars (Grisham, Johnson et al. 2010, Afonso, Guerra et al. 2017, Ramamoorthy, Samiappan et al. 2022), or various stages of disease infection (Gold, Townsend et al. 2020, Gold, Townsend et al. 2020). The classification process can be further delineated into hard or soft categories. When each observation is assigned to a singular, discrete class, it is referred to as a ‘hard’ classification. Conversely, ‘soft’ classification also termed probabilistic or fuzzy classification, involves attributing each observation to one or more categories while accompanied by associated probabilities or confidence scores (Kuhn and Johnson 2013, Rashidi, Tran et al. 2019, Grandini, Bagli et al. 2020).

Authentication analysis can be performed when only the class of interest is known (e.g., healthy tissue profile) (Pereira, Milori et al. 2010, Atanassova, Nikolov et al. 2019). This classification approach verifies the legitimacy of a sample centered on its chemical and phenotypical characteristics, usually applying different chemometric models (based on multivariate data analysis). Authentication enables us to ascertain whether a sample is genuine or corresponds to a known reference, using solely the target category in the training stage (not needing the model’s training in non-conform samples), constituting a second-order advantage (Rodionova, Titova et al. 2016). The establishment of different categories in classification models can be a challenging task. Furthermore, producers may only need to confirm if a crop has a biotic disorder or not to support their decision-making, authentication may be an interesting, and simplified approach.

### 6.2.2. Predictive model techniques and algorithms

Predictive models, both classification and regression approaches, can be divided into parametric and non-parametric techniques. The first type captures spectral relationships based on a mathematical formula with a fixed set of parameters, which are usually sensitive to specific biophysical variables (circa 71% of the articles analyzed). In contrast, the second type does not rely on predefined parameters or assumptions about the underlying data. It generally comprises advanced techniques that search for patterns (relationships) directly on spectral data without assuming a specific functional form (e.g., the majority of ML techniques, mentioned in 51% of the articles) (Verrelst, Rivera et al. 2016).

Another type of categorization that can be applied to predictive models is the division between linear and non-linear approaches. The first type involves a statistical technique used to describe the linear relationship between a dependent variable and one or more independent variables (mentioned in 54% of the articles). A non-linear approach, in turn, does not assume a linear relationship between the dependent and independent variables and can capture complex connections and learn intricate patterns in data (Ouyang, Guo et al. 2019). It was found to be used in at least 63% of the scientific articles analyzed.

According to our research, several machine- and deep-learning models have been computed for assessing pests and diseases on several plants. Our results demonstrate that the most applied techniques were machine learning based, namely Support Vector Machines (SVMs) (used in 20% of the scientific articles, mostly in a non-linear and non-parametric form), Discriminant Analysis (several forms, including Linear, Stepwise, Quadratic, Flexible, and Fisher) (22%, the majority of them are in a linear, parametric form), k-Nearest Neighbor (KNN) (and variations) (16%, non-linear and non-parametric), Partial Least Squares (13%, linear and non-parametric), and the PLS-Discriminant Analysis (PLS-DA) version (16%, linear and parametric) (Table 2, 3, 4). Neural Networks (11%, non-linear and parametric), Soft independent modeling by class analogy (SIMCA, linear and parametric) (9%), Decision Trees (DT) (7%, non-linear and non-parametric), and Principal Component analysis (PCA) (7%, linear, non-parametric) were also found to be used (Table 2, 3). The details of these methods and the guidelines to efficiently use these for predicting data capture from optical and spectral sensors are beyond the scope of this article (further information can be found in different literature, such as (Liakos, Busato et al. 2018)).

### 6.2.3. Model development and evaluation

To perform applied predictive modeling, data is usually split into training, validation, and test sets. Modeling approaches are typically developed using a training dataset and evaluated through validation and testing on separate sets. This procedure prevents overfitting, which occurs when a model becomes excessively attuned to the intricacies of the training data. Overfitting can lead the model to incorporate not only the underlying structured patterns within the data but also the inherent noise and random fluctuations. Consequently, an overfitted model might struggle to generalize effectively to new, independent observations, diminishing its ability to offer consistent and reliable predictions (Kuhn and Johnson 2013). Furthermore, validation sets can be useful for promoting the model's tuning hyperparametrization, leading to more efficient solutions (Kuhn and Johnson 2013).

Cross-validation (CV) approaches were also performed to provide a reliable estimate of a model's generalization performance (Grisham, Johnson et al. 2010, Mahlein, Rumpf et al. 2013, Yeh, Chung et al. 2013, Abdulridha, Ehsani et al. 2016, Herrmann, Vosberg et al. 2018, Ranulfi, Senesi et al. 2018, Susič, Žibrat et al. 2018, Anderegg, Hund et al. 2019, Mandrile, Rotunno et al. 2019, Sinha, Khot et al. 2019, Gold, Townsend et al. 2020, Owomugisha, Nuwamanya et al. 2020, Furlanetto, Nanni et al. 2021, Khan, Liu et al. 2021, Cen, Huang et al. 2022, Reis-Pereira, Tosin et al. 2022) (Table 2, 3, 4). They can be in the form of k-fold CV, repeated k-fold CV (e.g., (Reis-Pereira, Tosin et al. 2022)), and leave one out (e.g., (Yeh, Chung et al. 2013, Khan, Liu et al. 2021)).

After the model development, its performance for plant disease diagnosis can be evaluated using distinct metrics. In our analysis, regression algorithms were mainly appraised according to their coefficient of regression (R2), and root mean square error (RMSE) (Mahlein, Steiner et al. 2010, Oerke, Fröhling et al. 2011, Zhao, He et al. 2012, Feng, Shen et al. 2016, Anderegg, Hund et al. 2019, Yao, Lei et al. 2019, Khan, Liu et al. 2021, Atta, Saleem et al. 2023, Zhu, Su et al. 2023) (Table 2, 3). Other metrics like Residual Predictive Deviation (RPD) (Yao, Lei et al. 2019) Standard Error of Calibration (SEC) (Atta, Saleem et al. 2023), and Standard Error of Prediction (SEP) (Mahlein, Steiner et al. 2010, Atta, Saleem et al. 2023) were also used. In turn, classification approaches usually are ranked through the determination of their confusion matrix (CM), accuracy, precision, recall, sensitivity, specificity, kappa, f1-measure, and Receiver Operating Characteristic (ROC) curve (Rumpf, Mahlein et al. 2010, Lu, Ehsani et al. 2017, Ranulfi, Senesi et al. 2018, Mandrile, Rotunno et al. 2019, Wang, Vinson et al. 2019, Gold, Townsend et al. 2020, Barthel, Dordevic et al. 2021, Cen, Huang et al. 2022, Reis-Pereira, Tosin et al. 2022). Error of Classification (EC) and Error of Omission (EO)

(Calamita, Imran et al. 2021). Overall Accuracy was the most applied classification evaluation metric, and its values ranged from 71% to 99.5% (Table 2, 3, 4). The Kappa coefficient has been frequently utilized (% of the classification models) to indicate models' overall accuracy, but profound criticisms exist regarding its appropriateness as a model performance metric (Foody 2020).

Spectral Mapping was also a strategy identified to be applied in evaluating the individual wavelength contribution and/ or in defining the spectral regions relevant to the prediction process (i.e., for the plant disease diagnosis). Ultimately, this process allows the evaluation of the biological significance of the selected spectral features. Several approaches can be used with this aim, namely several FeS and DR algorithms such as PCA (where is made the analysis of the principal components loadings) (Yao, Lei et al. 2019), PLS (latent vectors/variables) (Gold, Townsend et al. 2020), LDA (Linear Discriminants) (Reis Pereira, Santos et al. 2023), among others. Nevertheless, several articles did not present this biological analysis, being mainly data-driven approaches.

## 7. Main conclusions and perspectives

This review highlights the potential of several innovative proximal sensing techniques to diagnose different plant diseases early (prior to symptom appearance) in laboratory, greenhouse, and in-field experiments. It showed that most of the literature on the topic reports the first sensing experiments and modeling approaches in the theme. Therefore, most techniques present a relatively low technology readiness level (TRL) and are only specific to a location, and plant (i.e., species and cultivar)-pathogen interaction. The main outcomes also demonstrate the suitability of assessing different lesions promoted by pathogens in different scales, ranging from cellular to the macroscopic plant (canopy) levels. The most common crops used were tomato, wheat, sugar beet, and soybean. In turn, from the types of pathogens studied, fungi (53%) and bacteria were the most considered. The experimental conditions observed in the search results demonstrated that most of the examined scientific articles (69%) described assays conducted in laboratory conditions.

In terms of sensing techniques, the outcomes from forty-six scientific articles screened demonstrated the feasibility of various sensors for the diagnosis of plant disease, namely Fluorescence Spectroscopy, Hyperspectral Spectroscopy, Laser-Induced Breakdown Spectroscopy, Nuclear Magnetic Resonance Spectroscopy, Raman Spectroscopy, RGB Imaging, Thermography, and X-Ray Fluorescence. Of these, hyperspectral spectroscopy (82% of the articles) was the most applied technique and the only one used in field assays (in uncontrolled environmental conditions).

The data sensed underwent, in the majority of the studies, preprocessing and a processing step. The first identified and handled missing values and outliers, along with denoising and smoothing methods. The Savitzky-Golay filter was one of the most used techniques with this aim. Regarding the second step, two main Feature Engineering approaches were found, namely Feature Selection and Dimensionality Reduction strategies. Of these, Vegetation Indices (29%) and Principal Component Analysis (PCA, 20%) were the most computed, referred to in almost 50% of the analyzed articles. Nevertheless, the VIs application can conduct information losses when narrow-band sensors are used since they only consider a reduced number of wavelengths. Moreover, since VIs are developed for unraveling biophysical information in site-specific conditions, they may present limited transferability between locations.

Data is then modeled through regression, classification, or authentication approaches. Specifically, classification was used in 80% of the articles screened, mostly following a binary categorization to distinguish between healthy and diseased tissues). A multi-class approach was also frequently employed, identifying samples collected into healthy, non-symptomatic, and symptomatic disease tissues). Machine Learning algorithms featured prominently across the literature were extensively used in the articles, and Support Vector Machines and Discriminant Analysis were used in 53% of the cases. Classification accuracies were mostly superior to 71% and coefficients of regression were superior to 61%.

The usage of proximal sensors allied with different modeling strategies seems to be a future path to be considered since these techniques allow for an accurate early disease diagnosis. Advantages related to in-situ, in-vivo conditions highlight the importance of these sensing devices in proximal-range works.

An additional development of high-resolution, cost-effective, and portable spectral sensors is suggested for enhancing the evaluation of plant diseases. By providing powerful tools for early, in situ, in vivo diagnosis of infections, these innovative methods will constitute an opportunity to perform efficient, personalized disease control. The possibility of coupling these devices in different ground-based platforms (e.g. robots, cranes, among others) to perform in-field disease mapping is another advantage to be explored, following a precision agriculture perspective. Fieldwork is currently a challenge due to non-structured environmental light conditions. The usage of light sources and nocturnal fieldwork can be studied to surpass this hurdle. Further research should also evaluate the development of multisensory / data fusion solutions for plant disease assessment, combined with equipment's sensitivity and resolution enhancements.

The goodness and reliability of the information extracted from the analysis of the data captured motivate the development and establishment of protocols for

measurements, preprocessing, and processing of collected data, that must consider the variability of the environmental conditions that arise during measurements. The development of metadata and data ontology to support efficient data sharing between researchers must be considered and improved. The development of appropriate platforms and websites for information sharing is desirable in the near future. Moreover, improvements in data analysis algorithms and models for specific spectra-disturbance assessment will need to be continually evaluated, upgraded, or even redefined to improve disease investigation. Most of the developed predictive modeling strategies were data-driven and did not consider plants' physiology. Thus, the development of new strategies based on the plant's physiology or metabolomics should be made. In this way, the smart-photonic sensing strategies presented in this work could be linked to other omic sciences to make plant protection measures more efficient and sustainable.

**Acknowledgments**

Mafalda Reis-Pereira was supported by the Fundação para a Ciência e Tecnologia (FCT) fellowship with the reference SFRH/BD/146564/2019.

**Funding**

This work is partially financed by National Funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots, with reference PTDC/ASP-HOR/1338/2021.

## Supplementary Material

## I. Extended research methods section

### I.1. Data analysis: search and selection strategies, and data collection methods

First, a full investigation of publications in scientific databases was performed. Numerous articles were listed for review. Thus, inclusion and exclusion criteria were applied to complement this process and fulfill an unbiased analysis of the listed publications. The online tool Parsifal (Freitas and Segatto 2021) was used to hold the ongoing review procedure, allowing an arrangement of the entire research process: procedure design, screening and removal of duplicated publications, quality evaluation, and data extraction. The present work appraised the primary indexed publications related to the usage of optical and spectroscopic sensors in plant disease assessment between 1971 and August of 2023. The duplicated articles were removed and the remaining publications were evaluated and selected according to i) The application of the sensor in plant disease detection/diagnosis; ii) the language in which it was written (the article must be written in English, French, or Portuguese); iii) The article must be published after 1970; iv) and, it must be a work that mention an experiment assay, and reports how this was performed or a demonstration of a new algorithm or method for crop disease detection/diagnosis. Review articles were considered to perform an independent analysis. After this process, the chosen studies were fully read and analyzed. Each one of them was evaluated according to its quality to confirm if the work fulfills the aims of the current review. For quality assessment, the following questions were regarded: i) 'Does the paper refer to the system configuration?'; ii) 'Are the sensor parameters presented in the publication?', iii) 'Are the measurement parameters presented in the publication'?; iv) 'Is the analyzed scenario applied in real-world tests'?; v) 'Are the results of real-world tests explained in the publication'?; vi) 'Is the application presented in the publication feasible with the current resources?'. All questions were answered according to three possibilities: 'No' (0.0), 'Partially' (0.5), and 'Yes' (1.0). Only the articles that scored a value higher or equal to 3.5 were, then, considered for data extraction.

The extraction procedure retrieved information related to the: i) crop(s) studied in the assay; ii) pest(s) or disease(s) in analysis, iii) plant part evaluated, iv) sensor/technology used, v) type of sensor (imaging or non-imaging), vi) light source system configuration (passive or active), vii) sensor parameters (wavelengths studied), viii) environmental conditions (indoor, greenhouse, or infield), ix) modeling approach applied, x) model metric results, xi) possibility of current application (Technology Readiness Level – TRL).

The analysis was performed on six databases: i) ACM Digital Library (Machinery 2023), ii) El Compendex / Engineering Village (Elsevier 2023), iii) IEEE Digital Library (IEEE 2022), iv) ISI Web of Science (Clarivate), v) ScienceDirect (Elsevier 2023), and vi) Scopus (Elsevier). The search in the databases applied the base string: ("plant disease" OR "crop disease" OR "plant pest" OR "crop pest") AND (detect* OR diagnos* OR identif* OR quantif*) AND ("proximal" OR "ground" OR "remote") AND ("spectroscopy" OR "hyperspectral" OR "multispectral") AND NOT ("satellite" OR "phone" OR "uav" OR "balloon"). Only, for the ScienceDirect database were used a different approach since it only allows fewer boolean connectors (maximum eight per field): ("plant disease") AND (detection) AND ("proximal" OR "remote") AND ("spectroscopy" OR "hyperspectral") AND NOT ("satellite" OR "uav").

The term “proximal sensing” was chosen to detect all the articles that apply proximal sensing (i.e., ground-based) techniques in plant disease detection. Spectroscopic sensors were identified as one important population for this work, thus, the words AND (spectrosc* OR “spectral sensor”) AND (sens* OR assess* OR measur* OR captur*) AND ("Plant disease") AND (detect* OR assess* OR diagnos* OR "classification" OR "regression") were set to identify the works of proximal sensors for plant disease analysis. A particular interest in classification and regression modeling methods was determined by the application of these terms.

### I.2. Search results

Publication assessment on the various databases using the selected key strings results on 1179 documents, published between 1970 and 2023. The majority of the papers (62 %) were registered in the ScienceDirect database, probably due to the less restrictive key string used. The Scopus library registered the second higher ratio (21 %), which can be related to it being a more general database, responsible for reporting more publications from diverse publishers. The third greater percentage belongs to ISI Web of Science results (10 %), which is, similarly, a non-publisher-specific library, but has more specific indexation criteria. El Compendex, occupied the fourth place (4 %), despite also being non-publisher specific, most likely because it can present scope specification. The minor percentages corresponded to ACM Digital Library and to IEE Digital Library (2 and 1 %), which reported a lower number of findings. Publications’ time series in these databases showed an increasing interest tendency in plant disease diagnosis through the usage of proximal devices.

Based on these results, a PRISMA approach was followed to determine the inclusion or exclusion of publications for the present systematic review (Page, Moher et al. 2021). In brief, we identified all the publications retrieved, compared them, and

removed the ones that were duplicated. After, we screened all the documents, taking into consideration their title, abstracts, and figures to infer the interest of the publication. Manuscript withdrawal occurred according to the exclusion criteria earlier defined. Then, the remaining manuscripts were fully analyzed to assess their quality and evaluate if they were suitable for responding to the hypothesis we are considering. Of the 1179 manuscripts retrieved, only 322 publications were screened, allowing the retrieval of 46 publications (Figure S1).

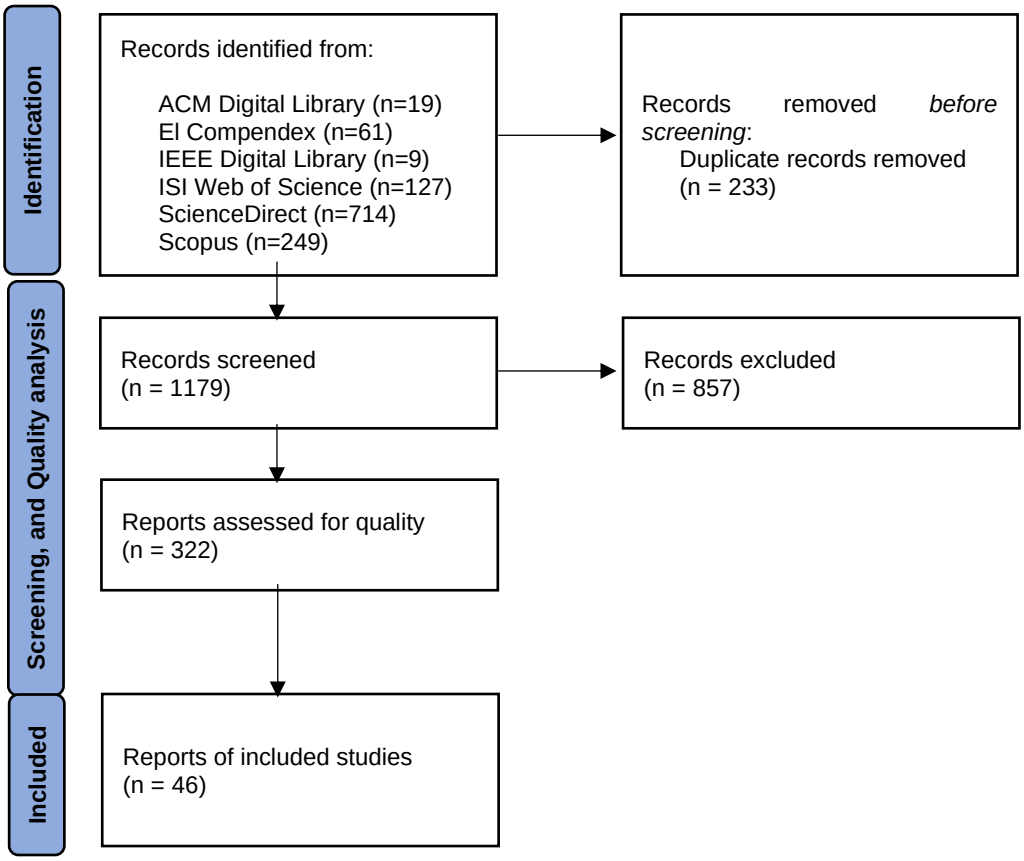



**Figure S1** Flow diagram of PRISMA technique for this systematic review aiming the analysis of the main proximal sensing technologies used for early plant disease diagnosis (i.e., before visual macroscopic symptom appearance).

## II. Extended description of sensing techniques in plant diseases

### II.1. Sensor's characteristics and usability

*RGB and Thermal Imaging*

RGB imaging (Table 1) is a technique that blends three primary colors – Red, Green, and Blue (RGB) – to capture and exhibit images. RGB digital cameras are cost-effective, adaptable, simple to manage, and suitable for both fixed and mobile setups. Their calibration is, furthermore, relatively easy to perform, allowing their usage in non-structured light conditions such as the ones occurring in field works. Nowadays, the

devices used for image collection and the programs applied to image analysis also acquire hundreds of images per hour. This data can be analyzed with great automation (Díaz-Lago, Stuthman et al. 2003) and stored to create a historical archive of crop status for a possible future application (Mirik, Michels Jr et al. 2006).

Thermography (or Thermal Imaging, TI) (Table 1) is another identified imaging technique, which involves the acquisition, processing, and interpretation of data acquired primarily in the thermal infrared (TIR, 3 to 14 μm) region of the electromagnetic spectrum (Ishimwe, Abutaleb et al. 2014). In general, thermal sensors can be thermographic or infrared cameras, capable of detecting emitted infrared radiation in the TIR spectrum region and converting the information captured into false-color images, where each image pixel contains the temperature value of the measured object (Li, Zhang et al. 2014). TI can be very useful when applied in the measurement (distribution) of the plant's thermal radiation, which is correlated with changes promoted by host-pathogen interactions. In brief, diseased plants presented modifications in the transpiration rate and in the water flow of plant organs or even the entire plant, as well as local temperature changes due to pathogen infection or defense mechanisms (Oerke, Steiner et al. 2006, Oerke, Fröhling et al. 2011). All these modifications lead to deviations in the plant's spectral behavior in the TIR, which these sensors can detect.

*UV-VIS and NIR Spectroscopy*

Multispectral and hyperspectral Spectroscopy are two UV-VIS-NIR Spectroscopy approaches that offer a rapid, typically non-invasive, and highly specific method for crop disease diagnosis (Table 1). These devices can sample and record radiation, in one or more regions of the electromagnetic spectrum, that is reflected, emitted, or transmitted from one surface. They have the advantage of being suitable for capturing both visible and non-visible wavelengths. These sensors are generally categorized based on their spectral resolution (i.e., the number and width of measured wavebands), spatial scale, and the type of detector (i.e., SSPO or MSPO). Multispectral sensors are capable of capturing data from several discrete spectral bands. The main sensor manufacturers produce devices that acquire between 3 and 25 bands, which may not be continuous, including VIS, NIR, or a set of custom bands (Perez-Sanz, Navarro et al. 2017). Usually, they are smaller in volume, lighter in weight, need fewer internal components for working, and are less expensive than hyperspectral instruments, that work with wider wavelength ranges (Cotrozzi 2022). In turn, hyperspectral solutions provide high spectral resolution data, assessing thousands of contiguous narrow spectra bands (hundreds or thousands covering 5−20 nm each) from one sample, being more sensitive to subtle variations in measured radiation and providing more data complexity and information content. They

may cover the main spectral regions of the electromagnetic spectrum, including VIS, NIR, SWIR, MWIR (mid-wave infrared), and LWIR (longwave infrared). The data obtained by hyperspectral sensors provide a spectral signature of a unique sample that can be characterized and identify any given material (Ben-Dor, Schläpfer et al. 2013). Nonetheless, it is important to recognize that spectral information captured with this equipment may present high dimensionality, since contiguous bands may present redundant information (Caicedo, Verrelst et al. 2014, Rivera, Verrelst et al. 2014). Thus, further data processing methods may be required to improve the quality of data extracted.

*Fluorescence Spectroscopy*

Fluorescence Spectroscopy (FS, also known as fluorometry or spectrofluorometry) (Table 1), is also a sensitive, non-invasive, and non-destructive type of UV-VIS-NIR technique. It assesses fluorescence from a sample after excitation with a beam of light (usually UV spectra, wavelength ranging from 10 to 400 nm). FS is interesting in plant studies since these organisms possess different pigments and structural pigment components which make them capable of emitting two different types of fluorescence, namely blue-green fluorescence (400–600 nm), and chlorophyll fluorescence (650–800 nm) (Belasque, Gasparoto et al. 2008). Chlorophyll fluorescence occurs when light is re-emitted by the chlorophyll molecules when it returns from the excited to the non-excited stage. Blue and red fluorescence contain complementary information on plant phenotyping traits and should be considered simultaneously.

*Emission-based techniques*

Similarly to Thermography and Fluorescence Spectroscopy, three other technologies also measure the amount of radiation emitted by a sample. Laser-induced breakdown spectroscopy (LIBS) is one of them, known for being a laser-based solution found in the screened scientific articles (Table 1). It is an atomic emission spectroscopy technique for simple, fast, and in situ analysis that uses highly energetic laser pulses to provoke optical sample excitation. This method enables the acquisition of both qualitative and quantitative information regarding sample composition within a singular spectrum, employing a quasi-non-destructive approach. LIBS generates an emission spectrum produced by focusing a high-energy laser pulse on a sample, which generates plasma by rapidly heating and vaporizing the material. The excited atoms in the plasma emit light as they return to their lower energy states, and the resulting spectrum is analyzed to identify and quantify the elemental composition of the sample (Ferreira, Anzano et al. 2009, Nicolodelli, Cabral et al. 2019). LIBS system configuration allows, nevertheless,

the creation of portable equipment suitable for in situ measurements with increased stability and sensitivity of the measurements, characteristics that make it attractive for analyses in the field (Wainner, Harmon et al. 2001, Fortes and Laserna 2010). LIBS is suitable for tracking changes in the standard composition of the major macro-and micronutrients in plants, allowing the differentiation between healthy and diseased individuals, even at non-symptomatic stages (Pereira, Milori et al. 2010, Ranulfi, Senesi et al. 2018).

X-ray Fluorescence (XRF) is another spectroscopic emission approach applied in plant disease studies. Similar to LIBS, XRF is mainly used to determine the elemental composition of different tissues. It is a well-established, non-destructive analytical method for qualitative and quantitative multielement evaluation. It involves minimal or no sample preparation, allowing in vivo studies. XRF work principle starts with a sample irradiation with X-rays, causing its inner-shell electrons to be ejected. As electrons from outer shells fall to fill the vacancies, characteristic X-ray photons are emitted (Rodrigues, Gomes et al. 2018). Thus, as the emitted radiation is typical for each chemical compound it can be applied as a fingerprint that makes elemental composition possible (Van Grieken and Markowicz 2001). XRF peak area can additionally provide quantitative information since the number of emitted photons is directly proportional to the number of emitting atoms (Rodrigues, Gomes et al. 2018).

*Vibrational Spectroscopy*

Raman Spectroscopy (RaS), like NIR Spectroscopy, (Table 1) is a vibrational-based technique that has been previously used in plant disease studies due to its non-invasive, non-destructive analytical character. RaS is known for its capacity to provide information about the sample's molecular vibrations and structure without requiring prior sample preparation. This technique uses a laser at a well-defined wavelength in the VIS or NIR range (frequently 532, 785, or 1064 nm) to excite a sample, producing inelastic scattering of light through interaction with the molecular vibrational modes of the sample (Sylvain and Cecile 2018). A portion of this scattered light has a wavelength distinct from that of the exciting laser since an energy exchange happens between the incident photon and the molecule. RaS is then applied to assess these modifications occurring through energy exchange between the incoming photons and the molecule that scatter light. RaS has, hence, unique advantages in practical applications since it provides crucial narrow and sharp characteristic peaks attributed to specific or several substances, specifically the fingerprint characteristics. Furthermore, RaS has the desired portability, low labor, and cost requirements (Weng, Hu et al. 2021).

Nuclear Magnetic Resonance (NMR) Spectroscopy (Table 1) was also a vibrational approach identified during the article screening process as useful for plant studies. It is based on measuring the resonances of magnetic nuclei (such as 1H, 13C, and 15N) that interact with an external magnetic field. NMR allows non-invasive structural assessment of metabolites making in vivo analysis possible. This approach generates unique spectra for each single compound and can be applied in identifying metabolites of biological origin of which no previous knowledge is needed (Fan and Lane 2008). NMR spectroscopy can also give quantitative information about a sample, as the signal intensity is directly proportional to the molar concentration (Pauli, Jaki et al. 2005). It also has the advantage of being highly reproducible and high throughput in sample analysis.

*Emerging techniques*

Two additional techniques were identified during the screening process as emerging due to their increasing interest, namely Biophoton Emission and Volatile Organic Compounds (VOCs) analysis (Table 1).

Biophoton spectroscopy (BS) (Table 1) is an electromagnetic emission-based technique, that assesses ultraweak photon emissions produced by living organisms, resulting from chemically excited molecules produced in cellular biochemical reactions, in various metabolic processes without photoexcitation. The photon emission is related to the interaction between constituents of living materials (such as lipids, protein, and DNA) with reactive oxygen species (ROS) and/or free radicals (Cifra, Pospíšil et al. 2014). The intensity of the biophoton emission is usually 3–6 orders lower than the light intensity that is visible to the naked human eye, but the wavelengths of emissions normally extend over the VIs region. To detect biophoton emissions is, hence, necessary to use a highly precise device with enough sensitivity to detect low levels of light, such as the state of a single photoelectron event (Creath, And et al. 2005, Nukui, Inagaki et al. 2013, Kobayashi and Biology 2014).

With a different type of operating fundamental system, VOCs assessment is also an interesting technique presenting suitability for non-destructive plant disease diagnosis. VOCs are biomolecules and metabolites emitted by plants into their immediate surroundings, presenting essential functions in growth, communication, defense, and survival processes. They usually are in the gaseous phase under standard temperature and pressure, generally at ultra-low concentrations, below the human olfactory threshold (Martinelli, Scalenghe et al. 2015, Buja, Sabella et al. 2021). VOC profile is influenced by the plant's growth and stage of development, as well as by abiotic and biotic (i.e., pathogens, insects, animals, and herbs). The type of VOC synthesized

and released by plants when attacked by harmful phytopathogens can be specific, allowing the determination of so-called VOC signatures (Lopez-Gresa, Maltese et al. 2010, Schlaeppi, Abou-Mansour et al. 2010, Agarrwal, Bentur et al. 2014, Ninkovic, Rensing et al. 2019).

# Chapter III |
# Case Studies

# Case Study 2

**Reis-Pereira, M.**; Tosin, R.; Verrelst, J.; Caicedo, J.; Tavares, F.; Santos, F. N. d.; Cunha, M. Plant disease diagnosis based on hyperspectral sensing: comparative analysis of parametric spectral indices and machine learning Gaussian process classification approaches.

Paper submitted

Classification according to journal: Original Research Article

# Plant disease diagnosis based on hyperspectral sensing: comparative analysis of parametric spectral vegetation indices and nonparametric Gaussian process classification approaches

Mafalda Reis-Pereira[1,2], Jochem Verrelst[3], Renan Tosin[1,2], Juan Pablo Rivera Caicedo[4], Fernando Tavares[5,6], Filipe Neves dos Santos[2], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal

[3] Image Processing Laboratory (IPL), University of Valencia, C/Catedrático José Beltrán 2, 46980 Paterna, Valencia, Spain

[4] CONACYT-UAN, Secretary of Research and Postgraduate, Autonomous University of Nayarit, Tepic C.P. 63155, Nayarit, Mexico

[5] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[6] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

* Corresponding author: E-mail address: mccunha@fc.up.pt (Mário Cunha)

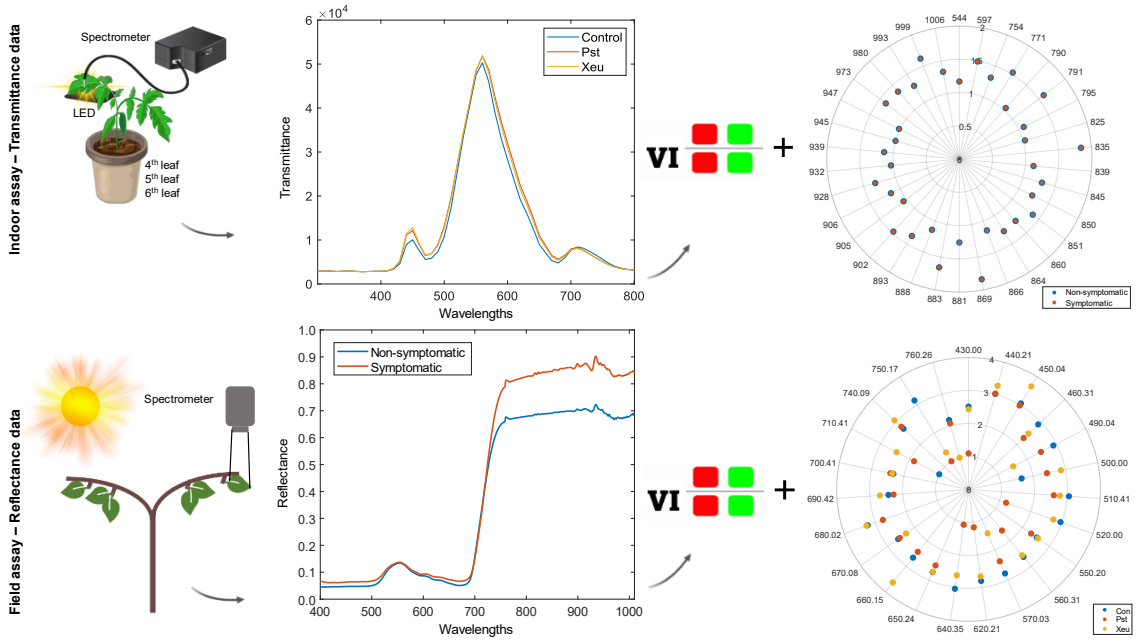
**Highlights**

- Hyperspectral biophysical health status classification accuracy is enhanced by wise feature selection.
- An automated Gaussian Process Classification Band Analysis Tool (GPC-BAT) is presented.
- GPC-BAT sequentially eliminated the minimum relevant spectral wavelength.
- Two models discriminated spectra collected in healthy and diseased tomato tissues.
- Two models identified spectra collected in symptomless and symptomatic kiwi leaves in field.

**Abstract**

Early and accurate disease diagnosis is pivotal for effective phytosanitary management strategies in agriculture. Hyperspectral sensing has emerged as a promising tool for early disease detection, yet challenges remain in effectively harnessing its potential. This study compares parametric spectral Vegetation Indices (VIs) and a non-parametric Gaussian Process Classification based on an Automated Spectral Band Analysis Tool (GPC-BAT) for diagnosing plant bacterial diseases using hyperspectral data. The study conducted experiments on tomato plants in controlled conditions and kiwi plants in field settings to assess the performance of VIs and GPC-BAT. VIs, known for their simplicity in extracting biophysical information, successfully distinguished healthy and diseased tissues in both plant species. The overall accuracy achieved was 63% and 71% for tomato and kiwi, respectively. However, limitations were observed, particularly in differentiating specific disease infections accurately. On the other hand, GPC-BAT, after feature reduction, showcased enhanced accuracy in identifying healthy and diseased tissues. The overall accuracy ranged from 70% to 75% in the tomato and kiwi case studies. Despite its effectiveness, the model faced challenges in accurately predicting certain disease infections, especially in the early stages. Comparative analysis revealed commonalities and differences in the spectral bands identified by both approaches, with overlaps in critical regions across plant species. Notably, these spectral regions corresponded to the absorption regions of various photosynthetic pigments and structural components affected by bacterial infections in plant leaves. The study underscores the potential of hyperspectral sensing in disease diagnosis and highlights the strengths and limitations of VIs and GPC-BAT. The identified spectral features hold biological significance, suggesting correlations between bacterial infections and alterations in plant pigments and structural components. Future research avenues could focus on refining these approaches for improved accuracy in diagnosing diverse plant-pathogen interactions, thereby aiding early disease detection and management in agricultural settings.

## Graphical Abstract




## Keywords

Plant disease; Tomato; Kiwi; Hyperspectral Spectroscopy; Gaussian process classification, Wavelength selection; ARTMO

## 1. Introduction

Plant diseases are a major threat to worldwide agriculture, causing substantial yield losses and impacting food security and quality (Ristaino, Anderson et al. 2021). Timely and accurate disease diagnosis is crucial for implementing effective management strategies in sustainable agriculture. These practices aim to contribute to more effective and precise plant protection measures due to more customized phytosanitary treatments regarding time, location, product used, and dose. However, traditional diagnostic methods often fail to detect diseases before visible symptoms emerge, limiting their effectiveness in proactive disease management (Martinelli, Scalenghe et al. 2015, Dyussembayev, Sambasivam et al. 2021). Innovative plant disease monitoring and diagnosis methods involving different state-of-the-art sensing approaches have recently been explored for precise and early in-vivo and in-situ disease assessment. Recent strides in innovative sensing techniques, particularly hyperspectral spectroscopy (HS), offer promising avenues for precise and early disease diagnosis (Martinelli, Scalenghe et al. 2015, Zhang, Huang et al. 2019). However, despite the potential of HS in plant disease diagnosis, challenges persist in harnessing its full potential due to the complexity of hyperspectral data and the need for efficient processing methodologies to extract relevant information (Thomas, Kuska et al. 2018, Galieni, D'Ascenzo et al. 2021).

Addressing these challenges is crucial to unlocking the full potential of HS in improved disease diagnosis and management strategies.

HS is known for acquiring data in narrow wavebands (<10 nm), with high precision and resolution and being able to capture detailed information from the electromagnetic spectrum (Galieni, D'Ascenzo et al. 2021). Nevertheless, despite this evident benefit, the measurement of this large number of variables (i.e., features, wavelengths) results in the data's high dimensionality, which increases the complexity of its processing to produce relevant information. Furthermore, the spectral data assessed in near-contiguous variables likely present similar or overlapping information. This potential data redundancy also increases the complexity of its analysis interpretation and the chance of overfitting occurrence (Szymańska 2018). Dimensionality reduction methods were developed to mitigate the effects of high dimensionality and collinearity, mostly based on identifying and extracting the most relevant and distinctive spectral features (without losing relevant information) (Thomas, Kuska et al. 2018).

The computation of spectral Vegetation Indices (VIs) is one of the most widespread Feature Selection (FS) approaches for retrieving crop biophysical information, especially due to their intrinsic simplicity. It consists of a user-defined mathematical combination of two or more wavelengths that improves crop biophysical information extraction from data, i.e., identifying spectral relationships that unravel specific plant properties. Hence, VIs are considered as parametric, physiological-driven methods. Nonetheless, it is important to note that when narrowband hyperspectral data is used, VIs can be a restrictive formulation since they only use some of the available wavelengths, failing to leverage the complete wealth of information in the continuous spectral data (Verrelst, Malenovský et al. 2019). Besides that, some of the VIs that have already been developed were designed to estimate specific vegetation traits (e.g., plant biomass and photosynthetic pigments research), which might not entirely suit the assessment of plant disease. The ones developed for studying specific plant-pathogen interactions (e.g., (Mahlein, Rumpf et al. 2013)) are usually only applicable in analyzing that specific pathosystem (usually in similar environmental conditions), mostly in symptomatic conditions, and are unsuitable for generalized disease assessment. Disease studies are usually modelled as a classification approach, which adds difficulty to the application of the index.

Another emerging strategy, recently employed for exploring hyperspectral data, is applying different advanced techniques (e.g., machine learning algorithms) that search for relationships between spectral data and biophysical variables (also known as non-

parametric, data-driven methods). They mostly consider all the spectral features measured by the hyperspectral sensors, which constitutes an important benefit compared to the VIs (Verrelst, Rivera et al. 2015). These methods can be based on linear or non-linear predictive methods.

Furthermore, automated band analysis tools have been developed in the domain of machine learning classification algorithms (MLCAs). Following a band selection method earlier introduced in regression (Verrelst, Rivera et al. 2016), this paper introduces an automated spectral band analysis tool (BAT) based on Gaussian process classification (GPC) for the spectral analysis of bacterial plant diseases. Briefly, starting from using all bands, GPC-BAT sequentially removes the least informative band in GPC until one band is left. By tracking the accuracy of statistics, GPC-BAT allows (1) to identify the most informative bands relating spectral data to a classification problem, and (2) to find the least number of bands that preserve optimized accurate classification tasks.

Hence, despite the development and availability of diverse methods for extracting meaningful spectral information in the context of plant bacterial disease diagnosis, it is necessary to address their suitability and performance when leading with different pathosystems. Therefore, the objectives of the present work aimed to: i) explore the suitability of different VIs for extracting relevant spectral features for performing plant bacterial disease diagnosis, using both reflectance and transmittance hyperspectral data (physiological driven approach); ii) investigate the potential of a GPC-BAT for performing plant bacterial disease diagnosis using reflectance and transmission hyperspectral data (data-driven approach); iii) compare and contrast the performance of VIs and GPC-BAT in discerning spectral features crucial for differentiating between healthy and diseased plant tissues; iv) and uncover the biological significance of the identified spectral features concerning specific plant-pathogen interactions and their implications for early disease diagnosis. Two case studies were conducted on tomato (controlled conditions) and kiwi (field conditions) plants, aiming to explore the capabilities of the developed approaches for performing bacterial disease diagnosis in distinct species in different environmental conditions.

## 2. Materials and Methods

The present analysis focuses on two case studies: one in controlled environmental conditions using hyperspectral transmittance sensing data and the second in field conditions using hyperspectral reflectance sensing data. The first case consisted of collecting spectral data in healthy tomato leaflets' tissues, along with measurements in inoculated (diseased) tissues with *Pseudomonas syringae* pv. *tomato*

(responsible for the bacterial speck disease of tomato), and tissues inoculated with *Xanthomonas euvesicatoria* (responsible for the bacterial spot disease of tomato). The second case assessed spectral data in non-symptomatic and symptomatic kiwi leaf tissues affected by bacterial canker of kiwi caused by *Pseudomonas syringae* pv. *actinidiae*. In both case studies, multiple spectral samples were gathered within an experimental setup at various time intervals encompassing all the plants involved in the study.

The hyperspectral data was then used in two modelling approaches involving a physiologically driven parametric approach based on VIs and a non-parametric approach based on a Gaussian Process Classification Banda Analysis Tool.

### 2.1. Case studies – experimental design for kiwi and tomato

#### 2.1.1. Tomato bacterial diseases – Indoor assay

An indoor assay was performed in a walk-in growth chamber (temperature of 25 to 27ºC, humidity of 60% approximately, photoperiod of 12/12h, and light intensity of 30W) with nine tomato (*Solanum lycopersicum* L.) plants of the variety Cherry in 200 mL pots with a commercial potting substrate. Groups of three plants were formed and physically separated from each other to avoid cross-contamination; and one group was sprayed with distilled water (Control, healthy class), the second group with a bacterial suspension ($1\text{x}10^8$ cells/mL) of *Pseudomonas syringae* pv. *tomato* DC 3000 (Pst), and the last group with a suspension ($1\text{x}10^8$ cells/mL) of *Xanthomonas euvesicatoria* (Xeu), following a previously developed protocol (Reis Pereira, Santos et al. 2023). Plant phenotypical observations were performed daily to assess symptom development for 10 days (Table 1).

The success of artificial bacterial inoculation was assessed by the performance of a viability assay and through a colony polymerase chain reaction (PCR), as stated in (Reis Pereira, Santos et al. 2023). The growth of Pst and Xeu in their appropriate selective (KB and YDC, respectively) media demonstrated that bacteria were viable at the moment of inoculation. PCR results proved the infection success, where the formation of bacteria-specific bands for each pathogen species, namely a 200-base pair (bp) fragment for Pst, and a 713 bp fragment for Xeu were observed. No PCR amplification was observed from samples collected from Control samples, assuring they were healthy until after the last spectral measurement. The first macroscopic lesions were detected in Pst inoculated samples 3 days after inoculation (DAI) and in Xeu inoculated samples at 8 DAI.

Hyperspectral transmittance point-of-measurement (POM) data was captured inside a walk-in chamber using a setup comprised of a mini spectrometer (Hamamatsu Photonics K.K. TM Series C11697MB) with a wavelength range of 200-1100 nm, and a spectral resolution of 0.6 nm. This setup includes a transmission optical fiber bundle (FCR-7UVIR200-2-45-BX, Avantes, Eerbeek, The Netherlands), a laptop for data storage and processing, and a white LED spanning from 390 to 800 nm. A specialized evaluation software (SpecEvaluationUSB2.exe, Hamamatsu Photonics K.K., Japan) was used for data acquisition. Further details about the setup can be found in previous work (Reis Pereira, Santos et al. 2023). Subsequently, a resampling technique of approximately 10 nm was employed to minimize data redundancy. A dataset comprising 2,346 samples (spectral observations) encompassing 51 wavelength features (spectral variables) was selected for subsequent analysis. The spectral measurements were later classified according to the leaflets' plant treatment group, including the classes: i) Control (healthy); ii) inoculated with Pst; iii) and inoculated with Xeu (Table 1). This dataset can be find in (Reis Pereira, Tavares et al.).

**Table 1** Spectral data characterization of the measurements randomly performed on tomato leaflets (walk-in chamber - controlled conditions, transmittance) and kiwi leaves (in-field conditions, reflectance), showing the number of assessment dates, plants, observations (classified according to visual phenotyping observations).

| | *Nº Dates* | *Nº Plants** | *Nº NS* | *Nº S* | *Total* |
|---|---|---|---|---|---|
| *Walk-in assay* | | | | | |
| *Tomato* | 8 | 9 | 1365 | 981 | 2346 |
| Con | | 3 | 809 | --- | 809 |
| Pst | | 3 | 93 | 634 | 727 |
| Xeu | | 3 | 463 | 347 | 810 |
| *In-field assay (2 sites)* | | | | | |
| *Kiwi* | 9 | 20 | 281 | 223 | 504 |

Control (healthy), Pst – Inoculated with *Pseudomonas syringae* pv. *tomato*, Xeu – Inoculated with *Xanthomonas euvesicatoria*, NS – Non-Symptomatic, S – Symptomatic. *Several measurements were taken over time on different leaflets on each plant

### 2.1.2. Kiwi bacterial diseased – Field assay

An assay was performed in field conditions in commercial orchards of kiwi (Actinidia deliciosa) of the cultivar 'Bo.Erika', located in Guimarães, Portugal. Macroscopic signs (i.e., symptoms visual to the human eye) of bacterial canker caused by *Pseudomonas syringae* pv. *actinidiae* (Psa) were assessed in feminine plants. Plant

visual phenotyping was performed, classifying leaves into non-symptomatic (NS, when no macroscopic visual symptoms were present) or symptomatic (S, when macroscopic symptoms were visible) as described in (Reis-Pereira, Tosin et al. 2022) (Table 1).

Hyperspectral reflectance measurements were collected in situ, in vivo leaves, using a portable spectroradiometer (ASD FieldSpec® HandHeld 2, ASD Instruments, Boulder, CO, USA) with a wavelength range of 325 to 1075 nm, spectral resolution of 1 nm, and field-of-view conical angle of 25º. The detailed procedures followed during the spectral acquisition assay can be found in previous research (Reis-Pereira, Tosin et al. 2022). In brief, three leaves were chosen per plant, and their spectral signatures were collected at different time points, resulting in 504 samples (spectral observations) and 751 spectral features (spectral variables). Binary classification of leaves' spectra was made according to the phenotype of the leaves resulting in the binary classes NS and S (Table 1). This dataset can be find in (Reis Pereira, Tavares et al.).

### 2.2. Modeling approaches

#### 2.2.1. Parametric approach – Vegetation Indices (VIs)

Hyperspectral data, including both transmittance and reflectance spectra, usually have an overlapping nature and multi-scale interference (Tosin, Martins et al. 2022).To address this issue, a selection of 33 spectral VIs, encompassing 42 distinct wavelength combinations, was computed to identify the most relevant wavelengths or bands for discriminating healthy and disease-biological tissues (Table 2). This selection process aimed to integrate VI formulations commonly used to assess different crop traits as well as crops' physiological conditions. The variables used in each formula corresponded to default values explicitly mentioned in the formula (Table 2) or values chosen by the authors, namely: 450 nm (representing the Blue region of the electromagnetic spectrum), 550 nm (Green), 680 nm (Red), 700 nm (Red Edge), and 800 nm (NIR). The feature representing the Blue was elected due to being related to pigment absorption features (~450 nm, e.g. chlorophylls and carotenoids) (Verdebout, Jacquemoud et al. 1994, Asner 1998) and a Blue fluorescence maximum (Lang, Stober et al. 1991). The 550 nm wavelength was selected because reflectance data corresponds to the green peak (or green edge), where reflectance values can be more than twice the surrounding wavelengths (Hennessy, Clarke et al. 2020, Moriya, Imai et al. 2023). This value is also sensitive to chlorophyll content and has been explored to detect plant stress-induced changes and pigment content variations (Blackmer, Schepers et al. 1994, Gitelson and Merzlyak 1994). Instead, 680 nm was chosen because it corresponds to the reflectance minimum in the Red region (Hennessy, Clarke et al. 2020, Moriya, Imai et al. 2023). The

Red-Edge value (700 nm) was used because it is highly sensitive to changes in chlorophyll-a absorption and is used to detect subtle changes related to plant physiological status and growth stage transitions (Gitelson and Merzlyak 1994, Lichtenthaler, Gitelson et al. 1996). The 800 nm spectral feature was chosen because it is related to the influence of changes in leaf structure and density, but it is not sensitive to pigment level changes (Haboudane, Miller et al. 2004). Furthermore, all these wavelengths have been extensively used in formulating multiple VIs, as seen in Index Data Base (IDB), a database for remote sensing indices (Henrich, Götze et al. 2009, Henrich V., Krauss G. et al. 2023).

A Flexible Discriminant Analysis (FDA), applied predictive modeling with a built-in Feature Selection (FS), was then performed to evaluate the most significant VIs used to discriminate between spectral data measured in i) healthy tomato tissues (Control, Con), diseased tomato tissues inoculated with Pst, and diseased tomato tissues infected with Xeu; ii) non-symptomatic (i.e., without macroscopic lesions, NS) and symptomatic (S) kiwi tissues. The datasets, encompassing both tomato and kiwi cases, were split according to the holdout method (Lantz 2019), which involved partitioning into a training set comprising 70% of the data and a testing set with the remaining 30% of the observations (Kuhn and Johnson 2013).

Model evaluation was employed through a resampling strategy involving repeated 10-fold cross-validation to estimate accuracy. A more detailed explanation can be found in (Kuhn and Johnson 2013, Lantz 2019, Reis-Pereira, Tosin et al. 2023). Model performance was then evaluated by assessing different classification model metrics, including the confusion Matrix (CM), accuracy score, kappa coefficient, and F1-Score (Reis Pereira, Santos et al. 2023).

The CM is a 2D-matrix representation of the actual classes of the collected spectral samples in one dimension and the predicted class values in the other. When the predicted class values are equal to the actual value, they are considered correct classifications and localized in the CM's diagonal. The remaining matrix cells correspond to incorrect classification predictions, where the predicted value is not coincident with the actual value. The class of interest is considered positive, while the other(s) are considered negative. When the predicted class is correctly classified as the class of interest, it is considered a true positive (TP) case. When the predicted class is accurately classified as not belonging to the class of interest, it is called a true negative (TN). When the predicted class is wrongly classified as the class of interest, it is called a false positive

(FP), and when incorrectly classified as not fitting the class of interest, it is classified as a false negative (FN).

The accuracy score (also known as Success Rate) corresponds to the number of rightfully classified prediction cases divided by the total number of predictions (Eq. 1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Kappa coefficient (also called Cohen's kappa) amends the accuracy score by considering the probability of an accurate prediction occurring by chance alone (Lantz 2019) (Eq. 2). Its value can range from zero, indicating an imperfect agreement, to one, the perfect agreement between models' predictions and true values. Kappa values (in percentage) can be interpreted as follows: when less than 20%, it is considered a poor agreement; 20% to 40% a fair agreement; 40% to 60% a moderate agreement; 60% to 80% a good agreement; 80% to 100% a very good agreement (Lantz 2019). Kappa coefficient can be estimated through the following formula where Pr(a) represents the proportion of actual agreement and Pr(e) refers to the expected agreement between the classifier and the true values, under the hypothesis that they were chosen randomly (Eq. 2).

$$Kappa\ coefficient = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (2)$$

The F1-score (also called F-measured) combines the proportion of positive cases that are truly positive (Precision) with the number of TP over the total number of positives (Recall, which measures how complete the results are) into a single number using the harmonic mean (Eq. 3).

$$F - score = \frac{2 \times precision \times recall}{recall + precision} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3)$$

Sensitivity was evaluated, indicating the models' ability to predict the TP of each available class (Eq. 4).

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

The Specificity metric was also calculated since it indicates the models' suitability for predicting TN of each available class (Eq. 5).

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

All these computation analyses were made in the software R (Team 2021) using the packages 'caret' (Kuhn 2015), and 'earth' (Milborrow 2019).

**Table 2** List of the Spectral Vegetation Indices (VIs) computed in this work, mentioning its formula and reference (when available).

| ***Vegetation Indices*** | ***Formula*** | ***Ref.*** |
|---|---|---|
| Ashburn Vegetation Index (AVI) | $2.0 \times \mathrm{NIR} - \mathrm{RED}$ | (Ashburn 1979, Bannari, Morin et al. 1995) |
| Anthocyanin reflectance index (ARI) | $\frac{1}{GREEN} - \frac{1}{RED}$ | (Gitelson, Merzlyak et al. 2001) |
| Blue Green Pigment Index (BGI) | $\frac{BLUE}{GREEN}$ | - |
| Browning Reflectance Index (BRI) | $\frac{\frac{1}{GREEN} - \frac{1}{RED}}{NIR}$ | (Merzlyak, Gitelson et al. 2003) |
| Blue/Red Pigment Index (BRI2) | $\frac{450\ nm}{690\ nm}$ | (Zarco-Tejada, Berjón et al. 2005) |
| Canopy Chlorophyll Content Index (CCI) | $\frac{\frac{NIR - RED\ EDGE}{NIR + RED\ EDGE}}{\frac{NIR - RED}{NIR + RED}}$ | (Barnes, Clarke et al. 2000, Clarke, Moran et al. 2001, Pinter, Hatfield et al. 2003, El-Shikha, Barnes et al. 2008, Herrmann, Karnieli et al. 2010, Henrich, Krauss et al. 2011) |
| Chlorophyll Green (Chlgreen) | $\left(\frac{NIR}{GREEN}\right)^{(-1)}$ | (Gitelson, Keydan et al. 2006) |
| Coloration Index (CI) | $\frac{RED - BLUE}{RED}$ | (Escadafal, Belghith et al. 1994) |
| Chlorophyll Index Green (CIgreen) | $\frac{NIR}{GREEN} - 1$ | (Gitelson, Viña et al. 2003, Ahamed, Tian et al. 2011, Hunt Jr, Daughtry et al. 2011) |
| Chlorophyll Index Red Edge (CIrededge) | $\frac{NIR}{RED\ EDGE} - 1$ | (Gitelson, Viña et al. 2003, Ahamed, Tian et al. 2011, Hunt Jr, Daughtry et al. 2011) |
| Chlorophyll vegetation index (CVI) | $NIR \times \frac{RED}{GREEN^2}$ | (Datt, McVicar et al. 2003) |
| Double Difference Index (DD) | $(749nm - 720nm) - (701nm - 672nm)$ | (Le Maire, François et al. 2004, Main, |

| | | |
|---|---|---|
| | | Cho et al. 2011) |
| Enhanced Vegetation Index (EVI) | $2.5 \times \frac{NIR - RED}{(NIR + 6RED - 7.5BLUE) + 1}$ | (Huete, Didan et al. 2002, Hunt Jr, Daughtry et al. 2011) |
| Green atmospherically resistant vegetation index (GARI) | $\frac{NIR - (GREEN - (BLUE - RED))}{NIR - (GREEN + (BLUE - RED))}$ | (Gitelson, Kaufman et al. 1996, Gitelson, ViÃ±a et al. 2003) |
| Green-Blue NDVI (GBNDVI) | $\frac{NIR - (GREEN + BLUE)}{NIR + (GREEN + BLUE)}$ | (Wang, HUANG et al. 2007) |
| Global Environment Monitoring Index (GEMI) | $\left(n \times (1 - 0.25n) - \frac{RED - 0.125}{1 - RED}\right)$ $n = \frac{2 \times (NIR^2 - RED^2) + 1.5 \times NIR + 0.5 \times RED}{NIR + RED + 0.5}$ | (Pinty and Verstraete 1992) |
| Simple Ratio Greenness Index (GI) | $\frac{GREEN}{RED}$ | (Zarco-Tejada, Miller et al. 2001, Main, Cho et al. 2011) |
| Green Normalized Difference Vegetation Index (GNDVI) | $\frac{NIR - GREEN}{NIR + GREEN}$ | (Ahamed, Tian et al. 2011, Hunt Jr, Daughtry et al. 2011) |
| Tasselled Cap – vegetation (GVI) | $-0.2848 \times Blue - 0.2435 \times Green - 0.5436 \times Red + 0.7243 \times NIR + 0.0840 \times SWIR - 0.1800 \times SWIR$ | (Schlerf, Atzberger et al. 2005, Lee, Alchanatis et al. 2010) |
| Infrared percentage vegetation index (IPVI) | $\frac{\frac{NIR}{NIR + RED}}{2} \times (NDVI + 1)$ | (Crippen 1990, Kooistra, Leuven et al. 2003) |
| Log Ratio (LogR) | $log\left(\frac{NIR}{RED}\right)$ | - |
| Misra Green Vegetation Index (MGVI) | $-0.386 \times GREEN - 0.530 \times RED + 0.535 \times REDEDGE + 0.532 \times NIR$ | (Misra, Wheeler et al. 1977, Bannari, Morin et al. 1995) |
| Modified NDVI (mNDVI) | $\frac{NIR - RED}{NIR + RED - 2 \times BLUE}$ | (Huete, Liu et al. 1997, Main, Cho et al. 2011) |
| Modified Simple Ratio (mSR) | $\frac{NIR - BLUE}{RED - BLUE}$ | (Kooistra, Leuven et al. 2003, Main, Cho et al. 2011) |
| Modified Simple Ratio 2 (mSR2) | $\left(\frac{NIR}{RED}\right) - \frac{1}{\sqrt{\left(\frac{NIR}{RED}\right) + 1}}$ | (Chen 1996) |
| Normalized Difference NIR / Red Normalized Difference Vegetation Index (NDVI) | $\frac{NIR - RED}{NIR + RED}$ | (Thenkabail, Smith et al. 2002, Zarco-Tejada and |

| | | |
|---|---|---|
| | | Sepulcre-CantÃ³ 2007) |
| Normalized Green (NG) | $\frac{GREEN}{NIR + RED + GREEN}$ | (Sripada, Heiniger et al. 2006) |
| Normalized Near Infrared (NNIR) | $\frac{NIR}{NIR + RED + GREEN}$ | (Sripada, Heiniger et al. 2006) |
| Hyperspectral perpendicular VI (PVIhyp) | $\frac{NIR - a \times 807 - b}{(1 + a^2)^{0.5}}$ $a = 1.17, b = 3.37$ | (Schlerf, Atzberger et al. 2005) |
| Plant Senescence Reflectance Index (PSRI) | $\frac{RED - BLUE}{NIR}$ | (Sims and Gamon 2002, Apan, Held et al. 2003) |
| Reflectance at the inflection point (Rre) | $\frac{RED + NIR}{2}$ | (Clevers, De Jong et al. 2002) |
| Red-Edge Stress Vegetation Index (RVSI) | $\frac{718 + 748}{2} - 733$ | - |
| Structure Intensive Pigment Index (SIPI) | $\frac{NIR - BLUE}{NIR - RED}$ | (Zarco-Tejada, Miller et al. 2001, le Maire, Francois et al. 2004) |
| Simple Ratio (SR) NIR/ RED | $\frac{NIR}{RED}$ | - |

## 2.3. ARTMO software

The Automated Radiative Transfer Models Operator (ARTMO) is a modular MATLAB GUI toolbox initially developed for automating the simulation of radiative transfer models (RTMs) (Verrelst, Romijn et al. 2012). This comprehensive toolbox integrates various leaf and canopy RTMs alongside essential tools for semi-automated retrieval of biophysical and biochemical variables. ARTMO is connected to a relational SQL database management system (MySQL, version 5.5 or 5.6; local installation required) for storing all generated data (i.e., simulations, statistical results) and trained models along with metadata, enabling the re-execution of earlier models or simulations. An initial version of the machine learning classification algorithm (MLCA) toolbox was introduced in version 3.29, and this functionality has been expanded in subsequent releases.

The current official version (v.1.02) of the MLCA toolbox incorporates 20 supervised MLCAs belonging to the principal families of supervised classifiers, predominantly affiliated with machine learning methodologies. Note that this initial version is limited to pixel-based classifiers, implying that object-based sub-pixel-based or scene-based deep learning classifiers have not been incorporated. Nevertheless, pixel-based classifiers enable the learning and characterization of intricate spectra.

Supervised classifiers are traditionally classified into parametric and non-parametric methods. Parametric methods are grounded in probabilistic theories, modeling the decision boundaries between classes from a fixed number of parameters, independent of the number of samples, employing global criteria for classification (Hubert-Moy, Cotonnec et al. 2001). By contrast, non-parametric methods guide the class grouping based on the digital number (single band/image) or spectral data (multi- and hyperspectral reflectance or transmittance). The spectral value distribution is independent and focused on the local data structure, requiring a substantial set of samples for the classification process (Phiri and Morgenroth 2017).

Arguably, one of the most promising non-parametric classifiers is the Gaussian process (GP) classification. GPs are stochastic processes where each random variable follows a multivariate normal distribution (Rasmussen and Williams 2006). The goal of GP classification is to learn a mapping from the input data (e.g., spectral reflectance or transmittance values) to their corresponding classification label (e.g., plant health group type), which can then be used on new, unseen spectral measurements. When the GP is developed with kernel methods (Schölkopf, Smola et al. 1998), it allows mapping the original data into a possibly infinite-dimensional space (Bishop and Nasrabadi 2006). In this space, the input-output relationship can be better estimated as the GP can consider more complex and flexible functions than the linear models. This enables the GP to capture intricate relationships between the spectral data and the health crop phenotype, leading to more accurate classification results. Due to its probabilistic framework, the GP provides uncertainty estimation per sample. This means that for each spectral measurement, the GP can provide a measure of how confident it is in its classification prediction. This uncertainty information can inform decision-making, allowing users to be more or less confident with the inferred classification label (e.g. see (Verrelst, Malenovský et al. 2019)).

### 2.3.1. Machine Learning Approaches – Gaussian process classification (GPC-BAT)

The GP has another advantage of being capable of using more sophisticated kernel functions than the standard linear kernel or the radial basis function (RBF) kernel equation (Eq. 6), which can be optimally tuned through likelihood maximization.

$$k_{RBF}(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

where $x_i$ and $x_j$ represent two spectra, σ is the variance, and $\|x_i - x_j\|$ is the Euclidean distance between the two spectra $x_i$ and $x_j$.

In the classification case, the output values of $k_{RBF}$ are discrete (±1); this causes the likelihood function to be non-Gaussian, and then some approximations should be performed (Aghababaei, Ebrahimi et al. 2022). We chose the Laplace approximation which performs well and is robust. One notable kernel function is the automatic relevance determination (ARD) kernel equation (Eq. 7),

$$K_{ARD}(x_i, x_j) = \exp\left(-\frac{1}{2}(x_i, x_j)^{\top} \sum^{-1} (x_i, x_j)\right) (7)$$

where Σ is a diagonal matrix, whose diagonal tries are constituted by $\{\sigma_1^2, \ldots, \sigma_d^2\}$ parameters to weight each input dimension. This kernel covariance function requires one parameter per input feature; it can be optimized under that framework, and it allows providing a band ranking based on their optimal values.

Following the rationale as presented in Verrelst et al., (2016) (Verrelst, Rivera et al. 2016) for GP regression, a GPC-based band ranking feature has been implemented into a so-called band analysis tool (i.e., GPC-BAT). In short, we employ a simplified and general iterative backward greedy algorithm to identify the most informative bands. This algorithm assesses the impact of each band on the prediction error in the context of the remaining bands. At each iteration, the algorithm removes the band with the highest uncertainty σb, thereby re-training the GPC model with the remaining bands. This is referred to as sequential backward band removal (SBBR). The SBBR algorithm is analogous to recursive feature elimination (RFE), a technique earlier presented with support vector machines or random forests. In RFE, the feature with the lowest ranking score is eliminated, iteratively removing insignificant features until only the most relevant ones remain (e.g., (Bazi and Melgani 2006, Archibald and Fann 2007, Pal and Foody 2010, Zhang and Yang 2020)). This SBBR approach allows us to pinpoint the bands that most strongly influence the prediction of our target classes. These bands provide valuable insights into the spectral characteristics that best capture the sensitivity of the classes of interest, e.g., healthy, and diseased groups.

A principal application of GPC-BAT is that the algorithm identifies how many bands are minimally needed in order to retain robust results and informs us about the most sensitive wavelengths. Accordingly, the output GUI delivers the following band analysis outputs: (1) overall accuracy (OA) statistics as a function of #bands plotted over the sequentially removed bands until only two bands are left, (2) associated wavelengths selected by the tool (Figure 1).

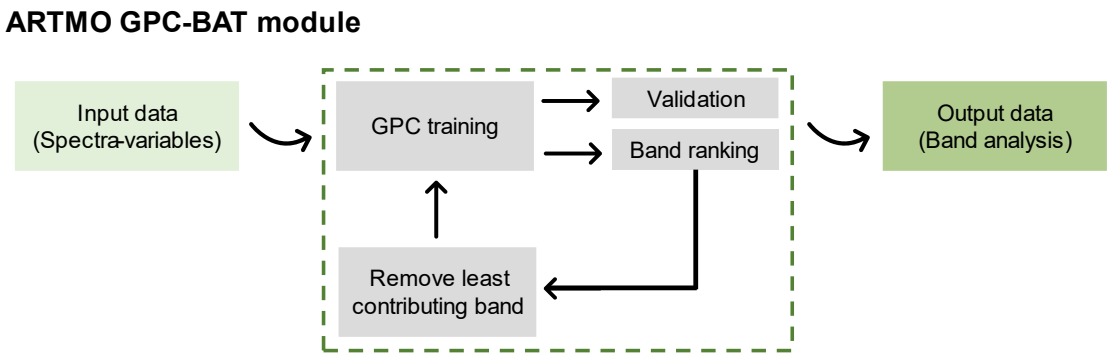



**Figure 1** Schematic flow diagram of GPC-BAT within ARTMO's MLCA toolbox adapted from (Verrelst, Rivera et al. 2016).

## 3. Experimental Results

### 3.1. Spectral vegetation indices – parametric approach

This section presents the predictive classification results of the approach combining the calculation of different VIs (Table 3) followed by the computation of the FDA model, which allowed i) the classification of tomato leaflet spectral samples collected on healthy (Control, Con), and both inoculated diseased tissues with Pst, and Xeu bacteria; ii) and the classification of kiwi leaf spectral samples measured on non-symptomatic (NS), and symptomatic (S) diseased tissues.

#### 3.1.1. Tomato disease in walk-in chamber

Table 2 presents the results of an overall accuracy of 63.30% (proportion of correctly classified instances), and a kappa coefficient of 44.70 (which indicates agreement between the predicted and actual classes beyond random occurrence) for the validation dataset (Table 3, detailed information about the training results is present in Supplementary Materials Table S1). The model metric analysis per class revealed that samples inoculated with Pst bacteria presented good precision (75.66%), sensitivity (65.60%), specificity (90.52%), and F1-score (70.27%) (Table 3). These metrics indicate the model's suitability for accurately performing correct predictions for both classes (healthy vs. inoculated), correctly identifying instances of these classes, distinguishing samples that do not belong to these classes, and a good balance in capturing true positives and avoiding false positives (Table 3). The control class presented a higher sensitivity, highlighting the model's efficiency in correctly identifying healthy samples. In contrast, the Xeu class had a reduced precision (52.44%), sensitivity (48.56%), and F1-score (50.43%) which translates into a higher likelihood of false positives in Xeu prediction, a more probable miss of a considerable proportion of positive instances of this class, and an imbalance in assessing true positives and refrain false positives (Table

3). Therefore, although the FDA model exhibits good capabilities in differentiating samples measured in healthy (Control) and Pst diseased tissues, there is potential for improving the identification of spectra captured on Xeu inoculated tissues (Table 3).

A fraction of these false positives may be accounted for by the sensitivity gap between digital and visual phenotyping methods. These results are based on visual phenotyping, which only allows samples with visible symptoms to be identified. However, Xeu-inoculated plants may present early changes in their optical properties long before the appearance of symptoms, which could justify some of the false positive cases recorded.

CM results indicate that predictions of samples collected on tissues inoculated with Xeu were more challenging to the model when compared to the healthy and inoculated with Pst ones, presenting a higher number of wrong classifications than in the remaining classes studied (Table 4, for both training and validation sets). In fact, the model only correctly identified 49% of the Xeu samples. The majority of the remaining Xeu samples were wrongly inputted to the Control class. This can be related to the fact that macroscopic symptoms only appeared 8 DAI, resulting in a high number of non-symptomatic samples (presenting a phenotype similar to the healthy ones) whose spectral signature is more similar to healthy samples than Pst diseased ones. In contrast, approximately 74% of the healthy (Control) samples were accurately classified, and 70% inoculated with Pst (Table 4).

### 3.1.2. Kiwi bacterial canker disease in the field

In turn, when the IVs-based modeling approach was applied in the binary classification of non-symptomatic (NS in Table 3) and symptomatic (S in Table 3) samples of the validation set taken on kiwi leaves in field conditions, the overall accuracy achieved was 71.33%. The Kappa value of 41.73% demonstrates the model's effectiveness in classifying the two classes (results for the training set can be seen in Supplementary Materials Table S1). The model metrics for the spectra collected in non-symptomatic and symptomatic tissue's spectra revealed that the model acceptably identifies a significant proportion of true positive samples of these classes, classifies samples belonging to both classes, and has a good ability to correctly identify instances that do not belong to the class in analysis, along with a good balance between finding positive cases and avoiding false positives (Table 3).

The CM values show that the model has less difficulty predicting non-symptomatic samples correctly classifying 78% of the total samples compared to the symptomatic samples (67%) (Table 4).

**Table 3** Cross-validation statistics of the Flexible Discriminant Analysis (FDA) using the validation set of the Vegetation Indices (VIs) computed in the hyperspectral data collected on tomato leaflet tissues (transmittance) and kiwi leaf tissues (reflectance). Model metrics by class are also provided.

| | *Overall classes* | | *Metrics per class* | | | |
|---|---|---|---|---|---|---|
| *Class* | *Accuracy (%)* | *Kappa (%)* | *Precision (%)* | *Sensitivity (%)* | *Specificity (%)* | *F1 (%)* |
| *Tomato (703 observations)* | | | | | | |
| Con | 63.33 | 44.77 | 63.67 | 76.03 | 77.22 | 69.30 |
| Pst | | | 75.66 | 65.60 | 90.52 | 70.27 |
| Xeu | | | 52.44 | 48.56 | 76.74 | 50.43 |
| *Kiwi (504 observations)* | | | | | | |
| NS/S | 71.33 | 41.73 | 67.69 | 66.67 | 75.00 | 67.18 |

Con – Control (healthy), Pst – Inoculated with *Pseudomonas syringae* pv. *tomato*, Xeu – Inoculated with *Xanthomonas euvesicatoria*, NS – Non-Symptomatic, S - Symptomatic

**Table 4** Confusion Matrix results of the Flexible Discriminant Analysis (FDA) using the Vegetation Indices (VIs) computed in the hyperspectral data collected on tomato leaflet tissues (transmittance) and kiwi leaf tissues (reflectance). The predicted samples of each class (column) that were correctly classified for each true class (row) for the spectral data collected on tomato leaflets tissues (left) and kiwi leaf tissues (right) are shown. The classes used in the tomato case study were control samples (healthy, Con), and samples inoculated with *Pseudomonas syringae* pv. *tomato* (Pst), and samples inoculated with *Xanthomonas euvesicatoria* (Xeu). In turn, the binary classes Non-Symptomatic (NS), and Symptomatic (S) were applied to the kiwi case study.

| | | *Predicted Class - Tomato* | | | | | | | *Predicted Class - Kiwi* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Training* | | | *Validation* | | | | *Training* | | *Validation* | |
| | | Con | Pst | Xeu | Con | Pst | Xeu | | NS | S | NS | S |
| *True Class* | Con | 416 | 10 | 141 | 184 | 4 | 54 | NS | 157 | 40 | 63 | 21 |
| | Pst | 52 | 345 | 112 | 22 | 143 | 53 | S | 50 | 107 | 22 | 44 |
| | Xeu | 193 | 90 | 284 | 83 | 42 | 118 | | | | | |

### 3.2. GPC-BAT performance with original training data and further validation

This section presents the predictive classification results of the approach using the ARTMO GPC-BAT tools, which also allowed i) the classification of tomato leaflet spectral samples collected on healthy (Control, Con), Pst inoculated, and Xeu inoculated

tissues; ii) and the classification of kiwi leaf spectral samples measured on non-symptomatic (NS), and symptomatic (S) tissues.

### 3.2.1. Tomato diseases in walk-in chamber

For the selected 23 wavelengths (from the 51 available) for the tomato case study and 577 wavelengths (from the 611) for the kiwi case study, the models presented the best classification metrics (Table 5). The prediction for discriminating the different classes defined for the tomato dataset achieved a maximum overall accuracy of 70.46% and kappa of 55.60%. Furthermore, metric evaluation per class provides insights into the model's performance in distinguishing between healthy (Con in Table 5) and diseased instances (Pst and Xeu in Table 5). In terms of Precision (80.72 vs. 75.24%), Specificity (88.75 vs. 89.14%), and F1-values (73.87 vs. 77.31%), the Control and Pst inoculated classes presented good metric levels. These results, highlight the model's accuracy for identifying positive predictions for both classes, distinguishing samples that do not belong to these respective classes and achieving good balance in capturing true positives and avoiding false positives (Table 5). The Xeu inoculated class, compared with Pst, showed lower values of Precision (55.93 vs. 75.24%) and F1-Score (60.04 vs. 71.31%), indicating a higher likelihood of false positives in their prediction and an imbalance in assessing true positives and refraining from false positives. This class had a good Specificity value, similar to the other two, suggesting the model’s performance in distinguishing spectral measurements that do not belong to it (Table 5). Regarding Sensitivity, the Pst class presented the higher level, indicating the model's effectiveness in correctly identifying instances of this class. However, the values for the two remaining classes were lower, implying that the model may miss a notable proportion of positive instances in the Control and Xeu classes (Table 5).

The CM also demonstrates that 81% of the healthy and 75% of Pst inoculated samples were correctly classified. Similar to the previous approach (based on VIs), the GPC-BAT also faced more difficulty in accurately classifying the inoculated Xeu samples, with only predicting 56% of the cases correctly (Table 6). Hence, while the model demonstrates strong capabilities in distinguishing samples collected in Control (healthy) tissues and those measured on Pst diseased tissues, it needs to be enhanced to accurately identify spectra collected on Xeu diseased tissues (Table 5).

### 3.2.2. Kiwi bacterial canker disease in the field

The binary classification performed for the kiwi leaves hyperspectral reflectance measurements achieved a maximum overall accuracy of 75.40% and a kappa of 49.95%. The model proved effective in both class predictions, allowing the distinction between

non-symptomatic (NS in Table 5) and symptomatic (S in Table 5) assessments collected in kiwi leaves in field conditions. The model metrics for the spectra collected in non-symptomatic and symptomatic tissue's spectra revealed that the model effectively identifies a significant proportion of true positive samples of these classes, a good ability to classify instances that do not belong to them correctly, and a well-balanced between identifying positive instances and avoiding false positives (Table 5).

**Table 5** Cross-validation statistics of the Gaussian Process Classification (GPC) models developed using hyperspectral data collected on tomato leaflet tissues (transmittance) and kiwi leaf tissues (reflectance).

| | *Overall classes* | | *Metrics per class* | | | |
|---|---|---|---|---|---|---|
| *Class* | *Accuracy (%)* | *Kappa (%)* | *Precision (%)* | *Sensitivity (%)* | *Specificity (%)* | *F1 (%)* |
| *Tomato* | | | | | | |
| Con | 70.46 | 55.60 | 80.72 | 68.09 | 88.75 | 73.87 |
| Pst | | | 75.24 | 79.51 | 89.14 | 77.31 |
| Xeu | | | 55.93 | 64.81 | 78.32 | 60.04 |
| *Kiwi* | | | | | | |
| NS | 75.40 | 49.95 | 79.36 | 77.16 | 73.02 | 78.25 |
| S | | | 70.40 | 73.02 | 77.16 | 71.69 |

Con – Control (healthy), Pst – Inoculated with *Pseudomonas syringae* pv. *tomato*, Xeu – Inoculated with *Xanthomonas euvesicatoria*, NS – Non-Symptomatic, S - Symptomatic

The CM demonstrates that 79% of the non-symptomatic samples were accurately classified, along with 70% of the symptomatic samples (Table 6).

The GPC model did not perform best for both case studies when all the spectral bands were applied. In the tomato case study, the overall accuracy and kappa values when all spectral features were used were lower, reaching 69.06% and 53.48%, respectively. Likewise, the kiwi case study's values were 61.90% and 22.50% for accuracy and kappa, respectively. Furthermore, for the kiwi case study, the outcomes were more unstable when compared to the model developed with the wavelengths chosen by BAT, presenting a higher standard deviation (SD), and processing time (Table 7). In the tomato case study, the SD value was lower when all the features were used, but the processing time was almost 40% superior (Table 7).

In terms of selected sensitive wavelengths, when the spectral data collected on kiwi was used, GPC profusely selected wavelengths greater than 800 nm (26 in 34 wavelengths), these wavelengths (> 800 nm) did not prove to be important in the construction of the selected VIs. Only PVIhyp (800, 1000 nm) selected wavelengths at

800 nm, but this VI has a very modest representation (24.46%) in the distinction between the non-symptomatic and symptomatic classes.

**Table 6** Confusion Matrix of the GPC model results show the predicted samples of each class (column) correctly classified for each true class (row) for the spectral data collected on tomato leaflets' tissues and kiwi leaf tissues. The classes used in the tomato case study were Control samples (healthy, Con), samples inoculated with *Pseudomonas syringae* pv. *tomato* (Pst), and Samples inoculated with *Xanthomonas euvesicatoria* (Xeu). The binary class Non-Symptomatic (NS), and Symptomatic (S) were applied to the kiwi case study.

| | | *Predicted Class - Tomato* | | | | | | | *Predicted Class - Kiwi* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Training* | | | *Validation* | | | | *Training* | | *Validation* | |
| | | Con | Pst | Xeu | Con | Pst | Xeu | | NS | S | NS | S |
| *True Class* | Con | 595 | 48 | 221 | 653 | 52 | 254 | NS | 232 | 32 | 223 | 66 |
| | Pst | 31 | 512 | 89 | 38 | 547 | 103 | S | 21 | 169 | 58 | 157 |
| | Xeu | 102 | 95 | 419 | 118 | 128 | 453 | | | | | |

**Table 7** Cross-validation statistics of the GPC models developed using hyperspectral data collected on tomato leaflet tissues (transmittance) and kiwi leaf tissues (reflectance) were used.

| *SD* | *Time (s)* | *FS* | *Wavelengths (nm)* |
|---|---|---|---|
| *Tomato* | | | |
| 2.75 | 819.20 | 51 | All wavelengths (300.09 to 800.34 nm) |
| 3.32 | 491.08 | 23 | 430.00, 440.21, 450.04, 460.31, 490.04, 500.00, 510.41, 520.00, 550.20, 560.31, 570.03, 620.21, 640.35, 650.24, 660.15, 670.08, 680.02, 690.42, 700.41, 710.41, 740.09, 750.17, 760.26 |
| *Kiwi* | | | |
| 4.49 | 171.86 | 611 | All wavelengths (400 to 1010 nm) |
| 3.23 | 14.80 | 34 | 544, 597, 754, 771, 790, 791, 795, 825, 835, 839, 845, 850, 851, 860, 864, 866, 869, 881, 883, 888, 893, 902, 905, 906, 928, 932, 939, 945, 947, 973, 980, 993, 999, 1006 |

FS – Feature Selection, SD – Standard Deviation

### 3.3. Comparing the performance of VIs and GPC-BAT

#### 3.3.1. Models' metrics-based comparison

The tested approach combining the VIs with the FDA model allowed the identification of the most relevant VIs analyzed for class identification for both the tomato and kiwi case studies (Table 8). When the tomato spectral data was used, the selected

wavelength combinations present in the computed VIs considered features in the Blue (450.04 nm), Green (500.00, and 550.20 nm), Red-Edge (680.02, 690.42, 700.41, 730.04, and 750.17), and NIR (800.34 nm). Similar wavelengths were selected when the data collected on kiwi leaves were used and were mainly located in the Blue (400, and 450 nm), Green (530, 553, and 554 nm), Red-Edge (670, 677, 700, 705, 730, and 750 nm), and NIR (780, 800, 994, and 1000 nm) (Table 8). For both species, equal wavelengths were identified by the approach, such as 450 (Blue), 550 (Green), 700, 730, 750 (Red-Edge), and 800 nm (NIR).

For the tomato case study, the FDA selected five VIs whose formula integrated three wavelengths: PSRI, CCCI, EVI, SIPI, and GARI (Table 8). In contrast, in the kiwi case study, all the chosen VIs presented only two wavelengths (Table 8).

**Table 8** Vegetation Index (VI) importance for classification according to Flexible Discriminant Analysis (FDA). The importance value corresponds to the t-statistic value scaled to the maximum.

| *Case study - Tomato* | | | *Case study - Kiwi* | | |
|---|---|---|---|---|---|
| *VI* | *Wavelengths (nm)* | *Importance (a.u.)* | *VI* | *Wavelengths (nm)* | *Importance (a.u.)* |
| mSR2 | 700.41, 750.17 | 100.00 | Chlgreen | 553, 800 | 100.00 |
| BRI2 | 450.04, 690.42 | 70.30 | mSR2 | 705, 750 | 67.15 |
| GEMI | 680.02, 800.34 | 42.92 | CI | 450, 700 | 52.94 |
| PSRI | 550.20, 680.02, 750.17 | 31.69 | GI | 554, 677 | 44.45 |
| CCCI | 550.20, 700.41, 800.34 | 27.59 | BRI2 | 450, 690 | 40.55 |
| EVI | 450.04, 680.02, 800.34 | 22.41 | AVI | 400, 994 | 33.71 |
| SIPI | 450.04, 680.02, 800.34 | 16.50 | PVIhyp | 800,1000 | 24.46 |
| Chlgreen | 550.20, 730.04 | 10.71 | Chlgreen | 530, 730 | 19.65 |
| SIPI | 500.00, 690.42, 800.34 | 6.66 | Rre | 670, 780 | 16.46 |
| GARI | 450.04, 550.20, 680.02, 800.34 | 0.00 | | | |

AVI—Ashburn Vegetation Index, BRI2 — Blue/Red Pigment Index, CCCI - Canopy Chlorophyll Content Index, Chlgreen—Chlorophyll Green, CI—Coloration Index, EVI - Enhanced Vegetation Index, GEMI - Global Environment Monitoring Index, GI—Simple Ratio Greenness Index, mSR2—Modified Simple Ratio, PSRI - Plant Senescence Reflectance Index, PVIhyp—Hyperspectral perpendicular VI, Rre—Reflectance at the inflection point, and SIPI - Structure Intensive Pigment Index

The spectral wavelengths identified by GPC-BAT as the most sensitive for performing bacterial plant disease diagnosis in tomato (in controlled environmental conditions) were mainly located in the Blue (430.00, 440.21, 450.04, and 460.31 nm), Green (510.41, 520.00, 550.20, and 560.31 nm), Red (640.35, 650.24, and 660.15 nm), Red-Edge (670.08, 680.02, 690.42, 700.41, 710.41, 740.09, 750.17, and 760.26 nm) (Table 5, Fig. 2). In turn, the model identified the most sensitive features for discriminating in field diseased kiwi leaves infected with bacterial canker (caused by Psa)

as mainly occurring in the Green (544, and 597 nm), Red-Edge (754 nm), and NIR (771, 790, 791, 795, 825, 835, 839, 845, 850, 851, 860, 864, 866, 869, 881, 883, 888, 893, 902, 905, 906, 928, 932, 939, 945, 947, 973, 980, 993, 999, and 1006 nm) (Table 5, Fig. 3). Similar features were only selected for the two species in the Green (550.20 vs. 544 nm), and Red-Edge (750.17 vs. 754 nm) spectral regions. In contrast, the Blue and Red spectral regions were only considered in the tomato case study, and the NIR was considered only in the kiwi case study (since only the spectral sensor used in this assay had wavelengths assessed in these region) (Table 5).

Both approaches, the one using VIs and FDA and the one using the GPC-BAT, for the tomato case study, selected coincident wavelengths in the Blue (450.04 nm), Green (550.20 nm), and Red-Edge (680.02, 690.42, 700.41, and 750.17 nm). The same consistency was not observed for the kiwi case study since different features were chosen. Nevertheless, these were similar in the Red-Edge (750 vs 754 nm), and NIR (795 vs 800 nm, 994 vs 993 nm, 999 vs 1000 nm) (Table 5).

In GPC-BAT the predictive power of each wavelength for the target variable is evaluated by the index sigma (σ). Accordingly, the lower the sigma value, the more important the feature is. Thus, the contribution of each spectral feature can be ranked through the quantification of this property. In the tomato case study, for the selected spectral features, it is possible to observe in Fig. 2 that for the identification of the Control (Con) class the 440.21, 450.00, 460.31, 490.04, 510.41, 520.00, 640.35, 680.02, and 750.17 nm were the more relevant wavelengths since they presented a lower sigma value (i.e., more weight in model) leading to a higher distance from the plots' center (blue dots). The most significant features for predicting samples made on tomato leaflet tissues inoculated with Pst were 440.21, and 450.04 nm (red dots). In classifying samples collected on leaflet tissues inoculated with Xeu the 440.21, 450.04, 660.15, and 680.02 nm (orange dots). In turn, in the kiwi dataset, from the selected wavelengths, the identification of samples belonging to both classes was more influenced by the 597, 771, 791, 835, 869, 883, 902, and 999 nm features (Fig. 3). The wavelengths selected as relevant for the classification of non-symptomatic samples are the same as those chosen to predict the opposing symptomatic class (Fig. 3). This tendency was expected since this is a binary classification task, where the differences between the two categories in the study are expected to occur in coincident spectral features.

### 3.3.2. Biological interpretation of sensitive wavelengths

These wavelengths chosen in both case studies present an interesting biological significance since they coincide with the spectral absorption regions of several

photosynthetic pigments, namely: i) chlorophylls, in the Blue region around 430 to 480 nm, and the Red, from 640 to 680 nm; ii) carotenoids, including xanthophylls, in the Blue-Green region nearby 400 to 550 nm; iii) flavonoids, in the UV-Blue wavelengths ranging from 315 to 500 nm, including anthocyanins whose absorption band is from 500 to 550 nm, iv) and pheophytins, whose absorption action is located in the Blue (430 to 480 nm) and Red (640 to 680 nm). Furthermore, the selected wavelength features also overlap the NIR spectral range associated with interactions between light and leaf water content and between light and leaf structural components (such as cellulose, and lignin). All these pigments and structural components are affected by the action of Psa, Pst, and Xeu bacteria in kiwi and tomato leaves, respectively.

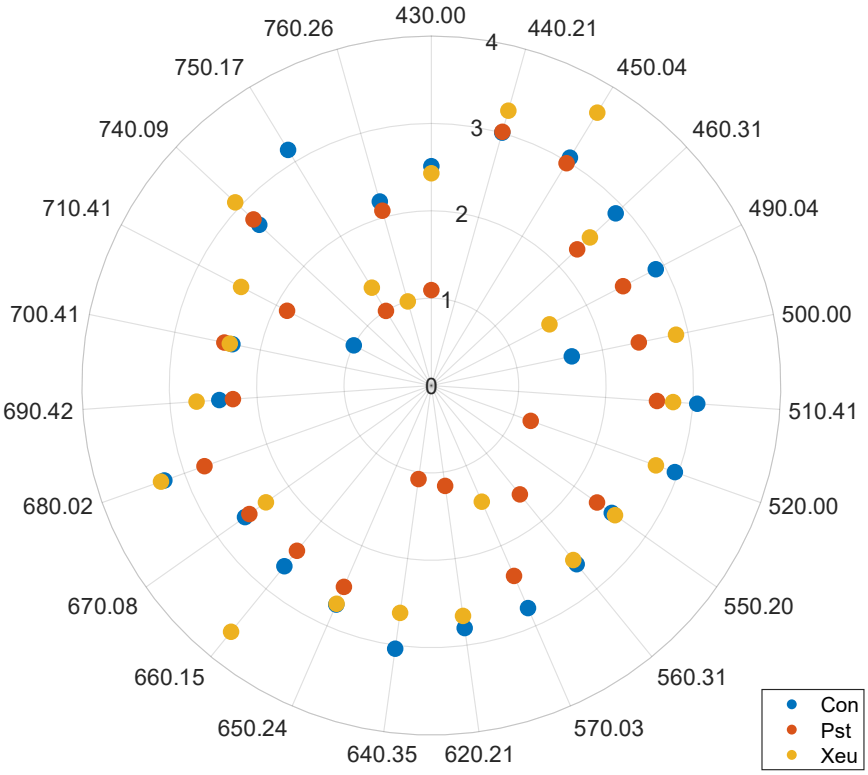



**Figure 2** Gaussian Process Classification sigma bands polar plot, representing the most significant wavelengths for each class in the study: Control samples (healthy, Con), samples inoculated with *Pseudomonas syringae* pv. *tomato* (Pst), and samples inoculated with *Xanthomonas euvesicatoria* (Xeu). The lower the sigma value, the greater the importance of the wavelength.

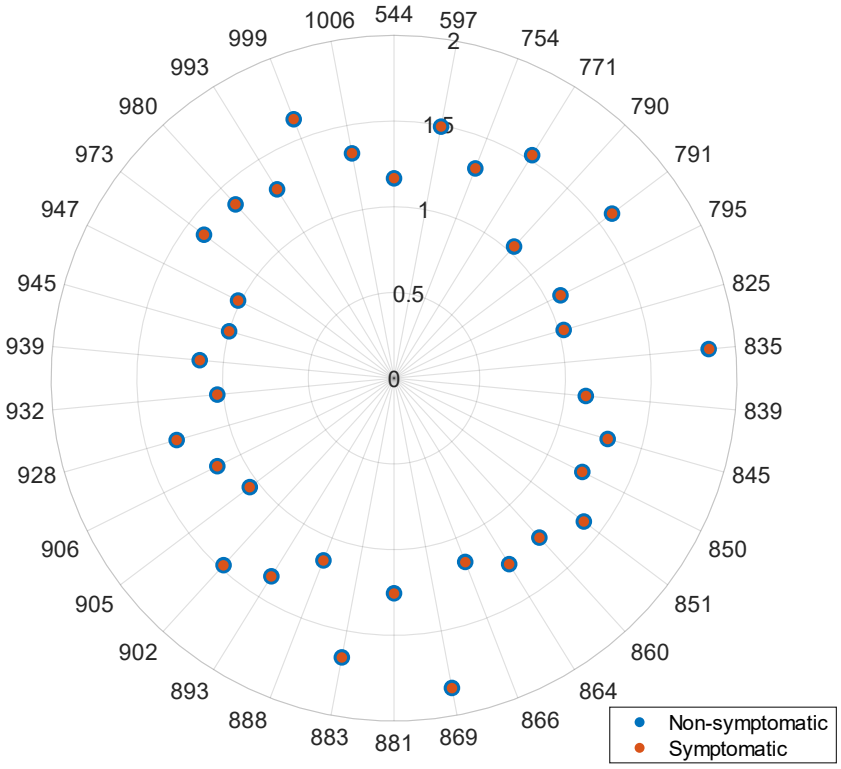



**Figure 3** Gaussian Process Classification sigma bands polar plot, representing the most significant wavelengths for the binary class in the study: Non-Symptomatic (NS), and Symptomatic (S). The lower the sigma value, the greater the importance of the wavelength.

## 4. Discussion

This work analyzed two methodologies for performing bacterial disease classification in tomato (assay performed in controlled environmental conditions, using a transmittance-based sensor) and kiwi (assay made in the field, using a reflectance-based sensor) plants. One approach combines the calculation of different VIs described in the literature and a machine learning algorithm with a built-in FS method (FDA). Another approach uses the two distinct hyperspectral datasets combined through the ARTMO GPC-BAT. In both approaches, the most relevant spectral wavelengths for class detection were identified and linked to their biological significance. The first approach uses VIs developed according to the physiological information of plants. In contrast, the second approach constitutes a data-driven approach. The tomato experiment (3 classes: Control vs. Pst and Xeu), compared to the kiwi (binary), represents a more complex classification model. The sensor used in the tomato experiment allows obtaining spectral information from the visible spectrum up to 800 nm, with wavelength resampling to 10 nm intervals (approximately), whereas the sensor used in the field experiment with kiwi

allows data from the visible spectrum up to 1000 nm, with a more precise step of 1 nm wavelength.

It is possible to observe that both approaches allowed the identification of the different classes in the study, using the tomato and kiwi datasets. Nevertheless, the strategy involving the application of VIs as an FS technique showed lower classification metrics than the methodology that used the GPC-BAT. This may be related to the fact that the VIs used, despite being well established in the literature, were developed for specific plant traits and situations differing from the bacterial plant disease diagnosing problem in the study. Furthermore, since they are calculated using only available spectral features, they may not use all the information in spectral narrowband, high-dimensional hyperspectral data [5]. In contrast, GPC-BAT considered all the available spectral features and performed a selection according to their relevance for identifying the class in the study.

The GPC-BAT, when applied to the analysis encompassing all wavelengths captured by the hyperspectral sensors, exhibited lower classification metrics than the VIs approach in both the tomato and kiwi case studies. For instance, in the kiwi study, where the hyperspectral data indicated a narrowband field of 1 nm, the model's performance using all available features was as poor as when only three wavelengths were utilized (data not displayed). This may be related to hyperspectral data being super-imposed in the recorded spectra at different interference scales (Tosin, Martins et al. 2022), (i.e., the data collected corresponds to several structural and metabolic plant compounds present in the area measured), and to the significant amount of redundant information embedded in contiguous wavelengths. As a result, only a few specific spectral variables are relevant to identify diseased plant tissues (Caicedo, Verrelst et al. 2014, Rivera, Verrelst et al. 2014).

In this regard, Feature Selection or spectral reduction techniques are, thus, recommended to overcome this hurdle. In this work, two approaches were analyzed namely an FDA algorithm and the BAT of ARMO. Given the selected wavelengths, both studied strategies (VIs-based vs. GPC-BAT) presented comparable results for both case studies, notably when dealing with the more complex tomato dataset especially when the tomato data set. Equivalent wavelengths were found in the Blue (450 nm), Green (550 nm), and Red-Edge (680, 690, 700, and 750 nm). VIs further highlighted the 800 nm wavelength in the NIR. In the kiwi case study, the features selected by the two algorithms were similar but not entirely coincident, namely in the Green (where VIs selected the 530, 553, and 554 nm, and the GPC-BAT chosen the 544, 597 nm), Red-

Edge (VIs identified the 670, 677, 700, 705, 730, and 750 nm as relevant, and GPC-BAT only considered the 754 nm), and NIR (VIs picked a lower amount of wavelengths, namely 780, 800, 994, and 1000 nm, when compared to the GPC-BAT which took into account the following wavelengths 771, 790, 791, 795, 825, 835, 839, 845, 850, 851, 860, 864, 866, 869, 881, 883, 888, 893, 902, 905, 906, 928, 932, 939, 945, 947, 973, 980, 993, 999, and 1006 nm). Only the VIs presented features in the Blue 400, and 450 nm. Thus, it is important addressing that despite the two modeling strategies applied to work with different types of spectra (reflectance vs. transmittance), having different spectral resolution (~10-10 nm vs. 1-1 nm) and presenting different pre-processing methods, they selected similar wavelengths.

These findings present biological significance since the relationship between the plant host and the pathogen causes changes in photosynthetic pigment content, water levels, and structural composition (e.g., cellulose and lignin levels) (Blancard 2012). This ultimately leads to modifications in the tissues' spectral behavior. In particular, the variance in spectral characteristics among diseased leaves infected by distinct bacteria could be linked to the generation of unique molecules by each pathogen, which may influence the spectral signature of the host. For instance, Pst bacteria produce a specific phytotoxin named coronatine that induces changes in chlorophyll fluorescence (by altering photosystem II – PSII), impacting tomato plant tissues' absorption and scattering of light (Zhang, He et al. 2021). Moreover, the host tomato plant can activate diverse defence responses upon encountering a pathogen, initiating a cascade of biochemical and molecular reactions that further contribute to spectral modifications in the visible wavelength ranges. Phytoalexins (e.g., flavonoids) serve as an example, with their production hypothesized to be linked to an increase in the spectral reflectance of plants in the VIS range (Leucker, Wahabzada et al. 2016).

Previous studies performed by our team also reported similar outcomes. In particular, that study developed a methodology for early diagnosing two bacterial diseases of tomato, caused by Pst and Xeu bacteria, using hyperspectral transmittance data and an applied predictive modeling approach (Reis Pereira, Santos et al. 2023). A total of 3478 spectral measurements were normalized and subjected to a Linear Discriminant Analysis (LDA) aiming to reduce data dimensionality. This algorithm highlighted similar relevant wavelengths in Blue, Green, and Red spectral regions. Furthermore, a modeling approach using a Support Vector Machine was applied for spectral classification. It achieved an accuracy of 100% for samples measured on tissues inoculated with Pst and 74% for tissues inoculated with Xeu when samples collected before symptom appearance were used (Reis Pereira, Santos et al. 2023). Likewise,

another study performed on a kiwi orchard allowed the identification of hyperspectral reflectance samples collected on non-symptomatic and symptomatic Psa disease leaf tissues. Several methodologies involving different Feature Selection techniques combined with different Machine Learning algorithms were explored, and the one combining a stepwise forward various selection (SFVS) approach followed by the computation of an SVM algorithm was selected, achieving an overall accuracy of 85%. Like the other strategies explored, the SFVS elected the Blue, Green, and NIR regions as the most relevant for sample classification.

Furthermore, other researchers reported similar classification findings to the ones found in the present work, namely when studying different tomato and kiwi diseases based on modelling hyperspectral spectroscopy data. The suitability of a portable hyperspectral spectrometer combined with various algorithms for FS and data modeling for early non-destructive diagnosis of tomato bacterial wilt disease (*Erwinia tracheiphila*) in leaves was explored (Cen, Huang et al. 2022). The model presenting higher evaluation metrics (overall accuracy of 90.70%) applied Genetic Algorithms for FS and SVM to predict classification. The Simple Ratio Pigment Index (SRPI) was the VI and found to have a higher contribution in the developed model. It considers 430 and 680 nm wavelengths and is sensitive to leaf nitrogen content and photosynthetic efficiency (and is similar to our findings) (Cen, Huang et al. 2022).

Another study using tomato plants explored the usage of a portable high-resolution spectroradiometer combined with VIs, Principal Component Analysis (PCA), and a classification model K-nearest neighbor (KNN) for the diagnosis of late blight (*Phytophthora infestans*), target (*Corynespora cassiicola*), and bacterial spot (*Xanthomonas euvesicatoria*) (Lu, Ehsani et al. 2018). They successfully identified the spectral samples collected on detached tomato leaflets with an accuracy reaching the 100% level even in non-symptomatic stages (Error Rate of 9.50%), when the 15 VIs selected by PCA in the first principal component (PC) were considered. Interestingly, it is possible to observe that when 30 VIs selected by PCA and belonging to the first PC were used, the model showed a lower accuracy value (65.20%) and a higher error rate (28.6). In terms of VIs, the ones selected presented similar features to the ones found in our study (such as the 680, and 800 nm used in the Normalized difference index and the Simple Ratio, Structure-intensive pigment index, among others) (Lu, Ehsani et al. 2018).

Hyperspectral VIS-NIR spectroscopy was, moreover, used for the non-destructive early diagnosis of tomato chlorosis virus (ToCV) (Morellos, Tziotzios et al. 2020). They used a Neighborhood component analysis (NCA) for performing FS and for

selecting the most relevant VIs in the study, along with two ML models for data modeling (XY-fusion network – XY-F – and Multilayer Perceptron with Automated Relevance Determination – MLP-ARD). The best overall accuracy (92.1% before outlier removal and 100% after outlier removal) was obtained using MLP-ARD. In terms of relevant VIs, is possible to observe that wavelengths like 550, 670, 700, 720, 740, and 800 nm, among others, were present in the most notable VIs formulae (such as Anthocyanin Reflectance Index – ARI – Pigment Specific Simple Ratio – PSSR –, Red Edge Inflection Point – REIP –, Simple Ratio – SR –, and Vogelmann Index – VOG). In turn, from the 15 wavelengths selected by the NCA, these were mostly located in the Blue (402.20 to 449.20 nm), green (556.40 to 566.40 nm), Red-Edge (676.40 to 726.30 nm), and NIR (862.10 nm). These outcomes coincided with our observations (Morellos, Tziotzios et al. 2020).

The feasibility of multispectral data for predicting kiwifruit decline (probably caused by *Phytophtora* spp. and *Phytopythium* spp.) in diseased orchards was also, tested (Savian, Martini et al. 2020). Multispectral data included the 550, 660, and 790 nm spectral features, and when combined with K-means clustering allowed the determination of kiwi plants' vigor affected or not by the disease with 73% (or more) Accuracy and 82% Precision. These results are, thus, in line with ours also identifying the Green, Red, and NIR as relevant for estimating plant biophysical traits (Savian, Martini et al. 2020).

The present outcomes demonstrate that hyperspectral transmittance and reflectance spectroscopy can identify healthy and diseased tissues, such as tomato (herbaceous) and kiwi (woody) crops, in laboratory or field conditions. Further research is advised to explore if specific host-pathogen interactions require customized modeling approaches to be predicted or if it is possible to elaborate a unified strategy that allows bacterial disease assessment. Nevertheless, it should be taken into consideration that model comparison may be challenging due to several factors: pathogen species in the study; the occurrence of specific host-pathogen interactions; the number of spectral points measured; the environmental conditions where the data is collected; and, the stage of the disease cycle where the spectral assessments are made, among others. Furthermore, hyperspectral spectroscopy sensors present a relatively low Technology Readiness Level (TRL), indicating that these sensors have a large margin to be improved. In this regard, developing and enhancing effective FS strategies or Dimensionality Reduction approaches may be conducted to identify specific spectral regions valuable for performing plant disease diagnosis, which may be incorporated in multispectral sensors involving lower production and data processing costs.

## 5. Conclusions

This study aimed to explore and compare two distinct modelling approaches, namely the parametric Spectral Vegetation Indices (VIs) and the Gaussian Process Classification based on an Automated Spectral Band Analysis Tool (GPC-BAT), for diagnosing bacterial diseases in plants using hyperspectral sensing. This comparative analysis was conducted across controlled conditions with tomato plants and field conditions with kiwi plants, highlighting performances and insights from each approach.

VIs demonstrated moderate success in differentiating healthy and diseased tissues in both tomato and kiwi plants. For tomato plants, VIs revealed good precision in distinguishing healthy tissues from those inoculated with Pst bacteria. However, the identification of Xeu-inoculated tissues showed limitations, possibly due to early spectral modifications before visible symptoms occurred later than in the Pst case. In kiwi plants, VIs performed reasonably well in discriminating between non-symptomatic and symptomatic tissues, although with slightly lower accuracy in the latter.

The feature reduction by GPC-BAT leads to enhanced accuracy in identifying healthy and diseased tissues in both tomato and kiwi plants. The model's precision in classifying healthy and Pst-inoculated tomato tissues was commendable. However, its performance in identifying Xeu-inoculated tissues required improvement. For kiwi plants, GPC-BAT displayed notable accuracy in distinguishing non-symptomatic and symptomatic tissues, though with a slight struggle in predicting symptomatic cases.

Both approaches demonstrated spectral bands in common across tomato and kiwi plants, particularly in the Blue, Green, Red-Edge, and NIR regions. VIs showed consistency in selecting specific wavelengths for differentiating healthy and diseased tissues in both plant species. GPC-BAT selected distinct wavelengths for each plant species, yet with overlaps in critical regions, indicating spectral sensitivity to disease presence.

The chosen wavelengths align with the absorption regions of various photosynthetic pigments and plant structural components. These spectral regions affected by bacterial infections align with alterations in chlorophylls, carotenoids, flavonoids, pheophytins, and interactions with water content and structural components in plant leaves.

The comparative analysis of VIs and GPC-BAT highlights their efficacy in diagnosing plant bacterial diseases through hyperspectral sensing. While VIs provides a simplistic yet moderately effective means of disease diagnosis, GPC-BAT, after feature

reduction, showcases improved accuracy. However, further refinements are necessary to enhance the identification of specific bacterial infections, especially in the early stages.

The identified wavelengths hold biological significance, suggesting a correlation between bacterial infections and alterations in photosynthetic pigments and leaf structural components. Future research could focus on refining and integrating these approaches to develop more robust and accurate diagnostic tools for various plant-pathogen interactions, thereby aiding in early disease detection and management in diverse agricultural settings.

**Acknowledgments**

Mafalda Reis-Pereira and Renan Tosin were supported by fellowships from Fundação para a Ciência e a Tecnologia (FCT) [grant references SFRH/BD/146564/2019, and SFRH/BD/145182/2019, respectively]. This work is partially financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots - OmicBots: High-Throughput Integrative Omic-Robots Platform for a Next Generation Physiology-based Precision Viticulture, with reference PTDC/ASP-HOR/1338/2021. Jochem Verrelst was funded by the European Research Council (ERC) under the project FLEXINEL (\#101086622). The views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

**Authors contributions**

Conceptualization, M.R.P., J.V., R.T., J.C., F.T., F.S., and M.C.; Methodology, M.R.P., J.V., R.T., J.C., and M.C.; Software, J.V., and J.C.; Validation, M.R.P., J.V., J.C., and M.C.; Formal Analysis, M.R.P, J.V., and J.C.; Investigation, M.R.P., J.V., R.T., J.C., F.T., F.S., and M.C.; Resources, J.V., J.C., F.T., F.S., and M.C.; Data Curation, M.R.P.; Writing – Original Draft Preparation, M.R.P., J.V., F.S., and M.C.; Writing – Review & Editing, M.R.P., J.V., R.T., J.C., F.T., F.S., and M.C.; Visualization, M.R.P., J.V., and J.C.; Supervision, F.T., F.S., and M.C.; Project Administration, F.T., F.S., and M.C.; Funding Acquisition, F.T., F.S., and M.C.

**Data availability Statement**

The data presented in this study are available in Zenodo at:

Reis Pereira, M., Tavares, F., Santos, F., & Cunha, M. (2024). Hyperspectral spectroscopic transmittance data collected in-vivo healthy and diseased tomato leaflets

in controlled conditions - dataset II [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10498473

Reis Pereira, M., Tavares, F., Santos, F., & Cunha, M. (2024). Hyperspectral spectroscopic reflectance data collected in-vivo non-symptomatic and symptomatic kiwi leaves in field conditions [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10498541

## Supplementary Material

**Table S1** Cross-validation statistics of the Flexible Discriminant Analysis (FDA) for the training hyperspectral data set collected on tomato leaflet tissues (transmittance) and kiwi leaf tissues (reflectance).

| *Class* | *Accuracy (%)* | *Kappa* | *Precision (%)* | *Sensitivity (%)* | *Specificity (%)* | *F1 (%)* |
|---|---|---|---|---|---|---|
| *Tomato (1643 observations)* | | | | | | |
| Con | 63.60 | 45.22 | 52.93 | 73.37 | 77.23 | 67.75 |
| Pst | | | 77.53 | 67.78 | 91.18 | 72.33 |
| Xeu | | | 52.89 | 50.09 | 76.49 | 51.45 |
| *Kiwi (354 observations)* | | | | | | |
| NS/S | 74.58 | 48.16 | 72.79 | 68.15 | 79.70 | 70.39 |

Con – Control (healthy), Pst – Inoculated with *Pseudomonas syringae* pv. *tomato*, Xeu – Inoculated with *Xanthomonas euvesicatoria*, NS – Non-Symptomatic, S – Symptomatic.

# Case Study 3

**Reis-Pereira, M.**; Tosin, R.; Martins, R.; Neves dos Santos, F.; Tavares, F.; Cunha, M. Kiwi plant canker diagnosis using hyperspectral signal processing and Machine Learning: detecting symptoms caused by *Pseudomonas syringae* pv. *actinidiae*. Plants 2022, 11, 2154. https://doi.org/10.3390/plants11162154

Paper published on 19th August 2022

Classification according to journal: Article

Special Issue: Detection and Diagnostics of Bacterial Plant Pathogens

Bibliometric indicators from the Journal Citation Report, Institute for Scientific Investigation

• Journal impact factor: 4.5

• Journal rank

*Plant Sciences: position 43/238; quartile Q1, JIF percentile 82.1*

# Kiwi plant canker diagnosis using hyperspectral signal processing and Machine Learning: detecting symptoms caused by *Pseudomonas syringae* pv. *actinidiae*

Mafalda Reis Pereira[1,2], Renan Tosin[1,2], Rui Martins[1,2], Filipe Neves dos Santos[2], Fernando Tavares[1,3,4], and Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, 4169-007 Porto, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Roberto Frias, 4200-465 Porto, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

* Correspondence: Mário Cunha mccunha@fc.up.pt

**Abstract**

*Pseudomonas syringae* pv. *actinidiae* (Psa) has been responsible for numerous epidemics of bacterial canker of kiwi (BCK), resulting in high losses in kiwi production worldwide. Current diagnostic approaches for this disease usually depend on visible signs of the infection (disease symptoms) to be present. Since these symptoms frequently manifest themselves in the middle to late stages of the infection process, the effectiveness of phytosanitary measures can be compromised. Hyperspectral spectroscopy has the potential to be an effective, non-invasive, rapid, cost-effective, high-throughput approach for improving BCK diagnostics. This study aimed to investigate the potential of hyperspectral UV–VIS reflectance for in-situ, non-destructive discrimination of bacte-rial canker on kiwi leaves. Spectral reflectance (325–1075 nm) of twenty plants were obtained with a handheld spectroradiometer in two commercial kiwi orchards located in Portugal, for 15 weeks, totaling 504 spectral measurements. Several modeling approaches based on continuous hyper-spectral data or specific wavelengths, chosen by different feature selection algorithms, were tested to discriminate BCK on leaves. Spectral separability of asymptomatic and symptomatic leaves was observed in all multi-variate and machine learning models, including the FDA, GLM, PLS, and SVM

methods. The combination of a stepwise forward variable selection approach using a support vector machine algorithm with a radial kernel and class weights was selected as the final model. Its overall accuracy was 85%, with a 0.70 kappa score and 0.84 F-measure. These results were coherent with leaves classified as asymptomatic or symptomatic by visual inspection. Overall, the findings herein reported support the implementation of spectral point measurements acquired in situ for crop disease diagnosis.

**Keywords**

Actinidia; Leaf bacterial canker; *Pseudomonas syringae*; Plant pathology; In-situ diagnosis; Hyperspectral spectroscopy; Feature selection; Support Vector Machine

## 1. Introduction

Bacterial canker of kiwi (BCK) is an emerging disease caused by the Gram-negative bacteria *Pseudomonas syringae* pv. *actinidiae* (Psa), which are responsible for several epidemics and important losses in kiwi production worldwide (Balestra, Mazzaglia et al. 2009, Scortichini, Marcelletti et al. 2012, Vanneste 2013, Kim, Kim et al. 2016). In the early stages of the disease, the Psa pathogen colonizes the surface of the host plant without causing significant lesions, but after systemic invasion, may cause severe damage and even death (Donati, Cellini et al. 2018, Saavedra, Abud et al. 2018, Donati, Cellini et al. 2020). Therefore, the early stage of Psa infection may pass unnoticed as the plant has no macroscopic manifestations of the disease (symptoms), jeopardizing the efficiency of phytosanitary procedures to contain the disease (Lowe, Harrison et al. 2017). In turn, advanced stages of the infection are more easily detectable since they present characteristic symptoms, consisting of brown leaf spots with chlorotic yellow haloes (Figure 1), necrotic discoloration of buds, cankers with exudate on trunks and twigs, and collapsed fruits (Balestra, Mazzaglia et al. 2009). This symptomatologic manifestation reveals that there is a microbial load that has probably already spread to other plants, making it difficult to implement control measures. Thus, it is crucial to develop an early and rapid in situ diagnostic tool for controlling the spread of Psa, through frequent and inexpensive monitoring.

Current diagnostic procedures usually focus on scouting and laboratory-based techniques. The first consists of the inspection of fields (generally visual) by specialized trained observers, to detect and identify infected plants based on the presence of disease symptoms (Parker, Shaw et al. 1995). It is subjective, error-prone (since symptoms alone are not entirely disease-specific), labor-intensive, time-consuming, and expensive (Sankaran, Mishra et al. 2010, Mahlein 2016, Khaled, Abd Aziz et al. 2018, Ali, Bachik

et al. 2019). Laboratory-based methods, in turn, include serological and molecular tests and are generally applied due to their sensitivity, accuracy, and effectiveness. The most common laboratory methods include the enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction (PCR). They entail detailed sampling procedures, which require several hours to be completed, and involve disruptive sample preparation, not allowing a follow-up of the disease progression nor its field mapping to support precision agriculture systems (e.g., site-specific management) (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015). Since these laboratory methods were designed to confirm the presence of pathogens, they do not have the necessary high throughput and speed required for supporting real-time agronomic decisions in field extensions. Moreover, they still present some diagnostic limitations, mainly in the asymptomatic and early stages of the disease infection process, due to the uneven spread of pathogens inside plants (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015).

(A) (B)

**Figure 1** (A) Median of the spectra of the 25% observations best classified as ‘asymptomatic’ (green) and ‘symptomatic’ (red) for the selected model, combining the SFVS with SVM with radial kernel and class weights (stepsvmrw); (B) Variance of the reflectance data measured by spectral wavelength and class (green line representing the variance in the mean spectra of ‘asymptomatic’ samples, and red line illustrating the variance in the mean data of ‘symptomatic’ leaves).

Innovative plant disease diagnostic tools are expected to provide additional information, namely related to plant–pathogen interactions and resulting changes in the host’s biochemical and biophysical behavior, to that currently generated by the conventional methods mentioned above and should be combined with them. Furthermore, these new techniques, namely spectroscopic approaches, must allow a

faster and earlier diagnosis of the disease, and ultimately its field mapping, contributing to more precise agricultural practices. Phytosanitary products can, thus, be applied in the exact area, moment, and dose as required, resulting in a reduction in chemical usage, and consequently in fewer expenses for the producer, residues in crop production, and environmental contamination (Zhang, Yang et al. 2020).

Hyperspectral spectroscopy (HS) is a non-invasive and high-throughput technology for measuring early indicators of BCK (Golhani, Balasundram et al. 2018). HS has been successfully applied in the assessment of a wide variety of plant structural, chemical, biophysical, and metabolic traits in living tissues (Thenkabail, Smith et al. 2000, Delalieux, van Aardt et al. 2007, Blackburn and Ferwerda 2008, Monteiro-Silva, Jorge et al. 2019, Martins, Barroso et al. 2022). HS also performed well in the detection of pests (Herrmann, Berenstein et al. 2017, Zhang, Wang et al. 2017) and phytopathogenic fungi (Yu, Anderegg et al. 2018, Skoneczny, Kubiak et al. 2020), bacteria (Bagheri, Mohamadi-Monavar et al. 2018), and viruses (Morellos, Tziotzios et al. 2020) affecting different crops, even at asymptomatic stages (Gold, Townsend et al. 2020). Through spectral measurements in the visible (VIS, 400–700 nm), and infrared (IR, 800–2500 nm), HS captures quantitative and qualitative changes in the optical properties of plant tissue, which derive from modifications in pigments, sugars, and water levels (among other constituents) (Curran 1989, Thenkabail, Gumma et al. 2014, Tosin, Pocas et al. 2021, Tosin, Martins et al. 2022). In a simplified way, plants' spectral behavior in VIS wavelengths is mainly related to pigment concentration and physiological processes (such as photosynthesis). In turn, in the IR region it is mainly correlated with leaf water levels, chemical composition (namely lignin and protein content), structure, and internal scattering processes (Hunt and Rock 1989, Jones and Vaughan 2010). This information is super-imposed in the recorded spectra at different scales of interference (Martins 2019, Martins, Barroso et al. 2022). Thus, the detection of BCK using spectral information can be based on the existence of a particular sequence of both metabolic and structural changes, promoted by host–pathogen interactions, which result in the development of characteristic symptoms and, consequently, in modifications in plants' spectral behavior in VIS–NIR.

HS data may contain a large amount of redundant information from adjacent bands, and only a few wavelength features might be interesting in classifying a diseased plant (Blackburn 2007, Caicedo, Verrelst et al. 2014, Rivera, Verrelst et al. 2014). Appropriate strategies usually involving statistical signal-processing approaches, mathematical combinations of different spectral bands, and predictive modeling techniques that can be applied to analyze spectral data and extract useful information

and contribute to dimensionality reduction and wavelength selection (Mahlein, Steiner et al. 2010, Mahlein, Rumpf et al. 2013, Thenkabail, Gumma et al. 2014, Ahmadi, Muharam et al. 2017, Thenkabail, Lyon et al. 2018, Saleem, Potgieter et al. 2019, Zhao, Fang et al. 2020). Machine learning (ML) algorithms have also been applied to handle the high dimensionality of hyperspectral information (Saha and Manickavasagan 2021). Several modeling approaches have been computed in previous studies to identify and classify plant stress and diseases from spectral data, using either direct spectral reflectance data or information with reduced dimensionality/features selected (Sankaran, Ehsani et al. 2012, Bajwa, Rupe et al. 2017, Gold, Townsend et al. 2020, Meng, Lv et al. 2020). The present research aims to explore the suitability and discrimination capability of different multi-variate and machine learning methods in the distinction of asymptomatic and symptomatic kiwi leaves affected by bacterial canker disease, using in-situ, ground-level, UV–VIS hyperspectral measurements. Modeling approaches evaluated the performance of the flexible discriminant analysis (FDA), general linear model (GLM), partial least squares (PLS) classification, and support vector machines (SVM, with different kernels and class weights) algorithms. The data gathered and the proposed workflow are expected to be a robust contribution to extend the HS approaches to plant disease diagnostics in field settings.

## 2. Materials and methods

### 2.1. Study area

The monitoring of kiwi plants (*Actinidia deliciosa*) was performed in two test sites, integrated in commercial orchards at Guimarães, Portugal, located in Caldas das Taipas (CT; 41°29′09.8″ N 8°21′54.3″ W) and Briteiros (BT; 41°30′53.3″ N 8°19′20.5″ W). In CT, where the orchard was 5 years old when the assay was performed (2020), twelve feminine kiwi plants of the variety Bo.Erika® were selected, marked with tape, and divided according to the presence or absence of visual symptoms characteristic of BCK (small greasy dark spots that become brown to black, that are distributed randomly on leaves, Figure 1). The same procedure was performed for the BT test site, whose orchard was 30-years-old, where eight plants of the same variety were selected to integrate the study.

Disease identification was accomplished by a visual assessment of BCK characteristic symptoms on the kiwi leaf's adaxial and abaxial sides (Figure 1). Samples were classified as asymptomatic (showing no BCK symptoms) or symptomatic (presenting at least one typical BCK chlorotic or necrotic spot). The monitoring of these two sites allowed the evaluation of the impact of different environmental and meso- and

microclimatic conditions, as well as the influence of different agricultural practices and plant age.

### 2.2. Spectral reflectance acquisition through ground measurements

Leaf hyperspectral data were obtained with a portable spectroradiometer (ASD FieldSpec® HandHeld 2, ASD Instruments, Boulder, CO, USA). Reflectance data were recorded in the wavelength range from 325 nm to 1075 nm, with 1 nm of spectral resolution. The spectroradiometer has a full conical field-of-view angle of 25°. During the data acquisition, the sensor was maintained 30 cm above the kiwi leaf, directed vertically downward (nadir view), giving a sampling footprint close to 13.3 cm. The leaf was placed upon a black card to reduce background noise. Prior to the hyperspectral acquisition, an internal dark calibration was performed, followed by a white calibration through a spectralon (white reference panel).

Measurements were acquired in the nadir position, in cloud-free conditions, between 11:00 and 14:00 h (local time), minimizing changes in the solar zenith angle. Weekly hyperspectral data on plant's reflectance were obtained between May and June 2020, which corresponded to the full development of Psa symptoms in kiwi plant leaves during the growing season. After, biweekly measurements were performed between July and August 2020. Three random leaves were chosen for each plant, and hyperspectral information was collected from one point, totaling 504 measurement points (Table 1). In each spectral measurement, 10 repetitions were performed and later averaged to minimize the noise effect.

The measurements were balanced regarding the test site and symptomatology (asymptomatic or symptomatic). Nearly 43% of the samples were collected in the BT region, presenting 59% of the typical symptoms of BCK. The remaining 57% of observations were collected in the CT region, where only 33% of them showed visual signs of the disease. In fact, differences in disease intensity were observed, with the BT test site being more severely affected by BCK than CT.

A multiplicative scatter correction log (MSC log) was applied in the hyperspectral reflectance according to (Martins, Barroso et al. 2022).

### 2.3. Modeling approaches

#### 2.3.1. Feature selection

Hyperspectral data are superimposed and result from multi-scale interference, resulting in an auto-correlated signal at various scales (Mariotto, Thenkabail et al. 2013,

Martins 2019, Martins, Barroso et al. 2022). The state-of-the-art enumerates several techniques useful for reducing the impacts of this high dimensional, redundant information (Thenkabail, Gumma et al. 2014). One approach consists of feature selection techniques applied to identify the most relevant bands and/or range of bands within hyperspectral data associated with the explaining variable. By directly choosing wavelengths, redundant information is removed, retaining only the more relevant discrimination features. If the removal of wavelengths is distributed, information is maintained with minimal loss since the spectrum is auto-correlated (Martins 2019, Martins, Barroso et al. 2022). In our study, the performance of different modeling approaches in BCK discrimination was assessed when (Figure 2): (i) all the 751 wavelengths predictors were considered (325–1075 nm), (ii) when built-in features selection models were computed, (iii) and, when different wavelength selection methods were applied, namely a sequential forward floating selection using Jeffries–Matusita distance, a stepwise forward variable selection method using Wilk's Lambda criterion, and a Lasso regularized generalized linear model. The main goal of feature selection was to capture systematic information, ensuring that the model description of data was optimal without under or overfitting.

*Sequential Forward Floating Selection Search Strategy and the Jeffries–Matusita (SFFS + JM) Distance*

A feature selection using the sequential forward floating selection search strategy and the Jeffries–Matusita (SFFS + JM) distance (Pudil, Novovičová et al. 1994) was computed to assess the spectral separability between the distributions of asymptomatic and symptomatic samples. This approach is an extension of the sequential forward selection algorithm. It comprehends a backward step that allows the variables included in the prior steps to be reconsidered, increasing the number of possible combinations evaluated. The Jeffries–Matusita (JM) distance was selected as a separability metric, whose value ranges from zero to two, with values above 1.9 being considered indicators of clear separability (Richards and Richards 1999). The JM distance among the distributions of the two classes $\omega_i$ and $\omega_j$ can be calculated by Equation (1) (Dalponte, Bruzzone et al. 2012):

$$JM_{ij} = \int_x \left[\sqrt{p_i(x|\omega_i)} - \sqrt{p_j(x|\omega_j)}\right]^2 dx \quad (1)$$

where p ($x/\omega_i$) and p ($x/\omega_j$) are the conditional probability density functions for the feature vector $x$, given the data classes $\omega_i$ and $\omega_j$, respectively. It can be rewritten according to the Bhattacharyya distance ($B_{ij}$):

$$JM_{ij} = \sqrt{2\left(1 - e^{-B_{ij}}\right)} \quad (2)$$

In hyperspectral remote sensing data, class distributions are often modeled as Gaussian distributions (Dalponte, Bruzzone et al. 2012). Under this hypothesis, the Bhattacharya distance can be mathematically written as Equation (3):

$$B_{ij} = \frac{1}{8}(\mu i - \mu j)^T \left(\frac{\sum i + \sum j}{2}\right)^{-1} (\mu i - \mu j) + \frac{1}{2} ln \left[\frac{1}{2}\frac{|\sum i + \sum j|}{\sqrt{|\sum i||\sum j|}}\right] (3)$$

where $\mu$i and $\mu$j represent the vector means of classes i and j, respectively, and ∑i and ∑j are the covariance matrices of the same classes.

JM distance was selected since it is an efficient method for class separation distances. The JM performs good feature ranking for two-class comparisons (Laliberte, Browning et al. 2012), and shows a saturated performance when the separability between the measured classes increases. When the saturation point is achieved, any further feature provided does not increase the separability (Dalponte, Bruzzone et al. 2012).

*Stepwise Forward Variable Selection Method Using Wilk's Lambda Criterion (SFVS)*

A stepwise forward variable selection (SFVS) approach was performed for feature selection within the initial 751 predictor candidates. This procedure is based on determining the predictive variables that most contribute to the model improvement in each step, compared to the model in the previous step. The choice is based on Wilk's Lambda criterion. This statistic measures distance based on scalar transformations of the covariance matrixes between and within groups (El Ouardighi, El Akadi et al. 2007).

*Lasso Regularized Generalized Linear Models (LASSO)*

Lasso regularized generalized linear models (LASSO) was also computed since this is considered an efficient procedure for fitting the entire Lasso regularization path for linear regression models via penalized maximum likelihood (Hastie and Qian 2014, Friedman, Hastie et al. 2021).

Computing models with built-in feature selection were also tested to compare their performance with the algorithms where the search routine for the right predictors is external to the model. These models generally work by pairing the predictor search algorithm with the parameter estimation and are usually optimized with a single objective function (e.g., error rates or likelihood) (Kuhn 2015). Generalized linear model with

stepwise feature selection (glmStepAIC) and the flexible discriminant analysis (FDA) were chosen to integrate this study.

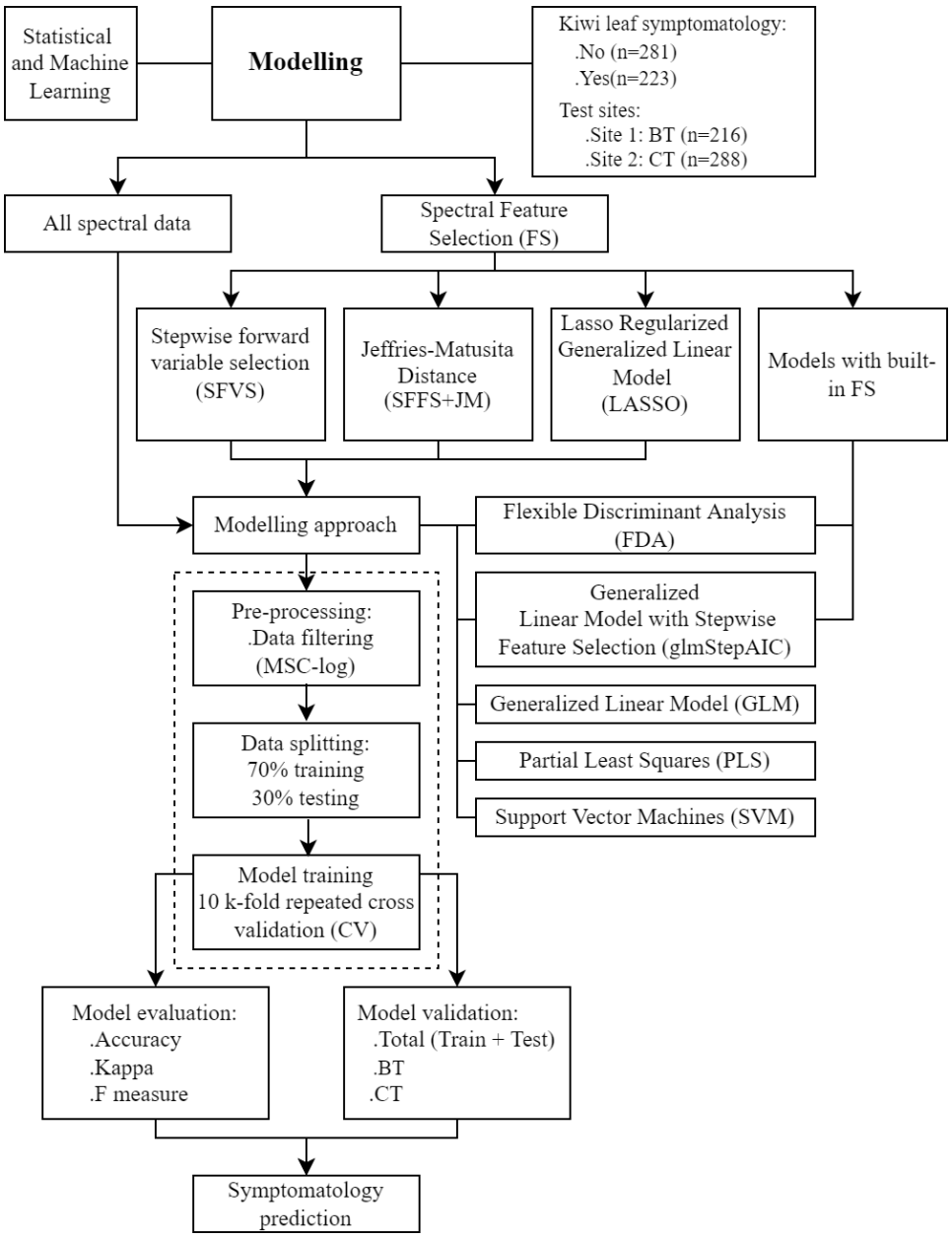



**Figure 2** Conceptual diagram for the predictive modeling approaches of bacterial canker of kiwi (BCK).

### 2.3.2. Predictive modeling in classification mode

Seven predictive modeling approaches were evaluated to detect the bacterial canker of kiwi disease (Figure 2). The leaf symptomatology was used as a binary variable in the models tested taking the values ‘No’ (asymptomatic) and ‘Yes’ (symptomatic). The algorithms computed included (i) flexible discriminant analysis (FDA); (ii) general linear model (GLM); (iii) partial least squares (PLS) classification; (iv) support vector machines

with linear kernel (SVM-L); (v) support vector machines with radial basis function kernel (SVM-R); (vi) linear support vector machines with class weights (SVM-LW); and (vii) radial support vector machines with class weights (SVM-RW).

*Flexible Discriminant Analysis (FDA)*

The FDA was selected since it is a multigroup nonlinear discrimination/classification and pattern-recognition method based on nonparametric regression followed by linear discriminant analysis (LDA). It uses optimal scoring to convert the response variable so that the data are better for linear separation, and multiple adaptive regression projections to generate the discriminant surface. FDA can be applied with standard linear regression, resulting in Fisher's discriminant vectors (Hastie, Tibshirani et al. 1994, Hastie, Tibshirani et al. 2009).

*Generalized Linear Model (GLM)*

GLM was chosen as a parametric, statistical approach that consists of an extension of linear models. GLM establishes the relationships between the explanatory factors and the responses through an estimated regression parameter via confidence intervals [78]. It evaluates the temporal variational pattern of signals instead of their absolute magnitude, being robust in many cases, including severe optical signal attenuations due to scattering or poor contact (Ye, Tak et al. 2009).

*Partial Least Squares (PLS) Classification*

PLS was computed as a multivariate statistic since it proved that PLS is a prominent modeling method capable of dealing with several, multicollinear variables, and in cases where the number of explanatory (number of wavelengths) variables is superior to the number of observations (Wold, Sjöström et al. 2001). It aims to minimize the sample prediction error, pursuing linear functions of the predictors that explain as much variation in each response as possible. Also, PLS aims to account for variation in the predictors, under the hypothesis that directions in the predictor space, which are well sampled, should offer an improved prediction for new observations when the predictors are highly correlated (Liu, Huang et al. 2007).

*Support Vector Machines (SVM)*

SVMs were used as a set of machine learning methods built on the concept of optimal separating hyperplane (Vapnik 1999), and they can be used for regression and classification tasks (Mosavi, Sajedi Hosseini et al. 2021). They are non-linear classifiers capable of finding the most extensive margin between two classes in feature space [84].

SVMs have several hyperparameters and different kernel types. The SVM methodology intends to reduce the error test and model complexity (Ballabio and Sterlacchini 2012). The kernel function transforms raw data inputs from the original user space into kernel space through a user-defined feature map. The kernel functions include linear, polynomial, and radial basis functions (RBF) (Patle and Chouhan 2013, Ding, Liu et al. 2021). Some SVMs approaches assign different weights to different data points such that SVM learns the decision surface according to the relative importance of the data points in the training set (Xulei, Qing et al. 2005).

*Model Development and Selection*

Symptomatology was then used as the response variable in modeling approaches, and the 751 wavelengths were considered predictor candidates. To run the predictive models, the dataset was divided into training data (70% of random observations) and validation data (30% of the remaining observations) (Kuhn and Johnson 2013), following a holdout method (Lantz 2019). The training and validation datasets integrate the pairs of concurrent measurements of the symptomatology and the corresponding values of the predicting variables (Figure 2).

For model evaluation criteria, a resampling strategy was considered following a repeated cross-validation strategy using a repeated 10-fold cross-validation to estimate accuracy. The dataset was split into 10 parts, trained in 9, and tested on 1. The process was repeated for all combinations of train–test splits. The final model accuracy was then taken as the mean from the number of repeats (Kuhn and Johnson 2013, Lantz 2019). This strategy allows the execution of verification steps by the model before the final verification is measured on the testing set, decreasing the possibility of overfitting (Berrar 2019, Valier 2020).

Different metrics were then considered to assess model performance and model selection, namely the confusion matrix (CM), accuracy score, kappa coefficient, and the F1-score (Figure 2).

The CM presented possible categories of predicted values in one dimension and the possible categories for actual values in the other. Correct classifications (when the predicted value was equal to the actual value) felt on the diagonal in the CM. The off-diagonal matrix cells corresponded to the incorrect predictions, where the predicted value diverges from the actual value. The class of interest was positive, while the other was identified as negative. The prediction was then classified as a true positive (TP) when it was correctly classified as the class of interest; true negative (TN) when it was properly categorized as not the class of interest; false positive (FP) when it was

incorrectly considered as the class of interest; and, false negative (FN) when it was mistakenly labeled as not the class of interest.

The accuracy can be considered as the number of correctly classified prediction instances divided by the total number of predictions. The accuracy (also known as success rate) can be calculated through the proportion of TP and TN in all evaluated cases with the confusion matrix results. Mathematically, this can be stated as presented in Equation (4) (Lantz 2019):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

The kappa statistic, or Cohen's kappa, corrects the accuracy by accounting for the possibility of an accurate prediction by chance alone (Lantz 2019). Its value can vary from zero to one. The interpretation of the kappa statistic may be different according to how a model is to be implemented. The value one indicates a perfect agreement between the model's predictions and the true values, and values lower than one indicate an imperfect agreement. Usually, kappa results can be interpreted as followed: less than 0.20—poor agreement; 0.20 to 0.40—fair agreement; 0.40 to 0.60—moderate agreement; 0.60 to 0.80—good agreement; and 0.80 to 1.00—very good agreement (Lantz 2019). The Kappa statistic can be calculated through the following formula, Equation (5):

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (5)$$

where Pr(a) represents the proportion of actual agreement and Pr(e) refers to the expected agreement between the classifier and the true values, under the hypothesis that they were chosen randomly.

F-measure (F1 score or F-score) was also used as an indicator of model performance that merged precision (proportion of positive cases that are truly positive) and recall (a measure of how complete the results are, which is computed as the number of TP over the total number of positives) into a single number using the harmonic mean, a type of average that is applied for levels of change, as represented mathematically by the formula in Equation (6):

$$\text{F} - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (6)$$

These metric scores were applied to the between model selection through a prediction process using the (i) total dataset (including training and test set), and (ii) site-

independent datasets (BT and CT observations). Between model selection was ultimately achieved through the evaluation of the mean and the coefficient of variation (CV) values for the different model metrics of the global (training and testing data), BT, and CT sets, being selected the model with an overall higher means and lower CV for the accuracy, kappa, and F-measure metrics.

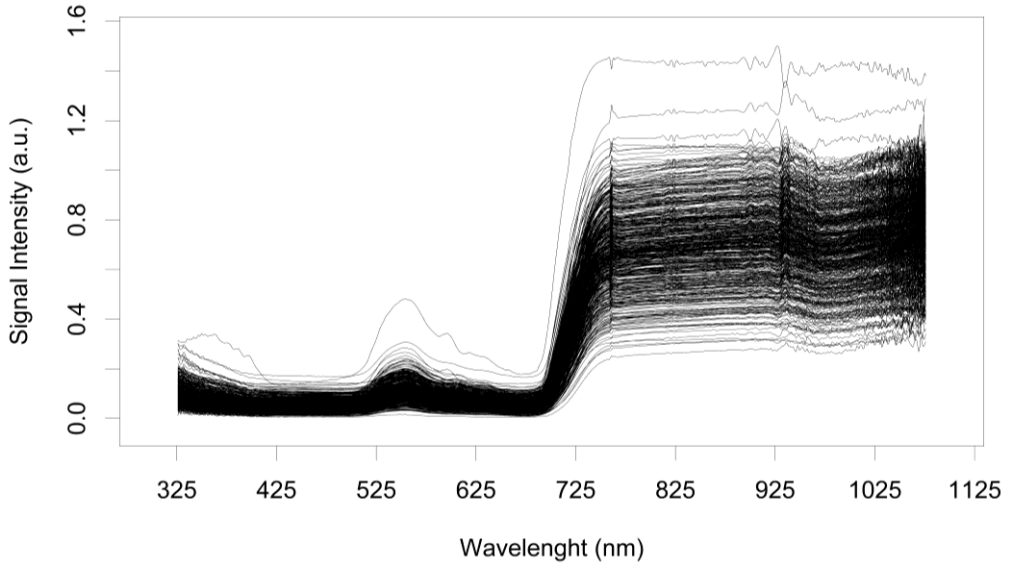
For the best model, the percentage of correct predictions was determined by dividing the number of cases where the model attributes the correct class to the prediction compared to the actual class through the total number of predictions performed. Also, the median of the spectra of the 25% predictions classified with higher probability as 'asymptomatic' and 'symptomatic' by the best model was computed.

All the computational analyses were performed in the software R (Team 2021) with the following packages: 'AppliedPredictiveModeling' (Kuhn, Johnson et al. 2013), 'caret' (Kuhn 2015), 'e1071′ (Meyer, Dimitriadou et al. 2019), 'earth' (Milborrow 2019), 'ggplot2' (Villanueva and Chen 2019), 'glmnet' (Friedman, Hastie et al. 2021), 'kernlab' (Karatzoglou, Smola et al. 2019), 'klaR' (Roever, Raabe et al. 2020), 'MASS' (Ripley, Venables et al. 2013), and 'mda' (Hastie 2020).

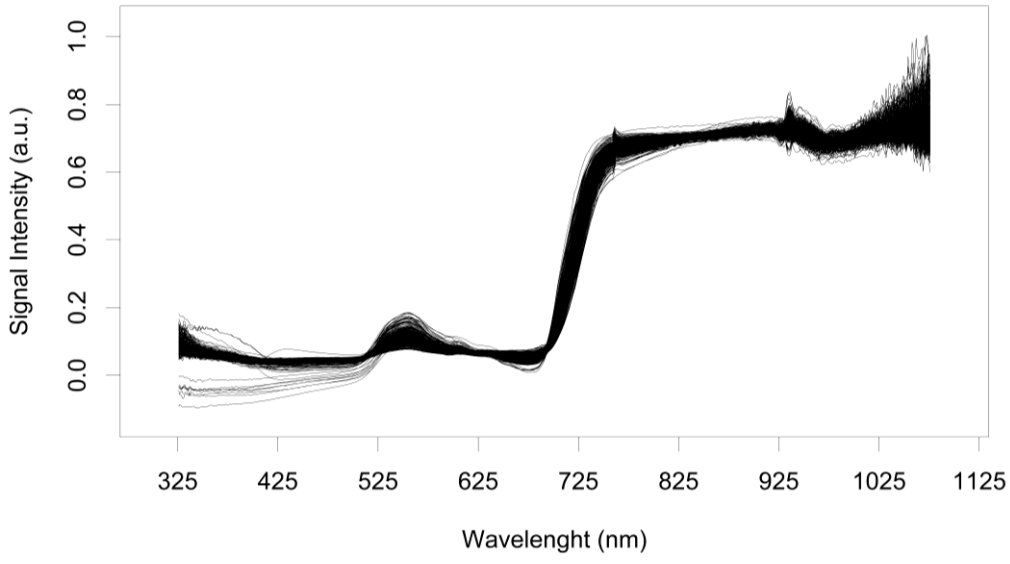
## 2. Results

### 2.1. Spectra filtering and feature selection

After data scatter correction using the MSC log algorithm (Figure 3), an SFFS + JM strategy was computed to assess separability between asymptomatic and symptomatic leaves as a function of the wavelength variables. From a total of 751 predictors in the VIS–NIR spectral region, the procedure selected 33 variables (Table 1) essentially involving wavelengths located in the blue (326–408 nm), green (562, 583 nm), and NIR (777–1068 nm) regions. The JM value was 1.41 indicating high separability between variables. An SFVS approach was also performed for feature choice within the initial 751 predictor candidates. The 35 wavelengths chosen are described in Table 2, including features belonging to the blue (388–446 nm), green (510–556), red (671–754 nm), and NIR (759–1070 nm) regions.




(A)




(B)

**Figure 3** Representation of the spectra collected (A), and after its filtering (B) using the MSC log algorithm.

With built-in feature selection, the FDA model only identified seven variables from the total predictors. They belonged to the blue region (424 and 464 nm), green (549 nm), red (719,753 nm), and NIR (759,935 nm) regions. In turn, GLM with the built-in stepwise feature selection sorted out 20 predictors, mainly localized in the blue (388–443 nm), green (510 nm), and NIR (759–1066 nm) regions.

The LASSO method recognized 22 predictors from the total 751 wavelengths available. These spectral features fitted the blue (329–375 nm), green (510, 536 nm), red (617, 671 nm), and NIR (771–1070 nm) regions.

All feature selection methodologies identified similar wavelengths and spectral bands important for discriminating BCK detection.

**Table 1** Selected discriminative wavelengths for model development.

| Method | Selected Discriminative Wavelengths (nm) |
|---|---|
| *SFFS + JM (n = 33)* | 326,327,329,330,335,336,352,359,360,364,365,408,562,583,762,777,778,779,786,828,897,908,923,995, 1018,1031,1038,1045,1057,1059,1061,1067,1068 |
| *SFVS (n = 35)* | 388,401,406,414,415,419,443,446,510,515,556,671,724,754,759,781,794,807,969,970,981,983,1009, 1027,1031,1032,1035,1045,1048,1049,1050,1053,1066,1068,1070 |
| *FDA (n = 7)* | 424, 464, 549, 716, 753,759, 935 |
| *glmStepAIC (n = 20)* | 388,414,415,419,443,510,759,794,970,981,982,1001,1031,1035,1045,1048,1049,1050,1053,1066 |
| *LASSO (n = 22)* | 329,369,375,510,531,536,617,671,771,772,778,903,932,959,969,970,1045,1048,1050,1052,1061,1070 |

SFFS + JM sequential forward floating selection using Jeffries–Matusita Distance; SFVS—Stepwise forward variable selection; glmStepAIC—Generalized linear model with stepwise feature selection; LASSO—Lasso regression (glmnet).

### 2.2. Model discrimination of Psa leaf symptom

Table 2 presents the metric values used to compare the model approaches computed to discriminate between asymptomatic and symptomatic kiwi leaves infected by the Psa pathogen, based on random sampling (with no temporal sequence correlated in the samples). Considering all of the available 751 predictors, the mean metrics of the three sets studied (total, BT, and CT data), including all the tested modeling approaches, presented mean values ranging from 0.71 to 0.82 for accuracy, 0.36 to 0.63 (fair to good agreement) for kappa, and 0.65 to 0.81 for the F-measure. In turn, CV ranged from 2.15 to 3.45, 2.62 to 10.16, and 4.57 to 15.18 for the same metrics.

**Table 2** Validation results for models classifying bacterial canker of kiwi (BCK) disease.

| Feature Selection | Model | Validation set | | | | | | Statistics of validation sets | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Total | | | BT | | | CT | | | Mean | | | CV | | |
| | | Acc | K | F1 | Acc | K | F1 | Acc | K | F1 | Acc | K | F1 | Acc | K | F1 |
| *None* | PLS | 0.7083 | 0.4047 | 0.6589 | 0.6806 | 0.3329 | 0.7356 | 0.7292 | 0.3536 | 0.5412 | 0.7060 | 0.3637 | 0.6452 | 3.4530 | 10.1605 | 15.1756 |
| N = 751 | SVM-L | 0.8274 | 0.6444 | 0.7883 | 0.8012 | 0.6154 | 0.8313 | 0.8403 | 0.6167 | 0.7262 | 0.8230 | 0.6255 | 0.7819 | 2.4209 | 2.6188 | 6.7574 |
| | SVM-LW | 0.8115 | 0.6274 | 0.8104 | 0.7917 | 0.5464 | 0.8421 | 0.8264 | 0.6324 | 0.7685 | 0.8099 | 0.6021 | 0.8070 | 2.1494 | 8.0180 | 4.5747 |
| | SVM-R | 0.7857 | 0.5628 | 0.7500 | 0.7593 | 0.5015 | 0.7969 | 0.8056 | 0.5435 | 0.6818 | 0.7835 | 0.5359 | 0.7429 | 2.9643 | 5.8482 | 7.7908 |
| *Built-in* | SVM-RW | 0.8056 | 0.6066 | 0.7822 | 0.7778 | 0.5367 | 0.8154 | 0.8264 | 0.6073 | 0.7368 | 0.8033 | 0.5835 | 0.7781 | 3.0356 | 6.9508 | 5.0708 |
| N = 7 | FDA | 0.7698 | 0.5339 | 0.7411 | 0.7546 | 0.4876 | 0.7969 | 0.7812 | 0.5013 | 0.6631 | 0.7685 | 0.5076 | 0.7337 | 1.7364 | 4.6856 | 9.1599 |
| N = 20 | glmStepAIC | 0.8147 | 0.6243 | 0.8342 | 0.7824 | 0.5471 | 0.7283 | 0.8392 | 0.6318 | 0.8814 | 0.8121 | 0.6011 | 0.8049 | 3.5081 | 7.8006 | 13.4507 |
| | Mean | *0.7890* | *0.5720* | *0.7552* | *0.7609* | *0.5034* | *0.8030* | *0.8015* | *0.5425* | *0.6863* | *0.7866* | *0.5456* | *0.7539* | *2.7431* | *5.9137* | *5.1895* |
| *SFVS* | GLM | 0.7937 | 0.5806 | 0.7636 | 0.7454 | 0.4754 | 0.7826 | 0.8299 | 0.6121 | 0.7380 | 0.7897 | 0.5560 | 0.7614 | 5.3686 | 12.8742 | 2.9395 |
| N = 35 | PLS | 0.7679 | 0.5249 | 0.7247 | 0.7685 | 0.527 | 0.7984 | 0.7674 | 0.4553 | 0.6215 | 0.7679 | 0.5024 | 0.7149 | 0.0717 | 8.1217 | 12.4302 |
| | SVM-L | 0.7619 | 0.5115 | 0.7143 | 0.7454 | 0.4942 | 0.7769 | 0.7708 | 0.4649 | 0.6292 | 0.7609 | 0.4902 | 0.7068 | 1.3715 | 4.8054 | 10.4888 |
| | SVM-R | 0.8512 | 0.6994 | 0.8344 | 0.8426 | 0.6773 | 0.864 | 0.8542 | 0.6667 | 0.7742 | 0.8485 | 0.6811 | 0.8242 | 0.8821 | 2.4494 | 5.5521 |
| | SVM-LW | 0.7897 | 0.583 | 0.7854 | 0.7778 | 0.5153 | 0.8322 | 0.8125 | 0.595 | 0.7404 | 0.7933 | 0.5644 | 0.7860 | 2.2226 | 7.6132 | 5.8401 |
| | **SVM-RW** | **0.8532** | **0.7035** | **0.8370** | **0.8472** | **0.6831** | **0.8716** | **0.8542** | **0.6753** | **0.7857** | **0.8515** | **0.6873** | **0.8314** | **0.4446** | **2.1187** | **5.1982** |
| | Mean | *0.8029* | *0.6004* | *0.7766* | *0.7882* | *0.5621* | *0.8210* | *0.8148* | *0.5782* | *0.7148* | *0.8020* | *0.5803* | *0.7708* | *1.6668* | *3.3257* | *6.9143* |
| *SFFS+JM* | GLM | 0.7202 | 0.4327 | 0.6831 | 0.7222 | 0.4109 | 0.7778 | 0.7500 | 0.4162 | 0.5955 | 0.7308 | 0.4199 | 0.6855 | 2.2794 | 2.7074 | 13.3009 |
| N = 33 | PLS | 0.7242 | 0.4355 | 0.6729 | 0.7407 | 0.4501 | 0.7926 | 0.7257 | 0.3209 | 0.4968 | 0.7302 | 0.4022 | 0.6541 | 1.2495 | 17.5938 | 22.7478 |
| | SVM-L | 0.7222 | 0.4253 | 0.6517 | 0.7593 | 0.4894 | 0.8074 | 0.7153 | 0.2849 | 0.4605 | 0.7323 | 0.3999 | 0.6399 | 3.2317 | 26.1576 | 27.1545 |
| | SVM-R | 0.7639 | 0.5117 | 0.7047 | 0.7639 | 0.5184 | 0.7935 | 0.8194 | 0.5618 | 0.6829 | 0.7824 | 0.5306 | 0.6270 | 4.0955 | 5.1256 | 31.9489 |
| | SVM-LW | 0.7381 | 0.4637 | 0.6887 | 0.7639 | 0.4984 | 0.8118 | 0.7188 | 0.2957 | 0.4706 | 0.7403 | 0.4193 | 0.6570 | 3.0567 | 25.8569 | 26.2985 |
| | SVM-RW | 0.8075 | 0.6057 | 0.7707 | 0.7824 | 0.5532 | 0.8127 | 0.8333 | 0.6022 | 0.7176 | 0.8077 | 0.5870 | 0.7670 | 3.1509 | 5.0002 | 6.2135 |
| | Mean | *0.7440* | *0.4747* | *0.6419* | *0.7460* | *0.4791* | *0.6453* | *0.7554* | *0.4867* | *0.7993* | *0.7539* | *0.4598* | *0.6718* | *0.9695* | *8.7409* | *17.3572* |
| *LASSO* | GLM | 0.7560 | 0.5056 | 0.7248 | 0.7176 | 0.4021 | 0.7732 | 0.7847 | 0.4973 | 0.6517 | 0.7528 | 0.4683 | 0.7166 | 4.4724 | 12.2796 | 8.5361 |
| N = 22 | PLS | 0.7560 | 0.5028 | 0.7172 | 0.7407 | 0.4501 | 0.7926 | 0.7674 | 0.437 | 0.5939 | 0.7547 | 0.4633 | 0.7012 | 1.7752 | 7.5177 | 14.3045 |
| | SVM-L | 0.7599 | 0.5127 | 0.7269 | 0.7361 | 0.4393 | 0.7897 | 0.7778 | 0.4725 | 0.6279 | 0.7579 | 0.4748 | 0.7148 | 2.7601 | 7.7407 | 11.4114 |
| | SVM-R | 0.8353 | 0.6654 | 0.8118 | 0.8009 | 0.5842 | 0.8352 | 0.8611 | 0.6774 | 0.7778 | 0.8324 | 0.6423 | 0.8083 | 3.6282 | 7.8933 | 3.5709 |
| | SVM-LW | 0.7639 | 0.523 | 0.7373 | 0.7269 | 0.4217 | 0.4807 | 0.7917 | 0.5213 | 0.6739 | 0.7608 | 0.4887 | 0.6306 | 4.2728 | 11.8692 | 21.1945 |
| | SVM-RW | 0.8373 | 0.6708 | 0.8178 | 0.8009 | 0.5828 | 0.8365 | 0.8646 | 0.6913 | 0.7914 | 0.8343 | 0.6483 | 0.8152 | 3.8307 | 8.8915 | 2.7795 |
| | Mean | *0.7847* | *0.5634* | *0.7560* | *0.7539* | *0.4800* | *0.7513* | *0.8079* | *0.5495* | *0.6861* | *0.7822* | *0.5310* | *0.7311* | *3.4659* | *8.4093* | *5.3430* |

CV—Coefficient of Variation; Acc—Accuracy; F1—F-measure; GLM—Generalized linear model; glmSte-pAIC—Generalized linear model with stepwise feature selection; FDA—Flexible discriminant analysis; K—Kappa; LASSO—Lasso regression (glmnet); PLS—Partial least squares; SFFS + JM—Sequential forward floating selection using Jeffries–Matusita distance; SFVS—Stepwise forward variable selection; SVM—Support vector machine (L—Linear kernel; LW—Linear kernel with class weights; R—Radial kernel; RW—Radial kernel with class weights).

Three independent feature selection methods were then applied and combined with the same models (except for FDA) to verify if selected wavelengths would improve model performance for the discrimination of Psa disease. For the SFVS approach, the

mean metric values of the three sets studied ranged from 0.76 to 0.85 for accuracy, 0.49 to 0.69 (moderate to good agreement) for kappa, and 0.71 to 0.83 for the F-measure. The CV scores ranged from 0.07 to 5.37 for accuracy, 2.12 to 12.87 for kappa, and 2.94 to 12.43 for the F-measure. For the SFFS + JM procedure, similar findings were observed, and the mean results covered the interval 0.73 to 0.81 for accuracy, 0.40 to 0.59 (moderate agreement) for kappa, and 0.63 to 0.77 for the F-measure. The CV numbers fluctuated from 0.97 to 4.10, 2.71 to 26.16, and 6.21 to 31.95 for accuracy, kappa, and the F-measure, respectively. These approaches, thus, generally showed higher relative dispersion of the data points in the datasets around the mean, for all the metrics. Lastly, for Lasso, the mean outcomes extended from 0.75 to 0.83, 0.46 to 0.65 (moderate to good agreement), and 0.63 to 0.82 for accuracy, kappa, and the F-measure, respectively. CV, for the same metrics, registered values of 1.78 to 4.48, 7.52 to 12.28, and 2.78 to 21.19.

Between models, the selection was achieved by determining the mean and the CV for the global (encompassing the training and testing data), BT, and CT datasets. The SFVS followed by an SVM algorithm with radial kernel and class weights (stepsvmrw) presented a higher mean (accuracy of 0.85, kappa of 0.69, and an F-measure of 0.83) and lower CV (0.45 for accuracy, 2.12 for kappa and 5.20 for the F-measure) for the different metrics. This model was, hence, selected.

Table 3 presents the confusion matrix for the selected model (stepsvmrw) for the three validation datasets. In the predictions using the total (training and validation set) data, the model correctly classified 190 (TP) spectra of the 223 spectra acquired over the symptomatic leaves (33 observations were wrongly classified — FN). The spectra acquired over the asymptomatic leaves allowed the correct classification of 240 (TN) of the 281 spectra (41 cases of FP) (Table 3).

Figure 4 presents the temporal prediction trend of correct classification as ‘asymptomatic’ in both test sites, based on the stepsvmrw model. According to dates and test sites, the percentage of cases where the stepsvmrw model attributed the correct classification as ‘asymptomatic’ to each observation ranged from 71% to 96% (Figure 4). The percentage of asymptomatic observations correctly classified decreased for the BT region over time but showed an inverse tendency for the CT site. The BT orchard presented more advanced symptoms of BCK and their growth was relatively stable throughout the measurement period. The lower values of correct asymptomatic class prediction of the last dates can be related to disease asymptomatic leaves showing a spectral signature more similar to symptomatic samples than healthy ones. In turn, for

the CT region, spectral measurements allowed complete surveillance from the appearance and development of the first signs of BCK to its full development throughout the time, coinciding with the visual separation between healthy and diseased leaves.

**Table 3** Confusion matrix for the selected model characterized by executing SFVS followed by an SVM algorithm with radial kernel and class weights (stepsvmrw) using the BT, CT, and complete dataset.

| | | **BT (*n* = 216)** | | | **CT (*n* = 288)** | | | **ALL (*n* = 504)** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Actual value | | | Actual value | | | Actual value | | |
| | | ‘No’ | ‘Yes’ | | ‘No’ | ‘Yes’ | | ‘No’ | ‘Yes’ |
| Predicted | ‘No’ | 71 | 15 | Predicted | ‘No’ | 169 | 19 | Predicted | ‘No’ | 240 | 33 |
| | ‘Yes’ | 18 | 112 | | ‘Yes’ | 23 | 77 | | ‘Yes’ | 41 | 190 |

‘No’ and ‘Yes’ correspond to asymptomatic and symptomatic leaves, respectively.

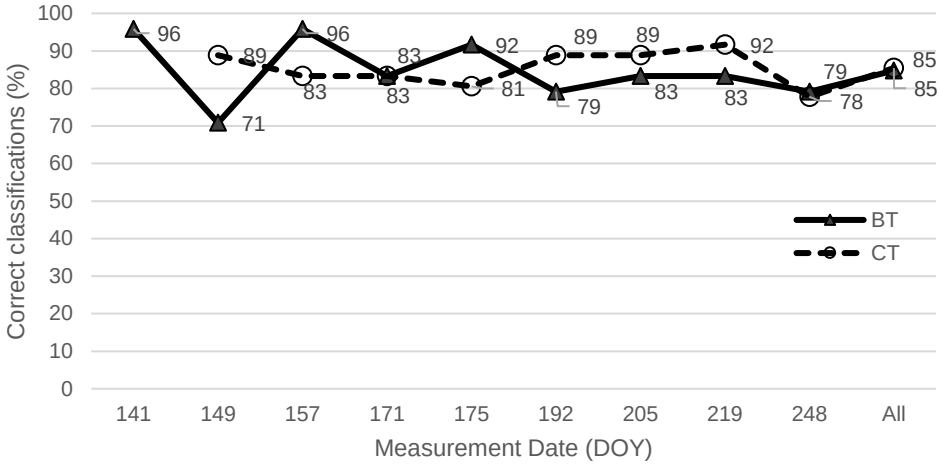
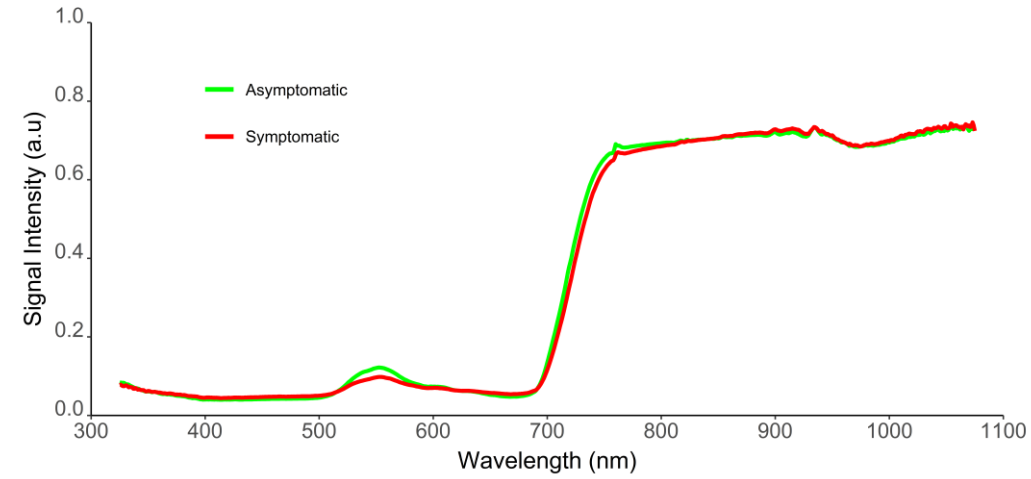


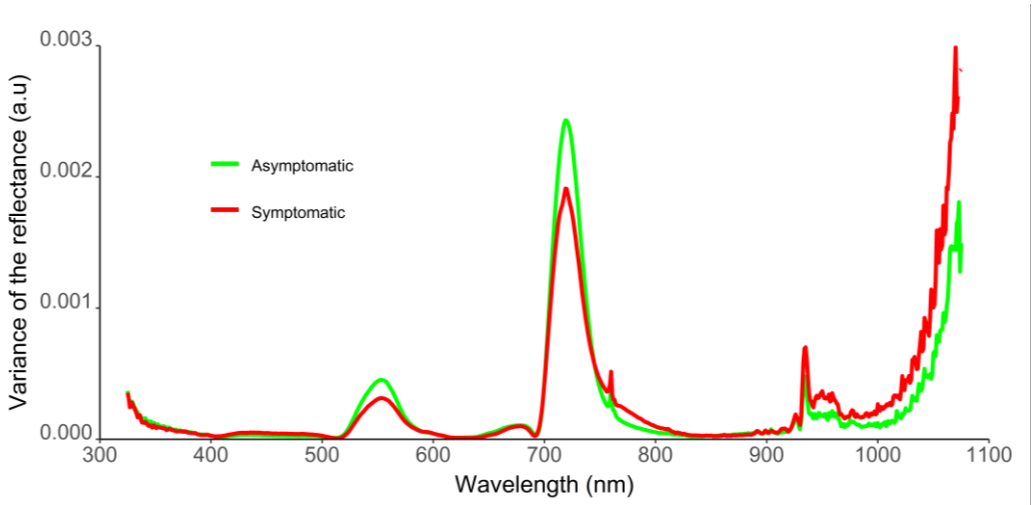

**Figure 4** Percentage of correct classification predictions as ‘asymptomatic’ by date and test site using the SFVS strategy, followed by an SVM algorithm with radial kernel and class weights (stepsvmrw model). Values of BT site are represented with triangles and CT with circles. DOY—Day of the year.

Figure 5A represents the median spectra of the 25% of observations classified with higher probability as ‘asymptomatic’ and ‘symptomatic’ by the predict function of the ‘caret’ package which was computed for the selected model. Reflectance curves of asymptomatic samples were characteristic of healthy green leaves, presenting lower reflectance values in the VIS spectral region, and a high reflectance level in the NIR region. In turn, symptomatic samples showed characteristic, divergent reflectance curves. Visual changes were observed between asymptomatic and symptomatic samples for wavelengths ranging from 515–650 nm (green–yellow–orange region), 651–714 nm (red region), and 715–850 nm (red-edge and NIR regions). Higher reflectance

values were observed for the blue region (450–520 nm) and most NIR regions (850–1075 nm) for symptomatic leaves compared to the asymptomatic ones. The opposite tendance was observed in the green, red-edge, and beginning of the NIR region (<850 nm). Nevertheless, spectral variance (Figure 5B) was reduced for wavelengths higher than 800 nm.




(A)




(B)

**Figure 5** (A) Median of the spectra of the 25% observations best classified as 'asymptomatic' (green) and 'symptomatic' (red) for the selected model combining the SFVS with SVM with radial kernel and class weights (stepsvmrw); (B) Variance of the reflectance data measured by spectral wavelength and class (green line representing the variance in the mean spectra of 'asymptomatic' samples, and red line illustrating the variance in the mean data of 'symptomatic' leaves).

## 3. Discussion

Proximal sensing techniques can be a useful tool for helping producers detect early crop diseases in situ. However, qualitative and/or quantitative differences between

the spectral information according to leaf symptomatology must be retrieved. In this regard, our study investigated the possibility of using different model approaches of hyperspectral data to correctly classify kiwi leaves according to the presence of characteristic symptoms of BCK disease. The analysis was performed in two kiwi orchards, where 504 spectral signatures were randomly acquired from symptomatic (diseased) and asymptomatic kiwi plant leaves over time (Table 4). Monitoring of these two kiwi orchards allowed the evaluation of the impact of different environmental and meso- and microclimatic conditions, and the influence of different agricultural practices and plant age on model development. A cross-validation strategy was applied to test the null hypothesis, which was assumed to occur when the training and validation sets are randomly sampled, resulting in similar predictions in both datasets. An n-series random sampling can, furthermore, be performed to assure a general evaluation of the error. Hence, cross-validation models can be derived from all datasets, taking the error of a predicted sample (Refaeilzadeh, Tang et al. 2009, Krstajic, Buturovic et al. 2014). Model transferability was later demonstrated by the results obtained in the modeling process.

**Table 4** Number of observations (leaves and plants) per test site and symptomatology.

| Test site | Sites | Dates | Plants | Asymptomatic Leaves | Symptomatic Leaves | Total Measurements |
|---|---|---|---|---|---|---|
| Briteiros (BT) | 1 | 9 | 8 | 89 | 127 | 216 |
| Caldas das Taipas (CT) | 1 | 8 | 12 | 192 | 96 | 288 |
| Total | 2 | 9 | 20 | 281 | 223 | 504 |

Hyperspectral data is acknowledged for containing many redundant adjacent features, prone to multicollinearity (Mariotto, Thenkabail et al. 2013), and suggested feature selection allows the identification of the most relevant information (Figure 2). Hyperspectral data may, in fact, hold limited useful information, reducing model performance due to overfitting, and increasing computational time (Morellos, Tziotzios et al. 2020). Thus, different feature selection techniques were applied to hyperspectral filtered data to identify relevant features having significance in the classification process, namely a sequential forward floating selection using Jeffries–Matusita distance (SFFS + JM), a stepwise forward variable selection method using Wilk's Lambda criterion (SFVS), and a Lasso regularized generalized linear model (LASSO). Furthermore, two models with built-in feature selection techniques were also computed, specifically the generalized linear model with stepwise feature selection (glmStepAIC) and the flexible discriminant analysis (FDA) (Figure 2).

All approaches (Figure 2) identified similar spectral wavelengths located mainly in the blue (350–500 nm), green (500–600 nm), red (600–750 nm), and NIR (>750 nm)

regions (Table 1). These results are coherent, presenting biological significance since the symptoms caused by Pseudomonas syringae pv. actinidiae (Psa) promote modifications in leaf biochemical and structural composition, as previously mentioned. These selected features for discriminating asymptomatic and symptomatic kiwi leaves are in line with those found for other crops with different diseases, namely: (i) for grapevine, where wavelengths near the green region of the visible (534, 576, 430, and 368 nm), and near-infrared spectra were selected by a stepwise-based approach (Naidu, Perry et al. 2009); (ii) also for grapevine, other wavebands also seem to have high discriminatory power, being mainly located at the green (520–550 nm), chlorophyll-associated wavelengths (650–670 nm), red edge (700–720 nm), beginning of near-infrared (800–900 nm) and shortwave infrared spectral regions (Junges, Almança et al. 2020); (iii) for soyabean, wavelengths in the green and red regions of the spectrum (top ten wavebands selected by: linear discriminant analysis—523, 535, 592, 658, 694, 700, 733, 766, 931, 1015; logistic discriminant analysis—400, 421, 427, 559, 571, 589, 679, 682, 688, 703; and linear correlation analysis—458, 461, 476, 479, 485, 494, 500, 626, 632, 686) similarly exhibit the best correlation with disease (Bajwa, Rupe et al. 2017); (iv) for wheat affected by Puccinia triticina, the relevant spectral characteristics corresponded to the wavelengths of 605, 695, and 455 nm, for various levels of the infection (Ashourloo, Mobasheri et al. 2014); (v) for oil palms diseased with ganoderma basal stem rot disease, the features with higher importance were found mainly in the green (from 550 to 560 nm), and in the red-edge (around 650 to 780 nm) regions (Ahmadi, Muharam et al. 2017); (vi) for rice, different levels of panicle blast could be differentiated at six different effective wavelengths, specifically 459, 546, 569, 590, 775, and 981 nm (WU, Cao et al. 2009).

In crop remote sensing studies, spectral vegetation indices (VIs) are still the most common approaches studied to identify and manage abiotic and biotic stresses in different crops [58–60]. VIs are composed of numerous combinations of different bands, providing spectral information with reduced dimensionality (Mahlein, Oerke et al. 2012, Oerke, Mahlein et al. 2014, Thenkabail, Gumma et al. 2014). Despite its extended usage and utility, it is not always clear if this plethora of VIs is sensitive to the variable of interest and, simultaneously, if they respond insensitively to confounding factors, namely variations of other leaf or canopy properties, background soil reflectance, solar illumination, and atmospheric composition, this may induce variability in the spectral properties of surfaces (Morcillo-Pallares, Rivera-Caicedo et al. 2019). In turn, feature selection methods may provide more robust and customized spectral information since they can identify the variables that are effective for modeling data class characteristics,

reducing the dimensionality of the original feature space by choosing only the best and minimum subset of features (Thenkabail, Lyon et al. 2018).

Data modeling was then performed using different statistical and machine learning approaches applied in the complete dataset and the wavelengths identified by the different feature selection approaches (Figure 2). The mean overall accuracy and coefficient of variation of the models allowed the identification of the combination of a stepwise forward variable selection with a support vector machine with radial kernel and class weights (stepsvmrw) as the best modeling approach among those evaluated (Table 2). In this model, the kernel trick reduced dimensions and provided the necessary class separation of non-linear features to the support vectors method [e.g., (e.g. Luts, Ojeda et al. 2010)]. However, kernels are not theoretically derived for spectroscopy (Martins, Barroso et al. 2022). This handicap may lead to non-optimal selection, that does not represent the relationship between spectral features and discrimination among symptomatic and asymptomatic leaves. This might explain the better performance of SVM models when combined with feature selection algorithms (e.g., stepwise feature selection; SFVS).

Stepsvmrw presented a classification accuracy of 85%, kappa score of 0.70 (good agreement), and f-measure of 0.84, when the total dataset (training and test sets) was used for prediction. It correctly classified 190 spectra of the 223 spectra acquired over the symptomatic leaves and classified 240 of 281 spectra belonging to asymptomatic observations. The percentage of asymptomatic observations correctly classified by this model ranged from 71% to 96% for both test sites, having decreased for the BT region over time but showing an inverse tendency for the CT region (where it increased) (Figure 4). The misclassification regarding the symptomatology of leaves in the early stages (Table 4) may indicate initial disease phases in the NIR domain of the spectrum when typical disease symptoms (e.g., chlorosis and necrosis) are not yet visually detectable by the human eye. In turn, for the CT region, spectral measurements allowed complete surveillance from the appearance and development of the first signs of BCK to its full development over time, coinciding with the visual separation between healthy and disease leaves.

Our results showed lower accuracies than those found by Lu et al. (Lu, Ehsani et al. 2017) for classifying strawberry leaves infected with *Colletotrichum gloeosporioides* using multitemporal indoor and in-field assessments. Their classification accuracy for indoor measurements varied from 81.6% to 89.7% for discriminant analysis (FDA), 84.2% to 93.1% for stepwise discriminant analysis (SDA), and 84.2% to 87.5 % for k-

nearest neighbor (KNN), corresponding the lower value to the classification accuracy for asymptomatic samples and the higher value to the accuracy of healthy plants. KNN misclassified healthy samples as asymptomatic. In-situ evaluations had lower accuracy scores ranging from 54.7% to 75.8% for FDA, 62.5% to 77.3% for SDA, and 15.4% to 90.6% for KNN. These poorer values obtained in in-field assessments were probably related to limitations in the dataset, namely the asymptomatic sample size being larger than the healthy and symptomatic sample, and uncontrolled environmental conditions acknowledged as the most important variations in sunlight during measurements. Zhao et. al. (Zhao, Fang et al. 2020) used three dimensionality reduction algorithms and three machine learning models to classify and identify powdery mildew (*Blumeria graminisf*. sp. *tritici*) on wheat under laboratory conditions. When applied to hyperspectral data, SVM achieved a classification accuracy of 88.0%. The best model combined principal component analysis (PCA), for dimensionality reduction, and SVM, having achieved an identification accuracy of 93.3% by cross-validation methods. The authors only assessed 75 picked leaves, with the number of diseased samples (60) being considerably higher than the number of healthy ones. Huang et al. (Huang, Ding et al. 2019) studied the wheat powdery mildew disease using 145 in-situ hyperspectral measurements (90 healthy and 55 diseased samples), different vegetation indices (alone and combined with each other), and three model classifiers. They obtained classification accuracies ranging from 74.5% to 94.8%. Despite our accuracy values being similar or slightly lower than these examples, their scores were generally obtained by performing indoor assessments (made under supervised, controlled conditions), and/or through modelling approaches developed with small datasets, where spectral noise and variability are low. Moreover, most models were only applied to a single test site, with restricted soil, climate conditions, and plant age, not being able to generalize to a practical application.

Model results were further supported by the empirical analysis of the spectral information of BCK disease. Asymptomatic leaves mostly revealed the typical spectral behavior of green and photosynthetically active vegetation (Figure 5a). In turn, spectral responses of symptomatic leaves registered variations in the VIS and NIR regions; having some spectral bands presenting a greater response to the BCK infection (Figure 5a,b). Overall, the mean spectral reflectance records of symptomatic leaves showed higher values of reflectance for the blue and the majority of the NIR regions (850–1075 nm), and lower values for the red-edge and beginning of the NIR regions (<850 nm), when compared to the asymptomatic cases. These results are consistent with the infection caused by Psa, since it results in necrotic leaf spots, which are related to membrane damage and cell death (Balestra, Mazzaglia et al. 2009). Modifications in the

content of chlorophyll and brown pigments, water, and structural components influence crop spectral behavior in these spectral regions (Asner 1998, Penuelas and Filella 1998). Other studies, performed on different crops, also reported an increase in diseased leaf reflectance in the VIS region (mainly in the green and red ranges of the spectrum), and a decrease in the NIR region, specifically: (i) sugar beet infected with Cercospora, in the VIS region from 550 to 700 nm and the NIR region from 700 nm to 850 nm (Mahlein, Rumpf et al. 2013); (ii) grapevine infected with leaf stripe disease (esca complex) in the green region (520–550 nm), and red region (650 nm) of the spectra (Junges, Almança et al. 2020); (iii) soybean affected by the soybean cyst nematode (SCN) and sudden death syndrome (SDS) (Bajwa, Rupe et al. 2017).

Our results are thus relevant for detecting and discriminating the bacterial canker disease of kiwi in leaves. Hyperspectral data provides a large amount of information, allowing the screening of samples based on their chemical composition rather than only their size, shape, and visible color (that RGB devices permit). Despite the promising findings supporting this proof-of-concept, this was a single season, in-field analysis (without control over agronomic, environmental, and infectious conditions). Future studies are thus needed, namely by analyzing the same leaf over time, to better understand the plant–pathogen interaction and its impact on host spectral behavior. Furthermore, supplementary laboratory assessments will be highly beneficial and allow more comprehensive knowledge about the disease caused by the Psa pathogen.

## 5. Conclusions

This study proposes the diagnostics of bacterial canker of kiwi (BCK) disease caused by *Pseudomonas syringae* pv. *actinidiae* (Psa), on kiwi leaves using hyperspectral in-field measurements. Asymptomatic leaves revealed the typical spectral behavior of green and photosynthetically active vegetation, while symptomatic leaves presented deviations in their spectral signature in the VIS and NIR regions. The different feature selection methods allowed the identification of several wavelengths as more important for BCK discrimination, being mainly located in the blue (350–500 nm), green (500–600 nm), red (600–750 nm), and NIR (>750 nm) regions. Spectral separability between asymptomatic and symptomatic observations were observed in the dataset, and a stepwise forward variable selection approach with an SVM algorithm with a radial kernel and class weights presented the best results in terms of disease discrimination. The model presented an overall accuracy of 0.85, with a 0.70 kappa score and 0.84 F-measure. Our findings allowed a rapid, non-destructive, in situ disease classification, supporting the implementation of spectral point measurements for crop disease

discrimination. Nonetheless, more research is necessary to better comprehend the plant–pathogen dynamics and their effects on host spectral behavior. Furthermore, feature selection approaches for disease diagnosis must be further explored to develop more economic, multiband sensors. Multi- and hyperspectral sensors can be coupled on different platforms, forming distinct functioning measurement systems. This results in more precise agronomic practices, such as mapping, monitoring, scouting, and treatment of crop diseases. Handheld sensors, terrestrial (e.g., robots) and aerial platforms (e.g., drones), and satellites can assess plant spectral behavior on different scales, including leaf, single-plant, canopy, plot, and farm levels.

**Funding**

The research leading to these results received funding from the European Union's Horizon 2020—The EU Framework Programme for Research and Innovation 2014–2020, under Grant Agreement No. 857202—DEMETER. Mafalda Reis-Pereira and Renan Tosin were supported by fellowships from Fundação para a Ciência e a Tecnologia (FCT) [grant references SFRH/BD/146564/2019, and SFRH/BD/145182/2019, respectively]. Rui C. Martins was supported by a research contract grant by Fundação para a Ciência e Tecnologia (FCT) [grant reference CEEIND/017801/2018].

# Case Study 4

**Reis Pereira, M.**; Santos, F.N.d.; Tavares, F.; Cunha, M. Enhancing host-pathogen phenotyping dynamics: early detection of tomato bacterial diseases using hyperspectral point measurement and predictive modeling. Frontiers in Plant Science. 2023. 14:1242201. https://10.3389/fpls.2023.1242201

Paper published on 16th August 2023

Classification according to journal: Original Research Article

Research Topic: Advanced AI Methods for Plant Disease and Pest Recognition

Bibliometric indicators from the Journal Citation Report, Institute for Scientific Investigation

• Journal impact factor: 5.6

• Journal rank

*Plant Sciences: position 27/238; quartile Q1, JIF percentile 88.9*

# Enhancing host-pathogen phenotyping dynamics: early detection of tomato bacterial diseases using hyperspectral point measurement and predictive modelling

Mafalda Reis Pereira[1,2], Filipe Neves dos Santos[2], Fernando Tavares[1,3,4], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

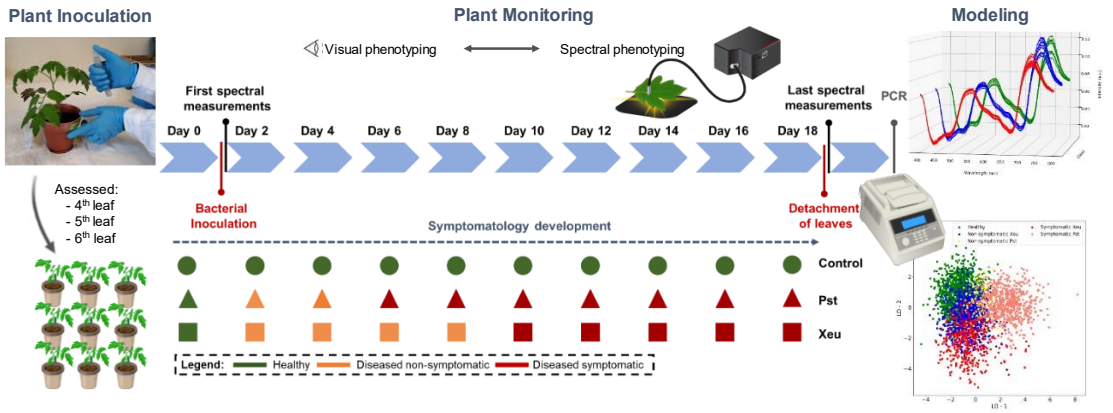
*** Correspondence:** Mário Cunha mccunha@fc.up.pt

**Abstract**

Early diagnosis of plant diseases is needed to promote sustainable plant protection strategies. Applied predictive modeling over hyperspectral spectroscopy (HS) data can be an effective, fast, cost-effective approach for improving plant disease diagnosis. This study aimed to investigate the potential of HS point-of-measurement (POM) data for in-situ, non-destructive diagnosis of tomato bacterial speck caused by *Pseudomonas syringae* pv. *tomato* (Pst), and bacterial spot, caused by *Xanthomonas euvesicatoria* (Xeu), on leaves (cv. cherry). Bacterial artificial infection was performed on tomato plants at the same phenological stage. A sensing system composed by a hyperspectral spectrometer, a transmission optical fiber bundle with a slitted probe and a white light source were used for spectral data acquisition, allowing the assessment of 3478 spectral points. An applied predictive classification model was developed, consisting of a normalizing pre-processing strategy allied with a Linear Discriminant Analysis (LDA) for reducing data dimensionality and a supervised machine learning algorithm (Support Vector Machine – SVM) for the classification task. The predicted model achieved classification accuracies of 100% and 74% for Pst and Xeu test set assessments, respectively, before symptom appearance. Model predictions were

coherent with host-pathogen interactions mentioned in the literature (e.g., changes in photosynthetic pigment levels, production of bacterial-specific molecules, and activation of plants' defense mechanisms). Furthermore, these results were coherent with visual phenotyping inspection and PCR results. The reported outcomes support the application of spectral point measurements acquired in-vivo for plant disease diagnosis, aiming for more precise and eco-friendly phytosanitary approaches.

## Graphical abstract




## Keywords

Plant disease diagnosis, Early diagnosis, Proximal sensing, Hyperspectral Spectroscopy, Point of Measurement, Applied Predictive Modeling, Linear Discriminant Analysis, Machine learning

## 1. Introduction

The tomato (*Solanum lycopersicum* L.) crop holds great importance worldwide due to its significant impact on agriculture, the economy, and human nutrition. This globally cultivated vegetable crop is very sensitive to diseases leading to dramatic yield and economic losses (Blancard 2012). Bacterial diseases of tomato plants caused by the Gram-negative bacteria *Pseudomonas syringae* pv. *tomato* (Pst, bacterial speck) and *Xanthomonas euvesicatoria* (Xeu) formerly known as *Xanthomonas campestris* pv. *vesicatoria*, bacterial spot, are two important etiological agents responsible for several plant outbreaks and considerable losses in tomato production worldwide. These two diseases are responsible for severe alterations in the host physiology, biochemistry, and structural composition, causing plant phenotype modifications (e.g., reduction of the photosynthetic capacity of diseased foliage, defoliation, flower abortion, and fruit lesions, among others). Ultimately, they result in yield reductions due to the damage caused to plants and fruits, which makes them unsuitable for the fresh market or processing. Control measures for these two crop diseases may be ineffective, especially when the

bacteria are well-established in a production site (medium to late stage of the disease infection process. Phytosanitary products, such as copper and antibiotics (Alves, Ribeiro et al. 2023), can be applied to mitigate the negative effects of the disease. Nevertheless, this approach can lead to bacteria tolerance to phytosanitary compounds (Blancard 2012), and conduct to considerable damage to the environment and food security due to non-targeted applications of these products (Zhang, Yang et al. 2020).

Nowadays, bacterial diseases are diagnosed essentially through scouting and 'wet lab' -based approaches. The first requires a careful and detailed inspection of crop fields (usually visual) by specialized trained observers. They must detect and identify diseased plants based on modifications to the characteristic phenotype of the crop, and the presence of disease symptoms (Parker, Shaw et al. 1995). Thus, it is subjective, error-prone (as symptoms alone are not entirely disease-specific, and can be promoted by other biotic and abiotic stresses), labor-intensive, time-consuming, and expensive (Mahlein 2016). In turn, laboratory-based techniques consist of serological and molecular assays, frequently applied due to their sensitivity, accuracy, and effectiveness. The most widespread lab methods include Enzyme-Linked Immunosorbent Assay (ELISA) and Polymerase Chain Reaction (PCR) methods. They involve comprehensive sampling procedures, which require several hours to be completed, and destructive sample preparation, precluding the accompaniment of disease development nor its field mapping to support precision agriculture systems (e.g. Site-Specific Management) (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015). Nevertheless, laboratory-based approaches lack appropriate high throughput and speed for supporting real-time agronomic precision decisions in-field since they were developed to verify the presence of pathogens. They also still have some diagnostic constraints, mostly in the non-symptomatic and early disease infection stages, related to the irregular spread of bacteria inside plants (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015).

Hyperspectral spectroscopy (HS) is one innovative approach that has been studied and successfully applied to assess different plant(host)-pathogen interactions in a fast, sensitive, standardized cost-effective, high-throughput, and non-invasive way (Golhani, Balasundram et al. 2018). Through spectral measurements in the visible (VIS, 400-700 nm) and infrared (IR, 800-2500 nm) regions, HS showed the capability of effectively assessing a wide variety of plant structural, chemical, biophysical, and metabolic traits in living tissues (Thenkabail, Smith et al. 2000, Delalieux, van Aardt et al. 2007). Changes in the typical spectral phenotype of a crop may indicate deviations in its health status, leading to an indirect method of diagnosing diseases. Plant-pathogen interactions shift plant metabolism and tissue composition, resulting in detectable

variations in the plant's optical behavior. In brief, these dynamics typically promote modifications in the VIS spectra of plants, due to changes in pigments' concentration and physiological processes. Furthermore, variations in the IR region may also occur and are essentially linked to leaf water levels, chemical compounds (namely lignin's and proteins content), structural elements, and internal scattering processes (Thenkabail, Gumma et al. 2014, Tosin, Martins et al. 2022).

Different types of pathogens, such as pests (Herrmann, Berenstein et al. 2017, Zhang, Wang et al. 2017), fungi (Yu, Anderegg et al. 2018, Skoneczny, Kubiak et al. 2020), bacteria (Bagheri, Mohamadi-Monavar et al. 2018), and viruses (Morellos, Tziotzios et al. 2020) affecting different crops have already been detected using the HS technique, mostly in symptomatic stages. Thus, this spectral phenotyping technique constitutes an interesting diagnosis method, allowing the distinction between the spectral signature of healthy and disease tissues, as well as between the spectral signature of diseased tissues infected with different pathogens.

HS holds great potential for early disease diagnosis, i.e., when plants are diseased but still don't manifest any visual symptoms of the infection (Gold, Townsend et al. 2020, Reis-Pereira, Tosin et al. 2022). However, the use of this approach for non-symptomatic plant disease diagnosis remains largely unexplored. Understanding host-pathogen specific interactions and overcoming technical challenges related to the biophysical status of infected plants, organ of the plant assessed, sensing technology, data processing, and modeling approaches is essential for the effective application of HS in vivo crop disease diagnosis (Mahlein, Kuska et al. 2018). Addressing these challenges is crucial for real-time monitoring of disease progression.

The most used sensing devices for plant disease detection are non-imaging (e.g., point-of-measurement, POM) and imaging sensors. In POM sensing, light usually enters the leaf, and undergoes internal reflections conditioned by tissue structures and composition status. Thus, this technique can indirectly infer certain internal tissue characteristics affected by the host-pathogen interaction. POM sensors are typically designed to measure specific parameters without being significantly affected by factors like lighting conditions or surface textures. This reduces the potential for external interference and ensures more accurate and consistent measurements. This allied with their higher spectral resolution, cost-effectiveness, compactness, and reduced data processing requirements, makes them an attractive option for plant studies (Martins, Barroso et al. 2022).

Spectral information provided by HS is extremely valuable, nonetheless, in biological tissues, it is super-imposed in the recorded spectra at different scales of interference (Barroso, Ribeiro et al. 2022, Tosin, Martins et al. 2022). Moreover, HS data can present substantial amounts of redundant information in contiguous wavelengths, and just some specific spectral features might be relevant to predict and classify diseased tissues (Caicedo, Verrelst et al. 2014, Rivera, Verrelst et al. 2014). Applied predictive classification modeling strategies can be developed to study spectral data and extract useful information. Diverse approaches of data correction and pre-processing (e.g., data scaling and normalization) can be computed to reduce undesired spectral effects, such as 'noise' and scattering effects. Additionally, modeling strategies, as well as feature selection (FS), feature extraction and dimensionality reduction techniques (DR), can be useful for determining the wavelength features which have more influence in disease discrimination (Mahlein, Steiner et al. 2010, Ahmadi, Muharam et al. 2017). In plant disease research, different predictive approaches using HS data have been explored to classify tissues affected by biotic stress, considering all the spectral features or only specific variables, designated by FS or DR techniques (Gold, Townsend et al. 2020, Meng, Lv et al. 2020). Nevertheless, there is a lack of standardized protocols for acquiring hyperspectral data from tomato leaves. Different studies employ various acquisition setups, lighting conditions, and preprocessing techniques, making comparing and integrating findings challenging.

This work addresses the main technological challenges for efficiently applying hyperspectral technologies in phenotyping to diagnose plant diseases. Conducting analysis for healthy and bacterial inoculated plants over time, this study aims i) to compare visual phenotyping against spectral phenotyping based on the hyperspectral point-of-measurement (HS-POM) for healthy and diseased tomato leaflets, ii) to evaluate the HS-POM ability to accurately classify samples at various stages of disease development, including those without any visible symptoms and iii) distinguish the etiological agents of distinct tomato bacterial diseases. The specific goals include developing an applied predictive modeling strategy (combining data pre-processing, dimensionality reduction, and a supervised machine learning algorithm) for tomato bacterial disease classification and establishing causal relationships between plant health status, specific spectra characteristics, and the physiological changes that occur during infection dynamics to advance theoretical knowledge and provide a foundation for further research.

## 2. Materials and methods

### 2.1. Bacterial inoculation and plant growth

#### 2.1.1. Inoculation on tomato leaflets

Tomato (*Solanum lycopersicum* L.) plants of the cultivar Cherry were grown in 200 mL pots containing a commercial potting substrate, in a walk-in plant growth chamber under controlled conditions (25-27 ºC, humidity of approximately 60%, photoperiod of 12 / 12 h and light intensity 30W). Plants were divided into three groups of three plants each (nine plants in total), being a) one group of plants inoculated with *Pseudomonas syringae* pv. *tomato* DC 3000 (Pst) bacteria, b) a second group of plants inoculated with *Xanthomonas euvesicatoria* LMG 905 (Xeu) bacteria, and c) a third group of plants was treated with sterile distilled water only (Control group) (Figure 1). Plants were physically separated to avoid cross-contamination.

Plants were inoculated in the laboratory, at the growth stage of 5-6 fully expanded leaves, by spraying until they became fully wet, and run-off occurred. The bacterial suspensions used for these inoculation assays consisted of 1 x $10^8$ cells / mL. They were prepared from 48-h-old cultures of Pst grown in KB medium (peptone, 20.0g; $K_2HPO^4$, 1.5g; $MgSO^4$, 1.5g; glycerol, 10 mL; agar, 15g; distilled water up to 1.0 liter), and of Xeu cultures grown in YDC medium (yeast extract, 10.0g; dextrose, 20.0g; $CaCO_3$, 20.0g; agar, 15.0g; distilled water up to 1.0 liter). The inoculated plants were then covered with transparent polythene bags for 48 h to increase the relative humidity that fosters bacterial entry into plant tissues through natural openings such as stomata (Lamichhane 2015). Plants were monitored daily for symptom development for 18 days (Figure 1).

During the inoculation period, to verify if the bacteria cultures used in these inoculation tests were viable, 20 µL of Pst solution and 20 µL of Xeu solution were cultured in Petri dishes containing KB and YDC media, respectively. After 48 h was possible to observe the bacteria growth in both nutrient media, proving that bacteria were viable at the moment of inoculation.

#### 2.1.2. Bacterial isolation from diseased leaflets

After the last spectral measurement, sample preparation for bacterial isolation was performed for all the leaflets. Leaflets were excised from plants using a sterile scalpel (Fernandes, Albuquerque et al. 2017). Bacterial isolation was carried out as defined by Fernandes et al. (2017, 2021). Briefly, each sample of excised leaflet tissue was disinfected by immersion in 70% ethanol followed by washing with sterile distilled water (SDW), and then macerated with SDW in extraction bags. The suspensions

obtained, and corresponding dilutions, were streaked on KB (samples inoculated with Pst bacteria), and on YDC medium (samples infected with Xeu pathogen). Characteristic colonies from these two bacteria species (milky white colonies in the case of Pst, and mucoid yellow colonies in the case of Xeu) were selected for growth on fresh nutrient agar medium to ensure purity.

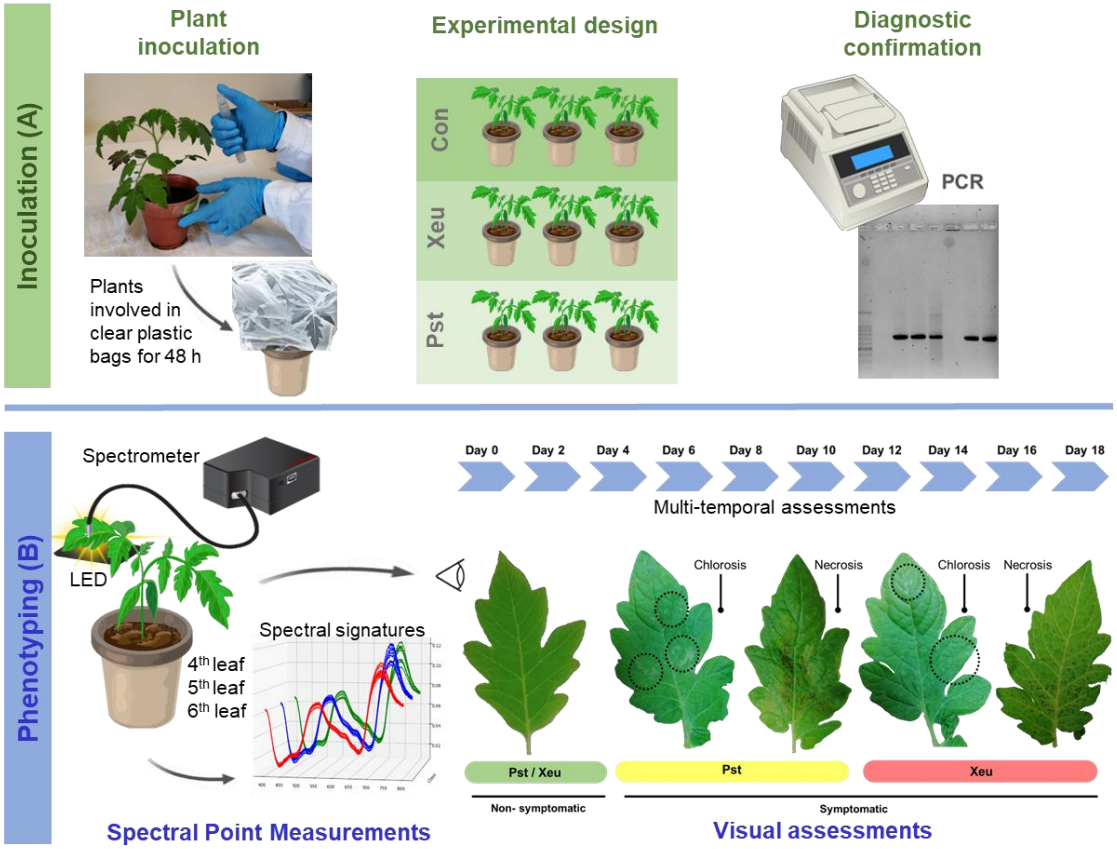



**Figure 1** Experimental setup of the bacterial inoculation assay performed on tomato leaves (A), and visual and spectral assessments (of the 4th, 5th, and 6th leaves) made in a dark room (B). Spectral measurements were performed on the adaxial side of leaflets, using a spectrometer combined with an optical fiber bundle with a reflection probe. A white LED was placed beneath each leaflet. Both visual and spectral assessments were made 18 Days After Inoculation (DAI), collecting leaflets' spectral signatures and registering modifications in their phenotype (e.g., the appearance of the first symptoms, both chlorosis and necrosis).

Pst characteristic symptoms resemble small greasy dark stains (circular or slightly angular), that become brown to black, and appear randomly on the leaflets (often on the youngest or the ones located at the edge of the canopy plant). These lesions may typically show a yellow halo of various sizes. They are about 2–3 mm and can develop and coalesce (especially in the presence of moisture), affecting large areas of the leaf,

that may later become necrotic and desiccate (Blancard 2012). In turn, Xeu characteristic symptoms comprise small, circle, or slightly angular, translucent, and water-soaked lesions, which turn brown with time. They appear randomly in leaflets, and eventually become necrotic spots, with light gray centers and dark margins, which also can become surrounded by a yellow hallow with time. Smaller lesions can coalesce into each other forming larger injuries, whose diameter can range from 2 to 3 mm. In severe cases, tissues in the center of a lesion become dry and fall out, leading to “shot-hole” symptoms (Ritchie 2000, Blancard 2012).

### 2.1.3. Colony PCR protocol

A colony PCR was performed to validate the presence of both bacteria species on tomato leaflets isolates. PST2 (Vieira, Mendes et al. 2007) and XV14 (Albuquerque, Caridade et al. 2012) were the chosen markers, for Pst and Xeu, respectively, with amplicon lengths of 200, and 713 bp, correspondingly. A 20 µL PCR reaction mix consisted of 1 × DreamTaq Buffer (ThermoFisher Scientific, Waltham, MA, USA), 0.2 mM of each deoxynucleotide triphosphate (dNTP) (Grisp, Porto, Portugal), 0.2 mM of each forward and reverse primers, 1 U of DreamTaq DNA Polymerase (ThermoFisher Scientific, Waltham, MA, USA) and 10 µL of DNA isolate solution. Sterile distilled water was used as the negative control. PCR cycling parameters were defined as stated by Vieira, Mendes et al. (2007) for Pst, and Albuquerque, Caridade et al. (2012) for Xeu. PCR products were then separated by electrophoresis on a 0.8% agarose gel (1 × TAE buffer) and visualized using Xpert Green DNA stain (Grisp, Porto, Portugal) with a Molecular Imager Gel Doc XR+ System (Bio-Rad, Hercules, CA, USA).

## 2.2. Spectral measurements in vivo tissue

### 2.2.1. Experimental setup for plant spectral acquisition

Figure 1 presents the main procedures for spectral measurements in the experimental setup. Hyperspectral point-of-measurements (HS-POM) were collected in vivo from the adaxial side of healthy and diseased leaflets of the nine tomato plants in the study, in a dark room. For each plant, spectral assessments were performed randomly on nine points of different leaflets, belonging to the $4^{th}$, $5^{th}$, and $6^{th}$ expanded leaflets.

Hyperspectral data were acquired using a Hamamatsu Photonics K.K. TM Series C11697MB spectrometer, which covers a wavelength range of 200-1100 nm with a spectral resolution of 0.6 nm. A transmission optical fiber bundle (FCR-7UVIR200-2-45-BX, Avantes, Eerbeek, The Netherlands) with a range of 200-2500 nm was used along

with a stainless-steel slitted reflection probe that was positioned 0.5 cm above the sample surface to capture the leaflet's spectral signal and direct it to the spectrometer's entrance lens. A white LED light was placed underneath the leaflet to provide uniform illumination to its entire abaxial surface. The spectral range of the LED emits light from 390 to 800 nm. Therefore, the LED spectra were used as a reference to the spectral range measured by the spectrophotometer and to check measurement and light emission stability (Figure 1 B). The hyperspectral data were collected using specialized evaluation software (SpecEvaluationUSB2.exe, Hamamatsu Photonics K.K., Japan).

### 2.2.2. Preprocessing hyperspectral data

The performance of the modeling approach in detecting bacterial diseases in tomato leaflets was assessed using only the spectral region of 400 to 800 nm, approximately. This decision was based on the spectral wavelength range of the light LED source used (where possible useful information could be retrieved) and due to the observation of spectral noise near the limits of the equipment's spectral range, which could negatively affect the performance of the classification process. Therefore, a total of 944 features (wavelength) were used in the development of the prediction modeling (Figure 2).

Preprocessing data was performed following spectra normalization (Figure 2, 3). This approach aimed to standardize the data to a common scale, enabling meaningful comparison and analysis across different scenes or datasets. Additionally, it aims to decrease spectral signal oscillations, related to measurement equipment specifications (including devices' internal noise), variations in data assessment conditions (comprising differences in global spectral trend, total energy, high-frequency noise, and/or local background) (Randolph 2006), associated to changes in environmental conditions or induced by the operator in the moment of assessment (e.g. variations in sample-sensor distance, uneven illumination conditions, choice of leaflets sample point location, appropriate scan parameters, spectral calibration, among others). This results in model abilities improvement by aiding in class separation (Randolph 2006, Guezenoc, Gallet-Budynek et al. 2019). Furthermore, this process enables the elimination of the spectral response of both the sensor and light source, making possible the transfer of the acquired classifier to a different sensing device. Spectral data retrieved from measurements in tomato leaflets $S(\lambda_n)_m^{raw}$ were normalized through their division by the white LED source spectral signature $S(\lambda_n)^{reference}$(considering the time of exposure of the spectral measurements), through the computation of the following forming (Equation 1):

$$S(\lambda_n)_m^{normalized} = S(\lambda_n)_m^{raw} / S(\lambda_n)^{reference} \quad (1)$$

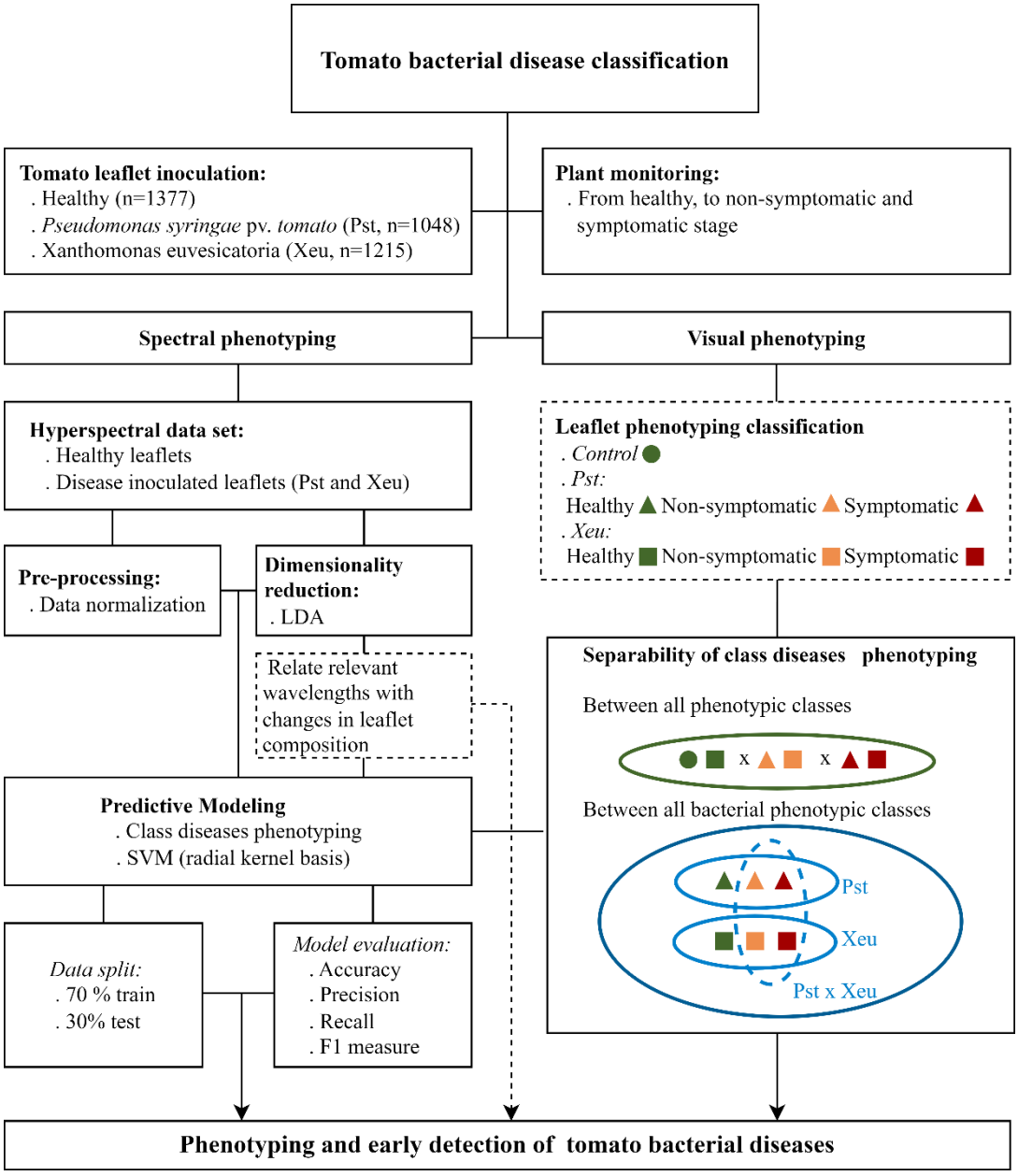



**Figure 2** Conceptual diagram for the applied predictive modeling approaches of bacterial tomato leaflet disease.

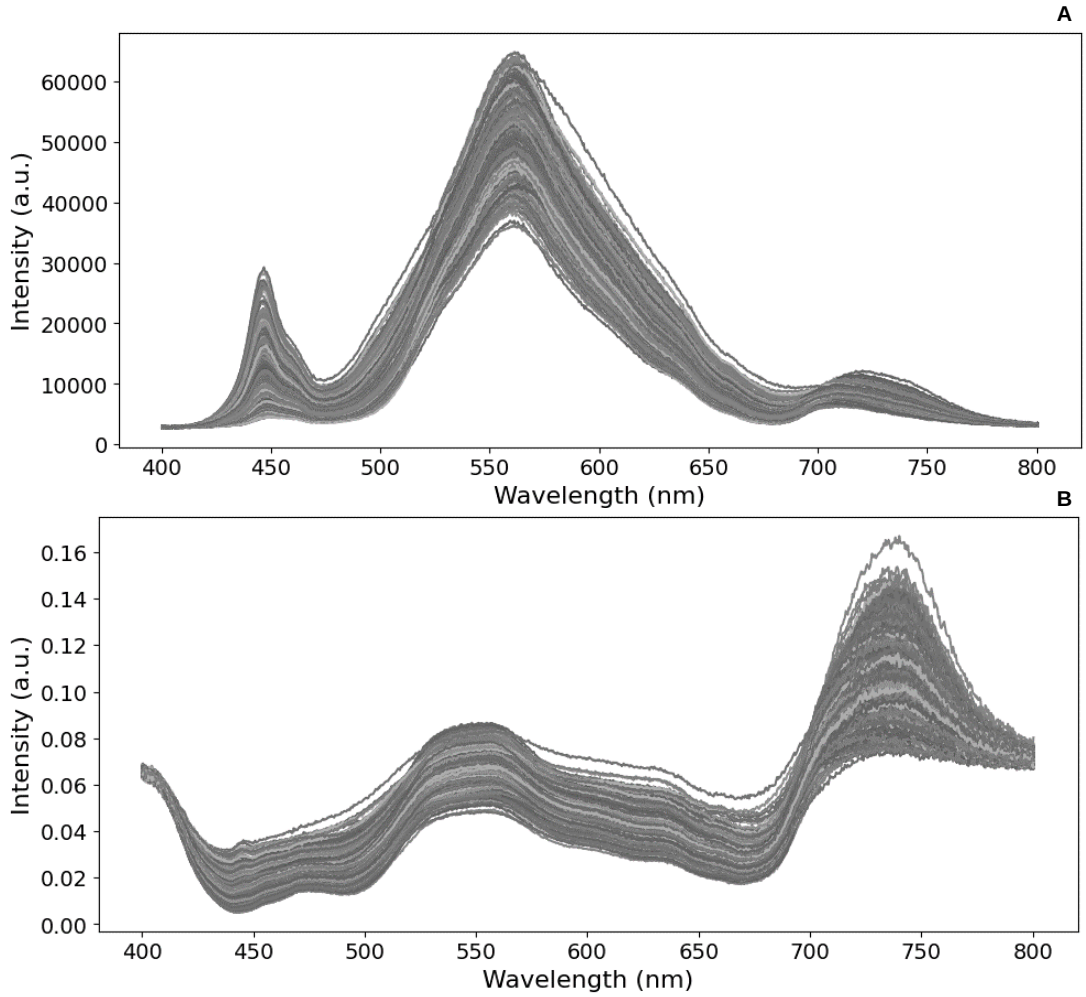



**Figure 3** Original (raw, A) and normalized (B) hyperspectral signatures assessed in tomato leaflets during the experimental assay.

### 2.3. Modeling leaflets symptomatology over time

#### 2.3.1. Data set structure

Seeking bacterial tomato disease classification, spectral signatures from leaflets were then grouped to perform an applied predictive modeling approach related to the plants' experimental condition. Leaflets were classified according to the plant treatment group and their health status, including the classes: i) healthy, including all the measurements which were performed before bacteria inoculation, and the remaining assessments that were made in non-inoculated plants considered as control plants; ii) non-symptomatic Pst; iii) non-symptomatic Xeu; iv) symptomatic Pst; and v) symptomatic Xeu. All the spectral data collected from tomato leaflets on different dates were unified in a single classification model (Figure 1, 2).

Data classification was, thus, performed seeking the unraveling of spectral phenotyping differences between i) healthy and non-symptomatic diseased tissues (early diagnosis), ii) healthy and diseased tissues (showing visual modifications due to changes in chemical and structural composition), ii) healthy and diseased tissues affected by different bacterial etiological agents (which present distinct host-pathogen specific interactions), iii) and diseased tissues infected by different bacteria species (responsible for causing similar visual symptoms but showing different pathogenic dynamics).

### 2.3.2. Dimensionality reduction of spectram data

Multi-scale interference in plants' tissue promotes superimposition on hyperspectral data, resulting in autocorrelations in their spectral signal at several scales (Martins, Barroso et al. 2022). To mitigate the effects of high dimensional, redundant information, several methodologies have been cited in the state-of-the-art, including dimensionality reduction (DR) approaches (Lapajne, Knapič et al. 2022, Reis-Pereira, Tosin et al. 2022). DR techniques are a class of predictor transformations. They can reduce data by creating a minor set of predictors that aim to retain most of the information contained in the original variables. Usually, these approaches generate new predictors which are functions of the original ones (signal extraction or feature extraction techniques) (Kuhn and Johnson 2013).

This study examined a DR approach called Linear Discriminant Analysis (LDA), generally computed as a pre-processing. It is a supervised learning algorithm used for classification tasks. LDA is usually applied as a feature extraction technique, performed to reduce the dimensionality of the data while maximizing the class separability. It projects the high-dimensional data onto a lower-dimensional space while preserving the discriminative information between classes. In brief, data is projected onto a linear subspace that maximizes the ratio of between-class variance to within-class variance. Thus, the projected data points are as far apart as possible in the new space, while the points belonging to the same class are as close as possible. Therefore, LDA contributes to reducing the problem's computational complexity and avoiding overfitting. It can also be useful for visualizing the data in a lower-dimensional space, helping interpret patterns in data (Tharwat, Gaber et al. 2017). Furthermore, this technique was applied since our dataset is not linearly separable, and LDA can organize it in another space with the maximum possible linear separability (Sachin 2015).

LDA feature space loadings (also called coefficients or weights) were additionally used to infer the most relevant wavelength variables, through the computation of the

interquartile range (IQR) for the weights. A threshold at 1.5 times the IQR beyond the upper quartile was established. This process aimed to increase sensitivity to the weight distribution, enabling the capture of outliers and extreme values. An applied predictive classification model was later computed to help deal with the non-linearity of the data.

### 2.3.3. Machine learning classification model

A Support Vector Machines (SVMs) algorithm was chosen to integrate this modeling strategy. This supervised machine learning algorithm performs classification based on the concept of optimal separating hyperplane (Vapnik 1999, Mosavi, Sajedi Hosseini et al. 2021). SVMs are nonlinear approaches that discover the most extensive margin between two classes in feature space. These approaches aim to decrease the error test and model complexity (Ballabio and Sterlacchini 2012). SVMs can present distinct hyperparameters and kernel forms, which convert raw data inputs from the original user space into kernel space through a user-defined feature map (Patle and Chouhan 2013, Ding, Liu et al. 2021). This study used a radial basis function (RBF) kernel was used since it allows SVMs to capture non-linear relationships between input features and target variables. It may also allocate distinct weights to different points since they learn the decision surface according to the relative importance of the data points in the training set (being well-suited for handling outliers and noisy data) (Xulei, Qing et al. 2005). More detailed information about the SVM algorithm, including relevant principles and calculation formulas, can be found in Ballabio et al (2012) and in Chang et al. (2011). The parameters of the SVM applied corresponded to the default values of the algorithm implemented in the ‘Scikit-learn’ machine learning library (Pedregosa, Varoquaux et al. 2011), which also can be found in Table 1.

The datasets were divided into training data (70% of random observations) and validation data (30% of the remaining observations) (Kuhn and Johnson 2013), following a holdout method (Lantz 2019). The training and validation sets combined the pairs of concurrent measurements of the group and health status and the corresponding values of the predicting variables. A resampling strategy was performed as stated in Reis-Pereira, Tosin et al. (2022) to reduce the possibility of overfitting (Berrar 2019, Valier 2020).

**Table 1** Default parameters of the SVM algorithm of 'Scikit-learn' library used in this study.

| Parameter | Value | Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|---|---|
| C | 1.0 | Probability | False | Verbose | False | Break ties | False |
| Kernel | rbf | Shrinking | True | Cache size | 200 | Tolerance | 1e-3 |
| Gamma | 'scale' | Class weight | None | Decision function shape | One-vs-rest (ovr) | Random state | None |
| | 1/(n_features *X.var()) | Max iteration | -1 | | | | |

### 2.3.4. Model performance evaluation

Different metrics were additionally retrieved to investigate model performance, namely the Confusion Matrix (CM), accuracy score (Equation 2), and F1-Score (Equation 3) whose description is detailed in Reis-Pereira, Tosin et al. (2022). Furthermore, precision (the fraction of correct positive predictions out of all positive predictions, Equation 4) and recall (fraction of correct positive predictions out of all observed positive samples, Equation 5) were also computed using the following formula, where true positive, false positive, false negative, and true negative values are denoted by TP, FP, FN, and TN, respectively:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_N + F_P} \quad (2)$$

$$F1\ score = \frac{2 * T_P}{2 * T_P + F_P + F_N} \quad (3)$$

$$Precision = \frac{T_P}{T_P + F_P} \quad (4)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (5)$$

All the computational analyses were performed in the Jupyter Notebook software using the libraries 'Matplotlib' (Ari and Ustazhanov 2014), 'numpy', 'pandas' (Betancourt, Chen et al. 2019), and 'Scikit-learn' (Pedregosa, Varoquaux et al. 2011).

# 3. Results

## 3.1. Observational-based phenotyping of leaflets symptomatology over time

### 3.1.1. PCR validation

Tomato plants were inoculated with Pst and Xeu bacteria, respectively. After spectral analysis, leaf samples from each treatment were tested for the presence of these bacteria. Proper controls from samples known to be positive and negative for Pst and Xeu bacteria were included to confirm the assay results. After the colony PCR reaction, the amplified products were separated by agarose gel electrophoresis and visualized under UV light. The PCR results showed bacteria-specific bands for each bacteria species, namely a 200-base pair (bp) fragment of Pst, and a 713 bp fragment for Xeu, indicating that Pst and Xeu bacteria were present in each inoculation treatment group. No PCR amplification was observed from samples collected from healthy leaves.

### 3.1.2. Visual and hyperspectral phenotyping

Tomato plants infected with Pst bacteria showed the first visual typical chlorotic symptoms mostly between four and five days after infection (DAI). These spots evolved into necrotic lesions at six to seven DAI. In turn, chlorotic lesions in samples inoculated with Xeu mainly developed among twelve to fifteen DAI, only evolving to the necrotic stage at seventeen to eighteen DAI. Pst-infected plants died 12 DAI (Figure 4).

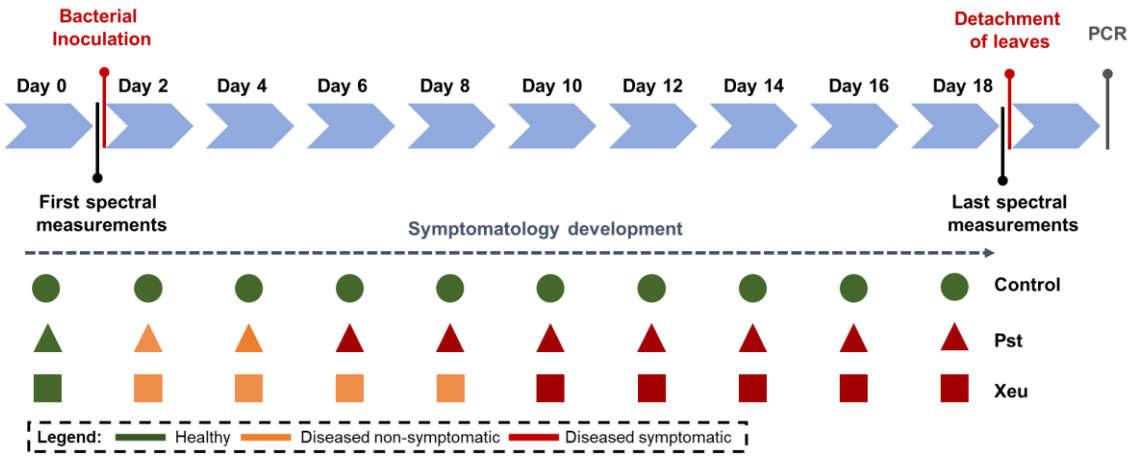



**Figure 4** Observational-based phenotyping of leaflet symptomatology over time. Spectral measurements were performed before bacteria inoculation (Day 0), until day 15 (*Pseudomonas syringae* pv. *tomato* diseased leaflets), and 18 days after infection (Control and *Xanthomonas euvesicatoria* diseased leaflets). In the last measurement date, tomato leaflets were detached from each diseased plant and isolated in different bags for later performing the bacteria isolation assay.

Table 2 presents the dataset structure used, composed of 3478 spectral point measurements, from which 1377 (39.6%) observations correspond to the healthy class. Of these, 1215 (34.9%) assessments belonged to Control leaflets, 81 to measurements performed on Pst leaflets before bacteria inoculation, and 81 to captures made on Xeu leaflets also before bacterial infection. Spectral records performed before symptom appearance reached the value of 844 (24.3%), where 101 (2.9%) measurements belonged to non-symptomatic leaflets inoculated with Pst, and 743 (21.4%) to leaflets inoculated with Xeu bacteria. Lastly, after symptom development, 1257 (36.1%) spectra were captured (866 – 24.90% – from symptomatic Pst leaflets, and 391 – 11.24% – from Xeu symptomatic tissue). Class imbalance is observed due to the disease infection process's dynamic, resulting in symptoms appearing throughout the measurements dates at different rates (Table 1). Spectral assessments were performed during 18 DAI for Control and Xeu leaflets. For Pst, the process was only made until 15 DAI because the plants presented high-stress levels, and leaf dehydration after this date, interfering with the spectral signal recording (Figure 1, Table 1).

Hyperspectral signatures captured in healthy leaflets showed the typical spectral behavior of healthy green tissues. On the other hand, spectral assessments belonging to disease leaflets (both with Pst and Xeu bacteria) presented deviations in their format (Figure 5). Thus, a more detailed analysis was performed for these measurements to evaluate the spectral modifications caused by the different bacteria, resulting in a higher number of classes in the study. Spectra signatures belonging to Pst inoculated samples had a more distinct spectral curve (for both, non-symptomatic and symptomatic stages) compared to the healthy measurements, showing higher intensity on the wavelength ranges of approximately 430 to 520 nm, and 560 to 680 nm. Nevertheless, the lower intensity was captured from 710 to 800 nm (Figure 6 A, B). Xeu-inoculated tissues also displayed modification in their spectral signature in these regions. The intensity measured in the first two spectral intervals was marginally higher than the one captured on healthy leaflets. However, a more evident variance was observed in the 710 to 780 nm range (Figure 6 A, C). When measurements belonging to disease samples were compared, the data showed differences between the samples infected with the different etiological agents. Pst measurements (for both non- and symptomatic stages) demonstrated greater intensity in the ranges of 430 to 520 nm, and 560 to 680 nm, but lower intensity in the 710 to 800 nm interval (Figure 6 A, D).

**Table 2** Spectral data characterization of the measurements randomly performed on tomato leaflets (healthy, diseased with *Pseudomonas syringae* pv. *tomato* – Pst –, and diseased with *Xanthomonas euvesicatoria* – Xeu), showing the number of assessments made by class and date.

| Days after Infection (DAI) | Non-inoculated | Inoculated classes | | | |
|---|---|---|---|---|---|
| | | Non-symptomatic | | Symptomatic | |
| | | Xeu | Pst | Xeu | Pst |
| 0 | 243* | 0 | 0 | 0 | 0 |
| 3 | 81 | 81 | 81 | 0 | 0 |
| 4 | 81 | 81 | 17 | 0 | 64 |
| 5 | 81 | 81 | 3 | 0 | 78 |
| 6 | 81 | 81 | 0 | 0 | 81 |
| 7 | 81 | 81 | 0 | 0 | 81 |
| 8 | 81 | 81 | 0 | 0 | 81 |
| 10 | 81 | 71 | 0 | 10 | 80 |
| 11 | 81 | 63 | 0 | 18 | 80 |
| 12 | 81 | 33 | 0 | 48 | 80 |
| 13 | 81 | 28 | 0 | 53 | 81 |
| 14 | 81 | 28 | 0 | 53 | 79 |
| 15 | 81 | 34 | 0 | 47 | 81 |
| 17 | 81 | 0 | -- | 81 | -- |
| 18 | 81 | 0 | -- | 81 | -- |
| **Total** | **1377** | **743** | **101** | **391** | **866** |
| **(n=3478)** | 39.6% | 21.4% | 2.9% | 11.2% | 24.9% |

* Including all plants. After day 0, only Control plants belong to this class.

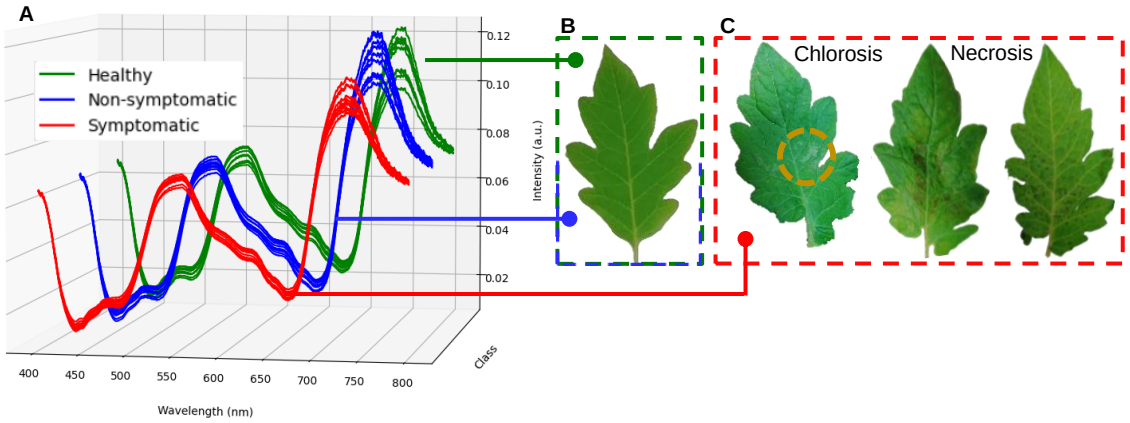



**Figure 5** Mean normalized spectra of healthy, non-symptomatic, and symptomatic leaflet measurements for the first ten measurements performed (12 DAI, A). Healthy and non-

symptomatic infected leaflets presented equal visual phenotype (B). With infection evolution over time, chlorotic symptoms started to appear and later turned into necrotic lesions (C).

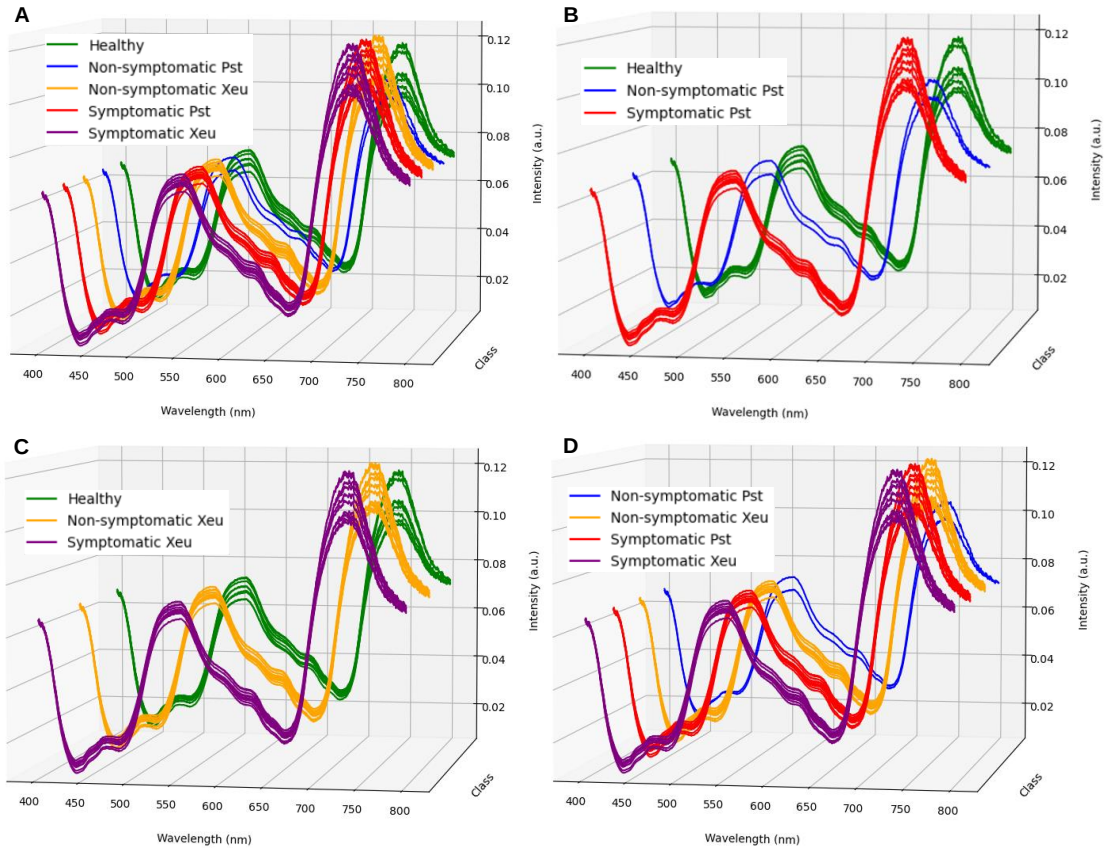



**Figure 6** Mean normalized spectra per class in study (i.e., healthy, non-symptomatic, and symptomatic *Pseudomonas syringae* pv. *tomato* – Pst – leaflet measurements, and non-symptomatic *Xanthomonas euvesicatoria* – Xeu – assessments) for the first ten measurements performed (12 DAI).

## 3.2. Hyperspectral sensing-based phenotyping of leaflets symptomatology over time

### 3.2.1. Reducing the spectral dataset dimensionality

A Linear Discriminant Analysis (LDA) was performed to reduce the dimensionality of the normalized dataset, organizing the spectral observations in a new space as the maximum linear separability possible. LDA results were plotted and showed spectral separability between the different classes studied (Figure 7 A). It was possible to see an evolution pattern through LD 1 for spectral data belonging to healthy, and Pst diseased leaflets regardless of whether they exhibit symptoms or not (Figure 7 A, B). In turn, healthy and Xeu-diseased leaflets (including, non- and symptomatic data) presented a

spectral separation gradient through LD2 (Figure 7 A, C). When data of diseased leaflets infected with distinct bacteria were compared, it was possible to observe a divergence gradient between the LD1 and LD2, especially at the symptomatic stage (Figure 7 A, D). Since data presented a non-linear characteristic, overlapping between classes was observed. Thus, these findings demonstrated the efficacy of the LDA technique for reducing the dataset dimensionality to the most important features. LDA's DR results were, then, applied in the following steps of the modeling process helping in the classification task and avoiding overfitting.

The most relevant wavelength variables for LD1 were assessed based on their coefficients, equaling 44 features. These variables were mostly located in the blue-green and red VIS regions of the electromagnetic spectrum (blue - 434.9, 435.72, 438.17, 438.58, 440.21, 441.44, 442.67, 443.08, 445.53, 445.94, 448.4, 448.81, 494.6 nm; green - 503.74, 508.74, 527.53 nm; red - 556.09, 562.0, 562.84, 590.37, 607.82, 609.1, 611.24, 618.5, 643.36, 650.24, 673.97, 680.02 nm), coinciding with the wavelength absorption range of chlorophylls (430 to 480 nm, and 640 to 700 nm), and carotenoids pigments, namely β-carotenes (whose primary and secondary absorption peaks are respectively located at 450 to 480 nm, and 600 to 650), and xanthophylls (520 to 580 nm) (Figure 8). This coincides with the action of Pst and Xeu bacteria on tomato leaves' levels of photosynthetic pigments during the infection process.

Other plants whose metabolites are affected by these two bacteria also have their absorption spectrum coinciding with the selected wavelengths of LD1, namely some phenolic compounds (e.g., flavonoids, 400 to 500 nm), and composts derived from chlorophylls decomposition, namely pheophytins (400 to 500 nm, and 600 to 700 nm) (Figure 8).

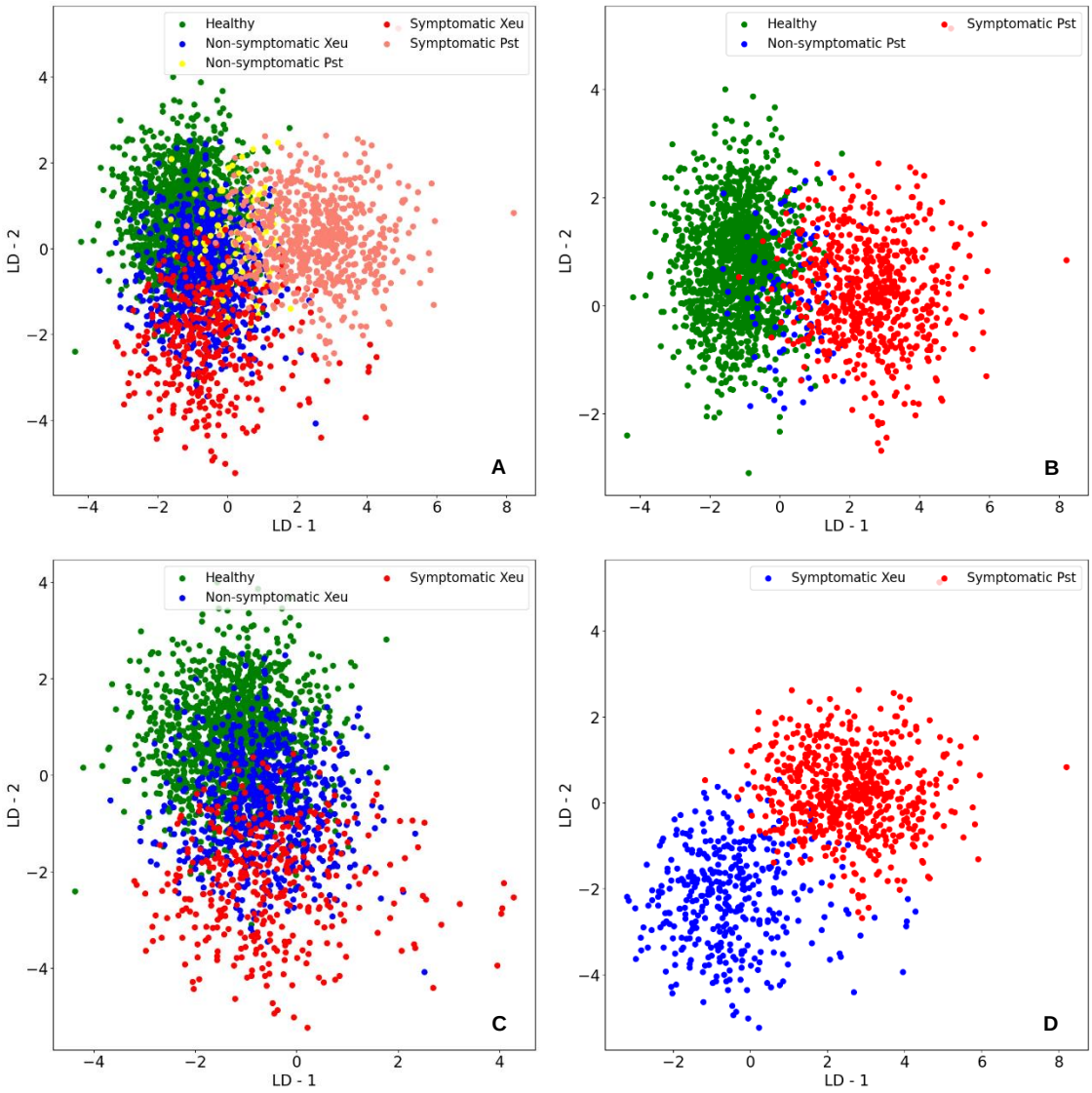



**Figure 7** Scatter plots of the outcomes of the application of Linear Discriminant Analysis on the normalized data, for Linear Discriminant 1 (LD1) and Linear Discriminant 2 (LD2).

Applied predictive classification modeling was, then, performed using the LDA-reduced normalized data (including all classes: i) healthy; ii) non-symptomatic diseased Pst leaflets; iii) non-symptomatic Xeu samples; iv) symptomatic inoculated Pst tissues; v) symptomatic Xeu observations) and an SVM algorithm with a Radial Basis Function (RBF) kernel. The model was trained using 70% (2434) of the spectral observations (randomly selected), and then, it was validated using the remaining 30% (1044) of the observations (test set), and the complete dataset. The test set comprised 413 healthy samples, 30 non-symptomatic Pst disease leaflets, 223 non-symptomatic Xeu, 260 symptomatic Pst observations, and 118 symptomatic Xeu.

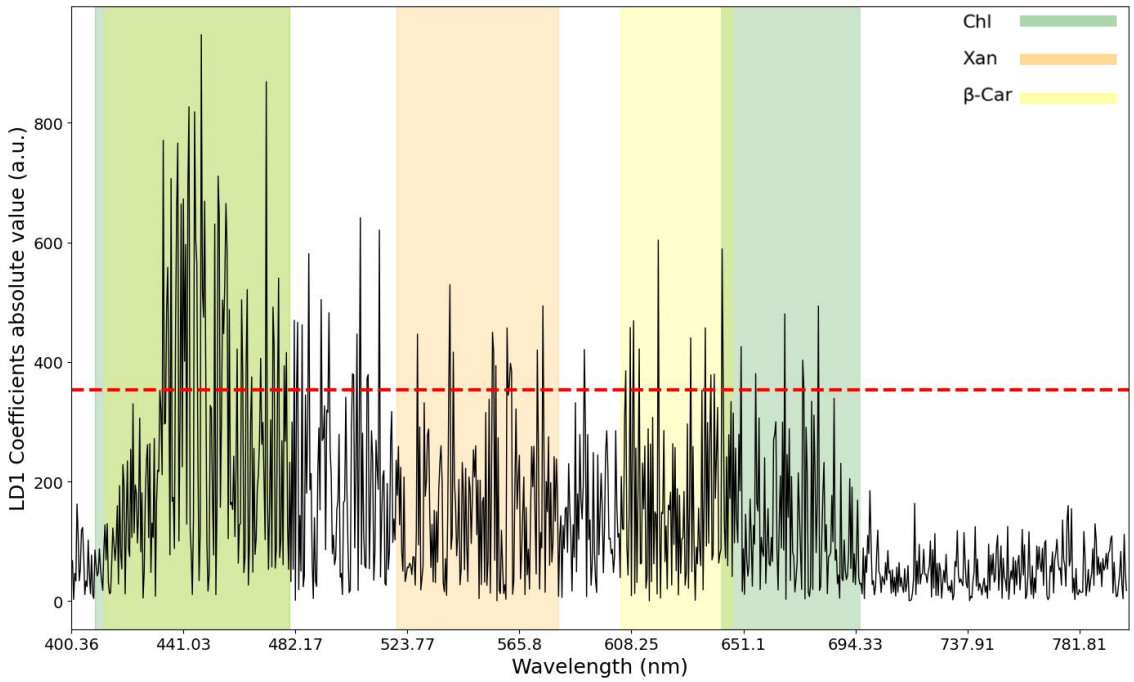



**Figure 8** Absolute values of the coefficients results of Linear Discriminant Analysis for Linear Discriminant 1. Forty-four spectral wavelengths were identified as relevant when variable weights were computed. These variables coincided with the absorption spectra of different photosynthetic pigments, namely chlorophylls (Chl, highlighted in green for chlorophyll), and carotenoids (β-carotenes, β-car, highlighted in yellow; and xanthophyll's, Xan, highlighted in orange).

The developed model performed well for both the test set and the complete dataset. The model achieved an accuracy of 0.85 for the test set and 0.86 for the complete dataset, indicating that it can correctly classify most of the measurements (Table 3, Figure 9). Furthermore, it demonstrated high metric values (precision, recall, and F1-score) for all the classes, indicating that it can identify both healthy and infected measurements. In detail, higher precision, recall, and F1-score values were found for the healthy and non-symptomatic Pst leaflets measurements (Table 3). This shows that the model more easily predicted spectral assessments belonging to these classes. Nevertheless, it showed more difficulties in classifying measurements of Xeu inoculated leaflets, especially those captured before symptom appearance (indicated by lower metric scores). It is important to note that the model's performance was consistent across both the test set and the complete dataset, indicating that the model is robust and can be used to classify new spectral samples accurately.

Model predictions for the non-symptomatic Pst class did not present any misclassification in the test set. In the complete dataset, the model accurately predicted

96% of the spectral measurements but missed 1% of the predictions, which it classified as assessments made on non-symptomatic leaflets infected by Xeu (Figure 9). Symptomatic spectral captures performed in Pst diseased leaflets were correctly categorized in 94% of the cases for both the test and complete sets. Nevertheless, the model mistakenly classified these assessments as non-symptomatic Xeu observations in 4% and 3% of the cases, and as healthy samples in 2% when the test set and complete dataset were used, respectively. Predictions of Xeu spectral assessments were more challenging to the model, presenting a higher number of wrong classifications in the non-symptomatic class than in the remaining classes studied. In fact, the model successfully classified 77% of the measurements of this class in the test set, and 78% when all data was used. However, it attributed 11% and 10% of the measurements as healthy, 5% and 8% as symptomatic diseased Xeu leaflets assessments, 3% as non-symptomatic inoculated Pst observations, and 2% as symptomatic Pst captures, when the test set and complete dataset were used, respectively. The model showed more efficacy in identifying symptomatic Xeu leaflets measurements, predicting 83% of these samples in the test and complete datasets. In terms of missed classifications, it predicted 6% and 5% of the assessments as non-symptomatic, 3% and 2% as healthy, 3% and 1% as non-symptomatic spectral captures of Pst infected leaflets, and 1% and 2% as symptomatic Pst, in the test set and complete dataset, respectively (Figure 9 A, B).

**Table 3** Performance metrics for the classification SVMs-based model using all the data (train and test set – All), only the train set (Trn) and only the test set (Test).

| Class of leaflets status | Precision | | | Recall | | | F1-score | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Trn* | *All* | *Test* | *Trn* | *All* | *Test* | *Trn* | *All* | *Test* | *Trn* | *All* | *Test* |
| *Healthy* | 0.86 | 0.85 | 0.84 | 0.89 | 0.89 | 0.88 | 0.88 | 0.87 | 0.86 | 0.89 | 0.85 | 0.88 |
| *N Sym. Pst* | 0.97 | 0.96 | 0.94 | 0.86 | 0.90 | 1.00 | 0.91 | 0.93 | 0.97 | 0.86 | 0.90 | 1.00 |
| *N Sym. Xeu* | 0.78 | 0.78 | 0.77 | 0.74 | 0.74 | 0.74 | 0.76 | 0.76 | 0.75 | 0.74 | 0.74 | 0.74 |
| *Sym. Pst* | 0.94 | 0.94 | 0.94 | 0.95 | 0.94 | 0.93 | 0.95 | 0.94 | 0.93 | 0.95 | 0.94 | 0.93 |
| *Sym. Xeu* | 0.83 | 0.83 | 0.83 | 0.80 | 0.79 | 0.77 | 0.81 | 0.81 | 0.80 | 0.79 | 0.79 | 0.77 |
| Weighted Avg ± s.d | 0.86 ± 0.07 | 0.86 ± 0.07 | 0.85 ± 0.07 | 0.86 ± 0.07 | 0.86 ± 0.08 | 0.85 ± 0.10 | 0.86 ± 0.07 | 0.86 ± 0.07 | 0.85 ± 0.08 | 0.86 ± 0.07 | 0.86 ± 0.08 | 0.85 ± 0.10 |

SVMs – Support Vector Machines, Trn – Train, N Symp. – Non- symptomatic, Sym. – Symptomatic, Avg. – Average, s.d. – standard deviation

For the complete dataset prediction, we investigated the number of misclassifications per class and date (Figure 10). As expected, the observed tendency for healthy spectral assessments showed a regular number of observations per date

(81). Nonetheless, the developed model categorized more samples than the true value per date, except for 7, 13, 17, and 18 DAI. On the other hand, the spectral model consistently underfit the infected Xeu leaflets, regardless of whether they exhibit symptoms or not (Figure 10 A).

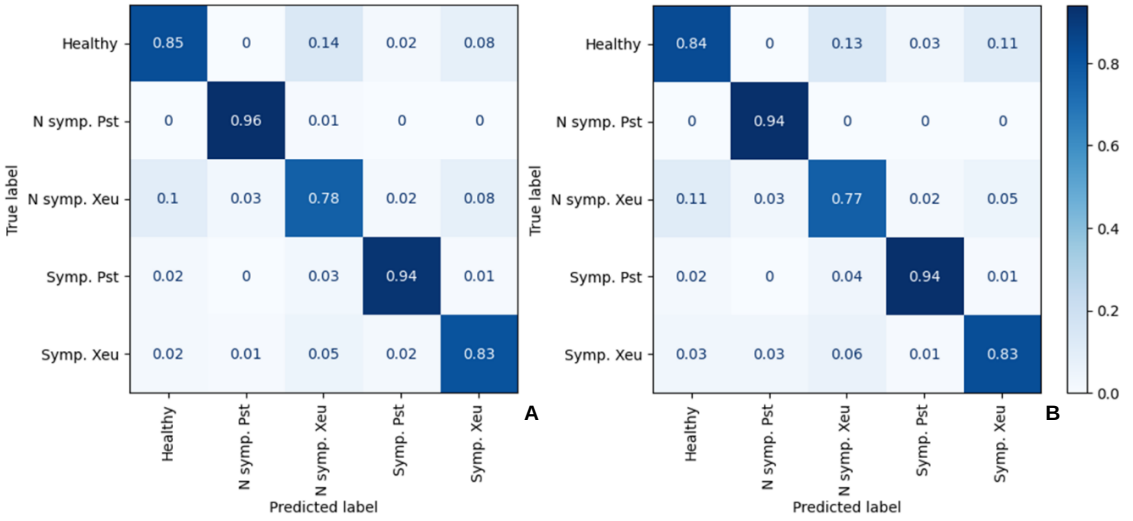



**Figure 9** Confusion Matrix of the percentage of predicted samples for each class (column) that were correctly classified for each true class (row), for the complete (a) and test (b) sets. (Legend: N Symp. – Non-symptomatic, Sym. – Symptomatic).

In plants inoculated with Xeu, discrepancies between observed and predicted classes are more evident in the non-symptomatic Xeu class in the observations recorded up to 10 days after infection. During this period, which included seven measurement dates of the non-symptomatic Xeu class, 53 observations were recorded below the predicted value of the developed model. In contrast, the healthy class accumulated 47 observations above the predicted value during the same period. Furthermore, according to the confusion matrix results (All data), 10% (148) of the non-symptomatic Xeu observations were misclassified as healthy. Considering the period up to 10 days after infection (data not shown), out of the 150 observations wrongly classified as healthy, 100 were from the non-symptomatic Xeu class. These results indicate that in the early stages of Xeu-induced disease infection, the symptoms developed in the plant leaflets are not strong enough for the developed model to distinguish them from healthy observations efficiently. Therefore, the non-symptomatic Xeu class, compared to other tested classes, exhibits the lowest model performance indicators (all data: accuracy 0.74, precision 0.78, recall 0.74, and F1-score 0.76). For the non-symptomatic stage, the actual observations presented a stable pattern until 8 DAI, and after a sharp drop was observed until 13 DAI, where the rate of infected leaflets increased up to 65%. A stable number of observations was maintained until 15 DAI, after which a period of exponential increase in observed

symptomatic spectral measurements was registered. After this day, all leaflets were symptomatic. The model was rigorous in discriminating non-symptomatic Xeu leaflet measurements after 9/10 DAI, presenting a percentage of error inferior to 10% (correctly classifying 64 of the 71 observations) when about 90% of the sampled leaflets (71 of the initial 81 assessments) still didn't show any typical symptoms of the disease (Figure 10 B).

For the prediction of the Pst disease samples, the non-symptomatic phase was very similar for both observed and predicted. Nevertheless, the prediction of the symptomatic phase showed irregularities between the five and seven days (corresponding to the dates were necrosis appeared). Is possible to observe that most of the Pst inoculated leaflets (79%) started to show the first symptoms of the disease 4 DAI. The number of symptomatic sampled leaflets increased until 6 DAI, where all the leaflets assessed were symptomatic (Figure 10 C).

## 4. Discussion

Plant infectious diseases are critical in agriculture and food security, impacting crop yields and quality. Understanding and effectively managing them is crucial for more sustainable agriculture, based on more preventive measures and early diagnosis.

The suitability of spectral phenotyping based on hyperspectral spectroscopy point-of-measurement (HS-POM) for diagnosing bacterial infectious diseases in tomato plants, namely bacterial speck and spot, was evaluated. In this approach, light penetrates the leaflet tissue and undergoes internal reflections, before ultimately being redirected to the spectrometer via a central fiber optics pinhole. This method ensures that all light reaching the sensor interacts with the leaf tissues, thereby maximizing the spectral information from all internal tissues, including any changes caused by the interaction between the host and bacteria.

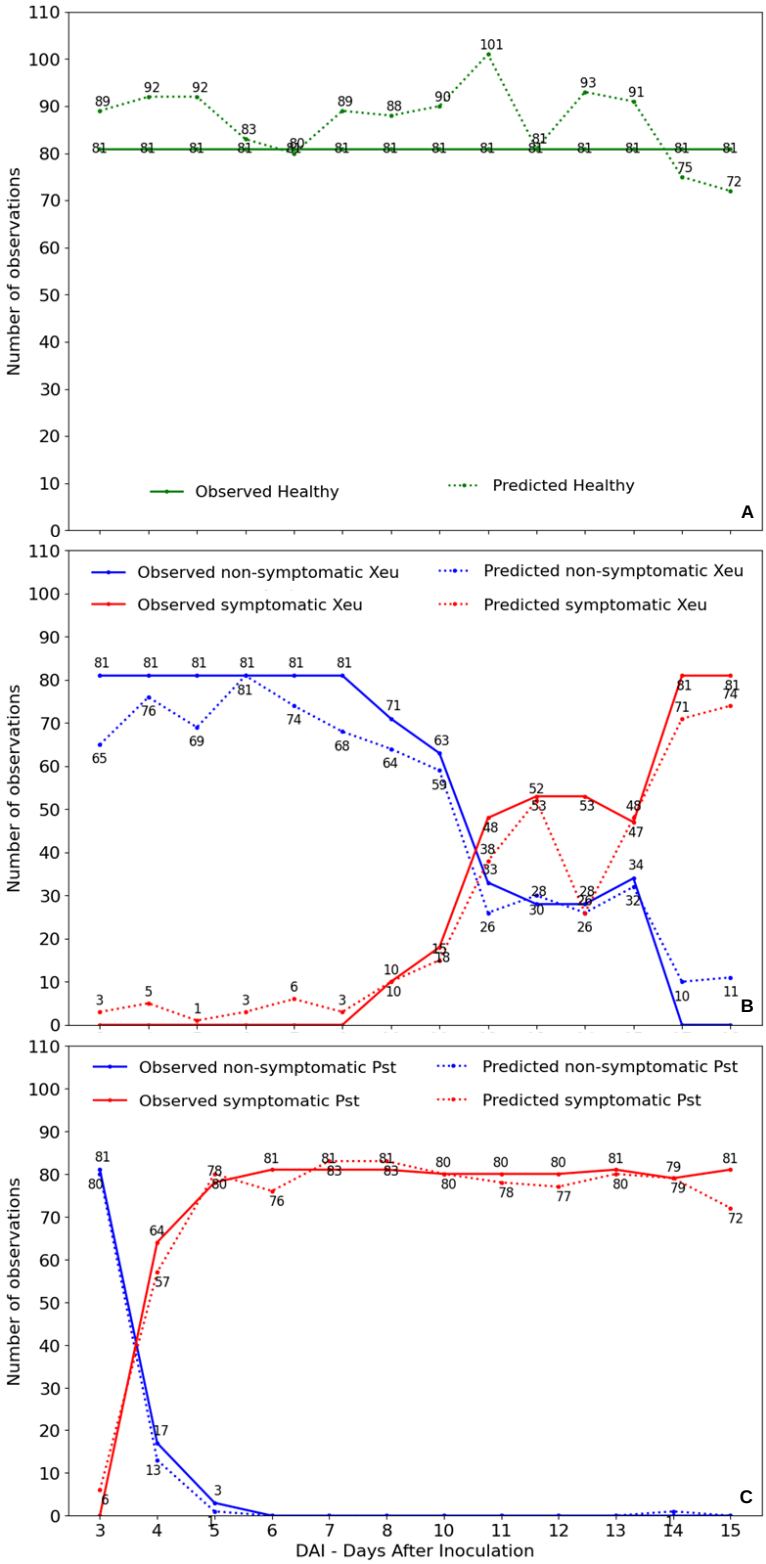



**Figure 10** Number of observed and predicted samples by date of measurement for healthy (A), *Xanthomonas euvesicatoria* diseased (B), and *Pseudomonas syringae* pv. *tomato* diseased (C) leaflets' assessments.

An applied predictive model integrating an SVM algorithm showed the capacity to accurately classify healthy and diseased tomato leaflets at various stages of disease development (specifically healthy, non-symptomatic diseased Pst, non-symptomatic disease Xeu, symptomatic Pst, and symptomatic Xeu). Even before symptom appearance, it showed a classification accuracy of 74% for Xeu and 100% for Pst diseased leaflets measurements, and a weighted average accuracy, precision, recall, and F1-score of 85%.This model was, thus, capable of categorizing healthy, disease (both non-symptomatic and symptomatic), and disease leaflet tissues infected with distinct bacteria species (both before and after symptom appearance), being coherent with visual phenotyping and PCR results. These outcomes, thus, demonstrate the suitability of this technique for performing an early disease assessment and class distinction (according to the phytosanitary health status, and type of pathogen responsible for the infection). This is extremely valuable since crops in the field are generally exposed to variable environmental and phytosanitary conditions and vulnerable to different types of abiotic and biotic stresses (which may cause similar visual lesions, difficult to distinguish by the naked eye). Also, bacterial spot and speck of tomato can develop in 6 to 14 days, depending on several factors (e.g., environmental conditions, pathogen strain, infection severity, inoculum concentration, and the susceptibility of the plants' variety) (Horst 2013, Borkar and Yumlembam 2016), and their spread among several plants in a production field is not immediate and may take time to occur. Thus, early diagnosis is crucial to prevent disease spread, promote preventive treatments, and lead to environmentally friendly practices, promoting precision agriculture principles.

LDA computation revealed spectral divergence between the different classes in study through LD1 and LD2 and uncovered relevant wavelengths for diagnosing the diseases caused by *Pseudomonas syringae* pv. *tomato* (Pst), and *Xanthomonas euvesicatoria* (Xeu). These were mostly located in the blue-green and red visible regions of the electromagnetic spectrum, corresponding to chlorophyll (mainly: 430 to 480 nm, and 640 to 700 nm) and carotenoid pigments' absorption spectra (i.e., 450 to 480 nm, 520 to 580 nm, and 600 to 650 nm), indicating modifications in the photosynthetic pigment's levels throughout the infection process. These findings are aligned with the impact of both bacteria species on host leaves' pigments values during infection, which start prior to symptoms appearance and became more pronounced with the formation of chlorotic and necrotic lesions. In this medium / late stages of infection, the breakdown of chlorophyll, in particular, can result in a subsequent accumulation of pheophytins (brown pigments, whose maximum absorption peak is located at 660-670 nm, and secondary

peak around 430-450 nm), which also affect plant spectral behavior (Bhandari, Wang et al. 2015). Also, spectral divergences in the 700 to 800 nm range may indicate that structural components of leaves are affected during the infection process, resulting in the degradation of leaf structures along disease development. Spectral divergence between diseased leaves infected by different bacteria may be related to the production of specific molecules by each pathogen, which may affect the host spectral signature. As an example, Pst produces a phytotoxin called coronatine which alters chlorophyll fluorescence (by modifying the photosystem II – PSII) and can affect the absorption and scattering of light by plant tissues, leading to modifications in the spectra (Zhang, He et al. 2021). In turn, the host plant can activate different defense responses when in contact with distinct pathogens, triggering a series of biochemical and molecular responses, which also promote spectral modifications in the visible wavelength ranges. An example are phytoalexins (e.g., flavonoids), whose production was hypothesized to be related to increased plants' spectral reflectance in the VIS range (Leucker, Wahabzada et al. 2016).

Hence, the present research findings demonstrate that HS-POM holds promise as an effective, fast, and cost-effective overtime method for early diagnosis of two bacterial infections caused by distinct pathogen species in vivo tomato plants, and for unraveling specific host-pathogen spectral dynamics. In the future, it is advisable to conduct further analysis, entailing the expansion of the dataset under study, test various values for SVM algorithm parameters, and enhance the modeling algorithms, among other potential approaches. This study corroborates previous research performed by our team using HS-POM for the early detection of bacterial tomato spot caused by Xeu bacteria. The spectral response properties of disease tomato leaves presented a divergent behavior when compared to healthy tissues, even before symptom appearance. This tendency was more evident in the absorption regions of photosynthetic pigments (namely, chlorophyll). A Principal Component Analysis (PCA) allowed the identification of relevant discriminative wavelengths at approximately 454-654 nm (Reis-Pereira, Martins et al. 2021), coinciding with the wavelengths identified by the LDA approach.

Other studies also demonstrated the potential of hyperspectral data and SVM-based classification modeling for disease diagnosis, presenting similar model evaluation metrics. As an example, Cen et al. (2022) studied the possibility of early detection of bacterial wilt in tomato by applying a portable hyperspectral spectrometer. Their model combined Genetic Algorithms and SVM and achieved overall accuracies (OA) of 90.7% in the distinction of healthy and symptomatic tissues. Tomaszewski et. al (2023) also

demonstrated the suitability of hyperspectral measurements and machine learning for the early detection of anthracnose, bacterial speck, early blight, late blight, and septoria leaf, using a temporally-aggregated approach. When all the data were analyzed, the researchers found that the best-quality classification approach (integrating a Ridge classifier) presented an F1 score ranging from 0.71 to 0.95 (0.84 average) for the period of the first two weeks from inoculation. Despite being possible to find research diagnosing different types of biotic stress agents in the same assay, they are usually more related to fungi identification. Scarce results can be retrieved for studies comparing the assessment of diseases caused by different types of bacterial species.

Besides tomato crop studies, hyperspectral measurements were also valuable to achieve disease diagnosis in several plant species with agronomic interest. For instance, Rumpf et al. (2010) studied the suitability of hyperspectral reflectance, SVM, and vegetation indexes (VIs) for detect and classify diseases on sugar beet leaves (namely, Cercospora leaf spot, leaf rust, and powdery mildew). Early differentiation between healthy and inoculated plants, as well as among specific diseases was achieved using SVM, registering accuracy values ranging from 65 to 90%. When data belonging to healthy and diseased leaves (including all the samples affected by the three different pathogens) was used, the classification model achieved an accuracy higher than 86%. Furthermore, Tian et. al (Tian, Xue et al. 2021) also proved the efficacy of spectroscopy and machine learning techniques for rice leaf blast infection from non-symptomatic to mild stages. An approach integrating an SVM algorithm achieved maximum classification accuracies of over 80% and 83% for the early infection stage of the 2018 and 2019 experiments.

The desirable possibility of applying hyperspectral data for in-field detection and classification of diseases was also proved. Deng et al. (2019) also demonstrated the possibility of applying hyperspectral reflectance in-field detection and classification of citrus Huanglongbing disease. They developed an SVM learner which achieved 90.8% accuracy in healthy, asymptomatic, and symptomatic discrimination. Our team, likewise demonstrated the capability of using HS to diagnose in situ bacterial canker disease, caused by another *Pseudomonas* pathovar, specifically *Pseudomonas syringae* pv. *actinidiae* (also known as Psa). asymptomatic and symptomatic leaves were successfully discriminated through the computation of several modeling approaches involving different feature selection techniques, as well as multivariate analysis methods and machine learning algorithms. The best predictive classification model for discriminating the bacterial kiwi canker disease showed an overall accuracy of 0.85, with an F1-score (Reis-Pereira, Tosin et al. 2022). These findings suggest that hyperspectral

data can be successfully used to predict plant diseases both indoor and infield conditions, caused by different etiological agents (e.g., fungi, bacteria, and virus), in both herbaceous and woody crops. Despite these encouraging findings, it is important to highlight that comparison between different research can be challenging due to the pathogens in study (e.g., generally disease detection using HS is mostly performed for fungal infections), host-pathogen specific interactions, number of samples used, number of classes analyzed, moment of disease assessment (before or after symptoms appearance, in a specific date or overtime), environmental and experimental conditions on the moment of data acquisition, among others. Thus, future studies using tomato plants should be performed to evaluate the efficacy of this approach.

In summary, point-of-measurement Hyperspectral Spectroscopy devices combined with applied predictive models seem to be suitable for spectral phenotyping of bacterial-infected tomato leaflets. Nevertheless, HS-POM approaches as plant disease diagnostic methods are still in a very initial phase of development, and their Technology Readiness Levels (TRLs) must be improved. Standardized protocols for hyperspectral data acquisition should be developed aiming to uniformize the diagnosis processes and reduce noise and undesired spectral interferences. Also, more research on different host-pathogen interactions must be performed. Classification models developed under controlled conditions can be highly effective and constitute an important step for improving and maturing the diagnosis process. In fact, these models usually can detect symptoms earlier than in field assays (since optimal conditions for bacteria development, dissemination, and infection can be recreated), making the process faster and specific to the host-pathogen in study. Hence, the more challenging in-field application of HS-POM for disease diagnosis, posing additional complexities due to sensing system configurations (e.g., light source, probe position, among others), can be established and improved.

Future studies must be conducted to complement these gaps and validate the application of this technique as a suitable tool for accurately predicting different host-pathogen interactions and their impact on the crops' spectral signature. Further methodological developments are necessary to address these challenges and enhance the suitability of HS-POM for real-time disease monitoring and precision agriculture systems. Moreover, the implementation of feature selection techniques and dimensionality reduction approaches can help identify relevant wavelengths for distinguishing crop diseases, making possible the development and production of more cost-effective multiband sensors. These devices can be integrated into different

platforms, enabling spectral data acquisition at different levels, such as leaf, single-plant, and canopy scales.

## 5. Conclusion

The present research explored the application of in-vivo POM hyperspectral spectroscopy combined with applied predictive modeling to classify bacterial leaf diseases in tomato crop, caused by *Pseudomonas syringea* pv. *tomato* and *Xanthomonas euvesicatoria*. Healthy leaves showed a characteristic spectral signature of green and photosynthetically active vegetation, while symptomatic leaves presented differences in their spectral signature in the VIS region. Spectral differentiation between healthy and diseased leaves was observed, even in the early stages of the infection process, when diseased samples didn't present any visual symptom (asymptomatic stage). Furthermore, plants inoculated with Pst bacteria also revealed a divergent spectral behavior from the ones infected with Xeu, indicating that this approach may be suitable for differentiating the etiological agents. Colony PCR also validated the effectiveness of the infection process for each sample group. The developed model revealed a classification accuracy for the test set of 100% for Pst disease leaflets without any visual symptom, and of 74% for Xeu disease leaflets also in a non-symptomatic stage of infection. The developed model achieved a weighted average accuracy, precision, recall, and F1-score of 85% for the test set. These findings strength the applicability of applied predictive classification modeling using HS-POM to early detect bacterial crop diseases. Nevertheless, complementary, and additional studies are recommended to unravel the host-pathogen interactions and their impact on the crop spectral signature. More economic, multiband devices can be developed hereafter considering the features selected for crop disease discrimination. Thus, different agronomic tasks (including mapping, monitoring, scouting, and treatment of plant diseases) can be performed more accurately with this methodology, fulfilling the precision agriculture concept. Spectroscopy sensors can also be mounted on diverse platforms, creating different functioning measurement systems, which can assess spectral data on distinct levels (namely, leaf, single-plant, and canopy scale).

## Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Funding**

This work is partially financed by National Funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots, with reference PTDC/ASP-HOR/1338/2021.

**Acknowledgments**

Mafalda Reis-Pereira was supported by the Fundação para a Ciência e Tecnologia (FCT) fellowship with the reference SFRH/BD/146564/2019.

# Case Study 5

**Reis-Pereira, M.**; Mazivila, S. J.; Tavares, F.; Santos, F. N. d.; Cunha, M. Hyperspectral POM sensing in combination with DD-SIMCA is adapted to the early diagnosis stage of plant diseases using control samples as a target class and MCR-ALS as a means to retrieve pure profiles of healthy and disease tissues.

Paper submitted on 22nd December 2023

Classification according to journal: Original Research Article

# Early plant disease diagnosis through hyperspectral point-of-measurement data coupled to DD-SIMCA as one-class classification and multivariate curve resolution

Mafalda Reis Pereira[1,2], Sarmento J. Mazivila[3], Fernando Tavares[4,5], Filipe Neves dos Santos[2], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal

[3] Centro de Investigação em Química da Universidade do Porto (CIQ-UP), DGAOT, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

[4] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

*** Corresponding author:** Mário Cunha mccunha@fc.up.pt

## Highlights

- DD-SIMCA as a one-class classifier was applied to Vis-NIR spectroscopic data for healthy tissue authentication.
- DD-SIMCA classified spectroscopic data measured on healthy tissues as target class members.
- DD-SIMCA identified biological data measured on diseased tissues as non-target class members.
- MCR-ALS successfully retrieved pure Vis-NIR spectral profiles of healthy and biological diseased tissues.

## Abstract

This contribution proposes a promising non-destructive methodology for the early diagnosis of bacterial diseases in tomato plants leveraging hyperspectral point-of-measurement (POM) data acquisition and chemometric processing tools utilizing Data-

Driven Soft Independent Modeling of Class Analogy (DD-SIMCA) and Multivariate Curve Resolution − Alternating Least-Squares (MCR-ALS). The present study aims to conduct a classification task to deal with the authentication problem leading to detaching the target class of control healthy plant leaflet tissues from a non-target class of plant leaflet tissues inoculated with the bacteria *Pseudomonas syringae* pv. *tomato*, Pst, and *Xanthomonas euvesicatoria*, Xeu at different stages of evolving bacterial disease. These plant-pathogen interactions result in a very broad alternative class and do not form a specific class, making it inappropriate to apply a discrimination task. Thus, hyperspectral POM data collected in healthy tomato leaflets were correctly identified throughout the rigorous DD-SIMCA optimization as members of the target class with 100% sensitivity in the training step. Subsequently, in the validation step, DD-SIMCA successfully differentiated these healthy samples from those inoculated with Pst or Xeu bacteria, even before the manifestation of macroscopic lesions associated with the diseases, detecting changes as early as 72 hours post-bacterial inoculation. The full distance that acts as a classification analytical signal calculated in the process of building the DD-SIMCA model was able to predict the distance value from which non-target class of samples are located farness the acceptance threshold. This classification result indicates a more advanced stage of bacterial infection, reflecting evident spectral modifications resulting from host-pathogen interactions, preceding phenotypical changes. On the other hand, non-target class of samples with higher proximity to the acceptance boundary suggested that they were at earlier stages of infection compared to more distant ones, presenting lower distance values. MCR-ALS constrained analysis allowed the description of the bacterial inoculation process, detecting the impact of Pst bacteria on diseased tissues was observed in the pure spectral bands between 430 and 475 nm, while the influence of Xeu was identified in the pure spectral range from 675 to 800 nm. These findings indicate that the hyperspectral POM technology is sufficiently sensitive to be used in acquiring biological data with suitable chemometric modeling for early disease diagnosis and prompt intervention, leading to sustainable agricultural practices, and ultimately enhancing crop yield and food security.

**Keywords**

Plant disease diagnosis, Hyperspectral spectroscopy, MCR-ALS, Healthy tissue authentication, One-class modeling

## 1. Introduction

Cutting-edge sensing technology for acquiring hyperspectral spectroscopic data (spectrum per sample) is now ubiquitously used to monitor and diagnose plant diseases

(Sankaran, Mishra et al. 2010, Mahlein 2016). An additional effort is being made to apply them at the early stages of pathogen infection, spanning from the incubation period, even when the symptoms are not visible to the human vision. This strategy allows more accurate and targeted plant protection measures (Cheshkova 2022). Currently, hyperspectral point-of-measurement (POM) sensing has already enabled a fast and non-destructive indirect diagnosis of plant diseases at an early stage of infection (Reis Pereira, Santos et al. 2023, Tomaszewski, Nalepa et al. 2023) in contrast to time-consuming and destructive direct diagnostic approaches based on biochemistry assays, which cannot be successfully applied in this early non-symptomatic phase of the infection process.

Hyperspectral spectroscopic datasets acquired throughout the early plant-pathogen interactions are unresolved spectral overlapping profiles (Atanassova, Nikolov et al. 2019), containing a complex mixture of metabolic changes in tissues with the characteristic reflection band at 550 nm (Lowe, Harrison et al. 2017). Different multivariate processing tools have been reviewed to decode useful information on hyperspectral spectroscopic data collected during host-pathogen interactions (Jackulin and Murugavalli 2022). However, the review paper (Jackulin and Murugavalli 2022) interchanged the meaning between classification and discrimination tasks, which has been recently clarified in the relevant literature (Pomerantsev and Rodionova 2021). Selected research reports on bioanalytical applications that couple a variety of cutting-edge sensing technologies with discrimination (Naidu, Perry et al. 2009, Liu, Gu et al. 2015, Fernández, Leblon et al. 2021) and classification (Pereira, Milori et al. 2010, Sankaran, Mishra et al. 2010, Atanassova, Nikolov et al. 2019) models for a reliable crop health-monitoring platform are briefly examined.

Conventional hard partial least-squares – discriminant analysis (PLS-DA) was applied to spectroscopic data containing healthy wheat (representing the control group without inoculation, denoted as class one) and another class of wheat inoculated with three predominant races of *Puccinia striiformis f.* sp. *tritici* at three concentrations level. This analysis was conducted subsequent to the manifestation of symptoms, aimed at differentiating leaves infected and uninfected by stripe rust pathogen (Liu, Gu et al. 2015). Prediction results suggested that when the ratio of the training set to the testing set was 4:1, the model had better recognition of test samples in their respective classes than other PLS-DA models (Liu, Gu et al. 2015). The reasonability of revealed PLS-DA performances in various training-to-testing set ratios (Liu, Gu et al. 2015) is attributed to the distinctiveness of one class, which exhibits a more consistent and compact grouping consisting of healthy wheat samples. Whereas another class is very broad and does not

form a specific class at all. This diverse class encompasses biological tissues damaged in different ways throughout plant-pathogen interactions (Liu, Gu et al. 2015), generating non-linear data that the linear PLS-DA model fails to discriminate effectively (Pomerantsev and Rodionova 2018). From an operational chemometric viewpoint, non-linear discriminant models are recommended when both classes are available in a non-linear case instead of using linear discriminant models like PLS-DA.

In this perspective, a non-linear discriminant model based on a support vector machine – discriminant analysis (SVM-DA) was applied to a dataset composed of tissue samples comprising 1000 healthy pixels and 1000 infected pixels, displaying visible symptoms (Fernández, Leblon et al. 2021) for disease diagnosis. Nevertheless, the effectiveness of a non-linear discriminant model is compromised during the initial phases of the infection, especially when external visual symptoms are absent. In such cases, where the healthy class (control group) lacks crucial information on the composition of another class of samples acquired from plant-pathogen interactions, the non-linear discriminant model faces challenges integrating their profiles during the training phase. Otherwise, the discriminant model will fail because it cannot properly assign new samples that do not belong to any predefined classes in the training phase (Rodionova, Titova et al. 2016).

Alternatively, a one-class classification (OCC) model (Rodionova, Titova et al. 2016) is a wise choice for a rigorous approach (Rodionova, Oliveri et al. 2016), i.e., one which only uses information regarding the target class (healthy samples) and does not utilize any information about the non-target classes (e.g., samples under plant-pathogen interaction), even when the data regarding such extraneous classes are available.

A classical soft independent modelling of class analogy (SIMCA) (Wold and Sjöström 1977) as an OCC was trained using only spectroscopic data of healthy plants (control group) as the target class (Pereira, Milori et al. 2010, Atanassova, Nikolov et al. 2019). The original SIMCA (Wold and Sjöström 1977) was developed in its simplest class-modeling version, where the decision rule relies on residual variance lately called orthogonal distance (Pomerantsev 2008) that belongs to the Q-statistics (Bro and Smilde 2014) further referred to as $q_i$ of a given object $i$ (Jouan-Rimbaud, Bouveresse et al. 1999). This metric is used to delineate an acceptance area, as detailed in the Pirouette_Toolbox guide (Infometrix 2014) used for creating SIMCA models. However, the developed classical SIMCA model (Pereira, Milori et al. 2010, Atanassova, Nikolov et al. 2019) overlooks the statistical leverage scores (Mazivila and Borges Neto 2021), usually represented as $h_i$ for a given object $i$ (Jouan-Rimbaud, Bouveresse et al. 1999)

and recently termed as score distance (Pomerantsev 2008). These approaches focused on the classical SIMCA that does not incorporate the advancements introduced in its subsequent four upgrade versions (Vitale, Cocchi et al. 2023), marking an evolution since its interception as the pioneering model which have evolved over time since the pioneering model (Wold and Sjöström 1977). SIMCA initially produced a description of the target class of objects and, subsequently, detected whether a new object resembles the target class (namely healthy or infected plants at a non-symptomatic stage) or diverges from it (plants exhibiting developed symptoms of the disease) (Pereira, Milori et al. 2010, Atanassova, Nikolov et al. 2019).

The present contribution extends prior studies based on the original SIMCA (Pereira, Milori et al. 2010, Atanassova, Nikolov et al. 2019), with a grounding-breaking perspective, including the application of the data-driven version of SIMCA (DD-SIMCA) (Rodionova, Titova et al. 2016) as a one-class classifier. In addition to the DD-SIMCA with decision rule (Pomerantsev and Rodionova 2020) based on a parallel data-driven estimation (Pomerantsev and Rodionova 2014), multivariate curve resolution – alternating least-squares (MCR-ALS) constrained analysis (Mazivila and Santos 2022) was applied. MCR-ALS bilinearly decomposed the contribution of the hyperspectral Vis-NIR responsive constituents to the entire signal. This facilitated the extraction of pure spectral signatures of either target class (specifically healthy plants or infected plants at a non-symptomatic stage) or not (plants with developed disease symptoms).

In this study, two hyperspectral POM datasets were acquired from two experimental biological assays to validate plant-pathogen interactions during infection and disease phenotyping. These hyperspectral POM datasets were collected in healthy tomato leaflets and diseased leaflets infected by *Pseudomonas syringae* pv. *tomato* (Pst) and *Xanthomonas euvesicatoria* (Xeu) during at least thirteen days after the artificial infection (Figure 1). DD-SIMCA optimization was performed by using only healthy samples (control group) in the training phase to establish a boundary around a class of interest (green line in Figure 1D) that helps to detach the target class of samples of the control (healthy tomato leaflets from other non-target classes (Pst diseased and Xeu-infected tomato leaflets) in the validation phase. Subsequently, MCR-ALS was subjected to each class assigned by DD-SIMCA, aiming to appraise pure profiles of samples, providing valuable additional information that might aid in developing real-time plant disease management strategies. Available results graphically summarized in Figure 1 provide a conceptual representation, anticipating that:

- The first hyperspectral POM data (Figure 1C), acquired from tomato leaflets before the bacteria inoculation assays (Figure 1A), were subjected to DD-SIMCA. This OCC model was trained using only a target class of healthy samples (Figure 1D), which classified all validation samples inside the decision area (green line in Figure 1E) as healthy samples with strong evidence demonstrated through the MCR-ALS retrieval of healthy tomato leaflet pure spectral profile (Figure 1F) that was found to be similar to the validation samples profile (Figure 1G). Outside the decision area in the validation phase, samples might be allocated non-target class of samples (plants inoculated with either Pst or Xeu), as will be demonstrated throughout this paper.

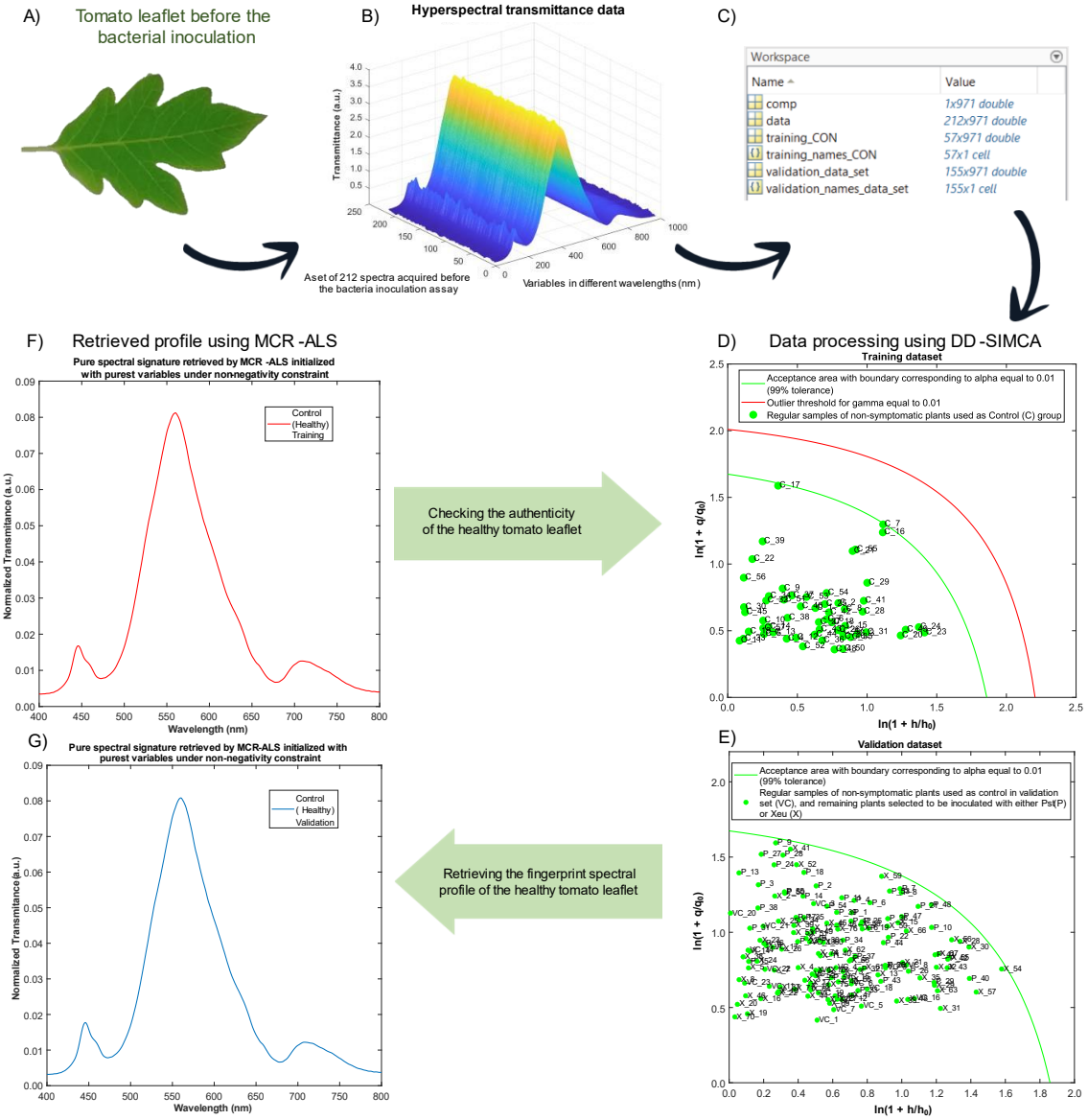



**Figure 1** Analytical flowchart showing that hyperspectral point-of-measurement (POM) was performed in tomato leaflet tissues before bacterial inoculation (A, B). A part of this biological data was uploaded in MATLAB environment as the training set (C), used as a

target class in a Data-driven Soft Independent Modelling of Class Analogy (DD-SIMCA) model to establish the acceptance boundary (D). The remaining healthy samples were used as the validation set to check authenticity, revealing that all samples were in the acceptance boundary (E). In DD-SIMCA each sample can be depicted in the coordinates $ln(1 + h_i/h_0) \quad vs \quad ln(1 + q/q_0)$, together with the two tolerance boundaries (acceptance area and outlier threshold). The fingerprint Vis-NIR spectral profile of the healthy tomato leaflet tissue spectra was then successfully retrieved in the training (F) and validation (G) datasets by Multivariate Curve Resolution − Alternating Least-Squares (MCR-ALS).

## 2. Experimental setup of plant growth and data acquisition

### 2.1. Plant growth and bacterial inoculation assays

In a walk-in plant growth chamber (temperature: 25-27 ºC, humidity: 60%, photoperiod: 12:12 h, light intensity: 30 W), eighteen tomato plants (*Solanum lycopersicum* L.) cv. 'Cherry' were grown in 200 mL pots containing a commercial potting substrate. The plants were separated into two equal sets. The second set was inoculated one week later than the first one, as depicted in Figure 2C. This approach aimed to assess the impact of plant age on infection dynamics, especially observing the evolving pace of the infection. In each set, the nine plants were grouped in three triads, which were physically separated. The first group was sprayed with only sterile distilled water and was used as a control (i.e., it was only composed of healthy plants). The second group was inoculated with Pseudomonas syringae pv. tomato DC 3000, the etiological agent responsible for causing the disease bacterial speck of tomato. The third one was infected with Xanthomonas euvesicatoria, the bacteria responsible for the tomato bacterial spot disease. The inoculation process was performed using our Lab's previously optimized protocol (Reis Pereira, Santos et al. 2023). Bacterial suspensions of each pathogen presented a concentration of $1 \times 10^8$ cells mL-1. The plants were subsequently covered in transparent polythene bags for 48 hours to increase the humidity levels, facilitating bacteria entry into plant tissues through natural openings (Lamichhane 2015).

The viability of the bacterial cultures used in the infection assays was tested by plating 20 µL of a $1 \times 10^8$ cells mL-1 aqueous solution of Pst and an aqueous solution of Xeu ($1 \times 10^8$ cells mL-1, 20 µL) in Petri dishes containing KB and YDC media, respectively. Bacterial growth was visible 48 hours after both nutrient media, confirming the bacterial viability at the moment of the inoculation (Fernandes, Albuquerque et al. 2017). Furthermore, on the last spectral measurement date, sample preparation for

bacterial isolation was made to verify the presence of the pathogen species inoculated in each group using a standard approach based on Colony PCR.

### 2.2. Hyperspectral POM data acquisition

Spectral phenotyping was performed in a dark room, through the assessment of hyperspectral POM of the adaxial side of in vivo tomato leaflets (both healthy and diseased) (Figure 2A). Hyperspectral POM data acquisition was made on the nine plants in the study on nine distinct random points of the 4th, 5th, and 6th expanded leaflets. The setup configuration included a Hamamatsu Photonics K.K. TM Series C11697MB mini-spectrometer with USB 2.0 interface connected to a PC (processor Intel(R) Core (TM) i7-10510U CPU @ 1.80 GHz  2.30 GHz, RAM 16.0 GB, graphic NVIDIA ® GeForce®) through an evaluation software (SpecEvaluationUSB2.exe). This portable mini-spectrometer encompasses a wavelength range spanning from 200 to 1100 nm, with a spectral resolution of 0.6 nm (Figure 2D). It comprises a transmission optical fiber bundle (FCR-7UVIR200-2-45-BX, Avantes, Eerbeek, The Netherlands) covering 200-2500 nm, a stainless-steel slitted reflection probe, and a white LED (390 to 800 nm). Notably, conversion factors delineating how to translate sensor pixel numbers into specific wavelengths were unavailable (Hamamatsu 2023). While the spectrometer acquires information across the 200 to 1100 nm range, the LED emits light solely within the Vis-NIR region, specifically from 400 to 800 nm, which was the spectral information used in the plant measurements. During the plant measurements the probe was positioned 0.5 cm above the leaflet surface. This arrangement allowed for evaluating the leaflet's spectral signal and promoting its direct transmission to the mini-spectrometer's entrance lens. The light source was positioned below the sample to supply uniform enlightenment to all the leaflets' abaxial sides (Figure 2A), measuring light transmittance. The hyperspectral POM data were directly exported to .CSV files and subsequently imported to the MATLAB environment for data processing.

### 2.3. Bioanalytical monitoring – disease evolving

Visual and spectral phenotyping procedures started 24 h before the bacteria inoculation procedures, for both process assays (Figure 2B, C). All plant leaflets were visually screened to assess noticeable phenotypic modifications (e.g., modifications in color, texture, or other visible traits) from the ones expected to be found in healthy tissues. Hyperspectral POM data acquired at this stage were divided into training and validation datasets. During the training phase, DD-SIMCA was optimized using healthy samples (green dots) from the training dataset, successfully classifying (for tolerance level) all samples within the acceptance area (green line in Figure 1D). Subsequently,

the trained model was able to allocate all samples from the validation dataset within the defined acceptance area intended exclusively for healthy samples as depicted by green dots in Figure 1E, thereby confirming their authentication. The retrieved pure spectral signatures from both training (Figure 1F) and validation (Figure 1G) datasets were similar spectral fingerprints, suggesting that MCR-ALS results might be useful for in-depth forensic analysis.

The measurements were restarted 48 hours after infection, after the plastic bag removal from the tomato plants. They were performed daily before the symptom's observation (homogeneous visual phenotyping) and after the observable lesion's appearance (heterogeneous visual phenotyping in Figure 2F). The data collected at different points of diseased leaflets were carefully authenticated based on DD-SIMCA. The fingerprinting profiles were measured for both target (healthy) and non-target class of samples (diseased measurements) (Figure 2E) with similar physical spectral characteristics, making it necessary to apply a proper chemometric model, as will be demonstrated in the next sections.

### 2.4. Plant diseases modeling strategies and software

Hyperspectral POM data (Figure 1B) acquired from the two data sets were subsequently imported into MATLAB R2022a workspace (Figure 1 C) for chemometric data analysis. A mathematical algorithm based on the Savitzky-Golay filter (Savitzky and Golay 1964) was computed for spectral smoothing using a third-degree polynomial with a fifteen-sensor window as a part of the data pre-processing step. Subsequently, an additional mathematical pre-processing algorithm relying on Standard Normal Variate (SNV) (Barnes, Dhanoa et al. 1989) was applied. The application of the pre-processing algorithms on hyperspectral POM data aimed at performing spectral signal correction, along with the minimization of dispersive effects (Figure 2D). Therefore, the chemometric model might focus on the biochemical/structural composition dynamic of the studied leaflets (healthy and infected) with higher efficiency, requiring less principal components (PCs in the case of principal component analysis PCA).

Multivariate data processing involving DD-SIMCA that operates as dual PCA/SIMCA (Pomerantsev and Rodionova 2014) was conducted using DDSGUI, a graphical user interface (Zontov, Rodionova et al. 2017) freely available (https://github.com/yzontov/dd-simca.git). The MCR-ALS using GUI (Jaumot, de Juan et al. 2015) freely available online at http://www.mcrals.info. PLS_Toolbox R9.2.1 (Eigenvector Research 2023) was employed for PCA bilinear data decomposition.

All chemometric tasks were performed in a MATLAB environment R2022a. A brief fundamental theory on how DD-SIMCA and MCR-ALS operate will be explained below in connection with specific examples involving (i) hyperspectral POM data acquired throughout the early disease detection that was defined as a stage when the bacteria had already inoculated in the plant; however, its symptoms were not visible to the human eye (Figure 2C) and (ii) hyperspectral POM data collected during the visual lesion's appearance (Figure 2C) to provide in-depth understanding on ground-breaking research herein proposed pedagogically.

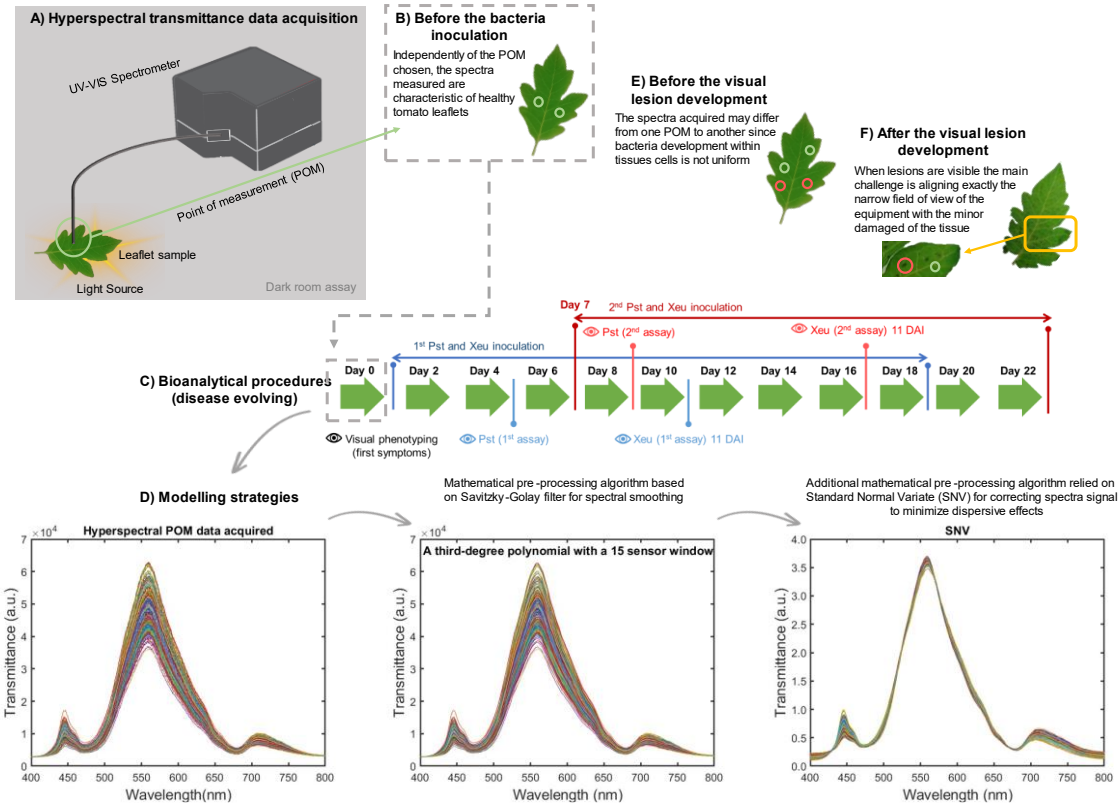



**Figure 2** Hyperspectral point-of-measurement (POM) was performed in the adaxial side of in vivo tomato leaflet tissues belonging to the 4th, 5th, and 6th leaves (A). A spectrometer combined with an optical fiber bundle with a reflection probe was used to acquire Vis-NIR spectroscopic data. A white LED was placed beneath each leaflet to provide uniform light to all the sampled surfaces. Spectral measurements were initiated 24 hours before bacteria inoculation in all the plants in the study when all tissues presented the characteristic phenotype of healthy leaflets (B). The bioanalytical procedures involved performing data acquisition and visual phenotyping in two bacterial inoculation assays, using two distinct groups of tomato plants, and initiated with one week difference (first assay represented in blue, and second assay in red, DAI corresponds to the designation 'Days after inoculation') (C). The host-pathogen interactions analyzed involved the usage

of *Pseudomonas syringae* pv. *tomato* (Pst) and *Xanthomonas euvesicatoria* (Xeu) belong to two different species and genera but are responsible for causing similar symptoms in tomato-diseased tissues. The hyperspectral data collected was pre-processed using an algorithm based on Savitzky-Golay filter for spectral smoothing and a Standard Normal Variate (SNV) to minimize dispersive effects (D). This procedure was performed over time, registering the appearance of the first macroscopic lesions caused by the bacteria in the study until their full development (E, F).

## 3. Results and discussion

### 3.1. Early disease detection at a stage when Pst and Xeu bacteria had already been inoculated in the plant leaflets non-symptomatic at visual inspection through hyperspectral POM data with DD-SIMCA and MCR-ALS

This subsection discusses the selected multivariate results on hyperspectral POM acquired in the two datasets (Figure 2C) and supplementary material. The follow-up on the evolution of tomato disease in the two sets of data allows for confirming the obtained results with a broader generalization of the acquired models. The first results correspond to the data obtained 72h after bacterial inoculation of tomato plants, since all leaflets were diseased but non-symptomatic. In turn, the results at 96h are also reported since they correspond to the date when Pst inoculated plants started to reveal the first macroscopic lesions (Figure 2C). Nevertheless, the Xeu inoculated plants remained phenotypically unchanged. Thus, these two dates were considered fundamental for the study of early bacterial disease diagnosis in tomato in the present study.

The multivariate samples involving hyperspectral POM data were acquired at a stage after the inoculation of Pst and Xeu bacteria into the plant leaflets, despite the absence of symptoms at the macroscopic level level (visual phenotyping). At this stage, the plant-pathogen interactions originated a non-linear data system (Figure 2C, E). These non-linearity systems result from the diverse damages inflicted on biological tissues during plant-pathogen interactions within the validation group, contrasting to the control group (healthy samples). Consequently, the application of DD-SIMCA, known for its quadratic approach (Pomerantsev and Rodionova 2018), became necessary to successfully models this non-linear data system. DD-SIMCA chemometric model under a rigorous approach (Rodionova, Oliveri et al. 2016) was developed by using only the target class composed of healthy samples herein denoted as the control group with acronym C (Figure 3B). This one-class modeling under rigorous approach employs a decision rule on the membership of the class that the threshold delineates the acceptance area, involving a two-step procedure.

The first step in the DD-SIMCA framework was the application of PCA (Rodionova, Kucheryavskiy et al. 2021) to the training dataset ($I \times J$) derived from the target class (C samples in Figure 3B) as a rigorous approach (Rodionova, Oliveri et al. 2016), containing a data matrix $\mathrm{D}$ with dimensions of $57 \times 971$ (Figure 3A). Where *I* refers to the number of target class of samples or objects (57) while *J* corresponds to Vis-NIR variables (971). These Vis-NIR spectroscopic data were mathematically pre-processed (Figure 2D) and subsequently processed according to Equation (1):

$$\mathrm{D} = \mathrm{T}\mathrm{P}^{\mathrm{T}} + \ \mathrm{E} \qquad (1)$$

where $\mathrm{T} = \{t_{ia}\}$ is the ($I \times A \ \ or \ 57 \times A$) scores matrix; $\mathbf{P} = \{p_{ja}\}$ is the ($J \times A \ \ or \ 971 \times A$) loadings matrix; $\mathrm{E} = \{e_{ij}\}$ is the ($I \times J \ \ or \ 57 \times 971$) matrix of residual; $A$ is the number of PCs required to explain a given data set, 3 PCs were required to explain $\mathrm{D}$.

In the second step of the DD-SIMCA framework, PCA results were employed to calculate two meaningful distances for each sample or object $i = 1, \ldots., I$ of the training set ($I \times J$). They were already referenced in the Introduction section as $h_i$ linked to score distance and $q_i$ associated with the orthogonal distance of a given object $i$. The score distance is more exactly $h_i$ refers to the position of a sample within the score multidimensional space (Rodionova, Oliveri et al. 2016), which can be computed as the squared Mahalanobis distance between the projection of the point and the multidimensional subspace origin (Rodionova, Kucheryavskiy et al. 2021) following Equation (2):

$$h_i = \mathrm{t}_i^t(\mathrm{T}^t\mathrm{T})^{-1}\mathrm{t}_i \ = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda a} \qquad (2)$$

where eigenvalues $\lambda_a$ associated with the eigenvectors, $a = 1, .., A$, are the diagonal elements of matrix $\mathrm{T}^t\mathrm{T}$.

The orthogonal distance is more exactly $q_i$ characterizes a sample distance to the score multidimensional space (Rodionova, Oliveri et al. 2016), which is calculated as the sum squared residuals according to Eq. 3. Chemometrically speaking, $q_i$ is the squared Euclidean distance between a data point related to the object, and the PC space computed in the original variable space (Rodionova, Kucheryavskiy et al. 2021).

$$q_i = \sum_{j=1}^{J} e_{ij}^2 \qquad (3)$$

The contribution from a sum of both distances $h$ and $q$, whose values follow a chi-square distribution (Pomerantsev 2008) are employed to estimate the full distance

(FD) as recently introduced in the relevant literature (Rodionova and Pomerantsev 2020) through Equation (4).

$$FD = N_h \frac{h}{h_0} + N_q \frac{q}{q_0} \quad (4)$$

where parameters $h_0$ and $q_0$ are the scaling factors while $N_h$ and $N_q$ are the number of the degrees of freedom (DoF). These scaling factors and DoFs are considered unknown a priori and subsequently estimated by using a parallel data-driven approach (Pomerantsev and Rodionova 2014) through a classic regression in our regular dataset without outliers (Figure 3B) instead of using a robust one, achieving $N_h = 4$ and $N_q = 5$ throughout the DD-SIMCA optimization.

One of the ground-breaking strategies of this contribution, when compared to the previous ones (Pereira, Milori et al. 2010, Atanassova, Nikolov et al. 2019), is the flexible way of defining the acceptance area established by inequality in a selected confidence level $\alpha$-value (for instance, 0.007262 in Figure 3B). The flexibility enables the establishment of the critical full distance ($FD_{critical}$) with a tolerance level $(1-\alpha)100\%$ depicted in Figure 3B, which is mathematically computed according to Equation (5):

$$FD_{critical} = x^{-2}\,(1-\alpha, N_h + N_q) \quad (5)$$

where $x^{-2}$ is the quantile of the chi-squared distribution. From the inequality, when a sample presents $FD \leq FD_{critical}$, it is classified as belonging to the target class in the case of rigorous OCC model. Whereas if a sample shows $FD \geq FD_{critical}$, the sample is not compatible with the profile characteristics of the target class. From an operational perspective, $FD$ statistics is connected to the so-called classification analytical signal (Pomerantsev, Vtyurina et al. 2023) in the DD-SIMCA, useful for multiple purposes.

It is worth highlighting the importance of the acceptance area in the DD-SIMCA performance in each confidence level that specifies a type I error in decision-making connected to a false negative decision. From this perspective, an evaluation of the acceptance area with a confidence level $\alpha$-0.007262 (green line) displayed in Figure 3B reveals that a training sample well-identified as “C_3” was classified as a regular one in the acceptance border. However, if the confidence level $\alpha$-value is changed (for instance) to $\alpha$-0.01, “C_3” might be classified as an extreme sample as graphically demonstrated in (Mazivila and Borges Neto 2021). Another tolerance boundary is the outlier threshold (O$\gamma$) introduced for a given $\gamma$-value (for instance, 0.01 in Fig. 3B), which can be adequately computed according to Equation (6):

$$O_\gamma = \left\{(h,q): FD > x^{-2}\,((1-\gamma)^{\frac{1}{I}}, \quad N_h + N_q)\right\} \quad (6)$$

where the position of the outlier border relies on $I$-size of the training set ($I \times J$), i.e., the greater $I$-size is, consequently the farther the outlier border will be.

Finally, DD-SIMCA training and validation results (Figure 3B, C) can be displayed by using a two-dimensional plot (Pomerantsev and Rodionova 2014), where each object (sample) is often shown in the coordinates $ln(1 + h_i/h_0) \quad vs \quad ln(1 + q/q_0)$, together with the two tolerance boundaries (acceptance area and outlier threshold).

A rigorous OCC was developed, meaning the DD-SIMCA optimization was completely based on a target class composed of healthy samples (non-symptomatic plants) in a total of 57 that were correctly attributed as members of the target class Figure 3B with a sensitivity of the training phase equal to 100%. An additional 'validation_data_set' with 186 samples loaded to MATLAB Workspace (Figure 3A) with a complex composition (24 samples of the target class or control group in the validation set with acronym VC, 81 samples inoculated with Pst bacteria with acronym P, and 81 samples inoculated with Xeu bacteria with acronym X) to help in the OCC model validation.

DD-SIMCA results depicted in Figure 3C revealed that OCC was able to delineate the target class (VC) consisting of healthy samples (non-symptomatic plants) located within this acceptance boundary (green line) that detaches its members from other non-target classes (P and X), meaning a successful authentication task. In this Figure 3C, the sample with higher distance ("P_67") revealed that its bacterial infection stage is likely more advanced when compared to the healthy samples, presenting a higher value than the FD statistics (Equation 4). In contrast to sample "P_67", samples "P_33", "P_42", "P_49", "P_63", and "X_31", "X_52", "X_54" barely surpassed the threshold established in the selected acceptance level (green line) suggesting that the disease phase of these leaflets was at an earlier stage. This enables to draw some findings on the sensitivity and specificity of the validation phase in the context of a rigorous OCC model built using only the target class as pointed out in the literature (Pomerantsev and Rodionova 2021), as well as taking into consideration that the real health-condition (non-symptomatic or symptomatic) of the alternative classes (P and X) included in 'validation_data_set' was unknown, as follows:

- Sensitivity can be properly calculated according to Equation (7) in the rigorous DD-SIMCA since throughout the optimization phase (Figure 3B) only employed a set of the target class of samples (healthy plants or control group with acronym C). Therefore, no specificity value might be found because the attribution of non-target class of symptomatic plant samples is unknown. Samples belonging to the

target class (VC) were correctly attributed to VC, corresponding to true positive in Equation (7), thereby achieving a test sensitivity of 100%. It is worth directing the readers to the pertinent literature (Pomerantsev and Rodionova 2021) for comprehensive guidance on computing figures of merit across various of approaches. For instance, within compliant OCC optimization, only specificity might be computable, while the remaining figures of merit necessitate the confusion matrix (Pomerantsev and Rodionova 2021). This requirement is applicable not only in intricate scenarios of multi-class classification but also in a simpler case of binary discrimination (Pomerantsev and Rodionova 2021).

- Although 72h after alternative classes (P and X) had been inoculated with Pst and Xeu bacteria, many samples were located below the acceptance threshold and categorized as belonging to the target class (green dots). These results indicate that, at this time span from inoculation, their health-condition is in a non-symptomatic state, similar to the control group identified as VC during the validation phase. Whereas a few numbers of samples of these alternative classes (P and X) are located beyond the acceptance boundary, being classified as belonging to the non-target classes (red dots), which indicates that the plant-pathogen interaction throughout the biological process was faster in these samples than others (green dots). The health condition of alternative classes (P and X) undergoes evolution. Among 162 samples within these alternative classes (P and X), the non-symptomatic state (green dots) indicates no bacterial effect, while the symptomatic state (red dots) showcases a substantial bacterial effect in 31 samples. Out of the symptomatic plants (red dots) with marked bacterial effect, only 6 samples (X) and 16 samples (P) depicted in Figure 3D were selected for bilinear decomposition aiming fingerprint profile retrieval. This selection was due to the noticeable overlap observed among their red dots (Figure 3C). This decision was made in conjunction with the well-documented health-condition of the validation group of control (VC). The aim was to address the question: “which is the biochemical composition of these alternative classes (P and X) classified either as target class or non-target class from a forensic standpoint?”

$$Sensitivity = \frac{Number\ of\ true\ positives}{Number\ of\ target\ samples} \quad (7)$$

This question can be answered through a bilinear decomposition through pure fingerprint profiles. The PCA bilinear decomposition using Equation (1) retrieved linear combination spectral profiles ($P^T$ in Fig. 3F) lacking physical meaning from a forensic

viewpoint when compared to the measured hyperspectral POM data (Figure 2D). This discrepancy arises from model's data decomposition adhering to orthogonal constraints, aiming to maximize the explained variance in each PC (Mazivila and Santos 2022). Alternatively, a bilinear data decomposition based on MCR-ALS was successfully initialized with pure variables under mathematical or natural constraints and computed according to Equation (8).

The MCR-ALS constrained analysis facilitated the extraction of distinct pure spectral signatures ($\mathrm{S}^{\mathrm{T}}$ in Fig. 3E) corresponding to the validation group of control (VC), Pst (P), and Xeu (X). These signatures provide supplementary forensic insights, reinforcing the robust authentication of the DD-SIMCA outcomes concerning the alternative classes (P and X) classified as either target or non-target classes.

$$\mathrm{D} = \mathrm{C}\mathrm{S}^{\mathrm{T}} + \mathrm{E} \quad (8)$$

where $\mathrm{D}$ is an $I \times J$ data matrix containing a group of $I$ samples individually acquired in a vectorial signal (hyperspectral POM) at $J$ measured variables in different wavelengths, as can be seen in Fig. 2D, $\mathrm{C}$ is an $I \times N$ matrix that contains the concentrations of $N$ species in the $I$ samples, $\mathrm{S}^{\mathrm{T}}$ is an $N \times J$ matrix containing the pure spectral signatures of $N$ intervening species at $J$ measured variables in different wavelengths, i.e., corresponding to the molar absorptivities of Beer-Lambert law (Mazivila and Santos 2022), and $\mathrm{E}$ is the error matrix.

Another advantage of this contribution is that curve resolution analysis provided a unique spectral profile for the above-raised question from a forensic perspective. In Figure 3E, the control samples belonging to the validation set (VC, blue profile) presented a spectral profile characteristic of healthy (non-symptomatic) leaflets, providing additional forensic evidence. This profile is statistically similar to most samples from the alternative classes (P and X) with green dots yet without typical bacterial symptoms. Nonetheless, in P samples (yellow profile) classified as symptomatic (red dots), the bacterial effect is visibly manifested in the band between 430 and 475 nm when compared to control samples (VC), while to the X samples (red dots), the bacterial impact is clearly expressed in the spectral range of 675-800 nm. This in-depth forensic analysis (Figure 3E) offers a complementary investigation to the conducted pure authentication task (Figure 3C) aimed at protecting the target class (non-symptomatic plants) against the non-target class (symptomatic plants). These findings demonstrate the feasibility of the methodology established for performing the early diagnosis of bacterial diseases in tomato plants since it allowed the detection of inoculated samples before they showed macroscopic lesions (i.e., detectable by the human eye). Through the analysis of the

spectroscopic profiles, the procedure identified microscopic modifications (i.e., not visible to the human eye) present in diseased samples, leading to their identification (red dots). Thus, agricultural practices can be performed early in the disease cycle, resulting in more effective, precise, and sustainable measures to control bacterial diseases.

The methodology was then applied to the spectral data collected 96h after bacterial inoculation. At this stage, the samples infected with Pst began showing initial macroscopic lesions, while the samples affected by Xeu did not exhibit any macroscopic sign of disease (Figure 2C, E, F). DD-SIMCA optimization was executed similarly, concentrating solely on a target class of healthy samples (non-symptomatic plants) out of 57. All 57 samples were correctly identified as belonging to the target class, as depicted in Figure 4B, resulting in a training phase sensitivity of 100%. Nevertheless, it is important to notice that “C_12” in Figure 4B might be classified as an extreme sample, since fluctuations at significance level $\alpha = 0.002453$ might impact its classification. The model validation was achieved by uploading an additional ‘validation_data_set’ with 184 samples to MATLAB Workspace (Figure 4A) composed of a single target class (control, VC) that represents 22 authentic samples, and an unlimited number of alternative classes that contain 162 non-authentic samples (81 measurements captured in tissues inoculated with Pst bacteria (P), and 81 measurements from tissues inoculated with Xeu bacteria (X)).

DD-SIMCA results demonstrated that the model was capable of prescribing a process that allows us to confirm that the target class (VC) is composed of healthy samples (non-symptomatic tissues) within the acceptance boundary (green line), effectively detaching them from members of other non-target classes (P and X) (Figure 4C). The OCC model accurately identified all samples (VC) of a single target class with test sensitivity equal to 100%. In turn, 98 samples of alternative classes infected with Pst (P) and Xeu (X) bacteria were contained in the acceptance threshold and linked to the target class (green dots).

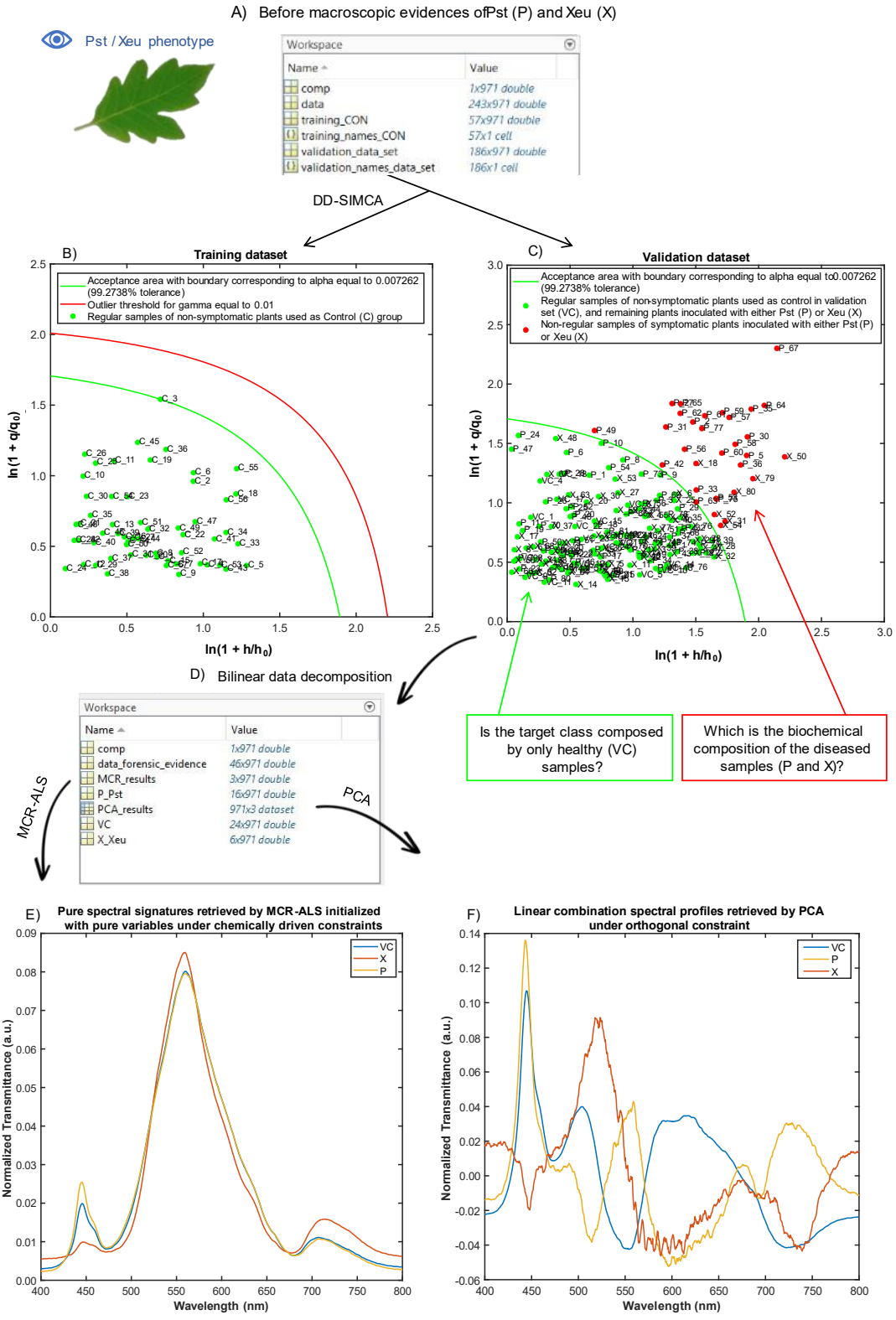



**Figure 3** Hyperspectral point-of-measurement (POM) was performed in vivo tomato leaflets before macroscopic evidence of the bacterial diseases caused by *Pseudomonas syringae* pv. *tomato* (Pst) and *Xanthomonas euvesicatoria* (Xeu) (72 hours after bacterial inoculation) (A). The spectroscopic data was then inserted into MATLAB, where a part of the measurements performed in healthy tomato leaflet tissues (C green dots in B) was

used as a training set, and the remaining healthy samples (VC in C) together with measurements made in inoculated non-symptomatic tissues (P for samples inoculated with Pst, and X for samples inoculated with Xeu in C) were used as validation set in the computation of a Data-driven Soft Independent Modelling of Class Analogy (DD-SIMCA) model. The training set was used to establish the acceptance boundary (green line) (B). In turn, the validation set was applied to demonstrate that the target class was composed of healthy (VC green dots) samples and spectral measurements were performed in non-symptomatic diseased tissues at earlier stages of the diseased process (P, X green dots) (C). In turn, samples in which microscopic lesions occurred were located out of the acceptance boundary (P, X red dots), indicating their disease stage was more advanced. A bilinear data decomposition was, then, performed (D) to retrieve the pure spectral signatures using Multivariate Curve Resolution − Alternating Least-Squares (MCR-ALS) initialized with pure variables under mathematical or natural constraints, and to retrieve the linear combination spectral profiles (F) by performing a Principal Component Analysis (PCA) under orthogonal constraint for comparison purpose.

Thus, their health state is believed to be considered as a non-symptomatic phase like the control group (VC) in the validation step, showing that host-pathogen interactions are still evolving as manifested also in the second experimental inoculation (Figure 2C) depicted in Figure S1C available at supplementary information. The remaining 64 samples, on the contrary, were found to be located outside the acceptance boundary, being identified as non-target classes (red dots) and subsequently subjected to bilinear decomposition (Figure 4D) to provide in-depth forensic evidence. The curve resolution analysis was conducted aiming at providing forensic evidence through unique spectral profiles, which belonged to the target class of control samples (VC, blue profile), non-target class of Pst (P, yellow profile), and non-target class of Xeu (X, orange profile) were provided (Figure 4E, Figure S1E) to confirm with high confidence that VC is declared to be as the target class (non-symptomatic plants).

### 3.2. DD-SIMCA and MCR-ALS applied to hyperspectral POM data acquired throughout plant-pathogen interactions with developed symptoms of the disease for leaflets authentication

In this subsection, biological data measured eleven days after the bacterial inoculation (Figure 2C) were considered for chemometric data modeling. This date was chosen since both Pst and Xeu inoculated samples showed, for the first time, macroscopic lesion characteristics of the bacterial speck and spot diseases,

respectively. This also happened in the second biological assay performed where the symptoms also appeared eleven days after infection (Figure 2C, E, F).

The optimization step of the rigorous DD-SIMCA model focused on the target class including just 57 healthy samples (non-symptomatic plants), which were correctly identified as members of the target class Figure 5C with 100% of training phase sensitivity (Eq. 7). Samples “C_49” and “C_37” in Figure 5C can be considered an extreme observation, given the potential impact of variations at a significant level $\alpha = 0.01$. A further ‘validation_data_set’ comprising 186 samples (Figure 5B) was utilized to evaluate the performance of the DD-SIMCA model. This dataset encompassed 24 authentic samples belonging to the target class of the control (healthy, VC) plants, along with 81 measurements from Pst inoculated samples (P) and another 81 from Xeu diseased samples (X). The evaluation of the DD-SIMCA model was centered on its accuracy in correctly classifying samples within the target class.

The OCC model proved that the target class of the control (healthy) plants (non-symptomatic measurements) contained in the acceptance boundary (green line), with test sensitivity of 100% denoting a share of correctly identified 24 control plants (healthy, VC) of the target class (Equation 7). It is worth pointing out that sensitivity (Equation 7) is computed separately in the training dataset (Figure 5C) and validation dataset (Figure 5D). DD-SIMCA differentiated these samples from other belonging to non-target classes (P and X) (Figure 5D). In the validation step, the model identified 67 samples of non-target classes (inoculated with Pst – P –, and Xeu – X) as part of the acceptance area (green dots) and associated them to the target class. These measurements are assumed to be collected in healthy tissues (non-symptomatic), like the validation step control group (VC). In turn, the remaining 119 samples were not contained in the acceptance boundary and were categorized as non-target classes (red dots) (Figure 5D). To present thorough forensic proof, a bilinear decomposition was made (Figure 5E). MCR-ALS allowed the determination of spectral profiles associated with control samples (VC, depicted in blue), non-target class of Pst (P, represented in yellow), and non-target class of Xeu (X, shown in orange) (Figure 5F, Figure S2F). This confirmation, achieved with a high level of confidence, demonstrates that VC is accurately identified as the target class, representing non-symptomatic plants. Consequently, a fully validated methodology could combine DD-SIMCA and MCR-ALS applied to hyperspectral POM data. This combination serves to confirm the farness (FD in Equation 4) of each visual phenotyping throughout the evolving bacterial disease (P and X, for instance, samples “P_31”, “P_8”, “X_14” in Fig. 5D) concerning the target class of the control (healthy) plants (VC) and pure biological profiles (yellow and orange in Figure 5F).

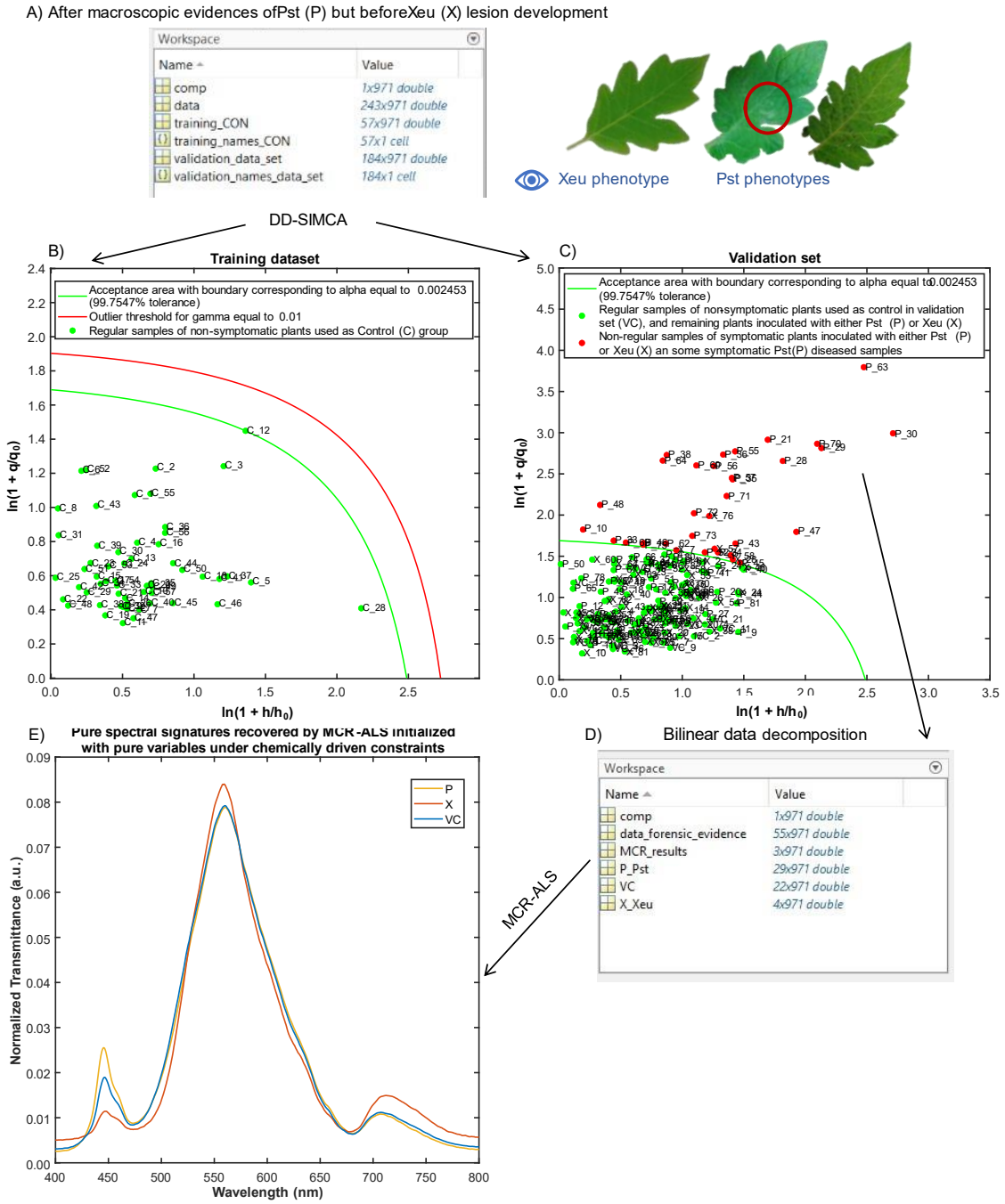



**Figure 4** Hyperspectral point-of-measurement (POM) was performed in vivo tomato leaflets after macroscopic evidence of the disease caused by *Pseudomonas syringae* pv. *tomato* (Pst) but before macroscopic evidence of the disease caused by *Xanthomonas euvesicatoria* (Xeu) (96 hours after bacterial inoculation) (A). The spectral data was then inserted into MATLAB, where a part of the measurements performed in healthy tomato leaflet tissues (C green dots in B) were used as training set, and the remaining healthy samples (VC in C) together with measurements made in inoculated tissues (P for samples inoculated with Pst, and X for samples inoculated with Xeu in C) were used as validation set in the computation of a Data-driven Soft Independent

Modelling by Class Analogy (DD-SIMCA) model. The training set was used to establish the acceptance boundary (green line) (B). In turn, the validation set was applied to demonstrate that the target class was composed of healthy (VC green dots) samples, and spectral measurements which were performed in non-symptomatic diseased tissues at earlier stages of the diseased process (P, X green dots) (C). In turn, samples that presented only microscopic (X red dots) or microscopic and macroscopic lesions (P red dots) were located outside the acceptance boundary, indicating their disease stage was more advanced. A bilinear data decomposition was, then, performed (D) to retrieve the pure spectral signatures (E) using Multivariate Curve Resolution − Alternating Least-Squares (MCR-ALS) initiated with − pure variables under mathematical or natural constraints

The combination of the hyperspectral POM analytical technique with suitable chemometric models relying on DD-SIMCA and MCR-ALS allowed the distinction of measurements collected in target class - non-symptomatic tissues (healthy samples recognized with green dots) and those from non-target class measurements (diseased samples identified with red dots). This distinction was evident not only during the early stages, preceding the formation of visible lesions (Figure 3), but also during the later phases of the disease cycle (Figure 5).

Despite the innovative results presented in this study, there remains a pressing need for the early identification of the non-target class of diseased samples. Achieving this goal would significantly enhance the precision, efficacy, and environmentally sustainable nature of phytosanitary measurements. Therefore, our primary operational message underscores the recommendation to utilize the model performance illustrated in Figure 3 for early diagnosis of tomato bacterial disease. This approach holds immense potential, enabling timely interventions by agronomists and farmers.

## 4. Conclusions

The present work developed a non-destructive methodology to diagnose two bacterial tomato diseases early. This approach combines hyperspectral point-of-measurement with chemometric approaches, leveraging DD-SIMCA as one-class classification assisted by MCR-ALS to provide forensic evidence through pure biological profiles. The groundbreaking nature of this study represents an advance concerning the previously published papers in the same field, particularly in contrast to the original version of SIMCA. This advancement lies in the comprehensive consideration of the full distance (FD) calculated in the process of building the rigorous DD-SIMCA. This model is instrumental in predicting distance values that reflect the evolving bacterial infection

within plant leaflet tissues. These predictions are then juxtaposed against the target class of control healthy plant leaflet tissues in a 2D plot, providing a novel perspective in disease assessment.

The main conclusions drawn from this contribution are that:

i) Vis-NIR spectroscopic data acquired from healthy tomato tissues were correctly identified as members of the target class, with 100% sensitivity. These healthy samples were effectively differentiated from those assessed on tissues inoculated with Pst or Xeu, a differentiation noticeable even before the appearance of characteristic macroscopic lesions associated with the disease, detectable as early as 72 hours following bacterial inoculation.
ii) Non-target class of samples located beyond the acceptance threshold indicated a more advanced of bacterial infection in these instances. This suggests that discernible spectral alterations resulting from host-pathogen interactions occurred even before noticeable phenotypical changes. Samples closer to the acceptance boundary are therefore, presumed to be at earlier stages of the infection process, contrasting with those farther away exhibiting a greater FD, indicating a more pronounced advancement in the infection stage.
iii) The unique Vis-NIR spectral profile obtained for each health group (control – healthy, inoculated with Pst, inoculated with Xeu), retrieved throughout the MCR-ALS optimization process, revealed specific bacterial effects on plant. The impact of the Pst bacterial infection is discernible within the spectral bands ranging from 430 to 475 nm, while the influence of Xeu was evident in the spectral range between 675 and 800 nm. These comprehensive findings, serving as forensic evidence, provided an additional layer of analysis beyond the primary authentication task. They significantly contributed to understanding host-pathogen interactions, particularly the evolving bacterial infection within the plant specimens.
iv) The modeling analysis of the data collected in the second biological assay reinforces the results obtained using the data of the first assay, despite the plants being inoculated with one week difference. In terms of bacterial evolution, plants start to develop symptoms in similar days: Pst between the 72h ($2^{nd}$ assay) and 96h ($1^{st}$ assay), and Xeu eleven days after inoculation (for both the 1st and 2nd biological assays). DD-SIMCA showed that, in both cases, the validation samples collected on the target class of healthy leaflet tissues were allocated inside the acceptance boundary. In contrast to non-target class of samples (measured in symptomatic diseased tissues) were located outside this acceptance threshold.

MCR-ALS also revealed that the retrieved fingerprinting spectral profiles for each class in study were analogous in the first and second biological assays.

These findings underscore the efficiency of the developed methodology for the early diagnosis of bacterial diseases in tomato plants. This approach facilitated the identification of subtle modifications at the microscopic level, indistinguishable from the human eye but evident in diseased samples. Consequently, it enables agricultural interventions, including disease monitoring and management and phytosanitary measures at earlier stages of the disease progression. This capability enhances the effectiveness and precision of interventions, aligning with the principles and goals of precision agriculture.

**Acknowledgements**

Mafalda Reis-Pereira was supported by a fellowship from Fundação para a Ciência e a Tecnologia (FCT) [grant reference SFRH/BD/146564/2019]. This work is partially financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots - OmicBots: High-Throughput Integrative Omic-Robots Platform for a Next Generation Physiology-based Precision Viticulture, with reference PTDC/ASP-HOR/1338/2021.

## Supplementary materials

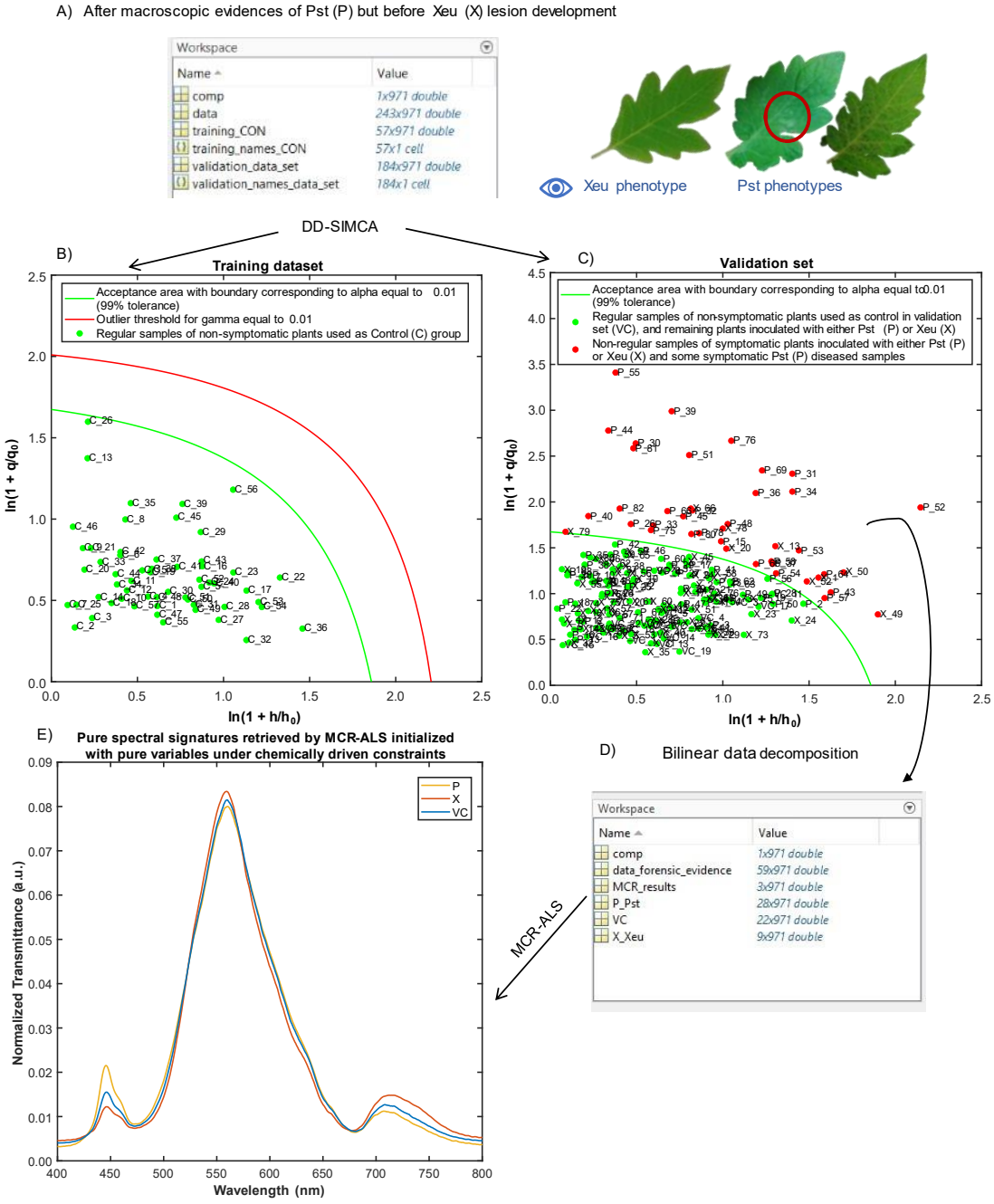



**Figure S1** Hyperspectral point-of-measurement (POM) was performed in a second assay where in vivo tomato leaflets after macroscopic evidence of the disease caused by *Pseudomonas syringae* pv. *tomato* (Pst) but before macroscopic evidence of the disease caused by *Xanthomonas euvesicatoria* (Xeu) (72 hours after bacterial inoculation) (A). The spectroscopic data was then inserted into MATLAB, where a part of the measurements performed in healthy tomato leaflet tissues (C green dots in B) were used as training set, and the remaining healthy samples (VC in C) together with measurements made in inoculated tissues (P for samples inoculated with Pst, and X for samples inoculated with Xeu in C) were used as validation set in the computation of a

Data-driven Soft Independent Modelling by Class Analogy (DD-SIMCA) model. The training set was used to establish the acceptance boundary (green line) (B). In turn, the validation set was applied to demonstrate that the target class was composed of healthy (VC green dots) samples, and spectral measurements which were performed in symptomless diseased tissues at earlier stages of the diseased process (P, X green dots) (C). Samples that presented microscopic or macroscopic lesions were located out of the acceptance boundary (P, X red dots), indicating their disease stage was more advanced. A bilinear data decomposition was, then, performed (D) to retrieve the pure spectral signatures (E) using Multivariate Curve Resolution − Alternating Least-Squares (MCR-ALS) initiated with pure variables under mathematical or natural constraints.

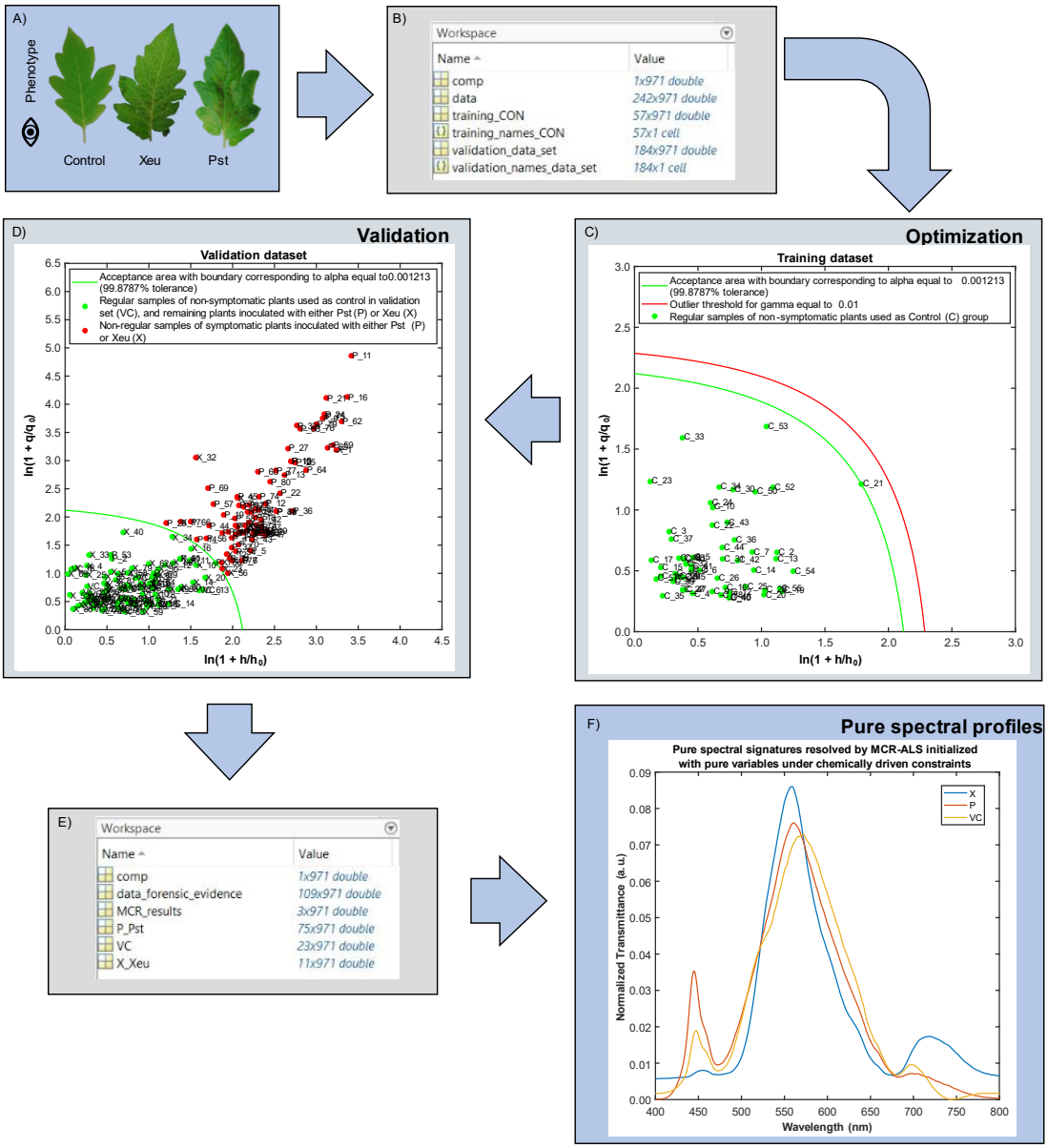



**Figure S2** Hyperspectral point-of-measurement (POM) was performed in vivo tomato leaflets after macroscopic lesions of the diseases caused by *Pseudomonas syringae* pv.

*tomato* (Pst) and *Xanthomonas euvesicatoria* (Xeu) (eleven days after bacterial inoculation, second assay) (A). The spectral data was then inserted into MATLAB (B), where a part of the measurements performed in healthy tomato leaflet tissues (C green dots in B) were used as training set, and the remaining healthy samples (VC in C) together with measurements made in inoculated tissues (P for samples inoculated with Pst, and X for samples inoculated with Xeu in C) were used as validation set in the computation of a Data-driven Soft Independent Modelling by Class Analogy (DD-SIMCA) model. The training set was used to establish the acceptance boundary (green line) (C). In the validation set was applied to demonstrate that the target class was composed of healthy (VC green dots) samples and spectral measurements were performed in symptomless diseased tissues at earlier stages of the diseased process (P, X green dots) (D). Samples in which microscopic or macroscopic signs were manifested were located out of the acceptance boundary (P, X red dots), indicating their disease stage was more evolved. A bilinear data decomposition was, then, performed (E) to retrieve the pure spectral signatures (F) using Multivariate Curve Resolution − Alternating Least-Squares (MCR-ALS) initiated with pure variables under mathematical or natural constraints.

# Case Study 6

## VIS-SWIR spectroscopy and microscope imaging fusion towards reagent less and in vivo diagnosis of bacterial infection in tomato plants *Pseudomonas syringae* pv. *tomato* and *Xanthomonas euvesicatoria*

In this case study, we briefly refer to an ongoing study currently under development. The collaborative efforts of researchers involved in this case study are as follows (in no particular authorship order):

- Doctoral Program Students: Mafalda Reis Pereira[1,2], Filipe Monteiro Silva[1,2]
- Senior Researchers: Filipe Neves dos Santos[2], Rui Martins[1,2]
- Associate Professors: Fernando Tavares[1,3,4], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, 4169-007 Porto, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Roberto Frias, 4200-465 Porto, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Cam-pus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

### Introduction

Crops are susceptible to various abiotic and biotic stressors throughout their agronomic process, damaging productivity, yields, nutritional quality, and aesthetic appeal. This multifaceted impact extends to both economic and consumer dimensions, resulting in financial losses attributable to crop damage, increased expenses linked to phytosanitary treatments, elevated prices, and constrained availability of plant-derived products. This ripple effect influences the sectors of food, feed, clothing, and building materials, ultimately affecting consumers in terms of quantity and quality (Oerke 2006, Flood 2010, Savary, Ficke et al. 2012).

Biotic stresses caused by pests and pathogens are responsible for losses in crop yields between 20% and 40% (Savary, Ficke et al. 2012). Phytopathogens, in particular, are estimated to cause annual economic losses of around $220 billion for the agricultural sector, causing problems with access to food for more than 800 million people (Mitra 2021). Contemporary agricultural methods contribute to plant disease epidemics' proliferation and pathogens' swift evolution. This trend is exacerbated by the widespread adoption of intensive monoculture across expansive regions, cultivating genetically uniform plant varieties, and establishing extensive global supply chains and logistical networks. Collectively, these practices create an environment conducive to the rapid transmission of diseases and the adaptive evolution of pathogens (Zhan, Thrall et al. 2015). Phytosanitary products can be applied to prevent and control crop biotic stresses, leading to considerable damage to the environment and affecting food quality and security (Bonner and Alavanja 2017, Zhang, Yang et al. 2020).

Early detection and precise identification of plant pathogens are crucial for mitigating their adverse effects and implementing effective phytosanitary measures. Conventional disease diagnosis techniques primarily hinge on the manifestation of observable symptoms, constituting what are commonly known as 'direct' methods. These approaches are mainly based on scouting or laboratory approaches. The first entails visual field inspection performed by specialized trained observers to detect and identify infected plants based on the presence of disease symptoms (Parker, Shaw et al. 1995), which is subjective, error-prone (since symptoms alone are not entirely disease-specific), labor-intensive, time-consuming, and expensive (Sankaran, Mishra et al. 2010, Mahlein 2016, Khaled, Abd Aziz et al. 2018, Ali, Bachik et al. 2019). The second includes serological and molecular tests, which are generally applied due to their sensitivity, accuracy, and effectiveness. The most common techniques consist of polymerase chain reaction (PCR), enzyme-linked immunosorbent assay (ELISA), fluorescence in situ hybridization (FISH), immunofluorescence (IF), and flow cytometry (FCM) (Mohammad-Razdari, Rousseau et al. 2022). Usually, these methods involve detailed sampling procedures, which take several hours to complete and require destructive sample preparation. They do not allow a follow-up of the disease progression nor field mapping to support Precision Agriculture systems (e.g. site-specific management). They can have limited diagnostic abilities, mainly in the asymptomatic and early stages of the disease infection process, related to the uneven spread of pathogens inside plants (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015). These approaches are also not suitable for supporting real-time agronomic decisions in-field, because they do not present the necessary high throughput and speed, since they have been developed only

to confirm the presence of pathogens in samples (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015).

More recently, strategies for phytopathogen diagnosis have been developed based on identifying the changes they cause on crops due to host-pathogen interaction (also called 'indirect' methods). Generally, they are focused on monitoring changes in morphological traits, transpiration rates, temperature, and volatile organic compounds (VOCs) released by disease plants (Mohammad-Razdari, Rousseau et al. 2022).

Visible near-infrared (Vis-NIR) hyperspectral spectroscopy (HS) has a high potential for point-of-measurement (POM) reagent-less crop disease diagnosis, promoted by fungi (Yu, Anderegg et al. 2018, Skoneczny, Kubiak et al. 2020), bacteria (Bagheri, Mohamadi-Monavar et al. 2018), and viruses (Morellos, Tziotzios et al. 2020) affecting different crops, even at asymptomatic stages of the disease (Gold, Townsend et al. 2020). This is an information-rich approach that captures both chemical and physical information about a sample, where their characteristics are distributed across several wavelengths. HS POM detects modifications in plants' tissue optical properties, which arise from variations in pigments, sugars, and water levels (amongst other components) (Curran 1989, Thenkabail, Gumma et al. 2014, Tosin, Pocas et al. 2021, Tosin, Martins et al. 2022). The dominant spectral information in plant tissues comes from highly absorbent compounds in VIS-NIR. Photosynthetic pigments, with chlorophylls being the most relevant, influence the spectral behavior in the visible region, whereas, the water levels, chemical and structural composition (including the action of lignins and proteins), and internal scattering processes affect the NIR range (Hunt and Rock 1989, Jones and Vaughan 2010). Spectral information of plant tissue is, thus, super-imposed in the recorded spectra at different scales of interference (Martins 2019, Martins, Barroso et al. 2022, Tosin, Martins et al. 2022).

Nevertheless, HS may present redundant information from adjacent wavebands, and only a few of these features may be important in classifying a diseased individual (Blackburn 2007, Caicedo, Verrelst et al. 2014, Rivera, Verrelst et al. 2014). Statistical signal processing, mathematical combinations of various spectral bands, and the implementation of predictive modeling techniques – such as Machine Learning algorithms – are commonly employed strategies for the analysis of hyperspectral data. These approaches aim to distill valuable information from the dataset, facilitating dimensionality reduction and the selection of pertinent wavelengths (Mahlein, Steiner et al. 2010, Mahlein, Rumpf et al. 2013, Thenkabail, Gumma et al. 2014, Ahmadi, Muharam et al. 2017, Thenkabail, Lyon et al. 2018, Saleem, Potgieter et al. 2019, Zhao, Fang et

al. 2020, Saha and Manickavasagan 2021). Previous research based on direct spectral data, information with reduced dimensionality, or selected waveband features evidence that different model approaches were effective for identifying and classifying several plant stress and diseases (Sankaran, Ehsani et al. 2012, Bajwa, Rupe et al. 2017, Gold, Townsend et al. 2020, Meng, Lv et al. 2020).

The present Case Study, hence, aimed to explore the suitability of hyperspectral transmittance point-of-measurement data for early bacterial disease diagnosis. Moreover, it investigated the potential of fusing this information with microscopic imaging collected in real time for providing more detailed information related to the host-pathogen interactions and effects, along with the monitoring of the infection progression and evolution over time. The study aimed to verify if i) bacterial infection promotes changes in the optical properties of the host plant, detectable by hyperspectral point-of-measurement sensors, ii) the possibility of discriminating healthy from diseased tissues, iii) the capacity of discriminating diseased tissues affected by different bacteria, iv) hyperspectral transmittance point-of-measurement data fused with RGB microscopic imaging enhances the diagnostic process, and v) data fusion allows the follow-up of the disease spectral and microscopically visual phenotypes.

## 2. Materials and methods

### 2.1. Experimental design

Tomato (*Solanum lycopersicum* L.) plants of the cultivar Cherry were grown in 200 mL pots containing a commercial potting substrate, in a walk-in plant growth chamber under controlled conditions (25-27 ºC, humidity of approximately 60%, photoperiod of 12 / 12 h and light intensity 30W). Plants were divided into three groups, one of them inoculated with *Xanthomonas euvesicatoria* LMG 905 (Xeu) bacteria, other with *Pseudomonas syringae* pv. *tomato* DC 3000 (Pst), and the last was treated with sterile distilled water only (Control group). Plants were inoculated in the laboratory, at the growth stage of 5-6 fully expanded leaves, by spraying until they became fully wet, and run-off occurred. The bacterial suspensions used for these inoculation assays consisted of 1 x 108 cells / mL. They were prepared from 48-h-old culture grown on KB medium (peptone, 20.0g; $K_2HPO^4$, 1.5g; $MgSO^4$, 1.5g; glycerol, 10 mL; agar, 15g; distilled water up to 1.0 liter), in the case of Pst Bacteria, and YDC medium (yeast extract, 10.0g; dextrose, 20.0g; CaCO3, 20.0g; agar, 15.0g; distilled water up to 1.0 liter) for Xeu bacteria. The inoculated plants were then covered with transparent polythene bags for 48 h to increase the relative humidity that fosters bacterial entry into plant tissues through

natural openings such as stomata (Lamichhane 2015). Plants were monitored daily for symptom development for 7 days.

At the same time, to verify if the bacteria cultures used in these inoculation tests were viable, 20 µL of Pst solution and 20 µL of Xeu solution were cultured in different Petri dishes containing KB and YDC media, respectively. After 48 h bacterial growth was observed in both nutrient media, proving that bacteria were viable at the moment of inoculation.

### 2.2. Visual and PCR-based confirmation of bacterial infection

#### 2.2.1. Visual confirmation (Scouting)

A visual search for typical symptoms of the infection caused by Pst and Xeu was made daily during the assay duration. Pst disease-specific symptoms usually consist of small, greasy dark stains that become brown to black and appear randomly. They vary between circular or slightly angular shapes and typically have a yellow halo of various sizes. In the first moment, lesions are about 2–3 mm, which can develop and coalesce (especially in the presence of moisture), affecting large leaflet areas, that may later become necrotic and desiccate (Blancard 2012). In turn, Xeu characteristic symptoms affecting leaves comprise small, brown, angular, and water-soaked lesions. Smaller lesions can coalesce into each other forming larger angular injuries, whose diameter can range from 1.6 to 6.4 mm. With time, they can evolve and form necrotic spots, presenting light gray centers with dark margins, which can become surrounded by a yellow hallow with time. In severe cases, tissues in the center of the lesion become dry and fall out, leading to “shot-hole” symptoms (Jones, Jones et al. 1991, Rudolph 1993, Stall, Beaulieu et al. 1994, Ritchie 2000, Dutta, Gitaitis et al. 2014, Teper, Girija et al. 2018).

#### 2.2.2. Bacterial isolation

Sample preparation for bacterial isolation was carried out for asymptomatic and symptomatic leaves at 24 hours and 48 hours. Leaves were excised from plants using a sterile scalpel (Fernandes, Albuquerque et al. 2017). Bacterial isolation was performed as described by Fernandes et al. (2017, 2021). Briefly, each sample of excised leaflet tissue was disinfected by immersion in 70% ethanol followed by washing with sterile distilled water (SDW) and then macerated with SDW in extraction bags. The suspensions obtained were streaked on KB (samples inoculated with Pst bacteria) and on YDC medium (samples infected with Xeu pathogen). Characteristic colonies from these two bacteria species (milky white colonies in the case of Pst, and mucoid yellow colonies in the case of Xeu) were selected for growth on fresh nutrient agar medium to ensure purity.

### 2.2.3. PCR validation

A colony PCR was performed to validate the presence of Pst and Xeu bacteria on tomato leaflet isolates. PST2 (Vieira, Mendes et al. 2007) and XV14 (Albuquerque, Caridade et al. 2012) were the chosen markers for Pst and Xeu, respectively, with correspondingly amplicon lengths of 200, and 713 bp. A 20 µL PCR reaction mix consisted of 1 × DreamTaq Buffer (ThermoFisher Scientific, Waltham, MA, USA), 0.2 mM of each deoxynucleotide triphosphate (dNTP) (Grisp, Porto, Portugal), 0.2 mM of each forward and reverse primers, 1 U of DreamTaq DNA Polymerase (ThermoFisher Scientific, Waltham, MA, USA) and 10 µL of DNA isolate solution. Sterile distilled water was used as the negative control. PCR cycling parameters consisted of an initial denaturation step of 5 min at 95ºC, followed by 35 cycles of 30 s at 95ºC, 30 s at 57ºC, 59ºC or 61 ºC, and 30 s at 72ºC with a final extension step of 10 min at 72ºC (Albuquerque, Caridade et al. 2012) for Xeu, and an initial denaturation step of 3 min at 95ºC followed by 35 cycles of 30 s at 95ºC, 30 s at 63ºC and 30 s at 72ºC and a final extension step of 10 min at 72ºC (Vieira, Mendes et al. 2007) for Pst.

PCR products were then separated by electrophoresis on a 0.8% agarose gel (1 × TAE buffer) and visualized using Xpert Green DNA stain (Grisp, Porto, Portugal) with a Molecular Imager Gel Doc XR+ System (Bio-Rad, Hercules, CA, USA).

### 2.3. Spectral measurements

Hyperspectral data were collected in vivo from the adaxial side of excised healthy and diseased tomato plant leaflets. Measurements were performed through a randomized process, on three leaves per plant, on three points per leaf, at different times (24, 48, 72, and 96 hours for all treatments, and a final measurement at 144 hours only for control and Xeu diseased plants since the ones inoculated with Pst were very weakened). The experimental design involved an in-house compact benchtop system, consisting of a laptop, a spectrometer (HR4000, Ocean Optics Inc., USA) with a 200-1100 nm range, a transmission optical fiber bundle (UV-VIS-NIR, FCR-7UVIR200-2-45-BX, Avantes, Eerbeek, The Netherlands) with a 200-2500 nm range, a stainless-steel slitted reflection probe (placed 1 cm above the sample surface, to conduct leaflet's spectral signal to the entrance lens of the spectrometer), and a white LED light (placed beneath the leaflet for provide homogeneous illumination to its entire abaxial surface). Specialized software (SpectraSuite, Ocean Optics Inc., USA) were used.

### 2.4. Microscope imaging acquisition

Leaflet samples of each plant were excised from the plant and readily analyzed: samples were laid between glass slides and placed on the microscopical setup for data

acquisition. A 5×5 mm spacing sampling grid was devised on each leaflet, comprising 25 data points of spectroscopic data; sampling grid initial coordinates were randomly set, yet assuring the avoidance of the leaflet main stem line. No further sample preparation was required or performed.

The hyperspectral microscopy system consisted of a Zeiss Axiovert.A1, in transmission mode, fitted with a UV-Vis-NIR fiber optic (I.D. 600 µm) connected to an Ocean Optics model HR4000 spectrometer. An adapter was developed to allow the alignment of the fiber optic with the microscope camera in order to allow the collection of detailed hyperspectral information of the acquired microscopic images. Microscopical imaging was performed using the “Best Fit” option for macro (Red, Green, Blue – RGB) images, whereas spectroscopic data acquisition was conducted under the same illumination conditions as for microscopical imaging and within the 200-1100 nm range and every 24 h, until the 4$^{th}$ day, and one final sampling moment at 144h, in a total of 5 sampling moments per group.

**2.4. Data modeling**

The measured hyperspectral data was modeled using a Principal Component Analysis (PCA). PCA is a multivariate data analysis approach applied to reduce the dimensionality of the hyperspectral data while preserving its structure by projecting the data into a new coordinate system. This methodology retains the overall variance of the dataset while minimizing mean square approximation errors. Principal Component Analysis (PCA) utilizes eigenvectors and eigenvalues to establish the reduced subspace, representing the original coordinate system. It generates principal components (PC), which are linear combinations of interrelated variables. PC1 encapsulates the highest proportion of variance information from the original dataset, as explained by the eigenvalue. Subsequent principal components (PC2, PC3, etc.) consecutively encapsulate the highest proportion of the unexplained residual variance (Lee, Alchanatis et al. 2010, Liu, Cheng et al. 2012).

Linear Discriminant Analysis (LDA) was also applied since it is a supervised learning algorithm usually applied in classification tasks using hyperspectral data since it can reduce the dimensionality of the data while maximizing the class separability. LDA projects the high-dimensional data onto a lower-dimensional space while maintaining the discriminative information between classes. Shortly, data is projected onto a linear subspace that maximizes the ratio of between-class variance to within-class variance. Therefore, the projected data points are as far apart as possible in the new space, while the points of the same class are as close as possible. Hence, it contributes to reducing

the classification computational complexity and avoiding overfitting. Moreover, LDA can also facilitate data visualization in a lower-dimensional space, helping interpret patterns (Sachin 2015, Tharwat, Gaber et al. 2017).

In turn, Partial Least Squares Discriminant Analysis (PLS-DA) is a statistical method that aims to maximize the covariance between the predictor variables and the class information, enabling effective discrimination between different groups or categories. PLS-DA extracts latent variables that capture the essential information for classification, making it particularly valuable in situations with multicollinearity and high-dimensional data (Lee, Liong et al. 2018).

## 3. Main Findings

### 3.1. Visual phenotyping timeline of healthy and inoculated leaflets

Control tomato plants maintain an identical phenotype from the first to the last visual and spectral measurement (Figure 1). On the contrary, tomato plants infected with Pst and Xeu bacteria showed the first visual typical symptoms of the disease between 48 and 72 h after inoculation (Figure 2, 3). Pst bacteria promote chlorotic spots on leaves at 48h, which evolve into necrotic tissues at 72 hours (Figure 2). Chlorotic lesions in samples inoculated with Xeu mostly appeared at 72 hours, only evolving to the necrotic stage at 144h (Figure 3).

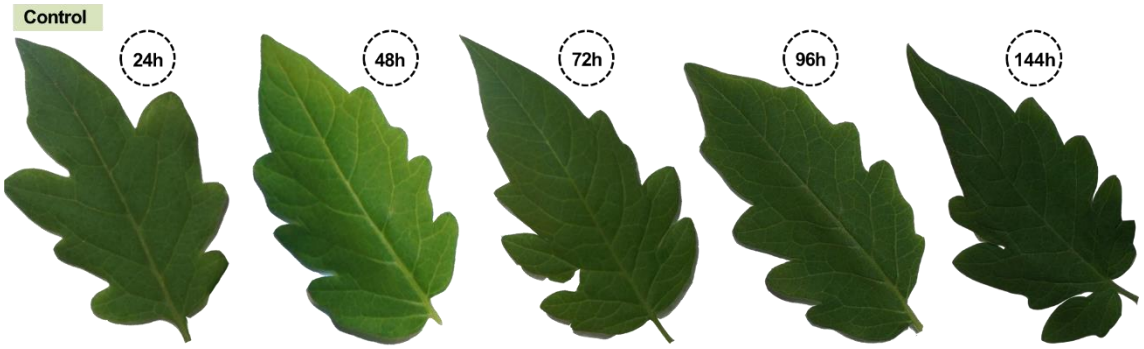



**Figure 1** Control leaflet phenotype evolution over time (from 24 to 144 hours).

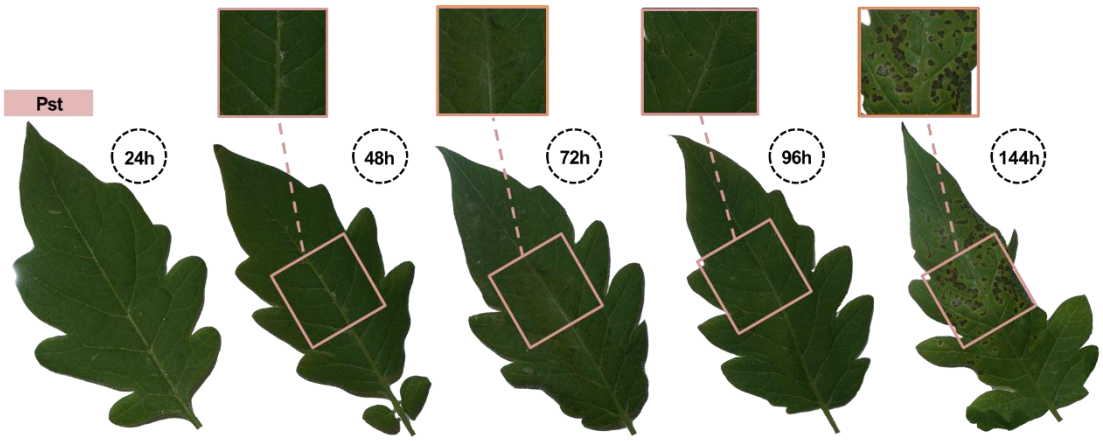



**Figure 2** *Pseudomonas syringae* pv. *tomato* (Pst) inoculated leaflet phenotype evolution over time (from 24 to 144 hours).

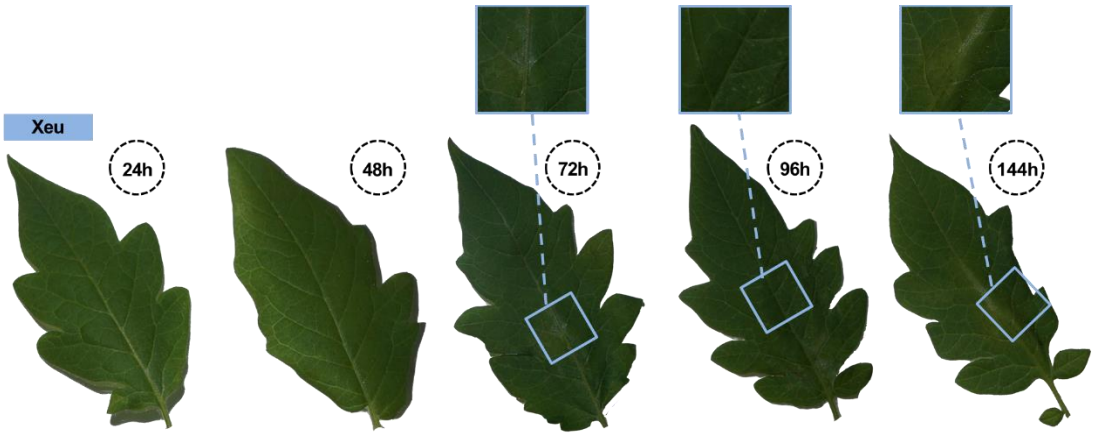



**Figure 3** *Xanthomonas euvesicatoria* (Xeu) inoculated leaflet phenotype evolution over time (from 24 to 144 hours).

### 3.2. PCR validation

After the phenotypical and spectral analysis, leaflet samples from each treatment were tested for the presence of these bacteria through a colony PCR. Bacteria-specific bands were amplified from samples, for each bacteria species (Figure 4). No corresponding DNA bands were amplified in PCR from samples collected from healthy leaves. These results indicated that Pst and Xeu bacteria were present in each inoculation treatment group (Figure 4).

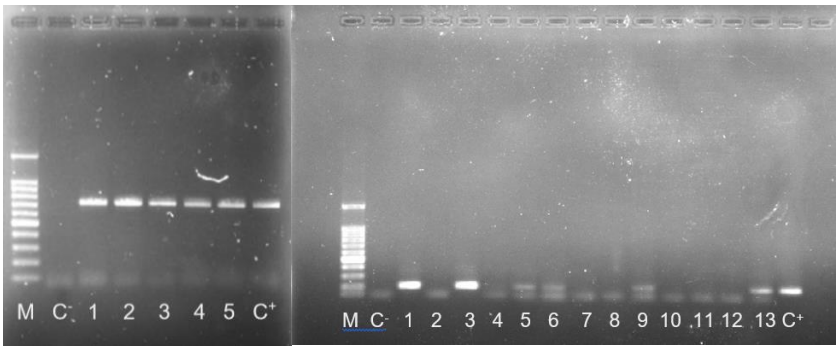



**Figure 4** A) Colony PCR of infected tomato leaflets with DNA markers XV14 to detect *Xanthomonas euvesicatoria* LMG 905. C- - Template sample (distilled water). 1, 2, 3 – DNA from different bacterial colonies obtained through a pathogen isolation assay performed 24 hours after infection (AI). 4, 5 – DNA samples from different bacterial colonies were obtained through a pathogen isolation assay performed 48 hours AI. C+ - *Xanthomonas euvesicatoria* DNA belonging to the laboratory bacterial collection. B) Colony PCR of infected leaflets with DNA markers PST2 to detect *Pseudomonas syringae* pv. *tomato*. C- - Template sample (distilled water). 1-6 – DNA from different bacterial colonies obtained through an isolation assay 24 hours AI. 7-13 – DNA from different colonies obtained through an isolation assay performed 48 hours AI. C+ - *Pseudomonas syringae* pv. *tomato* DNA belonging to the laboratory bacterial collection.

### 3.3. Hyperspectral point-of-measurement data analysis

LDA analysis allowed class distinction in hyperspectral samples collected on tomato leaflets only 24 hours after inoculation, especially between healthy (Control, green dots), and diseased tissues (red and blue dots) (Figure 5). At this point, no clear separation between the spectral data collected in tomato leaflet tissues inoculated with Pst, and the ones measured in Xeu-inoculated tissues. After 48 hours, class discrimination is even more evident, especially between the spectral measurements performed in tissues inoculated with different bacteria species. At 96 hours, all classes were fully discriminated against by the LDA algorithm (Figure 5). So, these results show that diseased tissues were detectable 24 hours after bacterial inoculation, and disease discrimination was achievable after 48 hours (Figure 5).

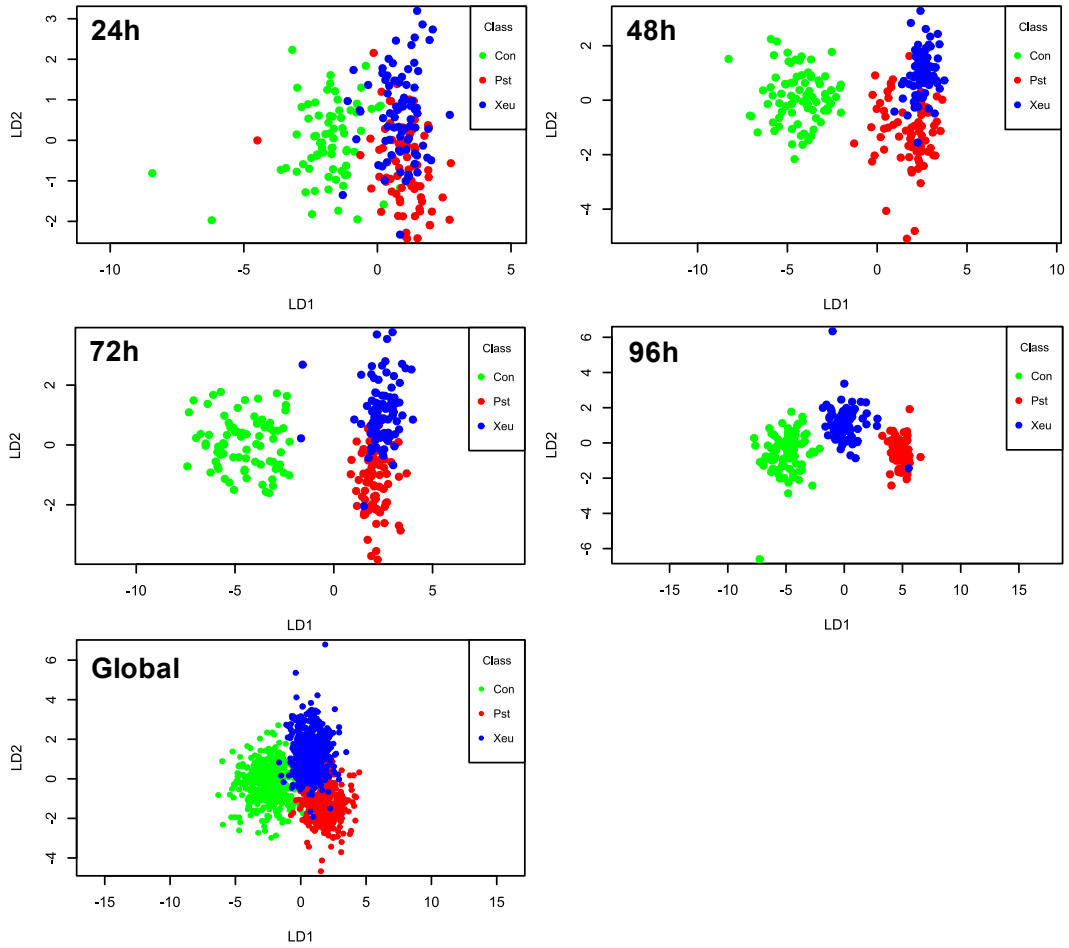



**Figure 5** LDA results using all the collected spectral measurements, and the spectra assessed only at 24, 48, 72, and 96 hours.

### 3.4. Hyperspectral microscopic tomography findings

Spectral differences between healthy (green line) and diseased (red and blue lines) at 96h were visible even when no macroscopic symptoms were visible in the RGB image, and then only slight variations in the green color were visible in the microscopic image (Figure 6)

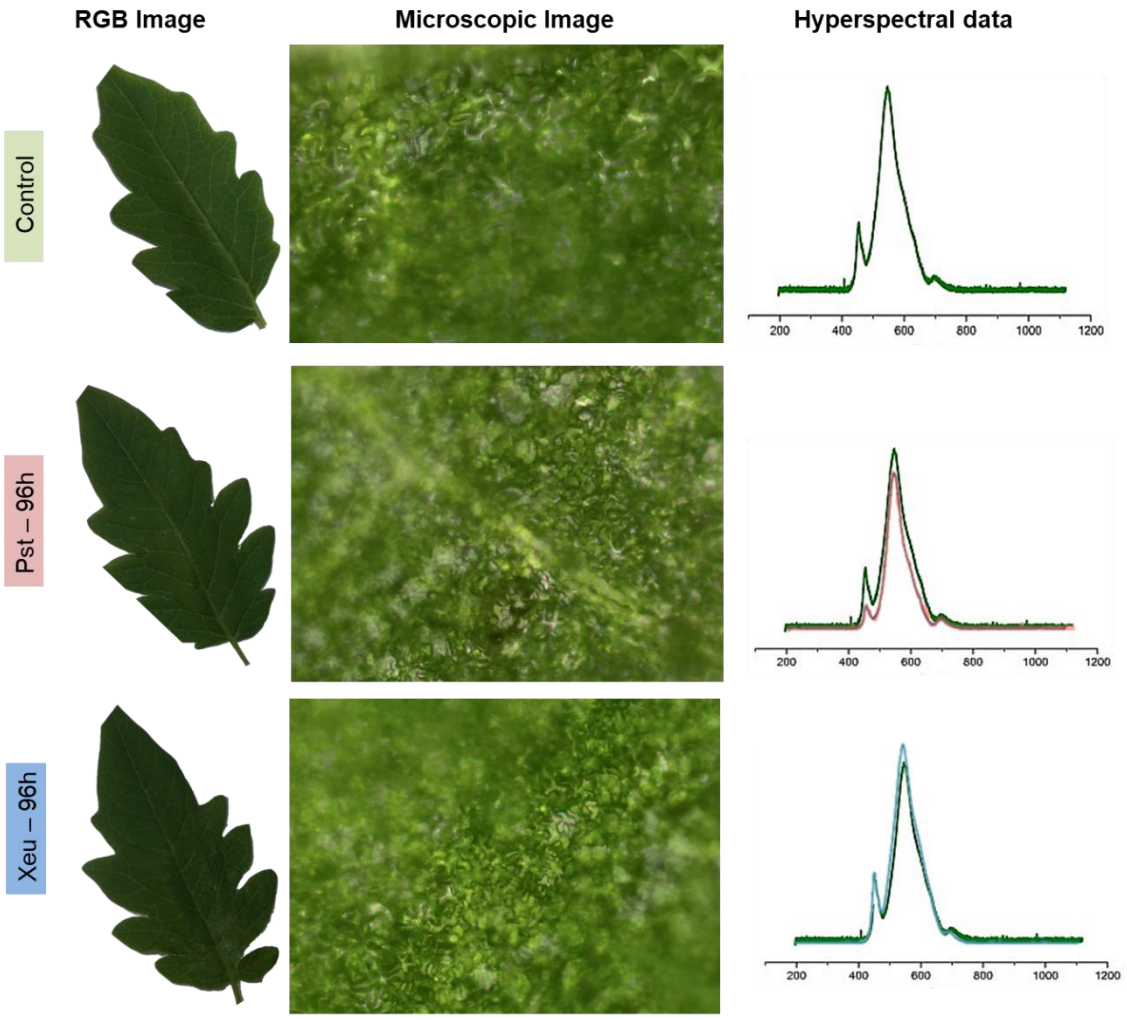



**Figure 6** Diagram showing the RGB image of the excised tomato leaflet and corresponding microscopic (200x) image and hyperspectral data (spectral curve in nanometers). Hyperspectral data represented in green corresponds to spectra measured at healthy (Control) leaflet tissues, in red to spectra collected on tissues inoculated with *Pseudomonas syringae* pv. *tomato* bacteria (Pst), and in blue to spectra assessed in leaflets inoculated with *Xanthomonas euvesicatoria* (Xeu), 96 hours after infection.

Results of a PCA applied to the hyperspectral data fused with microscopy imaging information showed that, at 144h after bacterial inoculation with Pst and Xeu, spectral variance can be spawned into three principal components (Figure 7). Thus, they can classify each category's spectra, namely Control (healthy), Pst, or Xeu. Moreover, in the third principal component, it is possible to observe a distinction between the spectral patterns from tissues inoculated with Pst from the ones infected with Xeu bacteria. This finding may be related to the fact that samples inoculated with Pst presented higher damage levels (related to a higher aggressiveness/virulence of the

bacteria), and the spectral patterns were indicative of necrosis. In turn, Xeu-inoculated leaflets showed more uniform damage over hyperspectral data (Figure 7).

A PLS-DA algorithm was, then, computed to estimate the probability of each bacteria at each node of the hyperspectral image (Martins, Santos et al. 2023). Spectral variance at each hyperspectral image was significant, due to both leaflet structures and compositional differences. Nevertheless, it did not influence the classification performance made by the linear classifier (Figure 7). PLS-DA predicted that the Control samples presented a small probability of infection, as expected (green, Figure 7). In contrast, PLS-DA predicted a high probability of infection for the samples inoculated with Pst and Xeu bacteria (Figure 7). These findings support the correspondent microscopy images, where at 144 hours all tissues are infected, not existing non-infected regions (Martins, Santos et al. 2023). Thus, the present outcomes indicate that microscopic analysis has the potential for early-stage diagnosis, confirming the results obtained by macroscopic tomography.

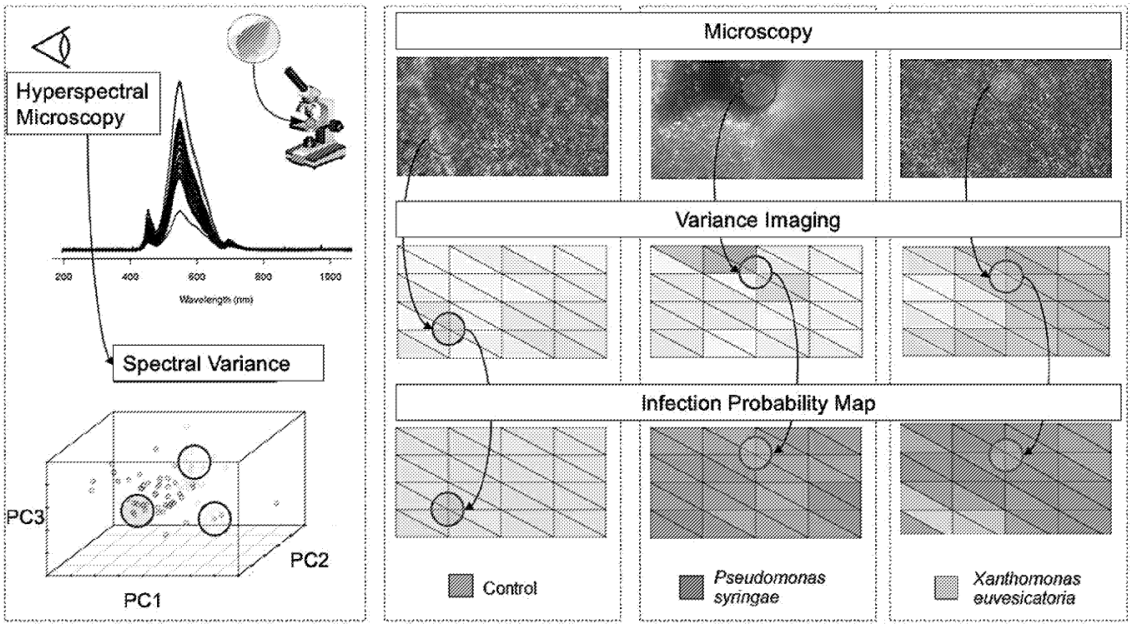



**Figure 7** Microscopic (200×) and hyperspectral data collected on tomato leaflets at 144 h after infection. The main outcome of Principal Component Analysis (PCA) is shown, along with the microscope images of healthy and diseased tomato leaflets, corresponding hyperspectral analysis by variance imaging, and the corresponding probability of infection determined by Partial Least Squares Discriminant Analysis (PLS-DA). The algorithm predicted that Control samples presented a small probability of infection, and a high probability of infection for the samples inoculated with *Pseudomonas syringae* pv. *tomato* and *Xanthomonas euvesicatoria* bacteria. Source:

WO2023126532 – Method and device for non-invasive tomographic characterization of a sample comprising a plurality of differentiated tissues (Martins, Santos et al. 2023).

## 4. SpecTOM Technology - A Spectroscopy-based Metabolomics Tomography Prototype System

These findings were integrated into the development of the SpecTOM project, a spectroscopy-based metabolomics tomography prototype system. It aimed to develop a non-invasive spectroscopy tomography system for providing compositional imaging of plant and animal tissues. This project was an upgrade of the MetBots (Metabolomic robots with self-learning artificial intelligence for precision agriculture) system (INESCTEC 2018). SpecTOM allowed the previous system to explore in detail plant internal structures and composition for a new omics approach, where molecular biology and plant physiology are key enablers of new diagnosis and agricultural practices.

SpecTOM is based on image reconstruction using latent hierarchical information fusion approaches to decode the recorded hyperspectral spectroscopy signal into its several components of the sampled tridimensional structures, i.e., the different plant tissues being assessed (e.g., tomato leaflet tissues). Hierarchical relations analyzed the parallax effect, which changes the spectral fingerprint depending on the point-of-view angle of the spectroscopy probe. This basic principle was used to develop the proof-of-concept of microscopic tomography for assessing spectral patterns that can be used to diagnose at very early stages different tomato bacterial infections, such as the ones caused by *Pseudomonas syringae* pv. *tomato* and *Xanthomonas euvesicatoria*. Proof-of-concept findings demonstrated the suitability of the technique.

SpecTOM won 2021 the BIP PROOF 20/21 (INESCTEC 2021) and in 2022, the project won the 8$^{th}$ edition of the Caixa Agrícola Entrepreneurship and Innovation Award, in the Agro-industry 4.0 category, which aimed to support digital technological solutions that promoted production optimization and efficient resource management (AGROTEC 2022, Negócios 2022, Silva 2022).

This project will also contribute to the development of the OmicBots project (High-Throughput Integrative Omic-Robots Platform for a Next Generation Physiology-based Precision Viticulture) (FCUP 2021, OMICBOTS 2024), which aims to explore the metabolic pathways of the grapevines to understand the physiology and metabolism of the vine in situ.

## 5. Patent WO2023126532 – Method and device for non-invasive tomographic characterization of a sample comprising a plurality of differentiated tissues

The described findings were furthermore used as a proof-of-concept for the development of a patent titled '*Method and device for non-invasive tomographic characterizations of a sample comprising a plurality of different tissues*'. The patent discloses a method and device for non-invasive tomographic characterization of a biological sample involving a plurality of differentiated tissues. In specific, it provides a computer-based approach for non-invasive tomographic metabolite characterization of a biological sample (from plant or animal base) (Martins, Santos et al. 2023).

**Acknowledgments**

Mafalda Reis Pereira and Filipe Silva were supported by fellowships from Fundação para a Ciência e a Tecnologia (FCT) with the references SFRH/BD/146564/2019 and DFA/BD/9136/2020, respectively. Rui C. Martins acknowledges Fundação para a Ciência e Tecnologia (FCT) research contract grant (CEEIND/017801/2018).

**Funding**

This research was supported by the project 'SpecTOM – Metabolomics Tomography Spectroscopy System', University of Porto, Fundação Amadeus Dias and Santander-Universities Grant. This work is partially financed by National Funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots, with reference PTDC/ASP-HOR/1338/2021.

# Case Study 7

Researchers collaborating in this case study (no specific authorship order):

- Doctoral Program Student: Mafalda Reis Pereira[1,2]
- Researchers: Filipe Neves dos Santos[2], Leonor Martins[1,3,4], Pedro Moura[2]
- Associate Professors: Fernando Tavares[1,3,4], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, 4169-007 Porto, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Roberto Frias, 4200-465 Porto, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

## 1. Contextualization

Due to the interesting findings regarding the usage of Thermography for early disease diagnosis presented in the review article mentioned in **Chapter I** of this thesis, and following the evidence that imaging information may carry important spectral information as related in the previous case study (**Case study 6**), two preliminary laboratory assays using tobacco plants (*Nicotidiana tabacum*) were conducted. They aimed to provide proof-of-concept of the suitability of Thermal and RGB Imaging for the early bacterial disease diagnosis, i.e., before the appearance of macroscopically visible lesions, at a non-symptomatic stage of the diseased plant. In this regard, further scientific experiments applying these two approaches for the early diagnosis of bacterial diseases are currently under development, whose outcomes are being prepared for later publication.

The tobacco species was chosen because it is currently considered a plant system model, presenting a high capacity to adapt itself to several environments, through the development of a broad range of morphological and chemical phenotypes (Baldwin 2001). Moreover, tobacco plants are natural allotetraploids that produce a million seeds per plant in three months after germination (Ganapathi, Suprasanna et al. 2004),

facilitating its maintenance and renewal. Moreover, in early disease diagnosis studies, the tobacco's leaf morphology facilitates the operability of the sensor used, along with the evaluation of its performance. The preliminary results shown here were partially communicated at two scientific conferences through oral presentations (**Appendix D** and **Appendix E**).

## 2. Tobacco plants bacterial infection assay

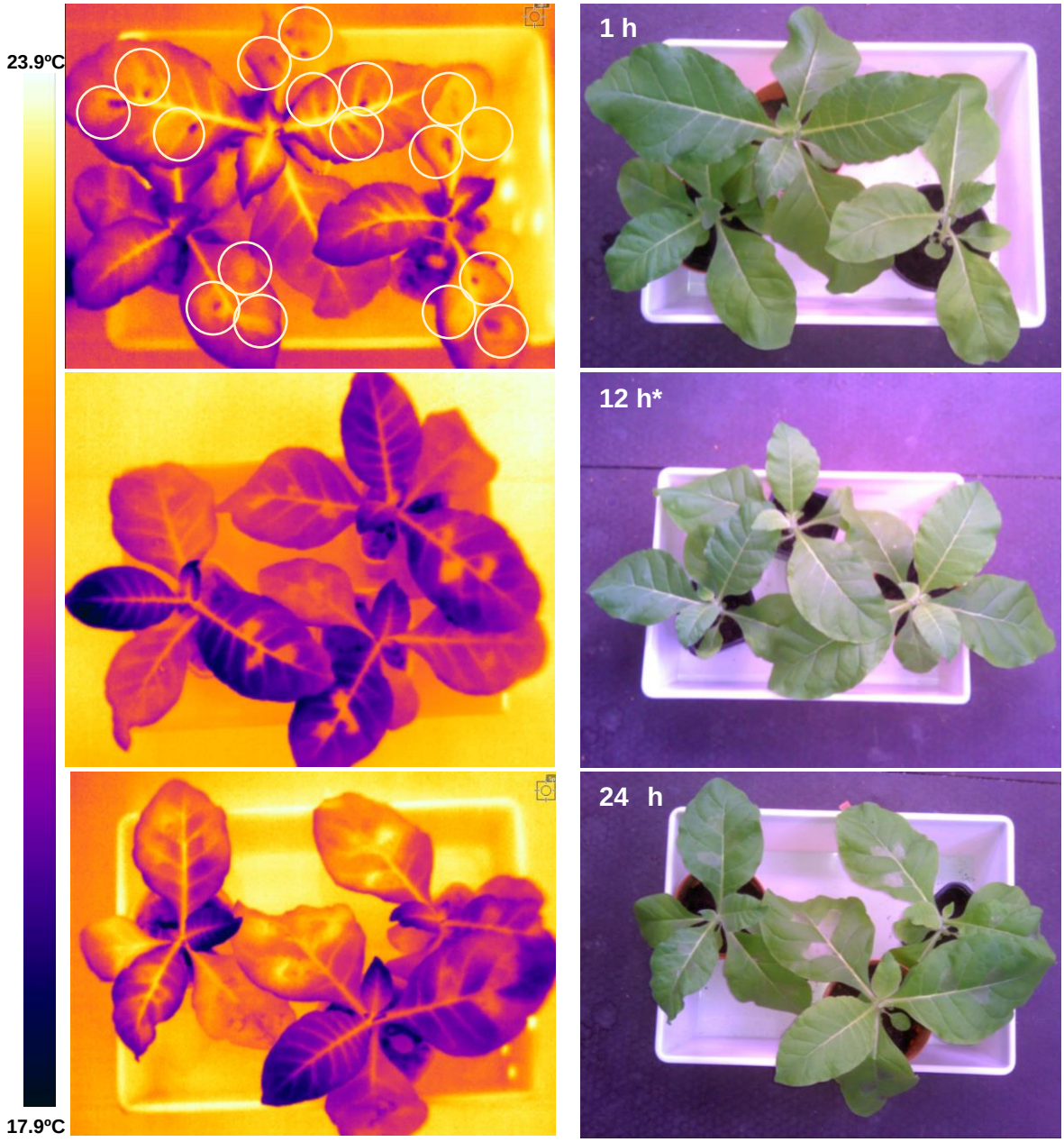
Tobacco plants were used in a hypersensitive reaction (HR) assay. Typically, HR is characterized by the rapid death of individual plant cells that come into contact with pathogenic organisms and is generally associated with disease resistance of the whole plant to the pathogen (Kiraly 1980, Klement 1982). In this HR analysis, the tobacco behavior after being inoculated with different bacteria was studied. With this purpose, 12 plants were selected and divided into four groups of 3 plants each. The plants were grown in 80 ml pots containing a commercial potting substrate in a plant Walk-in growth chamber at 25-27ºC, with a photoperiod of 12 / 12 hours, and humidity of 50 %.

Different bacterial strains, namely *Pseudomonas syringae* pv. *tomato* (Pst) DC 3000, *Xanthomonas arboricola* pv. *juglandis* (Xaj) CFBP 7179 (these two bacteria pathovars were considered the positive controls), Xaj CPBF 427, and Xaj CPBF 1521, were used in these inoculation assays. They were initially stored at -80ºC in the bacterial culture collection of the Microbial Diversity and Evolution (MDE) Group (CIBIO 2024). The bacterial growth occurred in Petri dishes incubated in a greenhouse at 28ºC, from 24 to 48 h, which was the time necessary to obtain isolated (pure) colonies. Xaj was cultured in Petri dishes containing the Nutrient Agar (NA) and Yeast extract-dextrose-CaCO3 media, and Pst DC 3000 was cultured in NA and King's B (KB) media. Later, an isolated (i.e., pure colony) of each one of these pathogens was suspended in liquid Lysogeny Broth (LB) nutritive medium, in Falcon tubes, and it was incubated at 28ºC, with agitation (220 rpm) overnight.

The plants were, then, inoculated by infiltration using blunt-ended syringes, in the $4^{th}$, $5^{th}$, and $6^{th}$ leaves, following the next arrangement: 3 plants were used as control plants (inoculated with purified water only); 3 plants were inoculated with a suspension of Pst DC 3000 in purified water at $10^8$ colony forming unit (CFU) $ml^{-1}$; 3 plants were inoculated with a suspension of Xaj CFBP 7179 in purified water at $10^8$ CFU $ml^{-1}$; 3 plants were inoculated with a suspension of Xaj CPBF 427 in purified water at $10^8$ CFU $ml^{-1}$; and, 3 plants were inoculated with a suspension of Xaj CPBF 1521 in purified water at $10^8$ CFU $ml^{-1}$. The first symptoms appeared 12-24 hours after the inoculation process, and they were fully developed 48 hours after the inoculation in every plant.

## 3. Thermography for the early assessment of the hypersensitive response in bacterial inoculated tobacco plants

"Convective temperature" changes undergo alterations in plants affected by non-symptomatic diseases, and these changes can be detected by using a portable thermal camera (FLIR 335, FLIR Systems, Sweden). This approach enables early detection of diseases in a non-contact and non-destructive manner. The assessment of the plant's emittance using this specialized equipment is crucial in achieving accurate results.




*Plants in this RGB image are in a slightly different arrangement than in the corresponding thermal image.

**Figure 1** Images captured with thermal camera of tobacco plants inoculated with *Pseudomonas syringae* pv. *tomato*. In the first hour after the inoculation, in the thermal image is possible to see the sites where the bacterial infiltration was performed, i.e., dark

blue spots (highlighted by white circles). These spots are, generally, surrounded by a higher temperature area, presenting a yellow color. The corresponding RGB image is on the left. At 12 hours yellow areas (higher temperature) can be observed in the diseased leaves, near the infiltration spots. At 24 hours, these yellow areas (of higher temperature) in the thermal image, are shown as lesioned leaf areas in the corresponding RGB image. The thermal image color scale is provided on the left side of the figure.

Measurements were performed in the first 72 hours that followed the inoculation process. The thermal images showed that after inoculation, the area near the infiltration site registered a higher temperature (represented with yellow in the thermal image, Figure 1), even when only one hour had passed. In these areas of higher temperature, disease symptoms later appeared, and originated necrotic lesions 24h after inoculation (Figure 1).

These findings seem to indicate that Thermography images may be suitable for early bacterial disease diagnosis. Nevertheless, modeling analyses are recommended to provide more robust evidence. Furthermore, it would be beneficial to verify the bacterial solution's temperature at the moment of inoculation to determine if it is inferior, equal, or higher than the temperature of the inoculated leaf. In that way, the origin of the thermal difference 'spots' around the local bacterial inoculation may be fully attributed to the host-pathogen interaction. Further assays may also consider the analysis of plants punctured with a syringe only, and plants inoculated with purified water only. Complementary research exploring different pathosystems and thermography sensors with different spectral resolutions is recommended.

## 4. RGB imaging captured under different LED light sources stimulation for the assessment of the hypersensitive response in bacterial inoculated tobacco plants

The results from **Chapter I**, which highlighted the use of RGB imaging for disease studies, prompted further investigation through exploratory assays employing this cost-effective and user-friendly system.

Image acquisition was made using a set-up including an RGB camera (Panasonic Lumix DC-FZ82) in combination with different light sources including individual Red, Green, Blue, and White LEDs, a UV LED floodlight, and an RGBW LED floodlight (Figure 2). This setup allowed the simultaneous imaging of 6 plants (3 control and 3 inoculated) (Figure 3).

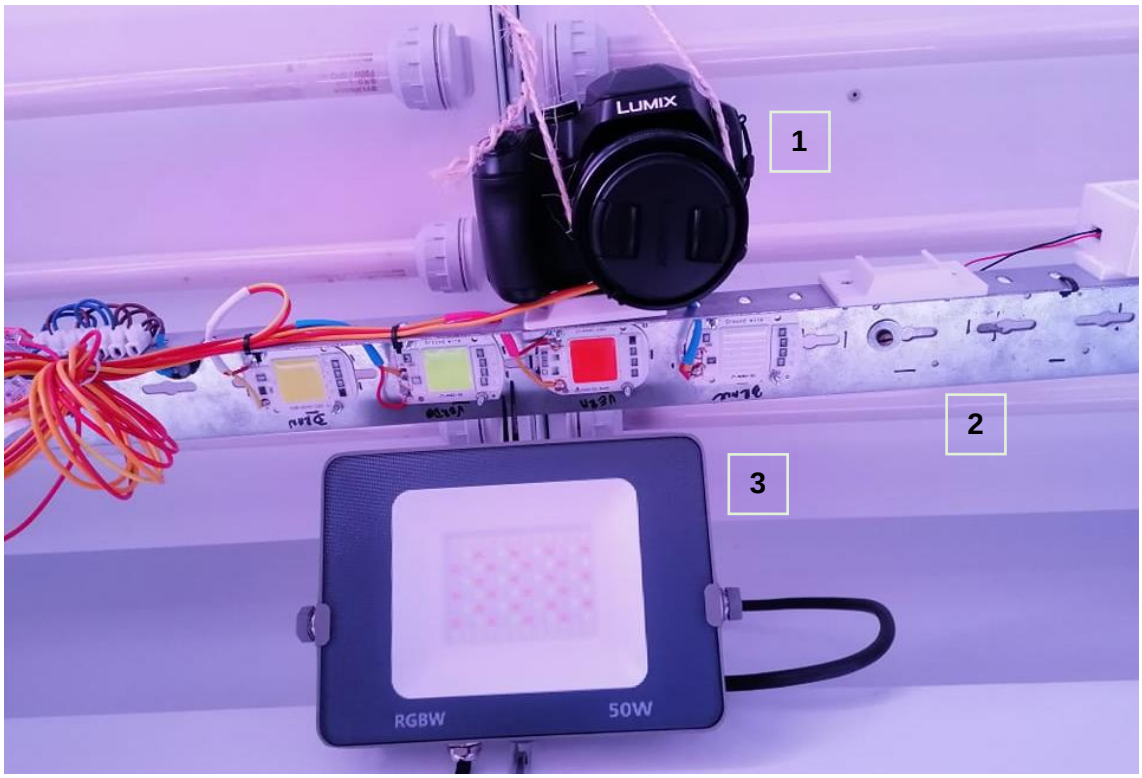



**Figure 2** Set-up composed of an RGB camera (1) in combination with different light sources including individual Red, Green, Blue, and White LEDs (2) mounted on an aluminum plate, and a floodlight (UV and RGBW) (3).

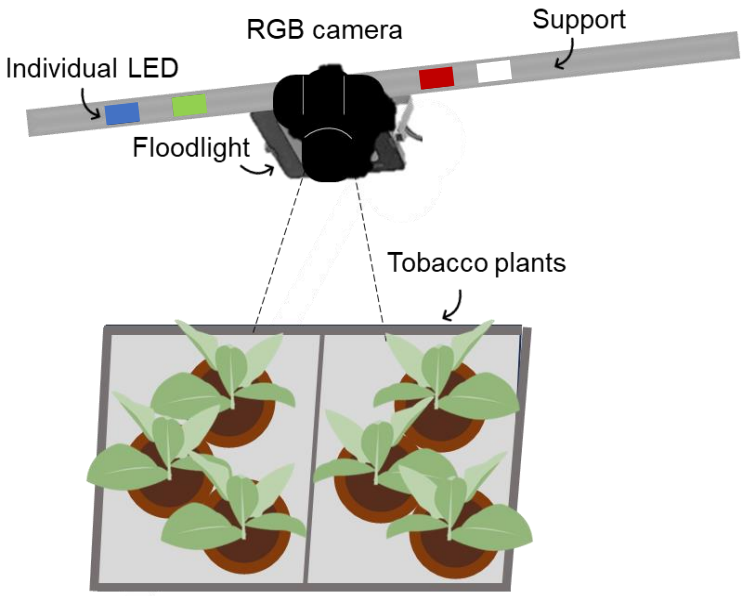



**Figure 3** Diagram showing the experimental conditions, the image acquisition set-up, and the plant arrangement assessed (three control and three inoculated plants).

The different artificial light sources were explored to investigate the distinct plant-light-sensor interactions and determine the efficacy of each illumination source in the diagnosis of bacterial diseases in plants. The outcomes may also justify the development

of an active RGB sensor-based system, using these LEDs, of reduced cost and high operability.

The photos were taken in a dark room, and the process started 4 hours before the inoculation process began, being repeated every 4 hours (between 09 and 20 hours). This assay was performed in the first 72 hours that followed the inoculation process. The data was then stored on an SD card and loaded on a computer.

A part of the images captured at 0, 12, and 24 hours after bacterial inoculation can be seen in Figure 4. RGB images of tobacco plants were taken using the Green, Red, Blue, and White individual LEDs. Of the six plants analyzed, three were inoculated with purified distilled water only and the remaining with a bacterial solution of Pst bacteria following the protocol described in the previous subsection 2.

At the moment, no significant visual enhancement could be observed in the outputs shown. Further image and spectral processing analysis (both qualitative and quantitative) are, hence, recommended to verify if there are non-visual differences between the images captured that could enhance the early bacterial diagnosis, providing more robust conclusions. Complementary experimental assays should also be repeated to validate the results, and different plant-pathogen interactions should be studied.

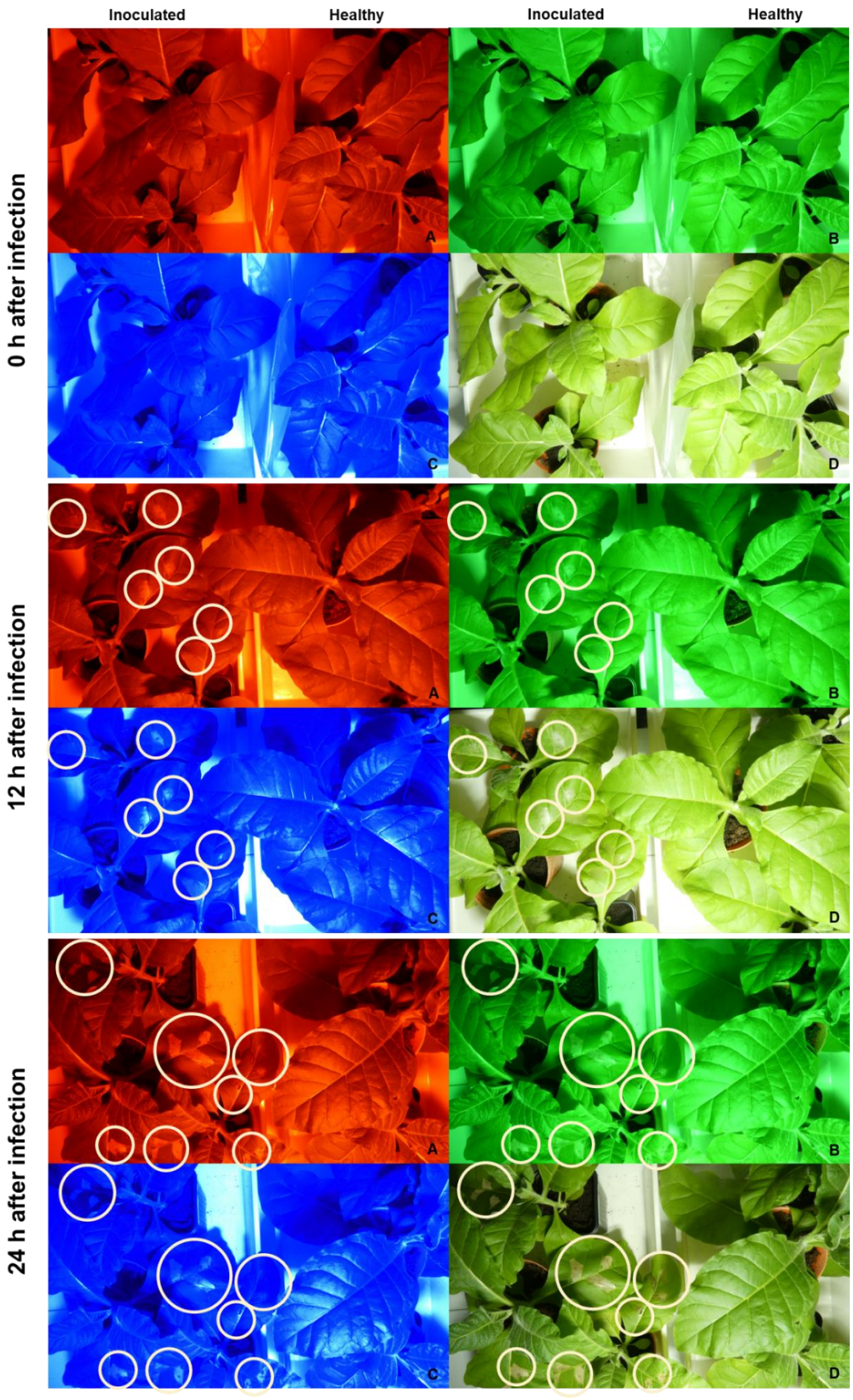



**Figure 4** Images captured with an RGB camera using stimulation with different individual LEDs source 0, 12, and 24 hours after the bacterial inoculation process in three tobacco

(left) plants using *Pseudomonas syringae* pv. *tomato* bacteria. Three control plants were also included in the image for comparison. For each time point (0, 12, 24 hours), the letters in each figure represent the type of LED radiation used to stimulate the plants: A) Red LED, B) Green LED, C) Blue LED, and D) White LED. Yellow pale circles highlight the hypersensitive response lesions.

# Chapter IV |
# General Discussion

# General Discussion

The main objective of this doctoral thesis was to investigate the suitability of using proximal optical sensed data for the early diagnosis of bacterial plant diseases in tomato (herbaceous, annual crop) and kiwi (woody, perennial crop). Moreover, several predictive modeling approaches were investigated for discriminating healthy and bacterial diseased tissues, as well as discriminating diseased tissues affected by different pathogen species.

A critical review (**Case Study 1)** exploring the scientific works regarding the application of proximal sensing for early disease diagnosis was made, aimed to identify the type of crop and pathogen studied, the environmental conditions (i.e., laboratory, greenhouse, and field) of the experimental assays, the sensor used, and the data handling and modeling procedures computed, and the performance of the predictive approaches for disease assessment.

## 1. Pathosystem dynamics and experimental environment: Unraveling the Disease Triangle Components

Tomato, wheat, sugar beet, and soybean were identified as the most studied crops in **Case Study 1**, mentioned in 16%, 13%, 9%, and 9% of the screened articles, respectively. These four species present high economic importance due to their widespread cultivation and consumption worldwide, especially wheat, which is furthermore considered a staple food source. Beyond their socio-economic importance, these crops exhibit straightforward cultivation and maintenance requirements. Their short life cycles contribute to their suitability for diverse research studies conducted within a relatively condensed timeframe. This combination of economic significance, global prevalence, and ease of study positions tomato, wheat, sugar beet, and soybean as valuable subjects for comprehensive agricultural investigations. In this thesis, tomato plants were used as herbaceous, annual species of interest in **Case Studies 2, 4, and 5** (**Supplementary Materials | Paper I**), along with tobacco plants used in **Case Studies 6,** and **7**. Furthermore, kiwi plants were also investigated as a woody, permanent species of interest in **Case Studies 2, 3** (and **Supplementary Materials | Paper II**), due to their great agronomic, and economic relevance in Portugal, and Europe. The studies also wanted to provide some new insights for this crop which was scarcely used in the literature regarding the usage of proximal sensing for bacterial disease diagnosis.

Regarding the etiological agents analyzed, fungi were the most used pathogen (referred to in 53% of the articles assessed in **Case Study 1**), followed by viruses (24%),

bacteria (approximately 18%), and pests (around 9%). The economic importance of these pathogens, global distribution, visibility, symptomatology, historical emphasis, pathogen collection availability, pathogen complexity, ecological significance, and resistance to phytosanitary products may be related to these outcomes. Also, fungal diseases frequently manifest themselves more conspicuously in plants compared to viral, bacterial, or pest-related diseases (even before symptom appearance). Thus, this may enable the understanding related to changes promoted by plant-pathogen interactions, motivating more research using these organisms.

In this study, bacteria were chosen as the etiological agent of interest due to being used in a lesser extension in the screened scientific literature, aiming to fill this gap, and also because their diagnosis poses an extra challenge due to the localized infection (more difficult to assess). The *Pseudomonas* spp. and *Xanthomonas* spp. were, moreover, selected to integrate the studies since they belong to different genera but affect a broad range of plant hosts, causing similar macroscopic symptoms in plant leaves (especially in the first visual lesions). Hence, the possibility of discriminating disease tissues affected by different bacteria could also be explored, complementing the distinction between healthy and diseased samples.

The majority of the aforementioned works were conducted in laboratory (control, 69% of the articles analyzed in **Case Study** 1) conditions, followed by greenhouse (22%), and field assays (20%). This may be related to laboratory conditions making possible a more detailed assessment of the host-pathogen interactions in the study due to the controlled environmental conditions which prevent the occurrence of any noise deriving from other types of biotic (e.g. action of other pathogens), and abiotic agents (e.g. effects of water, nutritional, light and temperature imbalances). Furthermore, these conditions also allow a more conducive environment for the use of sensors due to their structured light properties. On the contrary, field conditions allow the same study but in real circumstances, which are currently more complex and challenging to explore but also more interesting. They also allow the validation of previous studies made in controlled conditions. This thesis endeavored to encompass investigations spanning both controlled environments (**Case Studies 2, 4, 5**, and **Supplementary Material | Paper I**) and field conditions (**Case Studies 2, 3**, and **Supplementary Material | Paper II**). This comprehensive approach aimed at enhancing the generalizability of the results obtained.

## 2. Proximal Sensing technologies and instrumentation

In terms of proximal sensing technologies, Biophoton Emission, Fluorescence Spectroscopy, Laser-Induced Breakdown Spectroscopy, Multi- and Hyperspectral Spectroscopy, Nuclear Magnetic Resonance Spectroscopy, Raman Spectroscopy, RGB Imaging, Thermography, Volatile Organic Compounds (VOC) assessment, and X-ray Fluorescence Imaging were the main identified during the screening process made in **Case Study 1**. Hyperspectral spectroscopy emerged as the predominant technique, referenced in 82% of the evaluated articles, for conducting early disease diagnosis—detecting issues prior to the visible manifestation of macroscopic symptoms. Notably, this method demonstrated the highest Technology Readiness Level among the various approaches investigated. In this regard, it was investigated the suitability of two types of hyperspectral sensors for early bacterial disease diagnosis. The first was based on a passive reflectance spectroradiometer, measuring data between 325 and 1075 nm, and was mainly applied in studies involving kiwi plants (**Case Study 2, 3**, and **Supplementary Material | Paper II**). The second sensing technique was based on an active (i.e. using a proper light source for sample irradiation) point-of-measurement transmittance spectrometer, capturing data between 400 and 800 nm, and was mostly used in studies using tomato plants (**Case Study 2, 4, 5**, and **Supplementary Material | Paper II**). Transmittance data conveys information from the sample's surface and first tissue layers, similar to reflectance data**,** and also from more in-depth parts (where bacterial development and colonization usually happen). This type of data allowed the discrimination of not only healthy and diseased tissues but also between diseased tissues affected by distinct bacteria (**Case Study 2, 4, 5**), even before symptom appearance (**Case Study 4, 5**).

The potential of imaging-based techniques, as well as of spectral data fusion (between single and multi-point measurement devices) for early disease diagnosis identified in **Case Study 1** was further explored in **Case Study 6** and **Case Study 7**. Both strategies are currently under development, but the first insights indicate their suitability for the early diagnosis of bacterial diseases, namely when Thermography images or hyperspectral and imaging data fusion were used.

Overall, the outcomes showed that both hyperspectral transmittance and reflectance spectroscopic data can be used to identify healthy and diseased tissues and discriminate diseased tissues affected by different bacteria, both in laboratory and field conditions, in tomato (herbaceous) and kiwi (woody) crops. Nevertheless, only hyperspectral point-of-measurement transmittance data, captured in laboratory conditions, allowed the early diagnosis of bacterial diseases. The outcomes allowed us

to accomplish the main objectives and goals of the present thesis, as well as to respond to the research questions previously made in **Section I.**

## 3. Explore the modelling approaches

**Case Study** 1 also identified the main strategies employed after data collection in handling and modeling steps. Usually, data is sequentially subjected to four preparation steps: preprocessing, Feature Engineering (i.e., Feature Selection and/or Dimensionality Reduction techniques), modeling, and performance evaluation (Figure 1).

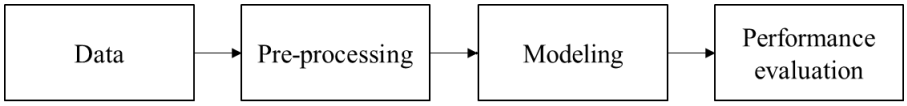



**Figure 1** Flowchart of the main steps for spectral data analysis.

The most used pre-processing techniques were the Normalization (used in 7% of the articles analyzed in **Case Study 1**), Standard Normal Variate (4%), Multiplicative Scattering Correction (4%), and Savitzky−Golay filter (4%). They dealt with missing values and outliers, as well as with denoising and smoothing data tasks. In the present thesis, Multiplicative Scattering Correction and normalization were used in **Case Studies 3 and 4,** and in **Case Study 5** a Savitzky-Golay filter and Standard Normal Variate were used.

In terms of Feature Engineering, it was mostly used for reducing spectral data high dimensionality, resulting from similar or even overlapping information presented in continuous variables. This redundancy increases the complexity of data analysis and increases the risk of overfitting occurrence when modeling strategies are later computed. Furthermore, it was used to deal with the super-imposed information present in spectral data collected at biological tissues, which occurs at different interference levels (Tosin, Martins et al. 2022). Spectral Vegetation Indices (VIs) were largely used in the articles screened (28% of the studies). In this regard, several Vegetation Indices and wavelength combinations were tested in this thesis in **Case Study 2** (**Supplementary Material | Paper 2**). Despite their suitability and relevant results, VIs only consider a limited number of wavelengths and may lead to information losses when hyperspectral narrowband sensors are used. Other important strategies identified in **Case Study 1**, included data condensation into 10 nm bands (that we later used in **Case Study 2**), Sequential Forward Floating Selection Search Strategy and the Jeffries–Matusita Distance, Stepwise Forward Variable Selection Method using Wilk's Lambda Criterion, and Lasso Regularized Generalized Linear Model (used in **Case Study 3**).

In turn, Dimensionality Reduction was computed to transform the original feature data space into a lower-dimensional representation. Principal Component Analysis (mentioned in approximately 20% of the publications presented in **Case Study 1**), Partial Least Squares (9%), Partial Least Squares Discriminant Analysis (7%), and Linear Discriminant Analysis (4%). In this regard, **Case Study 3** explored the effects of Linear Discriminant Analysis, and **Supplementary Material | Paper 1** tested Principal Component Analysis. All these data preparation techniques improve interpretability, simplify visualization, decrease the computational cost, help identify and improve useful spectral features, and enhance model performance.

Applied modeling techniques are usually then computed for performing predictive tasks. In instances where the variable in analysis was continuous, and represented mostly by numeric values, a regression technique was used (quantitative analysis). On the contrary, when the target variable was based on categorical values (e.g., classes or categories), classification models were computed (qualitative approach).

Classification approaches were mentioned in 80% of the analyzed scientific articles in **Case Study 1**. Our results demonstrate that the most applied techniques were Machine Learning based, namely Support Vector Machines (used in 20% of the studied scientific articles in **Case Study 1**), Discriminant Analysis (22%), k-nearest Neighbor (16%), Partial Least Squares (13%), and PLS-Discriminant analysis (16%). Hence, the potentialities of predictive classification models based on Machine Learning approaches were explored in **Case Studies 2, 3,** and **4**. **Case Study 2** explored Discriminant Analysis and an innovative Gaussian Process Classification Band Analysis Tool recently released in ARTMO's (Verrelst, Rivera et al. 2011) software. **Case Study 3** also studied Discriminant Analysis, as well as Generalized Linear Model, Partial Least Squares, and Support Vector Machines. In turn, **Case Study 4** investigated the potential of Support Vector Machines. In an innovative approach, a chemometrics-based classification model technique was tested in **Case Study 5**, namely authentication, aiming for the earlier identification of bacterial diseases in tomato. The Data-Driven Soft Independent Modelling by Class Analogy (DD SIMCA) was the model computed to perform this task.

All the strategies showed potential for discriminating healthy from diseased tissues (**Case Study 2, 4, 5**) and between symptomless and symptomatic samples (**Case Study 2, 3**), both in laboratory and field conditions, and using transmittance and reflectance hyperspectral data. Furthermore, Support Vector Machines and Data-Driven Soft Independent Modelling by Class Analogy using hyperspectral point-of-measurement transmittance data showed great potential for classifying samples even before the

macroscopic manifestation of disease symptoms, i.e., in the non-symptomatic stage (**Case Study 4, 5**). Thus, these two approaches are recommended to perform early bacterial diagnosis in both herbaceous and woody crops.

Despite the success of all the strategies explored in this thesis, authentication demonstrated a great potential for early distinguishing healthy from diseased tissues only requiring the spectral data of healthy samples to perform the classification task. This relevant finding streamlines the classification process, eliminating the need for prior plant inoculation in controlled environments or the identification of diseased plants in the field to establish their spectral behavior for use in predictive classification.

In general, the literature data was split into training and validation sets. The model was usually developed using the training data and later validated or tested on separate datasets. This procedure prevents overfitting, which occurs when a model becomes excessively attuned to the intricacies of the training data. Overfitting can lead the model to incorporate not only the underlying structured patterns within the data but also the inherent noise and random fluctuations. Consequently, an overfitted model might struggle to offer consistent and reliable predictions [123]. The present thesis used the same strategy in **Case Studies 2, 3, 4,** and **5**.

Cross-validation approaches were also performed to provide a reliable estimate of a model's generalization performance (used in approximately 35% of the articles screened in **Case Study 1**). This strategy was also followed in our thesis, in **Case Studies 2, 3,** and **4**).

Predictive models were, lastly, evaluated to assess the success or failure of their performance. In the literature review made in **Case Study 1**, regression models were mainly appraised according to their coefficient of regression ($R^2$), and root mean square error (RMSE) (mentioned in 20% of the scientific articles). Classification models, in turn, were ranked by determining their confusion matrix (CM), accuracy, precision, F1-score, and Kappa coefficient (mentioned in more than 20% of the articles). These metrics were also computed in this thesis.

## 4. Biophysical meaning associated with the predictions

Most of these predictive modeling strategies assessed in the literature review were data-driven and did not consider plants' physiology and the biological significance of the results. This thesis attempted to establish these connections in all the case studies performed, especially by determining which were the most relevant wavelengths contributing to class discrimination. In **Case Study 2**, the spectral wavelengths selected

by both modeling strategies were in the blue (450.04 nm), green (550.20 nm), and red edge (680.02, 690.42, 700.41, and 750.17 nm) spectral regions when the tomato dataset was used. In turn, when the kiwi dataset was analyzed the features were located in the blue (400, and 450 nm), green (530, 544, 553, 554, and 597 nm), red-edge (670, 677, 700, 705, 730, 750, and 754 nm), and NIR (Vegetation Indices picked 780, 800, 994, and 1000 nm, and Gaussian Process Classification Band Analysis Tool choose 771, 790, 791, 795, 825, 835, 839, 845, 850, 851, 860, 864, 866, 869, 881, 883, 888, 893, 902, 905, 906, 928, 932, 939, 945, 947, 973, 980, 993, 999, and 1006 nm). **Case Study 3** presented similar results for diagnosing bacterial diseases in kiwi leaves. Spectral wavelengths located mainly in the blue (350–500 nm), green (500–600 nm), red (600–750 nm), and NIR (>750 nm) regions were identified as relevant. The same finding was performed in **Case Study 4** where forty-four spectral features were considered important. They and were mostly located in the blue-green and red visible regions of the electromagnetic spectrum (blue - 434.9, 435.72, 438.17, 438.58, 440.21, 441.44, 442.67, 443.08, 445.53, 445.94, 448.4, 448.81, 494.6 nm; green - 503.74, 508.74, 527.53 nm; red - 556.09, 562.0, 562.84, 590.37, 607.82, 609.1, 611.24, 618.5, 643.36, 650.24, 673.97, 680.02 nm). Lastly, in **Case Study 5** the bacterial effect in Pst inoculated samples was visibly manifested in the band between 430 and 475 nm when compared to control samples, while in Xeu inoculated samples, the bacterial impact was clearly expressed in the spectral range of 675-800 nm.

These variables align with the absorption wavelength ranges of chlorophylls (430 to 480 nm and 640 to 700 nm) and carotenoid pigments, specifically β-carotenes, with primary and secondary absorption peaks located at 450 to 480 nm and 600 to 650 nm, respectively. Additionally, the spectral ranges of xanthophylls (520 to 580 nm), certain phenolic compounds (e.g., flavonoids, 400 to 500 nm), and composts derived from chlorophyll decomposition, namely pheophytins (400 to 500 nm and 600 to 700 nm), coincide with the observed findings.

Since plant-pathogen interactions induce changes in the photosynthetic pigment content, water levels, and structural composition of the host, changing its spectral behavior in the VIS-NIR region, these findings present important biological relevance (Blancard 2012).

## 5. Advancing early disease diagnosis

The integration of proximal sensors with diverse modeling strategies emerges as a promising avenue for advancing early disease diagnosis, occurring prior to the visible appearance of symptoms. The inherent advantages associated with in-situ, in-vivo

conditions underscore the significance of these sensing devices in close-range applications. The proposed methodologies contribute to more proactive and timely agricultural interventions, encompassing disease monitoring, management, and the execution of phytosanitary measures. This approach enhances the efficacy and precision of protective measures, facilitating the judicious selection of appropriate phytosanitary compounds, administered at the right dosage and timing. Consequently, it aids in minimizing product residues and mitigating pathogen resistance. In effect, these practices align with the principles and objectives of precision agriculture, fostering more sustainable agricultural practices.

Nevertheless, more research is recommended to unveil specific host-pathogen interactions, including the related metabolic, structural, and physiological changes. Since this is an indirect method of disease diagnosis, it is also important to discriminate the effect of these changes in the spectral signature of diseased plants from modifications occurring due to other types of stress (e.g. water deficit, nutritional imbalances, among others). Relating predicting modeling outcomes with the plant’s physiology and biological significance holds the potential to address this inquiry and substantially enhance the diagnostic method's utility. Additionally, incorporating authentication classification approaches that integrate the retrieval of the sample's pure spectral profiles can serve as forensic evidence regarding the sample's health status. This proactive measure helps prevent confusion arising from interferences caused by other biotic and abiotic agents.

Furthermore, Hyperspectral Spectroscopy sensors still present a technological maturity level with a high potential for progression, and further studies and improvements in these sensors are necessary. Additional development of high-resolution, cost-effective, portable, and easy-to-use spectral devices is suggested for enhancing the diagnosis of bacterial plant diseases, especially in greenhouse or field conditions. The spectral wavelengths highlighted as relevant for early disease diagnosis may contribute to this point. Also, spectral data acquisition and modeling protocols should be developed and standardized.

# Chapter V |
# Remarques and perspectives

# Final remarks and perspectives

Plant diseases greatly impact agricultural production worldwide, affecting crop yields and, consequently, leading to negative impacts on farmers' incomes, availability, and quality of agricultural products (affecting the supply of food, feed, clothing, and building materials), and resulting in higher prices of these products for consumers. Thus, it's urgent to develop innovative solutions for precocious disease diagnosis. Currently, professional visual scouting, biochemical analysis, and pathological analysis have been well investigated and explored. Non-invasive technologies have also been considered in recent years and have begun to be explored.

This work highlights the potential of hyperspectral spectroscopy combined with chemometrics or applied predictive modeling as a simple, quick, reliable, non-disruptive, reagent-less, cost-effective tool for diagnosing different bacterial plant diseases, both in controlled (laboratory) and in-field environmental conditions, as can be seen in **Case Study 2, 3, 4, and 5**. It describes several protocols for spectral data acquisition and modeling, which are effective for early discriminating healthy from bacterial-diseased leaves, in both herbaceous (tomato) and woody (kiwi) crops.

Nevertheless, it is important to address that the proposed protocols constitute an indirect method of diagnosis and may suffer from interferences with other types of stress, mostly with abiotic factors (e.g., meteorological, nutritional, and water conditions). Thus, additional studies in different plant conditions and stages might be relevant. Further research should also be carried out for different pathosystems, including the study of distinct bacteria species and even pathovars to see if the relevant spectral wavelengths for disease diagnosis are stable or if, on the contrary, they change. Also, the study of different types of pathogens (such as fungi, viruses, and pests) could fortify the suitability of hyperspectral data combined with different predictive models for biotic stress assessment.

Additional development of high-resolution, cost-effective, and portable spectral sensors is recommended for enhancing the evaluation of crop diseases. By providing powerful tools for early, in situ, in vivo diagnosis of infections, these innovative methods will constitute an opportunity to perform efficient, personalized disease control. Also, the possibility of coupling these hyperspectral sensors in robotic platforms will create the opportunity for continuous monitoring of plants.

These ambitions may be currently achievable, since more cost-effection sensors may be developed considering the most relevant spectral wavelengths identified in the

different case studies presented in this thesis. These sensors may, then, be coupled on different types of ground-based platforms, such as high throughput phenotypic platforms and robots, performing data collection and storage with inter-institutional standards and protocols (e.g., meta-data, ontology, semantics), which allows for efficient data exchange.

In the future, the currently used commercial sensors, as well as future design broad band devices, are intended to be added to the AgIoT modules (INESCTEC) developed by the author's team TRIBE - Laboratory of Robotics and IoT for Smart Precision Agriculture and Forestry (INESCTEC 2024). This solution is already equipped with other sensors and integrates robots and tractors to acquire data in field conditions. The information from here could be used for plant disease forecast, diagnosis, and for the application of several plant protection measurements, providing insights about the microclimate (at leaf, canopy level), phenology, canopy's biometric parameters (e.g. Leaf Area Index – LAI).

Concerning data storage, the information gathered can be gathered on the team's server http://vcriis01.inesctec.pt:1880/ui/#/0/ considering the FIWARE framework. The archived data may be, additionality, stored considering the 'meta formats' designed in the H2020 DEMETER project (DEMETER 2021), in which the team participated. Furthermore, the four datasets shared in Zenodo could also be used in new studies, regarding plant disease and pathology studies, also encouraging data sharing between researchers.

Furthermore, ground sensors, performing measurements at the leaf level, can be used as reference or training input for airborne-based platforms and scaled to field and plot levels, i.e., allowing the upscaling of the disease predictive models. The findings and deliverables of this study may now serve as the basis for further studies on this subject. Also, the potential of the robotic platforms coupled with different optic and environmental sensors can be used to map the differentiated risk of disease. This mapping can, then, be considered for differentiated phytosanitary treatments through spraying at a variable rate, even in preventive cases (i.e., before the development of macroscopic disease lesions), following a precision agriculture perspective. The use of this information to support plant protection decisions will make the treatments more efficient, either by reducing the cost of production or the environmental impact due to the lower amount of phytosanitary products used. The team's Weta ground-based robot (Figure 1) capable of transporting a sprayer may be an effective solution to be explored since it is able to perform precision spraying and apply UV-C treatments, avoiding the application of

fungicides (INESCTEC 2023, INESCTEC 2024) developed under the scope of SCORPION H2020 project (SCORPION 2022).


**Figure 1** WETA Agro Robot: a hardworking autonomous platform conceived for helping people in agriculture and forestry developed under the scope of the SCORPION H2020 project (SCORPION 2022, INESCTEC 2024).

Further research should also evaluate the development of multisensory / data fusion solutions for plant disease assessment, combined with enhancements in equipment's sensitivity and resolution. The goodness and reliability of the information extracted from the analysis of the data captured motivate the development and establishment of protocols for measurements, preprocessing, and processing of collected data, that must consider the variability of the environmental conditions that arise during measurements. Moreover, improvements in data analysis algorithms and models for specific spectra-disturbance assessment will need to be continually evaluated and upgraded or even redefined to improve disease investigation.

Data fusion should also explore the integration of the advanced optical technologies already described (that also can be known as smart photonics) and omics for advancing plant disease diagnosis, through the lens of systems biology. The different Omics sciences, such as genomics, transcriptomics, proteomics, and metabolomics, provide a comprehensive understanding of the molecular compounds and processes

within a biological system. In the context of plant disease diagnosis, omics technologies may be very relevant since they allow the identification of key biomarkers associated with different pathogens. The synergy between these two approaches may be particularly powerful since the recognition of specific molecular markers through omics can guide the design of targeted smart photonics sensors for precise diagnosis. The real-time, non-invasive nature of smart photonics complements the static snapshot provided by omics, allowing for dynamic monitoring of plant responses to pathogenic invasions. This would contribute to a holistic understanding of plant-pathogen interactions at various levels, from the molecular to the macroscopic.

Moreover, through the computation of several data analysis and model techniques, future studies can unravel intricate networks of molecular events associated with host-pathogen interactions and with the development of plant diseases, providing insights into the underlying mechanisms and potential targets for intervention, fulfilling the requisites of system biology.

This pathway is already being explored by the author's team in the Omicbots project (OMICBOTS 2024) a High-Throughput Integrative Omic-Robots Platform for Next Generation Physiology-based Precision Viticulture (Figure 2). It aims to integrate multi-omics, smart-photonics, and robotics into system biology. The final purpose is to increase crop productivity and nutritional quality, through precise inputs (e.g. water use efficiency), in terms of quantity and location, achieve crop resilience under different environmental conditions, and monitor biotic and abiotic stress (Figure 2).

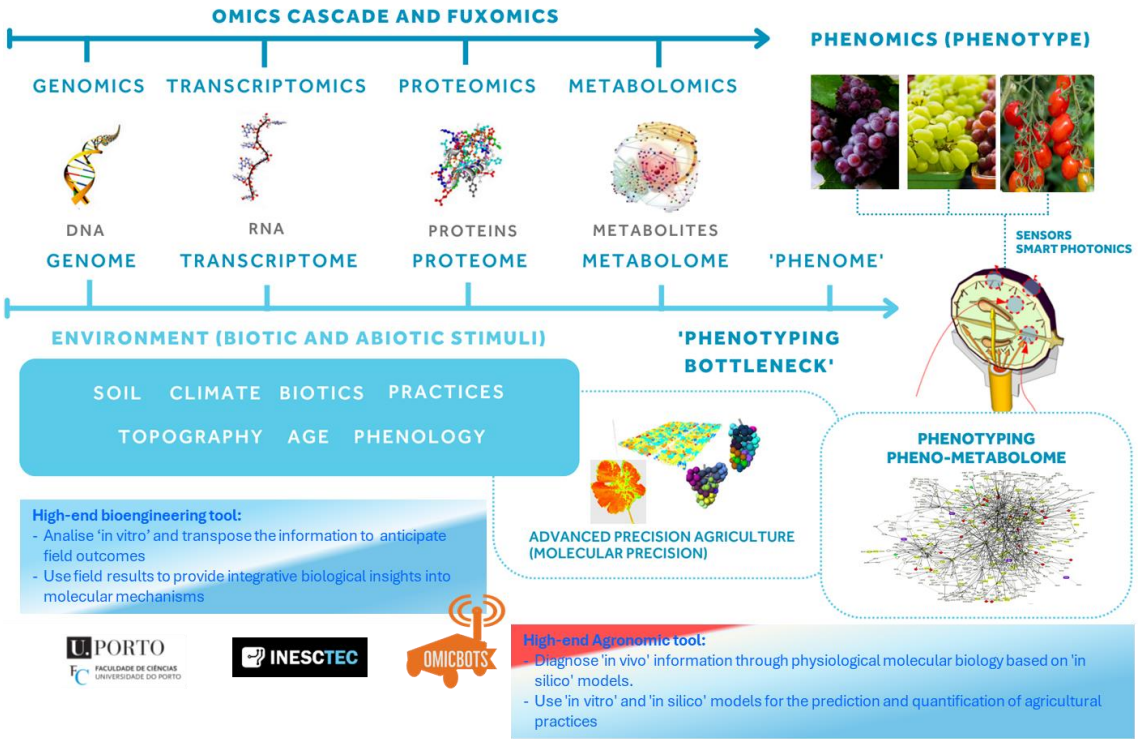



**Figure 2** Conceptual model of advanced precision agriculture (‘molecular precision’) combined omics, smart-photonics and system biology. Developed in the project Omicbots project: High-Throughput Integrative Omic-Robots Platform for a Next Generation Physiology-based Precision Viticulture (OMICBOTS 2024). Omics tools like systems biology and bioinformatics are currently available and allow the development of very thorough computer simulations of this omics cascade (fluxomics) and the respective production of in-silico models to connect the information between the genotype and the phenotype. These omic tools, combined with high-dimensional, high-throughput sensors, support the transfer of information to measure the plant's response at the cellular and metabolic level in the field, in a non-invasive way, thus enhancing the transition to a molecular precision agronomic model. Adapted from (Cunha et al. 2022)

# References

Abdel-Rahman, E. M., F. B. Ahmed and M. van den Berg (2010). "Estimation of sugarcane leaf nitrogen concentration using in situ spectroscopy." International Journal of Applied Earth Observation and Geoinformation **12**: S52-S57.

Abdulkhair, W. M. and M. A. Alghuthaymi (2016). "Plant pathogens." Plant Growth: 49.

Abdulridha, J., R. Ehsani and A. De Castro (2016). "Detection and differentiation between laurel wilt disease, phytophthora disease, and salinity damage using a hyperspectral sensing technique." Agriculture **6**(4): 56.

Afonso, A. M., R. Guerra, A. M. Cavaco, P. Pinto, A. Andrade, A. Duarte, D. M. Power and N. T. Marques (2017). "Identification of asymptomatic plants infected with Citrus tristeza virus from a time series of leaf spectral characteristics." Computers and Electronics in Agriculture **141**: 340-350.

Agarrwal, R., J. S. Bentur and S. Nair (2014). "Gas chromatography mass spectrometry based metabolic profiling reveals biomarkers involved in rice-gall midge interactions." Journal of Integrative Plant Biology **56**(9): 837-848.

Aghababaei, M., A. Ebrahimi, A. A. Naghipour, E. Asadi, A. Pérez-Suay, M. Morata, J. L. Garcia, J. P. Rivera Caicedo and J. Verrelst (2022). "Introducing ARTMO's Machine-Learning Classification Algorithms Toolbox: Application to plant-type detection in a semi-steppe Iranian landscape." Remote sensing **14**(18): 4452.

Agrios, G. (2009). "Plant pathogens and disease: general introduction.". 613-646.

Agrios, G. N. (2012). "Plant Pathology", Elsevier Science.

AGROTEC (2022). Smart Trap e Spectom: tecnologias digitais inovadoras na prevenção de doenças e pragas. AGROTEC. **43**.

Ahamed, T., L. Tian, Y. Zhang and K. Ting (2011). "A review of remote sensing methods for biomass feedstock production." Biomass and bioenergy **35**(7): 2455-2469.

Ahamed, T., L. Tian, Y. Zhang and K. C. Ting (2011). "A review of remote sensing methods for biomass feedstock production." Biomass & Bioenergy **35**(7): 2455-2469.

Ahmadi, P., F. M. Muharam, K. Ahmad, S. Mansor and I. Abu Seman (2017). "Early detection of Ganoderma Basal Stem Rot of oil palms using Artificial Neural Network spectral analysis." Plant Disease **101**(6): 1009-1016.

Albuquerque, P., C. M. R. Caridade, A. S. Rodrigues, A. R. S. Marcal, J. Cruz, L. Cruz, C. L. Santos, M. V. Mendes and F. Tavares (2012). "Evolutionary and experimental assessment of novel markers for detection of *Xanthomonas euvesicatoria* in plant samples." PLOS ONE **7**(5): e37836.

Ali, K., F. Maltese, A. Figueiredo, M. Rex, A. M. Fortes, E. Zyprian, M. S. Pais, R. Verpoorte and Y. H. Choi (2012). "Alterations in grapevine leaf metabolism upon inoculation with Plasmopara viticola in different time-points." Plant Science **191-192**: 100-107.

Ali, M. M., N. A. Bachik, N. A. Muhadi, T. N. T. Yusof and C. Gomes (2019). "Non-destructive techniques of detecting plant diseases: A review." Physiological and Molecular Plant Pathology **108**: 101426.

Almoujahed, M. B., A. K. Rangarajan, R. L. Whetton, D. Vincke, D. Eylenbosch, P. Vermeulen and A. M. Mouazen (2022). "Detection of fusarium head blight in wheat under field conditions using a hyperspectral camera and machine learning." Computers and Electronics in Agriculture **203**.

Alves, A., R. Ribeiro, M. Azenha, M. Cunha and J. Teixeira (2023). "Effects of exogenously applied copper in tomato plants' oxidative and nitrogen metabolisms under organic farming conditions." Horticulturae **9**(3): 323.

Anderegg, J., A. Hund, P. Karisto and A. Mikaberidze (2019). "In-field detection and quantification of *Septoria tritici* blotch in diverse wheat germplasm using spectral-temporal features." Frontiers in Plant Science **10**.

Apan, A., A. Held, S. Phinn and J. Markley (2003). Formulation and assessment of narrow-band vegetation indices from EO-1 hyperion imagery for discriminating sugarcane disease. 2003 Spatial Sciences Institute Conference: Spatial Knowledge Without Boundaries (SSC2003). Canberra, Australia**:** 1-13.

Archibald, R. and G. Fann (2007). "Feature selection and classification of hyperspectral images with support vector machines." IEEE Geoscience and remote sensing letters **4**(4): 674-677.

Arens, N., A. Backhaus, S. Doll, S. Fischer, U. Seiffert and H. P. Mock (2016). "Non-invasive presymptomatic detection of *Cercospora beticola* infection and identification of early metabolic responses in sugar beet." Frontiers in Plant Science **7**.

Ari, N. and M. Ustazhanov (2014). "Matplotlib in python." 2014 11th International Conference on Electronics, Computer and Computation (ICECCO), IEEE.

Ashburn, P. (1979). "The vegetative index number and crop identification." NASA. Johnson Space Center Proceedings of Technical Sessions, Vol. 1 and 2.

Ashourloo, D., M. R. Mobasheri and A. Huete (2014). "Developing two spectral disease indices for detection of wheat leaf rust (*Puccinia triticina*)." Remote Sensing **6**(6): 4723-4740.

Asner, G. P. (1998). "Biophysical and biochemical sources of variability in canopy reflectance." Remote Sensing of Environment **64**(3): 234-253.

Atanassova, S., P. Nikolov, N. Valchev, S. Masheva and D. Yorgov (2019). "Early detection of powdery mildew (*Podosphaera xanthii*) on cucumber leaves based on visible and near-infrared spectroscopy". AIP conference proceedings, AIP Publishing LLC. **2075:** 160014.

Atta, B. M., M. Saleem, M. Bilal, A. U. Rehman and M. Fayyaz (2023). "Early detection of stripe rust infection in wheat using light-induced fluorescence spectroscopy." Photochemical & Photobiological Sciences **22**(1): 115-134.

Bagheri, N., H. Mohamadi-Monavar, A. Azizi and A. Ghasemi (2018). "Detection of fire blight disease in pear trees by hyperspectral data." European Journal of Remote Sensing **51**(1): 1-10.

Bajwa, S. G., J. C. Rupe and J. Mason (2017). "Soybean disease monitoring with leaf reflectance." Remote Sensing **9**(2): 127.

Baldwin, I. (2001). "An ecologically motivated analysis of plant-herbivore interactions in native tobacco.". Plant physiology **127**(4): 1449-1458.

Balestra, G., A. Mazzaglia, A. Quattrucci, M. Renzi and A. Rossetti (2009). "Current status of bacterial canker spread on kiwifruit in Italy." Australasian Plant Disease Notes **4**(1): 34-36.

Ballabio, C. and S. Sterlacchini (2012). "Support vector machines for landslide susceptibility mapping: the Staffora River Basin case study, Italy." Mathematical geosciences **44**(1): 47-70.

Bannari, A., D. Morin, F. Bonn and A. Huete (1995). "A review of vegetation indices." Remote sensing reviews **13**(1-2): 95-120.

Barnes, E. M., T. R. Clarke, S. E. Richards, P. D. Colaizzi, J. Haberland, M. Kostrzewski, P. Waller, C. R. E. Choi, T. Thompson, R. J. Lascano, H. Li and M. S. Moran (2000). Coincident detection of crop water stress, nitrogen status and canopy density using

ground based multispectral data. Proceedings of the fifth international conference on precision agriculture, Bloomington, MN, USA (**1619**): 6.

Barnes, R. J., M. S. Dhanoa and S. J. Lister (1989). "Standard Normal Variate transformation and de-rrending of Near-Infrared diffuse reflectance spectra." Applied Spectroscopy **43**(5): 772-777.

Barroso, T. G., L. Ribeiro, H. Gregório, F. Monteiro-Silva, F. Neves dos Santos and R. C. Martins (2022). "Point-of-Care using Vis-NIR Spectroscopy for white blood cell count analysis." Chemosensors **10**(11): 460.

Barthel, D., N. Dordevic, S. Fischnaller, C. Kerschbamer, M. Messner, D. Eisenstecken, P. Robatscher and K. Janik (2021). "Detection of apple proliferation disease in Malus × domestica by near infrared reflectance analysis of leaves." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **263**: 120178.

Bazi, Y. and F. Melgani (2006). "Toward an optimal SVM classification system for hyperspectral remote sensing images." IEEE Transactions on geoscience and remote sensing **44**(11): 3374-3385.

Belasque, J. J., M. C. G. Gasparoto and L. G. Marcassa (2008). "Detection of mechanical and disease stresses in citrus plants by fluorescence spectroscopy." Applied Optics **47**(11): 1922-1926.

Bellow, S., G. Latouche, S. C. Brown, A. Poutaraud and Z. G. Cerovic (2012). "Optical detection of downy mildew in grapevine leaves: daily kinetics of autofluorescence upon infection." Journal of Experimental Botany **64**(1): 333-341.

Ben-Dor, E., D. Schläpfer, A. Plaza and T. Malthus (2013). "Hyperspectral Remote Sensing" Airborne measurements for environmental research: Methods and instruments**:** 413-456.

Bennett, M., M. Mehta and M. J. M. p.-m. i. Grant (2005). "Biophoton imaging: a nondestructive method for assaying R gene responses." Molecular plant-microbe interactions **18**(2): 95-102.

Bernstein, S. (2017). "The United Nations and the governance of sustainable development goals." Governing through goals: Sustainable Development Goals as governance innovation: 213-240.

Berrar, D. (2019). Cross-Validation. 542-545.

Betancourt, R., S. Chen, R. Betancourt and S. Chen (2019). "pandas Library." Python for SAS Users: A SAS-Oriented Introduction to Python: 65-109.

Bhandari, D. R., Q. Wang, W. Friedt, B. Spengler, S. Gottwald and A. Römpp (2015). "High resolution mass spectrometry imaging of plant tissues: towards a plant metabolite atlas." Analyst **140**(22): 7696-7709.

Bishop, C. M. and N. M. Nasrabadi (2006). "Pattern recognition and machine learning", Springer.

Blackburn, G. A. (2007). "Hyperspectral remote sensing of plant pigments." Journal of Experimental Botany **58**(4): 855-867.

Blackburn, G. A. and J. G. Ferwerda (2008). "Retrieval of chlorophyll concentration from leaf reflectance spectra using wavelet analysis." Remote Sensing of Environment **112**(4): 1614-1632.

Blackmer, T. M., J. S. Schepers and G. E. Varvel (1994). "Light reflectance compared with other nitrogen stress measurements in corn leaves." Agronomy Journal **86**(6): 934-938.

Blancard, D. (2012). "3 - Principal characteristics of pathogenic agents and methods of control." Tomato Diseases (Second Edition). D. Blancard. San Diego, Academic Press**:** 413-650.

Bock, C., G. Poole, P. E. Parker and T. Gottwald (2010). "Plant disease severity estimated visually, by digital photography and image analysis, and by Hyperspectral Imaging." Critical Reviews in Plant Sciences **29**.

Bonner, M. R. and M. C. Alavanja (2017). "Pesticides, human health, and food security", Wiley Online Library.

Borkar, S. G. and R. A. Yumlembam (2016). "Bacterial diseases of crop plants", CRC Press.

Bro, R. and A. K. Smilde (2014). "Principal component analysis." Analytical Methods **6**(9): 2812-2831.

Bruinsma, J. (2009). "The resource outlook to 2050: by how much do land, water and crop yields need to increase by 2050". Expert meeting on how to feed the world.

Buja, I., E. Sabella, A. G. Monteduro, M. S. Chiriaco, L. De Bellis, A. Luvisi and G. Maruccio (2021). "Advances in plant disease detection and monitoring: From traditional assays to in-field diagnostics." Sensors **21**(6).

Bürling, K., M. Hunsche and G. Noga (2012). "Presymptomatic detection of powdery mildew infection in winter wheat cultivars by Laser-Induced Fluorescence.". Applied Spectroscopy. **66**(12): 1411-1419.

Caicedo, J. P. R., J. Verrelst, J. Muñoz-Marí, J. Moreno and G. Camps-Valls (2014). "Toward a semiautomatic machine learning retrieval of biophysical parameters." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(4): 1249-1259.

Calamita, F., H. A. Imran, L. Vescovo, M. L. Mekhalfi and N. La Porta (2021). "Early identification of root rot disease by using Hyperspectral Reflectance: The case of pathosystem grapevine/Armillaria." Remote Sensing **13**(13): 2436.

Cardoza, Y. J., P. E. A. Teal and J. H. Tumlinson (2003). "Effect of peanut plant fungal infection on oviposition preference by *Spodoptera exigua* and on host-searching behavior by *Cotesia marginiventris*." Environmental Entomology **32**(5): 970-976.

Carlsen, L. and R. Bruggemann (2022). "The 17 United Nations' sustainable development goals: A status by 2020." International Journal of Sustainable Development & World Ecology **29**(3): 219-229.

Cellini, A., E. Biondi, S. Blasioli, L. Rocchi, B. Farneti, I. Braschi, S. Savioli, M. T. Rodriguez-Estrada, F. Biasioli and F. Spinelli (2016). "Early detection of bacterial diseases in apple plants by analysis of volatile organic compounds profiles and use of electronic nose." Annals of Applied Biology **168**(3): 409-420.

Cen, Y., Y. Huang, S. Hu, L. Zhang and J. Zhang (2022). "Early detection of bacterial wilt in tomato with portable Hyperspectral Spectrometer." Remote Sensing **14**(12): 2882.

Cerovic, Z. G., G. Samson, F. Morales, N. Tremblay and I. Moya (1999). "Ultraviolet-induced fluorescence for plant monitoring: present state and prospects." Agronomie 19.7 (1999): 543-578.

Chaerle, L., W. Van Caeneghem, E. Messens, H. Lambers, M. Van Montagu and D. Van Der Straeten (1999). "Presymptomatic visualization of plant-virus interactions by thermography." Nature Biotechnology **17**(8): 813-816.

Chakraborty, S. and A. C. Newton (2011). "Climate change, plant diseases and food security: an overview." Plant Pathology **60**(1): 2-14.

Chang, C.-C. and C.-J. Lin (2011). "LIBSVM: a library for support vector machines." ACM transactions on intelligent systems and technology (TIST) **2**(3): 1-27.

Chen, J. M. (1996). "Evaluation of vegetation indices and a modified simple ratio for boreal applications." Canadian Journal of Remote Sensing **22**(3): 229-242.

Cheshkova, A. (2022). "A review of hyperspectral image analysis techniques for plant disease detection and identif ication." Vavilov Journal of Genetics and Breeding **26**(2): 202.

Chew, W., M. Hashim, A. Lau, A. Battay and C. Kang (2014). "Early detection of plant disease using close range sensing system for input into digital earth environment". IOP Conference Series: Earth and Environmental Science, IOP Publishing.

Choi, Y. H., E. C. Tapias, H. K. Kim, A. W. M. Lefeber, C. Erkelens, J. T. J. Verhoeven, J. Brzin, J. Zel and R. Verpoorte (2004). "Metabolic discrimination of *Catharanthus roseus* leaves infected by phytoplasma using 1H-NMR spectroscopy and multivariate data analysis." Plant physiology **135**(4): 2398-2410.

CIBIO. (2024). "Microbial Diversity and Evolution - MDE." Retrieved 10.01.2024, from https://cibio.up.pt/en/groups/microbial-diversity-and-evolution-mde/.

Cifra, M., P. J. J. o. P. Pospíšil and P. B. Biology (2014). "Ultra-weak photon emission from biological samples: definition, mechanisms, properties, detection and applications.". Photobiology B: Biology **139**: 2-10.

Clarivate, A. (2022). "Web of science." Clarivate Analytics Retrieved 22.08.2022, from https://www.webofscience.com/wos/woscc/basic-search.

Clarke, T. R., M. S. Moran, E. M. Barnes, P. J. Pinter, Jr. and J. Qi (2001). "Planar domain indices: a method for measuring a quality of a single component in two-component pixels." Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International. **3:** 1279-1281.

Clevers, J., S. De Jong, G. Epema, F. Van Der Meer, W. Bakker, A. Skidmore and K. Scholte (2002). "Derivation of the red edge index using the MERIS standard band setting." International Journal of Remote Sensing **23**(16): 3169-3184.

Commission, E. (2020). "Farm to fork strategy: for a fair, healthy and environmentally-friendly food system." Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions **381**: 1-9.

Conrad, A. O., W. Li, D.-Y. Lee, G.-L. Wang, L. Rodriguez-Saona and P. Bonello (2020). "Machine Learning-based presymptomatic detection of rice sheath blight using spectral profiles." Plant Phenomics **2020**.

Cotrozzi, L. (2022). "Spectroscopic detection of forest diseases: a review (1970–2020)." Journal of Forestry Research **33**(1): 21-38.

Couture, J. J., S. P. Serbin and P. A. Townsend (2013). "Spectroscopic sensitivity of real-time, rapidly induced phytochemical change in response to damage." New Phytologist **198**(1): 311-319.

Couture, J. J., A. Singh, A. O. Charkowski, R. L. Groves, S. M. Gray, P. C. Bethke and P. A. Townsend (2018). "Integrating spectroscopy with potato disease management." Plant Disease **102**(11): 2233-2240.

Cozzolino, D. (2014). "Use of infrared spectroscopy for in-field measurement and phenotyping of plant properties: instrumentation, data analysis, and examples." Applied Spectroscopy Reviews **49**(7): 564-584.

Creath, K., G. And and Schwartz (2005). "What biophoton images of plants can tell us about biofields and healing." Journal of Scientific Exploration **19**.

Crippen, R. E. (1990). "Calculating the vegetation index faster." Remote Sensing of Environment **1**(34): 71-73.

Cunha, M. and R. Braga (2022). "Agricultura de precisão e sustentabilidade." INESC TEC Science&Society **1**(4).

Cunha, M., R. Martins and F. Neves dos Santos (2022). "Sustainable agriculture in the Era of field-omics. Let's improve AgronOmics." INESC TEC Science&Society **1**(4).

Curran, P. J. (1989). "Remote-Sensing of foliar chemistry." Remote Sensing of Environment **30**(3): 271-278.

Dalponte, M., L. Bruzzone and D. Gianelle (2012). "Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data." Remote Sensing of Environment **123**: 258-270.

Datt, B., T. R. McVicar, T. G. Van Niel, D. L. Jupp and J. S. Pearlman (2003). "Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes." IEEE Transactions on Geoscience and Remote Sensing **41**(6): 1246-1259.

Delalieux, S., J. van Aardt, W. Keulemans, E. Schrevens and P. Coppin (2007). "Detection of biotic stress (*Venturia inaequalis*) in apple trees using hyperspectral data:

Non-parametric statistical approaches and physiological implications." European Journal of Agronomy **27**(1): 130-143.

DEMETER. (2021). "DEMETER." Retrieved 09.01.2024, from https://h2020-demeter.eu/.

Deng, X., Z. Huang, Z. Zheng, Y. Lan and F. Dai (2019). "Field detection and classification of citrus Huanglongbing based on hyperspectral reflectance." Computers and Electronics in Agriculture **167**: 105006.

Deshmukh, R., P. Janse, S. Karungaru, N. Kalyankar and P. Koinkar (2018). "Hyperspectral Remote Sensing for Agriculture: A Review." International Journal of Computer Applications **172.7**: 30-34.

Díaz-Lago, J., D. Stuthman and K. J. J. P. D. Leonard (2003). "Evaluation of components of partial resistance to oat crown rust using digital image analysis." Plant Disease **87**(6): 667-674.

Ding, X., J. Liu, F. Yang and J. Cao (2021). "Random radial basis function kernel-based support vector machine." Journal of the Franklin Institute **358**(18): 10121-10140.

Donati, I., A. Cellini, G. Buriani, S. Mauri, C. Kay, G. Tacconi and F. Spinelli (2018). "Pathways of flower infection and pollen-mediated dispersion of Pseudomonas syringae pv. actinidiae, the causal agent of kiwifruit bacterial canker." Horticulture Research **5**.

Donati, I., A. Cellini, D. Sangiorgio, J. L. Vanneste, M. Scortichini, G. M. Balestra and F. Spinelli (2020). "*Pseudomonas syringae* pv. *actinidiae*: Ecology, Infection Dynamics and Disease Epidemiology." Microbial Ecology **80**(1): 81-102.

Dutta, B., R. Gitaitis, H. Sanders, C. Booth, S. Smith and D. B. Langston (2014). "Role of blossom colonization in pepper seed infestation by *Xanthomonas euvesicatoria*." Phytopathology **104**(3): 232-239.

Dyussembayev, K., P. Sambasivam, I. Bar, J. C. Brownlie, M. J. A. Shiddiky and R. Ford (2021). "Biosensor technologies for early detection and quantification of plant pathogens." Frontiers in Chemistry **9**.

Eigenvector Research, I. (2023). "PLS_Toolbox an Advanced Chemometrics Software for use with MATLAB®." Manson, WA, USA.

El-Shikha, D. M., E. M. Barnes, T. R. Clarke, D. J. Hunsaker, J. A. Haberland, P. J. Pinter, P. M. Waller and T. L. Thompson (2008). "Remote sensing of cotton nitrogen

status using the Canopy Chlorophyll Content Index (CCCI)." Transactions of the Asabe **51**(1): 73-82.

El Ouardighi, A., A. El Akadi and D. Aboutajdine (2007). "Feature selection on supervised classification using Wilk's Lambda statistic." Isciii '07: 3rd International Symposium on Computational Intelligence and Intelligent Informatics, Proceedings: 51-55.

Elsadek, M. and B. Liu (2021). "Effects of viewing flowering plants on employees' wellbeing in an office-like environment." Indoor and Built Environment **30**(9): 1429-1440.

Elsevier. (2023). "Engineering Village." Retrieved 09.01.2024, from https://www.engineeringvillage.com/search/quick.url?usageZone=evlogo&usageOrigin=header.

Elsevier. (2023). "ScienceDirect." Retrieved 09.01.2024, from https://www.sciencedirect.com/.

Elsevier, B. V. (2022). "SCOPUS." Retrieved 09.01.2024, from www.scopus.com/.

Escadafal, R., A. Belghith and H. Ben Moussa (1994). "Indices spectraux pour la dégradation des milieux naturels en Tunisie aride." 6eme Symposium international sur les Mesures Physiques et Signatures en Télédétection, ISPRS-CNES, France.

Fan, T. W.-M. and A. N. Lane (2008). "Structure-based profiling of metabolites and isotopomers by NMR." Progress in Nuclear Magnetic Resonance Spectroscopy **2**(52): 69-117.

Fan, X., P. Luo, Y. Mu, R. Zhou, T. Tjahjadi and Y. Ren (2022). "Leaf image based plant disease identification using transfer learning and feature fusion." Computers and Electronics in Agriculture **196**: 106892.

Fang, Y. and R. P. Ramasamy (2015). "Current and prospective methods for plant disease detection." Biosensors-Basel **5**(3): 537-561.

FAO (2018). "The future of food and agriculture: alternative pathways to 2050." Food and Agriculture Organization of the United Nations Rome.

FAO. (2020). "International year of plant health " Retrieved 18.01.2021, from http://www.fao.org/plant-health-2020/home/en/.

FAO. (2020). "Pesticides use." FAOSTAT Retrieved 12.21.2020, from http://www.fao.org/faostat/en/#data/RP.

FAO. (2021). "International Year of Fruits and Vegetables (IYFV)." Retrieved 14.01.2021, from http://www.fao.org/fruits-vegetables-2021/en/.

FCUP. (2021). "Projeto/Contrato PS:PTDC/ASP-HOR/1338/2021." Retrieved 11.01.2024, from 10.54499/PTDC/ASP-HOR/1338/2021.

Feng, W., W. Shen, L. He, J. Duan, B. Guo, Y. Li, C. Wang and T. Guo (2016). "Improved remote sensing detection of wheat powdery mildew using dual-green vegetation indices." Precision agriculture **17**: 608-627.

Ferentinos, K. P. (2018). "Deep learning models for plant disease detection and diagnosis." Computers and electronics in agriculture **145**: 311-318.

Fernandes, C., P. Albuquerque, N. Mariz-Ponte, L. Cruz and F. Tavares (2021). "Comprehensive diversity assessment of walnut-associated xanthomonads reveal the occurrence of distinct *Xanthomonas arboricola* lineages and of a new species (*Xanthomonas euroxanthea*) within the same tree." Plant Pathology **70**(4): 943-958.

Fernandes, C., P. Albuquerque, R. Sousa, L. Cruz and F. Tavares (2017). "Multiple DNA markers for identification of *Xanthomonas arboricola* pv. *juglandis* isolates and its direct detection in plant samples." Plant Diseases **101**(6): 858-865.

Fernández, C. I., B. Leblon, J. Wang, A. Haddadi and K. Wang (2021). "Detecting infected cucumber plants with close-range Multispectral Imagery." Remote Sensing **13**(15): 2948.

Ferreira, E. C., J. M. Anzano, D. M. B. P. Milori, E. J. Ferreira, R. J. Lasheras, B. Bonilla, B. Montull-Ibor, J. Casas and L. M. Neto (2009). "Multiple response optimization of Laser-Induced Breakdown Spectroscopy parameters for multi-element analysis of soil samples." Applied Spectroscopy **63**(9): 1081-1088.

Fetting, C. (2020). "The European green deal." ESDN report **53**.

Finlayson, C. (2018). "Perfect food: perspectives on consumer perceptions of fresh produce quality." Fennia-International Journal of Geography **196**(2): 168-186.

Flood, J. (2010). "The importance of plant health to food security." Food Security **2**(3): 215-231.

Foody, G. M. (2020). "Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification." Remote Sensing of Environment **239**: 111630.

ForceA (2019). Multiplex Research™ Fluorimètre Portable UV-Visible.

Fortes, F. J. and J. J. Laserna (2010). "The development of fieldable laser-induced breakdown spectrometer: No limits on the horizon." Spectrochimica Acta Part B: Atomic Spectroscopy **65**(12): 975-990.

Freitas, D., E. Carlos, M. Gil, L. Vieira and G. Alcantara (2015). "NMR-Based Metabolomic analysis of Huanglongbing-asymptomatic and-symptomatic citrus trees." Journal of agricultural and food chemistry **63**(34): 7582-7588.

Freitas, V. and W. Segatto. (2021). "Parsifal." Retrieved 08.08.2022, from https://parsif.al/.

Fried, G., B. Chauvel, P. Reynaud and I. Sache (2017). "Decreases in crop production by non-native weeds, pests, and pathogens." Impact of biological invasions on ecosystem services: 83-101.

Friedman, J., T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon and J. Qian (2021). "Package 'glmnet'." Journal of Statistical Software. 2010a **33**(1).

Furlanetto, R. H., M. R. Nanni, M. S. Mizuno, L. G. T. Crusiol and C. R. da Silva (2021). "Identification and classification of Asian soybean rust using leaf-based hyperspectral reflectance." International Journal of Remote Sensing **42**(11): 4177-4198.

Galieni, A., N. D'Ascenzo, F. Stagnari, G. Pagnani, Q. G. Xie and M. Pisante (2021). "Past and future of plant stress detection: An overview from Remote Sensing to Positron Emission Tomography." Frontiers in Plant Science **11**.

Ganapathi, T., P. Suprasanna, P. Rao and V. Bapat (2004). "Tobacco (*Nicotiana tabacum* L.)-A model system for tissue culture interventions and genetic engineering."

Gitelson, A. and M. N. Merzlyak (1994). "Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves." Journal of Photochemistry and Photobiology B: Biology **22**(3): 247-252.

Gitelson, A. A., Y. J. Kaufman and M. N. Merzlyak (1996). "Use of a green channel in remote sensing of global vegetation from EOS-MODIS." Remote Sensing of Environment **58**(3): 289-298.

Gitelson, A. A., G. P. Keydan and M. N. Merzlyak (2006). "Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves." Geophysical research letters **33**(11).

Gitelson, A. A., M. Merzlyak, Y. Zur, R. Stark and U. Gritz (2001). "Non-destructive and remote sensing techniques for estimation of vegetation status."

Gitelson, A., A. Viña, T. Arkebauer, D. Rundquist and G. Keydan (2003). "Remote estimation of leaf area index and green leaf biomass in maize canopies." Geophysical research letters **30**(5): 1248.

GmbH, U. (2014). Force A MULTIPLEX UV-Visible portable fluorometer.

Godfray, H. C., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas and C. Toulmin (2010). "Food security: the challenge of feeding 9 billion people." Science **327**(5967): 812-818.

Gold, K. M., P. A. Townsend, A. Chlus, I. Herrmann, J. J. Couture, E. R. Larson and A. J. Gevens (2020). "Hyperspectral Measurements Enable Pre-Symptomatic Detection and Differentiation of Contrasting Physiological Effects of Late Blight and Early Blight in Potato." Remote Sensing **12**(2).

Gold, K. M., P. A. Townsend, A. Chlus, I. Herrmann, J. J. Couture, E. R. Larson and A. J. Gevens (2020). "Hyperspectral measurements enable pre-symptomatic detection and differentiation of contrasting physiological effects of late blight and early blight in potato." Remote Sensing **12**(2): 286.

Gold, K. M., P. A. Townsend, I. Herrmann and A. J. Gevens (2020). "Investigating potato late blight physiological differences across potato cultivars with spectroscopy and machine learning." Plant Science **295**: 110316.

Golhani, K., S. K. Balasundram, G. Vadamalai and B. Pradhan (2018). "A review of neural networks in plant disease detection using hyperspectral data." Information Processing in Agriculture **5**(3): 354-371.

Grandini, M., E. Bagli and G. Visani (2020). "Metrics for multi-class classification: an overview." arXiv preprint arXiv:2008.05756.

Griffel, L. M., D. Delparte, J. Whitworth, P. Bodily and D. Hartley (2023). "Evaluation of artificial neural network performance for classification of potato plants infected with potato virus Y using spectral data on multiple varieties and genotypes." Smart Agricultural Technology **3**: 100101.

Grisham, M. P., R. M. Johnson and P. V. Zimba (2010). "Detecting Sugarcane yellow leaf virus infection in asymptomatic leaves with hyperspectral remote sensing and associated leaf pigment changes." Journal of virological methods **167**(2): 140-145.

Guezenoc, J., A. Gallet-Budynek and B. Bousquet (2019). "Critical review and advices on spectral-based normalization methods for LIBS quantitative analysis." Spectrochimica Acta Part B: Atomic Spectroscopy **160**: 105688.

Guyot, G. (1990). "Optical properties of vegetation canopies." Optical properties of vegetation canopies.: 19-43.

Haboudane, D., J. R. Miller, E. Pattey, P. J. Zarco-Tejada and I. B. Strachan (2004). "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture." Remote Sensing of Environment **90**(3): 337-352.

Hall, C. and M. Knuth (2019). "An update of the literature supporting the well-being benefits of plants: A review of the emotional and mental health benefits of plants." Journal of Environmental Horticulture **37**(1): 30-38.

Hamamatsu. (2023). "Mini-spectrometers." Retrieved October 2023, from https://www.hamamatsu.com/content/dam/hamamatsu-photonics/sites/documents/99_SALES_LIBRARY/ssd/mini-spectrometer_kacc9003e.pdf.

Haq, I. U. and S. Ijaz (2020). "Plant disease management strategies for sustainable agriculture through traditional and modern approaches", Springer International Publishing.

Hastie, M. T. (2020). Package 'mda'. CRAN R Project.

Hastie, T. and J. Qian (2014). "Glmnet vignette." Retrieved June **9**(2016): 1-30.

Hastie, T., R. Tibshirani and A. Buja (1994). "Flexible discriminant analysis by optimal scoring." Journal of the American statistical association **89**(428): 1255-1270.

Hastie, T., R. Tibshirani, J. H. Friedman and J. H. Friedman (2009). "The elements of statistical learning: data mining, inference, and prediction", Springer.

Heideman, M. T., D. H. Johnson and C. S. Burrus (1985). "Gauss and the history of the fast Fourier transform." Archive for history of exact sciences **34**(3): 265-277.

Hennessy, A., K. Clarke and M. Lewis (2020). "Hyperspectral classification of plants: A review of waveband selection generalisability." Remote Sensing **12**(1): 113.

Henrich, V., E. Götze, A. Jung, C. Sandow, D. Thürkow and C. Gläßer (2009). "Development of an online indices database: Motivation, concept and implementation." Proceedings of the 6th EARSeL imaging spectroscopy sig workshop innovative tool for scientific and commercial environment applications, Tel Aviv, Israel.

Henrich, V., G. Krauss, C. Götze and C. Sandow (2011). "The IndexDatabase." from https://www.indexdatabase.de/.

Henrich V., Krauss G., Götze C. and S. C. (2023). "Index data base a database for remote sensing indices." from https://www.indexdatabase.de/.

Herrmann, I., M. Berenstein, T. Paz-Kagan, A. Sade and A. Karnieli (2017). "Spectral assessment of two-spotted spider mite damage levels in the leaves of greenhouse-grown pepper and bean." Biosystems Engineering **157**: 72-85.

Herrmann, I., A. Karnieli, D. J. Bonfil, Y. Cohen and V. Alchanatis (2010). "SWIR-based spectral indices for assessing nitrogen content in potato fields." International Journal of Remote Sensing **31**(19): 5127-5143.

Herrmann, I., S. K. Vosberg, P. Ravindran, A. Singh, H. X. Chang, M. I. Chilvers, S. P. Conley and P. A. Townsend (2018). "Leaf and canopy level detection of *Fusarium Virguliforme* (Sudden Death Syndrome) in soybean." Remote Sensing **10**(3).

Hirshorn, S. and S. Jefferies (2016). "Final report of the NASA Technology Readiness Assessment (TRA) study team." No. HQ-E-DAA-TN43005.

Horst, R. K. (2013). "Westcott's plant disease handbook", Springer Science & Business Media.

Huang, L. S., W. J. Ding, W. J. Liu, J. L. Zhao, W. J. Huang, C. Xu, D. Y. Zhang and D. Liang (2019). "Identification of wheat powdery mildew using in-situ hyperspectral data and linear regression and support vector machines." Journal of Plant Pathology **101**(4): 1035-1045.

Hubert-Moy, L., A. Cotonnec, L. Le Du, A. Chardin and P. Pérez (2001). "A comparison of parametric classification procedures of remotely sensed data applied on different landscape units." Remote Sensing of Environment **75**(2): 174-187.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao and L. G. Ferreira (2002). "Overview of the radiometric and biophysical performance of the MODIS vegetation indices." Remote Sensing of Environment **83**(1-2): 195-213.

Huete, A. R., H. Q. Liu, K. Batchily and W. van Leeuwen (1997). "A comparison of vegetation indices over a global set of TM images for EOS-MODIS." Remote Sensing of Environment **59**(3): 440-451.

Hunt, E. R. and B. N. Rock (1989). "Detection of changes in leaf water content using Near- and Middle-Infrared reflectances." Remote Sensing of Environment **30**(1): 43-54.

Hunt Jr, E. R., C. Daughtry, J. U. Eitel and D. S. Long (2011). "Remote sensing leaf chlorophyll content using a visible band index." Agronomy journal **103**(4): 1090-1099.

Hunt Jr, E. R., C. S. T. Daughtry, J. U. H. Eitel and D. S. Long (2011). Remote Sensing Leaf Chlorophyll Content Using a Visible Band Index. Agronomy Journal. **103:** 1090-1099.

Hunt Jr, E. R. and B. N. Rock (1989). "Detection of changes in leaf water content using near-and middle-infrared reflectances." Remote sensing of environment **30**(1): 43-54.

IEEE. (2022). "IEEE Xplore." Retrieved 22.08.2022, from http://ieeexplore.ieee.org.L.

INESCTEC. (2018). "MetBots Metabolomic robots with self-learning artificial intelligence for precision agriculture." Retrieved 11.01.2024, from https://www.inesctec.pt/en/projects/metbots#intro.

INESCTEC (2021) "INESC TEC project among the winners of the BIP PROOF." Retrieved 18.01.2024 from https://www.inesctec.pt/en/news/inesc-tec-project-among-the-winners-of-the-bip-proof#about

INESCTEC. (2023). "INESC TEC robots already roam the Douro terraces." Retrieved 18.01.2024, from https://www.inesctec.pt/en/news/inesc-tec-robots-already-roam-the-douro-terraces#about.

INESCTEC. (2024). "AgIoT—IoT Solution for Agrifood Sector." Retrieved 09.01.2024, from https://agiot.inesctec.pt/

INESCTEC. (2024). "TRIBE - Laboratory of Robotics and IoT for Smart Precision Agriculture and Forestry." from https://www.inesctec.pt/en/laboratories/tribe-laboratory-of-robotics-and-iot-for-smart-precision-agriculture-and-forestry.

Infometrix, I. (2014). "Pirouette Multivariate Data Analysis Software." Bothell, WA.

Ishimwe, R., K. Abutaleb and F. Ahmed (2014). "Applications of Thermal Imaging in agriculture A review." Journal of Advances in Remote Sensing **Vol.03No.03**: 13.

ISPA, I. S. o. P. A. (2024). "Precision Ag definition." Retrieved 17.01.2024, from https://www.ispag.org/about/definition.

Iyozumi, H., K. Kato, C. Kageyama, H. Inagaki, A. Yamaguchi, K. Furuse, K. Baba, H. J. P. Tsuchiya (2005). "Plant defense activators potentiate the generation of elicitor-responsive photon emission in rice." Physiological and molecular plant pathology **66**(1-2): 68-74.

Iyozumi, H., K. Kato, T. J. P. Makino and photobiology (2002). "Spectral shift of Ultraweak Photon Emission from sweet potato during a defense response." Photochemistry and photobiology **75**(3): 322-325.

Jackulin, C. and S. Murugavalli (2022). "A comprehensive review on detection of plant disease using machine learning and deep learning approaches." Measurement: Sensors **24**: 100441.

Jacquemoud, S. and F. Baret (1990). "PROSPECT: A model of leaf optical properties spectra." Remote Sensing of Environment **34**(2): 75-91.

Jacquemoud, S. and F. J. R. s. o. e. Baret (1990). "PROSPECT: A model of leaf optical properties spectra." Remote sensing of environment **34**(2): 75-91.

Jain, N., S. S. Ray, J. P. Singh and S. Panigrahy (2007). "Use of hyperspectral data to assess the effects of different nitrogen applications on a potato crop." Precision Agriculture **8**(4-5): 225-239.

Jaumot, J., A. de Juan and R. Tauler (2015). "MCR-ALS GUI 2.0: New features and applications." Chemometrics and Intelligent Laboratory Systems **140**: 1-12.

Jensen, J. R. (2009). "Remote sensing of the environment: An earth resource perspective 2nde", Pearson Education India.

Jinendra, B., K. Tamaki, S. Kuroki, M. Vassileva, S. Yoshida and R. Tsenkova (2010). "Near infrared spectroscopy and aquaphotomics: Novel approach for rapid in vivo diagnosis of virus infected soybean." Biochemical and Biophysical Research Communications **397**(4): 685-690.

Jones, H. G. and R. A. Vaughan (2010). "Remote sensing of vegetation: principles, techniques, and applications", Oxford university press

Jones, J. B., J. P. Jones, R. E. Stall and T. A. Zitter (1991). "Compendium of tomato diseases", The American Phytopathological Society

Jouan-Rimbaud, D., E. Bouveresse, D. L. Massart and O. E. de Noord (1999). "Detection of prediction outliers and inliers in multivariate calibration." Analytica Chimica Acta **388**(3): 283-301.

Junges, A. H., M. A. K. Almança, T. V. M. Fajardo and J. R. Ducati (2020). "Leaf hyperspectral reflectance as a potential tool to detect diseases associated with vineyard decline." Tropical Plant Pathology **45**(5): 522-533.

Kang, X., C. Huang, L. Zhang, M. Yang, Z. Zhang and X. Lyu (2022). "Assessing the severity of cotton Verticillium wilt disease from in situ canopy images and spectra using convolutional neural networks." The Crop Journal.

Karatzoglou, A., A. Smola, K. Hornik and M. A. Karatzoglou (2019). "Package 'kernlab'." CRAN R Project.

Karthikeyan, L., I. Chawla and A. K. Mishra (2020). "A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses." Journal of Hydrology **586**: 124905.

Kawabata, R., T. Miike, H. Okabe, M. Uefune, J. Takabayashi, M. Takagi and S. J. J. j. o. a. p. Kai (2005). "Spectral analysis of ultraweak chemiluminescence from kidney bean leaf infested with *Tetranychus Kanzawai kishida*." Japanese journal of Applied Physics **44**(2R): 1115.

Kawabata, R., M. Uefune, T. Miike, H. Okabe, J. Takabayashi, M. Takagi and S. Kai (2004). "Biophoton Emission from kidney bean leaf infested with *Tetranychus Kanzawai Kishida*." Japanese Journal of Applied Physics **43**(8A): 5646-5651.

Khaled, A. Y., S. Abd Aziz, S. K. Bejo, N. M. Nawi, I. Abu Seman and D. I. Onwude (2018). "Early detection of diseases in plant tissue using spectroscopy - applications and limitations." Applied Spectroscopy Reviews **53**(1): 36-64.

Khan, I. H., H. Liu, W. Li, A. Cao, X. Wang, H. Liu, T. Cheng, Y. Tian, Y. Zhu, W. Cao and X. Yao (2021). "Early detection of powdery mildew disease and accurate quantification of its severity using Hyperspectral Images in wheat." Remote Sensing **13**(18): 3612.

Kim, G. H., K. H. Kim, K. I. Son, E. D. Choi, Y. S. Lee, J. S. Jung and Y. J. Koh (2016). "Outbreak and spread of bacterial canker of kiwifruit caused by *Pseudomonas syringae* pv. *actinidiae* biovar 3 in Korea." Plant Pathology Journal **32**(6): 545-551.

Kiraly, Z. O. L. T. A. N. (1980). "Defenses triggered by the invader: hypersensitivity.". Plant disease: An advanced treatise: How plants defend themselves: 201-224.

Klement, Z. (1982). "Hypersensitivity [Defense reaction of plants to pathogens]." Plant disease: An advanced treatise: How plants defend themselves: 201-224

Kobayashi, M., K. Sasaki, M. Enomoto and Y. Ehara (2007). "Highly sensitive determination of transient generation of biophotons during hypersensitive response to cucumber mosaic virus in cowpea." Journal of Experimental Botany **58**(3): 465-472.

Kobayashi, M. and P. B. Biology (2014). "Highly sensitive imaging for ultra-weak photon emission from living organisms." Journal of Photochemistry and Photobiology B: Biology **139**: 34-38.

Kooistra, L., R. S. E. W. Leuven, R. Wehrens, P. H. Nienhuis and L. M. C. Buydens (2003). "A comparison of methods to relate grass reflectance to soil metal contamination." International Journal of Remote Sensing **24**(24): 4995-5010.

Kopittke, P. M., N. W. Menzies, P. Wang, B. A. McKenna and E. Lombi (2019). "Soil and the intensification of agriculture for global food security." Environment International **132**: 105078.

Krstajic, D., L. J. Buturovic, D. E. Leahy and S. Thomas (2014). "Cross-validation pitfalls when selecting and assessing regression and classification models." Journal of Cheminformatics **6**(1): 10.

Kucheryavskiy, S. (2020). "mdatools–R package for chemometrics." Chemometrics and Intelligent Laboratory Systems **198**: 103937.

Kuhn, M. (2015). "Caret: classification and regression training." Astrophysics Source Code Library: ascl: 1505.1003.

Kuhn, M. and K. Johnson (2013). "Applied predictive modeling", Springer.

Kuhn, M. and K. Johnson (2013). "Data pre-processing". Applied Predictive Modeling. New York, NY, Springer New York**:** 27-59.

Kuhn, M., K. Johnson, M. M. Kuhn and I. CORElearn (2013). "Package 'AppliedPredictiveModeling'."

Laliberte, A. S., D. Browning and A. Rango (2012). "A comparison of three feature selection methods for object-based classification of sub-decimeter resolution UltraCam-L imagery." International Journal of Applied Earth Observation and Geoinformation **15**: 70-78.

Lamichhane, J. R. (2015). "Bacterial diseases of crops: Elucidation of the factors that lead to differences between field and experimental infections." Advances in Agronomy, Vol 134 **134**: 227-246.

Lamichhane, J. R. (2015). "Chapter Five - Bacterial diseases of crops: Elucidation of the factors that lead to differences between field and experimental infections". Advances in Agronomy. D. L. Sparks, Academic Press. **134:** 227-246.

Lang, M., F. Stober and H. K. Lichtenthaler (1991). "Fluorescence emission spectra of plant leaves and plant constituents." Radiation and Environmental Biophysics **30**(4): 333-347.

Lantz, B. (2019). "Machine learning with R: expert techniques for predictive modeling", Packt publishing ltd.

Lapajne, J., M. Knapič and U. Žibrat (2022). "Comparison of selected dimensionality reduction methods for detection of root-knot nematode infestations in potato tubers using Hyperspectral Imaging." Sensors (Basel) **22**(1).

Lay, L., H. S. Lee, R. Tayade, A. Ghimire, Y. S. Chung, Y. Yoon and Y. Kim (2023). "Evaluation of soybean wildfire prediction via Hyperspectral Imaging." Plants **12**(4): 901.

le Maire, G., C. Francois and E. Dufrene (2004). "Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements." Remote Sensing of Environment **89**(1): 1-28.

Le Maire, G., C. François and E. Dufrene (2004). "Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements." Remote sensing of environment **89**(1): 1-28.

Lee, L. C., C.-Y. Liong and A. A. Jemain (2018). "Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps." Analyst **143**(15): 3526-3539.

Lee, W. S., V. Alchanatis, C. Yang, M. Hirafuji, D. Moshou and C. Li (2010). "Sensing technologies for precision specialty crop production." Computers and Electronics in Agriculture **74**(1): 2-33.

Leucker, M., M. Wahabzada, K. Kersting, M. Peter, W. Beyer, U. Steiner, A.-K. Mahlein and E.-C. Oerke (2016). "Hyperspectral imaging reveals the effect of sugar beet quantitative trait loci on Cercospora leaf spot resistance." Functional Plant Biology **44**(1): 1-9.

Li, L., Q. Zhang and D. J. S. Huang (2014). "A review of imaging techniques for plant phenotyping." Sensors **14**(11): 20078-20111.

Liaghat, S., R. Ehsani, S. Mansor, H. Z. M. Shafri, S. Meon, S. Sankaran and S. H. M. N. Azam (2014). "Early detection of basal stem rot disease (Ganoderma) in oil palms based on hyperspectral reflectance data using pattern recognition algorithms." International Journal of Remote Sensing **35**(10): 3427-3439.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson and D. Bochtis (2018). "Machine Learning in agriculture: A review." Sensors **18**(8): 2674.

Lichtenthaler, H. K., A. Gitelson and M. Lang (1996). "Non-destructive determination of chlorophyll content of leaves of a green and an aurea mutant of tobacco by reflectance measurements." Journal of Plant Physiology **148**(3): 483-493.

Liu, H., B. Bruning, T. Garnett and B. Berger (2020). "Hyperspectral imaging and 3D technologies for plant phenotyping: From satellite to close-range sensing." Computers and Electronics in Agriculture **175**: 105621.

Liu, Q., Y. Gu, S. Wang, C. Wang and Z. Ma (2015). "Canopy spectral characterization of wheat stripe rust in latent period." Journal of Spectroscopy **2015**: 126090.

Liu, Z.-y., J.-f. Huang, J.-j. Shi, R.-x. Tao, W. Zhou and L.-l. Zhang (2007). "Characterizing and estimating rice brown spot disease severity using stepwise regression, principal component regression and partial least-square regression." Journal of Zhejiang University Science B **8**(10): 738-744.

Liu, Z. Y., J. A. Cheng, W. J. Huang, C. J. Li, X. G. Xu, X. D. Ding, J. J. Shi and B. Zhou (2012). "Hyperspectral discrimination and response characteristics of stressed rice leaves caused by rice leaf folder." Computer and Computing Technologies in Agriculture V, Pt Ii **369**: 528-+.

Lopez-Gresa, M. P., F. Maltese, J. M. Belles, V. Conejero, H. K. Kim, Y. H. Choi and R. Verpoorte (2010). "Metabolic Response of Tomato Leaves Upon Different Plant-Pathogen Interactions." Phytochemical Analysis **21**(1): 89-94.

López-Gresa, M. P., F. Maltese, J. M. Bellés, V. Conejero, H. K. Kim, Y. H. Choi and R. Verpoorte (2010). "Metabolic response of tomato leaves upon different plant-pathogen interactions." Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques **21**(1): 89-94.

Lowe, A., N. Harrison and A. P. French (2017). "Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress." Plant Methods **13**.

Lowe, A., N. Harrison and A. P. French (2017). "Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress." Plant Methods **13**(1): 80.

Lu, J., R. Ehsani, Y. Shi, J. Abdulridha, A. I. de Castro and Y. Xu (2017). "Field detection of anthracnose crown rot in strawberry using spectroscopy technology." Computers and Electronics in Agriculture **135**: 289-299.

Lu, J., R. Ehsani, Y. Shi, A. I. de Castro and S. Wang (2018). "Detection of multi-tomato leaf diseases (late blight, target and bacterial spots) in different stages by using a spectral-based sensor." Scientific Reports **8**(1): 2793.

Lu, J. Z., R. Ehsani, Y. Y. Shi, J. Abdulridha, A. I. de Castro and Y. J. Xu (2017). "Field detection of anthracnose crown rot in strawberry using spectroscopy technology." Computers and Electronics in Agriculture **135**: 289-299.

Luts, J., F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel and J. A. K. Suykens (2010). "A tutorial on support vector machine-based methods for classification problems in chemometrics." Analytica Chimica Acta **665**(2): 129-145.

Machinery, A. f. C. (2023). "ACM Digital Library." from https://dl.acm.org/.

Macleod, A., M. Pautasso, M. Jeger and R. Haines-Young (2010). "Evolution of the international regulation of plant pest and challenges for future plant health." Food Security **2**.

Magalhães, S. A., A. P. Moreira, F. N. d. Santos and J. Dias (2022). "Active perception fruit harvesting robots — A systematic review." Journal of Intelligent & Robotic Systems **105**(1): 14.

Mahlein, A.-K. (2011). "Detection, identification, and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques." Rheinischen Friedrich-Wilhelms-Universität Bonn: 4184.

Mahlein, A.-K. (2016). "Plant disease detection by imaging sensors – Parallels and specific demands for Precision Agriculture and plant phenotyping." Plant Disease **100**(2): 241-251.

Mahlein, A.-K., M. T. Kuska, J. Behmann, G. Polder and A. Walter (2018). "Hyperspectral sensors and imaging technologies in phytopathology: State of the Art." Annual Review of Phytopathology **56**(1): 535-558.

Mahlein, A.-K., E.-C. Oerke, U. Steiner and H.-W. Dehne (2012). "Recent advances in sensing plant diseases for precision crop protection." European Journal of Plant Pathology **133**(1): 197-209.

Mahlein, A.-K., T. Rumpf, P. Welke, H.-W. Dehne, L. Plümer, U. Steiner and E.-C. Oerke (2013). "Development of spectral indices for detecting and identifying plant diseases." Remote Sensing of Environment **128**: 21-30.

Mahlein, A., M. Kuska, S. Thomas, D. Bohnenkamp, E. Alisaac, J. Behmann, M. Wahabzada and K. Kersting (2017). "Plant disease detection by hyperspectral imaging: from the lab to the field." Advances in Animal Biosciences **8**(2): 238-243.

Mahlein, A. K. (2016). "Plant disease detection by Imaging Sensors - Parallels and specific demands for Precision Agriculture and plant phenotyping." Plant Disease **100**(2): 241-251.

Mahlein, A. K., U. Steiner, H. W. Dehne and E. C. Oerke (2010). "Spectral signatures of sugar beet leaves for the detection and differentiation of diseases." Precision Agriculture **11**(4): 413-431.

Main, R., M. A. Cho, R. Mathieu, M. M. O'Kennedy, A. Ramoelo and S. Koch (2011). "An investigation into robust spectral indices for leaf chlorophyll estimation." ISPRS Journal of Photogrammetry and Remote Sensing **66**(6): 751-761.

Main, R., M. A. Cho, R. Mathieu, M. Kennedy, A. Ramoelo and S. Koch (2011). "An investigation into robust spectral indices for leaf chlorophyll estimation." ISPRS Journal of Photogrammetry and Remote Sensing **66**(6): 751-761.

Maller, C., M. Townsend, A. Pryor, P. Brown and L. St Leger (2006). "Healthy nature healthy people: 'contact with nature' as an upstream health promotion intervention for populations." Health Promotion International **21**(1): 45-54.

Mandrile, L., S. Rotunno, L. Miozzi, A. M. Vaira, A. M. Giovannozzi, A. M. Rossi and E. Noris (2019). "Nondestructive Raman Spectroscopy as a tool for early detection and discrimination of the infection of tomato plants by two economically important viruses." Analytical Chemistry **91**(14): 9025-9031.

Mankins, J. C. (1995). "Technology readiness levels." White Paper, April **6**(1995): 1995.

Manolakis, D. G., R. B. Lockwood and T. W. Cooley (2016). "Hyperspectral Imaging Remote Sensing: Physics, sensors, and algorithms", Cambridge University Press.

Marín-Ortiz, J. C., N. Gutierrez-Toro, V. Botero-Fernández and L. M. Hoyos-Carvajal (2020). "Linking physiological parameters with visible/near-infrared leaf reflectance in the incubation period of vascular wilt disease." Saudi Journal of Biological Sciences **27**(1): 88-99.

Mariotto, I., P. S. Thenkabail, A. Huete, E. T. Slonecker and A. Platonov (2013). "Hyperspectral versus multispectral crop-productivity modeling and type discrimination for the HyspIRI mission." Remote Sensing of Environment **139**: 291-305.

Martinelli, F., R. Scalenghe, S. Davino, S. Panno, G. Scuderi, P. Ruisi, P. Villa, D. Stroppiana, M. Boschetti, L. R. Goulart, C. E. Davis and A. M. Dandekar (2015). "Advanced methods of plant disease detection. A review." Agronomy for Sustainable Development **35**(1): 1-25.

Martins, R., F. Santos, M. Cunha, F. Silva, R. Tosin, S. Magalhães and M. Reis Pereira (2023). "WO2023126532 - Method and device for non-invasive tomographic characterisation of a sample comprising a plurality of differentiated tissues." W. I. P. O. (WIPO)

Martins, R. C. (2019). "Unscrambling complex sample composition, variability and multi-scale interference in optical spectroscopy." Fourth International Conference on Applications of Optics and Photonics **11207**.

Martins, R. C., T. G. Barroso, P. Jorge, M. Cunha and F. Santos (2022). "Unscrambling spectral interference and matrix effects in *Vitis vinifera* Vis-NIR spectroscopy: Towards analytical grade 'in vivo' sugars and acids quantification." Computers and Electronics in Agriculture **194**: 106710.

Mauck, K. E., C. M. De Moraes and M. C. Mescher (2010). "Deceptive chemical signals induced by a plant virus attract insect vectors to inferior hosts." Proceedings of the National Academy of Sciences **107**(8): 3600-3605.

Mazivila, S. J. and W. Borges Neto (2021). "Detection of illegal additives in Brazilian S-10/common diesel B7/5 and quantification of Jatropha biodiesel blended with diesel according to EU 2015/1513 by MIR spectroscopy with DD-SIMCA and MCR-ALS under correlation constraint." Fuel **285**: 119159.

Mazivila, S. J. and J. L. M. Santos (2022). "A review on multivariate curve resolution applied to spectroscopic and chromatographic data acquired during the real-time monitoring of evolving multi-component processes: From process analytical chemistry (PAC) to process analytical technology (PAT)." TrAC Trends in Analytical Chemistry **157**: 116698.

Mellit, A., M. Benghanem, O. Herrak and A. Messalaoui (2021). "Design of a novel remote monitoring system for smart greenhouses using the internet of things and deep convolutional neural networks." Energies **14**(16): 5045.

Menesatti, P., F. Antonucci, F. Pallottino, S. Giorgi, A. Matere, F. Nocente, M. Pasquini, M. G. D'Egidio and C. Costa (2013). "Laboratory vs. in-field spectral proximal sensing

for early detection of Fusarium head blight infection in durum wheat." Biosystems Engineering **114**(3): 289-293.

Meng, R., Z. Lv, J. Yan, G. Chen, F. Zhao, L. Zeng and B. Xu (2020). "Development of spectral disease indices for southern corn rust detection and severity classification." Remote Sensing **12**(19): 3233.

Merzlyak, M., A. Gitelson, O. Chivkunova, A. Solovchenko and S. Pogosyan (2003). "Application of reflectance spectroscopy for analysis of higher plant pigments." Russian journal of plant physiology **50**: 704-710.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin and M. D. Meyer (2019). "Package 'e1071'." The R Journal.

Milborrow, M. S. (2019). "Package 'earth'." R Software package.

Mirik, M., G. Michels Jr, S. Kassymzhanova-Mirik, N. Elliott, V. Catana, D. Jones, R. J. C. Bowling and e. i. agriculture (2006). "Using digital image analysis and spectral reflectance data to quantify damage by greenbug (Hemitera: Aphididae) in winter wheat." **51**(1-2): 86-98.

Mishra, P., G. Polder and N. Vilfan (2020). "Close range spectral imaging for disease detection in plants using autonomous platforms: A review on recent studies." Current Robotics Reports **1**(2): 43-48.

Misra, P. N., S. G. Wheeler and R. E. Oliver (1977). "Kauth-Thomas brigthness and greenness axes". Contract NASA. **9-14350:** 23-46.

Mitra, D. (2021). "Emerging plant diseases: Research status and challenges." Emerging Trends in Plant Pathology. K. P. Singh, S. Jahagirdar and B. K. Sarma. Singapore, Springer Singapore**:** 1-17.

Moghadam, P., D. Ward, E. Goan, S. Jayawardena, P. Sikka and E. Hernandez (2017). "Plant disease detection using Hyperspectral Imaging." 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)

Mohammad-Razdari, A., D. Rousseau, A. Bakhshipour, S. Taylor, J. Poveda and H. Kiani (2022). "Recent advances in E-monitoring of plant diseases." Biosensors & Bioelectronics **201**.

Monteiro-Silva, F., P. A. S. Jorge and R. C. Martins (2019). "Optical sensing of nitrogen, phosphorus and potassium: A spectrophotometrical approach toward smart nutrient deployment." Chemosensors **7**(4): 51.

Mora-Romero, G. A., R. Félix-Gastélum, R. A. Bomberger, C. Romero-Urías and K. Tanaka (2022). "Common potato disease symptoms: ambiguity of symptom-based identification of causal pathogens and value of on-site molecular diagnostics." Journal of General Plant Pathology **88**(2): 89-104.

Morcillo-Pallarés, P., J. P. Rivera-Caicedo, S. Belda, C. De Grave, H. Burriel, J. Moreno and J. Verrelst (2019). "Quantifying the robustness of vegetation indices through global sensitivity analysis of homogeneous and forest leaf-canopy radiative transfer models." Remote Sensing **11**(20): 2418.

Morellos, A., G. Tziotzios, C. Orfanidou, X. E. Pantazi, C. Sarantaris, V. Maliogka, T. K. Alexandridis and D. Moshou (2020). "Non-destructive early detection and quantitative severity stage classification of Tomato Chlorosis Virus (ToCV) infection in young tomato plants using Vis–NIR Spectroscopy." Remote Sensing **12**(12): 1920.

Moriya, É., N. Imai, A. Tommaselli, E. Honkavaara and D. L. Rosalen (2023). "Design of Vegetation Index for identifying the mosaic virus in sugarcane plantation: A Brazilian case study." Agronomy **13**(6): 1542.

Mosavi, A., F. Sajedi Hosseini, B. Choubin, F. Taromideh, M. Ghodsi, B. Nazari and A. A. Dineva (2021). "Susceptibility mapping of groundwater salinity using machine learning models." Environmental Science and Pollution Research **28**(9): 10804-10817.

Nagler, P., C. Daughtry and S. Goward (2000). "Plant litter and soil reflectance." Remote Sensing of Environment **71**(2): 207-215.

Naidu, R. A., E. M. Perry, F. J. Pierce and T. Mekuria (2009). "The potential of spectral reflectance technique for the detection of Grapevine leafroll-associated virus-3 in two red-berried wine grape cultivars." Computers and Electronics in Agriculture **66**(1): 38-45.

Nations, F. A. O. U. (2020). "Fruit and vegetables – your dietary essentials: The International Year of Fruits and Vegetables, 2021, background paper", Food & Agriculture Organization

Nations, T. U. (2021). "Goal 2: Zero Hunger." Sustainable Development Goals Retrieved 14-01-2021, 2021, from http://www.un.org/sustainabledevelopment/hunger.

Nations, U. (2019). "World population prospects 2019." Population Division.

Nazarov, P. A., D. N. Baleev, M. I. Ivanova, L. M. Sokolova and M. V. Karakozova (2020). "Infectious plant diseases: Etiology, current status, problems and prospects in plant protection." Acta Naturae **12**(3): 46-59.

Negócios, J. d. (2022) "Inteligência artificial ao serviço da saúde das plantas." Journal de Negócios.

Nelson, E. B. (1994). "The disease triangle and the disease cycle." Interactions: comments and observations.

Nelson, R. (2020). "International plant pathology: Past and future contributions to global food security." Phytopathology® **110**(2): 245-253.

Nguyen, C., V. Sagan, M. Maimaitiyiming, M. Maimaitijiang, S. Bhadra and M. T. Kwasniewski (2021). "Early detection of plant viral disease using Hyperspectral Imaging and Deep Learning." Sensors **21**(3).

Nicolodelli, G., J. Cabral, C. R. Menegati, B. Marangoni and G. Senesi (2019). "Recent advances and future trends in LIBS applications to agricultural materials and their food derivatives: an overview of developments in the last decade (2010–2019). Part I. Soils and fertilizers." TrAC Trends in Analytical Chemistry **115**: 70-82.

Ninkovic, V., M. Rensing, I. Dahlin and D. Markovic (2019). "Who is my neighbor? Volatile cues in plant interactions." Plant Signaling & Behavior **14**(9): 1634993.

Nukui, H., H. Inagaki, H. Iyozumi and K. J. H. A. i. R. Kato, InTech Open Science, Rijeka, Croatia (2013). "Biophoton emissions in sulfonylureaherbicide-resistant weeds." InTech Open Science, Rijeka, Croatia, 219-235.

Oerke, E.-C., H.-W. Dehne, F. Schönbeck and A. Weber (2012). "Crop production and crop protection: estimated losses in major food and cash crops", Elsevier.

Oerke, E.-C., P. Fröhling and U. Steiner (2011). "Thermographic assessment of scab disease on apple leaves." Precision agriculture **12**(5): 699-715.

Oerke, E.-C., A.-K. Mahlein and U. Steiner (2014). "Proximal Sensing of plant diseases." Detection and Diagnostics of Plant Pathogens. M. L. Gullino and P. J. M. Bonants. Dordrecht, Springer Netherlands**:** 55-68.

Oerke, E. C. (2006). "Crop losses to pests." The Journal of Agricultural Science **144**(1): 31-43.

Oerke, E. C., U. Steiner, H. W. Dehne and M. Lindenthal (2006). "Thermal imaging of cucumber leaves affected by downy mildew and environmental conditions." Journal of Experimental Botany **57**(9): 2121-2132.

OMICBOTS. (2024). "Omicbots High-throughput integrative omic-robots platform for a next generation physiology-based Precision Viticulture." Retrieved 1.01.2024, from https://omicbots.fc.up.pt/.

Ouyang, F.-s., B.-l. Guo, L.-z. Ouyang, Z.-w. Liu, S.-j. Lin, W. Meng, X.-y. Huang, H.-x. Chen, H. Qiu-gen and S.-m. Yang (2019). "Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules." European Journal of Radiology **113**: 251-257.

Owomugisha, G., E. Nuwamanya, J. A. Quinn, M. Biehl and E. Mwebaze (2020). "Early detection of plant diseases using spectral data". Proceedings of the 3rd International Conference on Applications of Intelligent Systems

Page, M. J., D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting and J. E. McKenzie (2021). "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews." BMJ **372**: n160.

Pal, M. and G. M. Foody (2010). "Feature selection for classification of hyperspectral data by SVM." IEEE Transactions on Geoscience and Remote Sensing **48**(5): 2297-2307.

Parker, S. R., M. W. Shaw and D. J. Royle (1995). "The reliability of visual estimates of disease severity on cereal leaves." Plant Pathology **44**(5): 856-864.

Patle, A. and D. S. Chouhan (2013). "SVM kernel functions for classification." 2013 International Conference on Advances in Technology and Engineering (ICATE)

Pauli, G. F., B. U. Jaki and D. C. Lankin (2005). "Quantitative 1H NMR: development and potential of a method for natural products analysis." Journal of natural products **68**(1): 133-149.

Payne, W. Z. and D. Kurouski (2020). "Raman-Based Diagnostics of Biotic and Abiotic Stresses in Plants. A Review." Frontiers in Plant Science **11**.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.

Penuelas, J. and I. Filella (1998). "Visible and near-infrared reflectance techniques for diagnosing plant physiological status." Trends in plant science **3**(4): 151-156.

Pereira, F. M. V. and D. M. B. P. Milori (2010). "Investigation of the stages of citrus greening disease using micro synchrotron radiation X-ray fluorescence in association with chemometric tools." Journal of Analytical Atomic Spectrometry **25**(3): 351-355.

Pereira, F. M. V., D. M. B. P. Milori, A. L. Venâncio, M. d. S. T. Russo, P. K. Martins and J. Freitas-Astúa (2010). "Evaluation of the effects of *Candidatus Liberibacter asiaticus* on inoculated citrus plants using laser-induced breakdown spectroscopy (LIBS) and chemometrics tools." Talanta **83**(2): 351-356.

Perez-Sanz, F., P. J. Navarro and M. Egea-Cortines (2017). "Plant phenomics: an overview of image acquisition technologies and image data analysis algorithms." GigaScience **6**(11): 1-18.

Phiri, D. and J. Morgenroth (2017). "Developments in Landsat land cover classification methods: A review." Remote Sensing **9**(9): 967.

Pinter, P. J., J. L. Hatfield, J. S. Schepers, E. M. Barnes, M. S. Moran, C. S. T. Daughtry and D. R. Upchurch (2003). "Remote sensing for crop management." Photogrammetric Engineering and Remote Sensing **69**(6): 647-664.

Pinty, B. and M. Verstraete (1992). "GEMI: a non-linear index to monitor global vegetation from satellites." Vegetatio **101**: 15-20.

Pomerantsev, A. L. (2008). "Acceptance areas for multivariate classification derived by projection methods." Journal of Chemometrics **22**(11-12): 601-609.

Pomerantsev, A. L. and O. Y. Rodionova (2014). "Concept and role of extreme objects in PCA/SIMCA." Journal of Chemometrics **28**(5): 429-438.

Pomerantsev, A. L. and O. Y. Rodionova (2018). "Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial." Journal of Chemometrics **32**(8): e3030.

Pomerantsev, A. L. and O. Y. Rodionova (2020). "Popular decision rules in SIMCA: Critical review." Journal of Chemometrics **34**(8): e3250.

Pomerantsev, A. L. and O. Y. Rodionova (2021). "New trends in qualitative analysis: Performance, optimization, and validation of multi-class and soft models." TrAC Trends in Analytical Chemistry **143**: 116372.

Pomerantsev, A. L., D. N. Vtyurina and O. Y. Rodionova (2023). "Limit of detection in qualitative analysis: Classification Analytical Signal approach." Microchemical Journal **195**: 109490.

Pontes, J. G. M., W. Y. Ohashi, A. J. Brasil, P. R. Filgueiras, A. P. D. Espíndola, J. S. Silva, R. J. Poppi, H. D. Coletta-Filho and L. Tasic (2016). "Metabolomics by NMR spectroscopy in plant disease diagnostic: Huanglongbing as a case study." ChemistrySelect **1**(6): 1176-1178.

Pudil, P., J. Novovičová and J. Kittler (1994). "Floating search methods in feature selection." Pattern recognition letters **15**(11): 1119-1125.

Ramamoorthy, P., S. Samiappan, M. J. Wubben, J. P. Brooks, A. Shrestha, R. M. Panda, K. R. Reddy and R. Bheemanahalli (2022). "Hyperspectral reflectance and Machine Learning approaches for the detection of drought and root–knot nematode infestation in cotton." Remote Sensing **14**(16): 4021.

Randolph, T. W. (2006). "Scale-based normalization of spectral data." Cancer Biomarkers **2**(3-4): 135-144.

Rangarajan, A. K., R. L. Whetton and A. M. Mouazen (2022). "Detection of fusarium head blight in wheat using hyperspectral data and deep learning." Expert Systems with Applications **208**: 118240.

Ranulfi, A. C., G. S. Senesi, J. B. Caetano, M. C. Meyer, A. B. Magalhães, P. R. Villas-Boas and D. M. B. P. Milori (2018). "Nutritional characterization of healthy and Aphelenchoides besseyi infected soybean leaves by laser-induced breakdown spectroscopy (LIBS)." Microchemical Journal **141**: 118-126.

Rashidi, H. H., N. K. Tran, E. V. Betts, L. P. Howell and R. Green (2019). "Artificial Intelligence and Machine Learning in pathology: The present landscape of supervised methods." Academic Pathology **6**: 2374289519873088.

Rasmussen, C. E. and C. K. Williams (2006). "Gaussian processes for machine learning", Springer

Refaeilzadeh, P., L. Tang and H. Liu (2009). "Cross-Validation." Encyclopedia of Database Systems. L. Liu and M. T. ÖZsu. Boston, MA, Springer US**:** 532-538.

Reis-Pereira, M., R. C. Martins, A. F. Silva, F. Tavares, F. Santos and M. Cunha (2021). "Unravelling plant-pathogen interactions: Proximal optical sensing as an effective tool for early detect plant diseases." Chemistry Proceedings **5**(1): 18.

Reis-Pereira, M., R. Tosin, R. Martins, F. Neves dos Santos, F. Tavares and M. Cunha (2022). "Kiwi plant canker diagnosis using hyperspectral signal processing and Machine Learning: Detecting symptoms caused by *Pseudomonas syringae* pv. *actinidiae*." Plants **11**(16): 2154.

Reis-Pereira, M., R. Tosin, R. C. Martins, F. N. Dos Santos, F. Tavares and M. Cunha (2023). "Enhancing kiwi bacterial canker leaf assessment: Integrating hyperspectral-based Vegetation Indexes in predictive modeling." Engineering Proceedings **48**(1): 22.

Reis Pereira, M., F. N. d. Santos, F. Tavares and M. Cunha (2023). "Enhancing host-pathogen phenotyping dynamics: early detection of tomato bacterial diseases using hyperspectral point measurement and predictive modeling." Frontiers in Plant Science **14**.

Reis Pereira, M., F. Tavares, F. Santos and M. Cunha (2024). Hyperspectral spectroscopic reflectance data collected in-vivo non-symptomatic and symptomatic kiwi leaves in field conditions, Zenodo.

Reis Pereira, M., F. Tavares, F. Santos and M. Cunha (2024). Hyperspectral spectroscopic transmittance data collected in-vivo healthy and diseased tomato leaflets in controlled conditions - dataset II, Zenodo.

Research, E. and OpenAIRE. (2013). "Zenodo." Retrieved 14.01.2024, from https://www.zenodo.org/.

Richards, J. A. and J. Richards (1999). "Remote sensing digital image analysis", Springer.

Riefolo, C., I. Antelmi, A. Castrignanò, S. Ruggieri, C. Galeone, A. Belmonte, M. R. Muolo, N. A. Ranieri, R. Labarile, G. Gadaleta and F. Nigro (2021). "Assessment of the hyperspectral data analysis as a tool to diagnose *Xylella fastidiosa* in the asymptomatic leaves of olive plants." Plants **10**(4): 683.

Ripley, B., B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth and M. B. Ripley (2013). "Package 'mass'." Cran r **538**: 113-120.

Ristaino, J. B., P. K. Anderson, D. P. Bebber, K. A. Brauman, N. J. Cunniffe, N. V. Fedoroff, C. Finegold, K. A. Garrett, C. A. Gilligan, C. M. Jones, M. D. Martin, G. K. MacDonald, P. Neenan, A. Records, D. G. Schmale, L. Tateosian and Q. Wei (2021). "The persistent threat of emerging plant disease pandemics to global food security." Proceedings of the National Academy of Sciences **118**(23): e2022239118.

Ritchie, D. (2000). "Bacterial spot of pepper and tomato. The Plant Health Instructor" DOI: 10.1094/PHI-I-2000-1027-01. http://www.apsnet.org/education/LessonsPlantPath/BacterialSpot.

Rivera, J. P., J. Verrelst, J. Delegido, F. Veroustraete and J. Moreno (2014). "On the semi-automatic retrieval of biophysical parameters based on spectral index optimization." Remote Sensing **6**(6): 4927-4951.

Rodionova, O., S. Kucheryavskiy and A. Pomerantsev (2021). "Efficient tools for principal component analysis of complex data— a tutorial." Chemometrics and Intelligent Laboratory Systems **213**: 104304.

Rodionova, O. Y., P. Oliveri and A. L. Pomerantsev (2016). "Rigorous and compliant approaches to one-class classification." Chemometrics and Intelligent Laboratory Systems **159**: 89-96.

Rodionova, O. Y. and A. L. Pomerantsev (2020). "Detection of outliers in projection-based modeling." Analytical Chemistry **92**(3): 2656-2664.

Rodionova, O. Y., A. V. Titova and A. L. Pomerantsev (2016). "Discriminant analysis is an inappropriate method of authentication." TrAC Trends in Analytical Chemistry **78**: 17-22.

Rodrigues, E. S., M. H. F. Gomes, N. M. Duran, J. G. B. Cassanji, T. N. M. da Cruz, A. Sant'Anna Neto, S. M. Savassa, E. de Almeida and H. W. P. Carvalho (2018). "Laboratory Microprobe X-Ray Fluorescence in plant science: Emerging applications and case studies." Frontiers in Plant Science **9**.

Roever, C., N. Raabe, K. Luebke, U. Ligges, G. Szepannek, M. Zentgraf, M. U. Ligges and S. SVMlight (2020). "Package 'klaR'." ftp. rediris. org/mirror/CRAN/web/packages/klaR/klaR. pdf (Last viewed 10.06.2020).

Römer, C., K. Bürling, M. Hunsche, T. Rumpf, G. Noga and L. Plümer (2011). "Robust fitting of fluorescence spectra for pre-symptomatic wheat leaf rust detection with Support Vector Machines." Computers and Electronics in Agriculture **79**(2): 180-188.

Rudolph, K. (1993). Infection of the plant by Xanthomonas. Xanthomonas, Springer**:** 193-264.

Rumpf, T., A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne and L. Plümer (2010). "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance." Computers and electronics in agriculture **74**(1): 91-99.

Rumpf, T., A. K. Mahlein, U. Steiner, E. C. Oerke, H. W. Dehne and L. Plümer (2010). "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance." Computers and Electronics in Agriculture **74**(1): 91-99.

Rumpf, T., C. Römer, L. Plümer and A.-K. Mahlein (2010). "Optimal wavelengths for an early identification of *Cercospora beticola* with Support Vector Machines based on hyperspectral reflection data". 2010 IEEE International Geoscience and Remote Sensing Symposium, Ieee.

Saavedra, J., C. Abud, R. Cuevas and P. Gonzalez (2018). "Impact of plastic covers on the progression of *Pseudomonas syringae* pv. *actinidiae* and fruit productivity in a yellow-kiwifruit orchard." Ix International Symposium on Kiwifruit **1218**: 341-345.

Sachin, D. (2015). "Dimensionality reduction and classification through PCA and LDA." International journal of computer Applications **122**(17).

Saha, D. and A. Manickavasagan (2021). "Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review." Current Research in Food Science **4**: 28-44.

Sahoo, R., S. Ray and M. R (2015). "Hyperspectral remote sensing of agriculture." Current science **108**: 848-859.

Saleem, M. H., J. Potgieter and K. Mahmood Arif (2019). "Plant disease detection and classification by Deep Learning." Plants (Basel) **8**(11).

Sanchez, L., A. Ermolenkov, X.-T. Tang, C. Tamborindeguy and D. Kurouski (2020). "Non-invasive diagnostics of Liberibacter disease on tomatoes using a hand-held Raman spectrometer." Planta **251**(3): 64.

Sankaran, S., R. Ehsani, S. A. Inch and R. C. Ploetz (2012). "Evaluation of visible-near infrared reflectance spectra of avocado leaves as a non-destructive sensing tool for detection of laurel wilt." Plant disease **96**(11): 1683-1689.

Sankaran, S., A. Mishra, R. Ehsani and C. Davis (2010). "A review of advanced techniques for detecting plant diseases." Computers and Electronics in Agriculture **72**(1): 1-13.

Savary, S., S. Bregaglio, L. Willocquet, D. Gustafson, D. Mason D'Croz, A. Sparks, N. Castilla, A. Djurle, C. Allinne, M. Sharma, V. Rossi, L. Amorim, A. Bergamin, J. Yuen, P. Esker, N. McRoberts, J. Avelino, E. Duveiller, J. Koo and K. Garrett (2017). "Crop health and its global impacts on the components of food security." Food Security **9**(2): 311-327.

Savary, S., A. Ficke, J.-N. Aubertot and C. Hollier (2012). "Crop losses due to diseases and their implications for global food production losses and food security." Food Security **4**(4): 519-537.

Savian, F., M. Martini, P. Ermacora, S. Paulus and A.-K. Mahlein (2020). "Prediction of the kiwifruit decline syndrome in diseased orchards by Remote Sensing." Remote Sensing **12**(14): 2194.

Savitzky, A. and M. J. E. Golay (1964). "Smoothing and differentiation of data by simplified Least Squares procedures." Analytical Chemistry **36**(8): 1627-1639.

Schlaeppi, K., E. Abou-Mansour, A. Buchala and F. Mauch (2010). "Disease resistance of Arabidopsis to *Phytophthora brassicae* is established by the sequential action of indole glucosinolates and camalexin." Plant Journal **62**(5): 840-851.

Schlerf, M., C. Atzberger and J. Hill (2005). "Remote sensing of forest biophysical variables using HyMap imaging spectrometer data." Remote Sensing of Environment **95**: 177−8194.

Schlerf, M., C. Atzberger and J. Hill (2005). "Remote sensing of forest biophysical variables using HyMap imaging spectrometer data." Remote sensing of environment **95**(2): 177-194.

Schölkopf, B., A. Smola and K.-R. Müller (1998). "Nonlinear component analysis as a kernel eigenvalue problem." Neural computation **10**(5): 1299-1319.

Schumann, G. L. and C. J. D'Arcy (2006). "Essential plant pathology", American Phytopathological Society (APS Press).

SCORPION. (2022). "WETA Robot from SCORPION wins the IF Prize." Retrieved 18.11.2024, from https://scorpion-h2020.eu/weta-robot-from-scorpion-wins-the-if-prize/.

Scortichini, M., S. Marcelletti, P. Ferrante, M. Petriccione and G. Firrao (2012). "*Pseudomonas syringae* pv. *actinidiae*: a re-emerging, multi-faceted, pandemic pathogen", Wiley Online Library.

Shahid, M., A. Zaidi, M. S. Khan, A. Rizvi, S. Saif and B. Ahmed (2017). "Recent advances in management strategies of vegetable diseases". Microbial Strategies for Vegetable Production. A. Zaidi and M. S. Khan. Cham, Springer International Publishing**:** 197-226.

Sharma, N., Y. Khajuria, J. Sharma, D. K. Tripathi, D. K. Chauhan, V. K. Singh, V. Kumar and V. K. Singh (2018). "Microscopic, elemental and molecular spectroscopic investigations of root-knot nematode infested okra plant roots." Vacuum **158**: 126-135.

Sharma, N., Y. Khajuria, V. K. Singh, S. Kumar, Y. Lee, P. K. Rai and V. K. Singha (2020). "Study of molecular and elemental changes in Nematode-infested roots in

papaya plant using FTIR, LIBS and WDXRF Spectroscopy." Atomic Spectroscopy **41**(3): 110-118.

Sharma, P. D. (2006). "Plant Pathology", Alpha Science International.

Shuaibu, M., W. S. Lee, J. Schueller, P. Gader, Y. K. Hong and S. Kim (2018). "Unsupervised hyperspectral band selection for apple Marssonina blotch detection." Computers and Electronics in Agriculture **148**: 45-53.

Silva, R. (2022) "Projeto da U.Porto vence Prémio Empreendedorismo e Inovação da Caixa Agrícola." Notícias Universidade do Porto.

Sims, D. A. and J. A. Gamon (2002). "Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages." Remote Sensing of Environment **81**(2-3): 337-354.

Sinha, R., L. R. Khot, A. P. Rathnayake, Z. M. Gao and R. A. Naidu (2019). "Visible-near infrared spectroradiometry-based detection of grapevine leafroll-associated virus 3 in a red-fruited wine grape cultivar." Computers and Electronics in Agriculture **162**: 165-173.

Skoneczny, H., K. Kubiak, M. Spiralski and J. Kotlarz (2020). "Fire blight disease detection for apple trees: Hyperspectral analysis of healthy, infected and dry leaves." Remote Sensing **12**(13).

Sripada, R. P., R. W. Heiniger, J. G. White and A. D. Meijer (2006). "Aerial color infrared photography for determining early in-season nitrogen requirements in corn." Agronomy Journal **98**(4): 968-977.

Stall, R., C. Beaulieu, D. Egel, N. Hodge, R. Leite, G. Minsavage, H. Bouzar, J. Jones, A. Alvarez, A. Benedict and E. Microbiology (1994). "Two genetically diverse groups of strains are included in *Xanthomonas campestris* pv. *vesicatoria*." **44**(1): 47-53.

Steddom, K., M. McMullen, B. Schatz, C. Rush, Education and E. Team (2004). "Assessing foliar disease of wheat image analysis." Proceedings of the Summer Crops Field Day Sponsored by the Cooperative Research, Education & Extension Team (CREET'04), 32-38.

Strange, R. N. and P. R. Scott (2005). "Plant disease: A threat to global food security." Annual Review of Phytopathology **43**: 83-116.

Surico, G. (2013). "The concepts of plant pathogenicity, virulence/avirulence and effector proteins by a teacher of plant pathology." Phytopathologia Mediterranea **52**(3): 399-417.

Susič, N., U. Žibrat, S. Širca, P. Strajnar, J. Razinger, M. Knapič, A. Vončina, G. Urek and B. G. Stare (2018). "Discrimination between abiotic and biotic drought stress in tomatoes using hyperspectral imaging." Sensors and actuators B: Chemical **273**: 842-852.

Sylvain, T. and L.-G. Cecile (2018). "Disease identification: a review of vibrational spectroscopy applications." Comprehensive analytical chemistry **80**: 195-225.

Szymańska, E. (2018). "Modern data science for analytical chemical data – A comprehensive review." Analytica Chimica Acta **1028**: 1-10.

Team, R. C. (2021). "R: A language and environment for statistical computing."

Teper, D., A. M. Girija, E. Bosis, G. Popov, A. Savidor and G. J. P. p. Sessa (2018). "The *Xanthomonas euvesicatoria* type III effector XopAU is an active protein kinase that manipulates plant MAP kinase signaling." PLoS pathogens **14**(1): e1006880.

Tharwat, A., T. Gaber, A. Ibrahim and A. E. Hassanien (2017). "Linear discriminant analysis: A detailed tutorial." AI Communications **30**: 169-190.

Thenkabail, P. S., M. K. Gumma, P. Teluguntla and I. A. Mohammed (2014). "Hyperspectral Remote Sensing of vegetation and agricultural crops." Photogrammetric Engineering and Remote Sensing **80**(8): 697-709.

Thenkabail, P. S., J. G. Lyon and A. Huete (2018). "Hyperspectral Indices and image classifications for agriculture and vegetation", CRC press.

Thenkabail, P. S., R. B. Smith and E. De Pauw (2000). "Hyperspectral vegetation indices and their relationships with agricultural crop characteristics." Remote Sensing of Environment **71**(2): 158-182.

Thenkabail, P. S., R. B. Smith and E. De Pauw (2002). "Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization." Photogrammetric Engineering And Remote Sensing **68**(6): 607-621.

Thomas, S., M. T. Kuska, D. Bohnenkamp, A. Brugger, E. Alisaac, M. Wahabzada, J. Behmann and A.-K. Mahlein (2018). "Benefits of hyperspectral imaging for plant disease detection and plant protection: a technical perspective." Journal of Plant Diseases and Protection **125**(1): 5-20.

Tian, L., B. Xue, Z. Wang, D. Li, X. Yao, Q. Cao, Y. Zhu, W. Cao and T. Cheng (2021). "Spectroscopic detection of rice leaf blast infection from asymptomatic to mild stages

with integrated machine learning and feature selection." Remote Sensing of Environment **257**: 112350.

Tomaszewski, M., J. Nalepa, E. Moliszewska, B. Ruszczak and K. Smykała (2023). "Early detection of *Solanum lycopersicum* diseases from temporally-aggregated hyperspectral measurements using machine learning." Scientific Reports **13**(1): 7671.

Tosin, R., R. Martins, I. Pôças and M. Cunha (2022). "Canopy VIS-NIR spectroscopy and self-learning artificial intelligence for a generalised model of predawn leaf water potential in Vitis vinifera." Biosystems Engineering **219**: 235-258.

Tosin, R., F. Monteiro-Silva, R. Martins and M. Cunha (2023). "Precision maturation assessment of grape tissues: Hyperspectral bi-directional reconstruction using tomography-like based on multi-block hierarchical principal component analysis." Biosystems Engineering **236**: 147-159.

Tosin, R., I. Pocas, H. Novo, J. Teixeira, N. Fontes, A. Graca and M. Cunha (2021). "Assessing predawn leaf water potential based on hyperspectral data and pigment's concentration of Vitis vinifera L. in the Douro Wine Region." Scientia Horticulturae **278**.

Türker-Kaya, S. and C. W. Huck (2017). "A review of mid-infrared and near-infrared imaging: principles, concepts and applications in plant tissue analysis." Molecules **22**(1): 168.

Turner, A., S. Martin and J. Camberato (2004). "Image analysis to quantify foliage damage to turfgrass." fromm http://virtual. clemson. edu/groups/turfornamental/sctop/turfsec/plpanem/plpanem6.htm.**2**: 2005.

Ustin, S. L., R. J. Zomer, M. Garcia, D. A. Roberts and R. O. J. P. S. Green (1999). "Remote sensing methods monitor natural resources." Photonics Spectra **33**(10): 108-111.

Valier, A. (2020). "The cross validation in automated valuation models: A proposal for use." Computational Science and Its Applications – ICCSA 2020, Cham, Springer International Publishing.

Vallejo-Pérez, M. R., J. A. Sosa-Herrera, H. R. Navarro-Contreras, L. G. Álvarez-Preciado, Á. G. Rodríguez-Vázquez and J. P. Lara-Ávila (2021). "Raman Spectroscopy and Machine-Learning for early detection of bacterial canker of tomato: The asymptomatic disease condition." Plants **10**(8): 1542.

van der Werf, H. M. G. (1996). "Assessing the impact of pesticides on the environment." Agriculture, Ecosystems & Environment **60**(2): 81-96.

Van Grieken, R. and A. Markowicz (2001). "Handbook of X-ray Spectrometry", CRC press.

Vanneste, J. (2013). "Recent progress on detecting understanding and controlling Pseudomonas syringae pv actinidiae a short review." New Zealand Plant Protection **66**: 170-177.

Vapnik, V. (1999). "The nature of statistical learning theory", Springer science & business media.

Venbrux, M., S. Crauwels and H. Rediers (2023). "Current and emerging trends in techniques for plant pathogen detection." Frontiers in Plant Science **14**.

Verdebout, J., S. Jacquemoud and G. Schmuck (1994). "Optical properties of leaves: Modelling and experimental studies." Imaging Spectrometry — a Tool for Environmental Observations. J. Hill and J. Mégier. Dordrecht, Springer Netherlands**:** 169-191.

Verrelst, J., G. Camps-Valls, J. Muñoz-Marí, J. P. Rivera, F. Veroustraete, J. G. Clevers and J. Moreno (2015). "Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–A review." ISPRS Journal of Photogrammetry and Remote Sensing **108**: 273-290.

Verrelst, J., Z. Malenovský, C. Van der Tol, G. Camps-Valls, J.-P. Gastellu-Etchegorry, P. Lewis, P. North and J. Moreno (2019). "Quantifying vegetation biophysical variables from Imaging Spectroscopy data: A review on retrieval methods." Surveys in Geophysics **40**(3): 589-629.

Verrelst, J., J. Rivera, L. Alonso and J. Moreno (2011). "ARTMO: An Automated Radiative Transfer Models Operator toolbox for automated retrieval of biophysical parameters through model inversion". Proceedings of the EARSeL 7th SIG-Imaging Spectroscopy Workshop, Edinburgh, UK.

Verrelst, J., J. P. Rivera, A. Gitelson, J. Delegido, J. Moreno and G. Camps-Valls (2016). "Spectral band selection for vegetation properties retrieval using Gaussian processes regression." International Journal of Applied Earth Observation and Geoinformation **52**: 554-567.

Verrelst, J., J. P. Rivera, F. Veroustraete, J. Muñoz-Marí, J. G. Clevers, G. Camps-Valls and J. Moreno (2015). "Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods–A comparison." ISPRS Journal of Photogrammetry and Remote Sensing **108**: 260-272.

Verrelst, J., E. Romijn and L. Kooistra (2012). "Mapping vegetation density in a heterogeneous river floodplain ecosystem using pointable CHRIS/PROBA data." Remote Sensing **4**(9): 2866-2889.

Veys, C., F. Chatziavgerinos, A. AlSuwaidi, J. Hibbert, M. Hansen, G. Bernotas, M. Smith, H. Yin, S. Rolfe and B. Grieve (2019). "Multispectral imaging for presymptomatic analysis of light leaf spot in oilseed rape." Plant Methods **15**(1): 4.

Vieira, J., M. Mendes, P. Albuquerque, P. Moradas-Ferreira and F. Tavares (2007). "A novel approach for the identification of bacterial taxa-specific molecular markers." Letters in applied microbiology **44**(5): 506-512.

Villanueva, R. A. M. and Z. J. Chen (2019). "ggplot2: elegant graphics for data analysis", Taylor & Francis.

Vitale, R., M. Cocchi, A. Biancolillo, C. Ruckebusch and F. Marini (2023). "Class modelling by Soft Independent Modelling of Class Analogy: why, when, how? A tutorial." Analytica Chimica Acta **1270**: 341304.

Wainner, R. T., R. S. Harmon, A. W. Miziolek, K. L. McNesby and P. D. French (2001). "Analysis of environmental lead contamination: comparison of LIBS field and laboratory instruments." Spectrochimica Acta Part B: Atomic Spectroscopy **56**(6): 777-793.

Walsh, K. B., J. Blasco, M. Zude-Sasse and X. Sun (2020). "Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use." Postharvest Biology and Technology **168**: 111246.

Wang, D., R. Vinson, M. Holmes, G. Seibel, A. Bechar, S. Nof and Y. Tao (2019). "Early detection of tomato spotted wilt virus by hyperspectral imaging and outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN)." Scientific reports **9**(1): 1-14.

Wang, F.-M., J.-f. HUANG, Y.-l. TANG and X.-z. WANG (2007). "New vegetation index and its application in estimating leaf area index of rice." Rice Science **14**(3): 195-203.

Wang, X., X. Zhang and G. Zhou (2017). "Automatic detection of rice disease using near infrared spectra technologies." Journal of the Indian Society of Remote Sensing **45**(5): 785-794.

Weng, S., X. Hu, J. Wang, L. Tang, P. Li, S. Zheng, L. Zheng, L. Huang and Z. Xin (2021). "Advanced application of Raman Spectroscopy and surface-enhanced Raman

Spectroscopy in plant disease diagnostics: A review." Journal of Agricultural and Food Chemistry **69**(10): 2950-2964.

Wold, S. and M. Sjöström (1977). "SIMCA: A method for analyzing chemical data in terms of similarity and analogy." Chemometrics: Theory and Application, AMERICAN CHEMICAL SOCIETY. **52:** 243-282.

Wold, S., M. Sjöström and L. Eriksson (2001). "PLS-regression: a basic tool of chemometrics." Chemometrics and Intelligent Laboratory Systems **58**(2): 109-130.

WU, D., F. Cao, H. Zhang, L. FENG and Y. HE (2009). "Study on disease level classification of rice panicle blast based on visible and near infrared spectroscopy." Spectroscopy and Spectral Analysis **29**(12): 3295-3299.

Xia, C., T.-S. Chon, Z. Ren and J.-M. Lee (2015). "Automatic identification and counting of small size pests in greenhouse conditions with low computational cost." Ecological informatics **29**: 139-146.

Xulei, Y., S. Qing and A. Cao (2005). "Weighted support vector machine for data classification." Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.

Yao, Z., Y. Lei and D. He (2019). "Early visual detection of wheat stripe rust using visible/near-infrared hyperspectral imaging." Sensors **19**(4): 952.

Ye, J. C., S. Tak, K. E. Jang, J. Jung and J. Jang (2009). "NIRS-SPM: Statistical parametric mapping for near-infrared spectroscopy." NeuroImage **44**(2): 428-447.

Yeh, Y.-H. F., W.-C. Chung, J.-Y. Liao, C.-L. Chung, Y.-F. Kuo and T.-T. Lin (2013). "A comparison of machine learning methods on hyperspectral plant disease assessments." IFAC Proceedings Volumes **46**(4): 361-365.

Yu, K., J. Anderegg, A. Mikaberidze, P. Karisto, F. Mascher, B. A. McDonald, A. Walter and A. Hund (2018). "Hyperspectral canopy sensing of wheat *Septoria tritici* blotch disease." Frontiers in Plant Science **9**(1195).

Yu, K., G. Leufen, M. Hunsche, G. Noga, X. Chen and G. Bareth (2014). "Investigation of leaf diseases and estimation of chlorophyll concentration in seven barley varieties using Fluorescence and Hyperspectral Indices." Remote Sensing **6**(1): 64-86.

Zarco-Tejada, P. J., A. Berjón, R. López-Lozano, J. R. Miller, P. Martín, V. Cachorro, M. R. González and A. de Frutos (2005). "Assessing vineyard condition with hyperspectral

indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy." Remote Sensing of Environment **99**(3): 271-287.

Zarco-Tejada, P. J., J. R. Miller, T. L. Noland, G. H. Mohammed and P. H. Sampson (2001). "Scaling-up and model inversion methods with narrow-band optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data." IEEE Transactions on Geoscience and Remote Sensing **39**: 1491−1507.

Zarco-Tejada, P. J. and G. Sepulcre-Cantã (2007). "Remote Sensing of vegetation biophysical parameters for detecting stress condition and land cover changes." Estudios de la Zona No Saturada del Suelo **8**.

Zhan, J., P. H. Thrall, J. Papaïx, L. Xie and J. J. Burdon (2015). "Playing on a pathogen's weakness: using evolution to guide sustainable plant disease control strategies." Annual review of phytopathology **53**: 19-43.

Zhang, F. and X. Yang (2020). "Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection." Remote Sensing of Environment **251**: 112105.

Zhang, J., X. Jing, X. Y. Song, T. Zhang, W. A. Duan and J. Su (2023). "Hyperspectral estimation of wheat stripe rust using fractional order differential equations and Gaussian process methods." Computers and Electronics in Agriculture **206**.

Zhang, J. C., Y. B. Huang, R. L. Pu, P. Gonzalez-Moreno, L. Yuan, K. H. Wu and W. J. Huang (2019). "Monitoring plant diseases and pests through remote sensing technology: A review." Computers and Electronics in Agriculture **165**.

Zhang, J. C., N. Wang, L. Yuan, F. N. Chen and K. H. Wu (2017). "Discrimination of winter wheat disease and insect stresses using continuous wavelet features extracted from foliar spectral measurements." Biosystems Engineering **162**: 20-29.

Zhang, N., G. Yang, Y. Pan, X. Yang, L. Chen and C. Zhao (2020). "A review of advanced technologies and development for hyperspectral-based plant disease detection in the past three decades." Remote Sensing **12**(19): 3188.

Zhang, Y. G., M. Migliavacca, J. Penuelas and W. M. Ju (2021). "Advances in hyperspectral remote sensing of vegetation traits and functions." Remote Sensing of Environment **252**.

Zhang, Z., B. He, S. Sun, X. Zhang, T. Li, H. Wang, L. Xu, A. J. Afzal and X. Geng (2021). "The phytotoxin COR induces transcriptional reprogramming of photosynthetic, hormonal and defence networks in tomato." Plant Biology **23**: 69-79.

Zhao, D. L., K. R. Reddy, V. G. Kakani, J. J. Read and G. A. Carter (2003). "Corn (*Zea mays* L.) growth, leaf pigment concentration, photosynthesis and leaf hyperspectral reflectance properties as affected by nitrogen supply." Plant and Soil **257**(1): 205-217.

Zhao, J., Y. Fang, G. Chu, H. Yan, L. Hu and L. Huang (2020). "Identification of leaf-scale wheat powdery mildew (*Blumeria graminis* f. sp. *Tritici*) combining Hyperspectral Imaging and an SVM classifier." Plants (Basel) **9**(8).

Zhao, Y., Y. He and X. Xu (2012). "A novel algorithm for damage recognition on pest-infested oilseed rape leaves." Computers and electronics in agriculture **89**: 41-50.

Zhou, R., S. i. Kaneko, F. Tanaka, M. Kayamori and M. Shimizu (2014). "Disease detection of cercospora leaf spot in sugar beet by robust template matching." Computers and electronics in agriculture **108**: 58-70.

Zhu, F., Z. Su, A. Sanaeifar, A. Babu Perumal, M. Gouda, R. Zhou, X. Li and Y. He (2023). "Fingerprint spectral signatures revealing the spatiotemporal dynamics of Bipolaris Spot Blotch progression for presymptomatic diagnosis." Engineering **22**: 171-184.

Zontov, Y. V., O. Y. Rodionova, S. V. Kucheryavskiy and A. L. Pomerantsev (2017). "DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach." Chemometrics and Intelligent Laboratory Systems **167**: 23-28.

# Appendix |

# Supplementary materials

# Appendix A | Paper I

Reis-Pereira, M.; Martins, R.C.; Silva, A.F.; Tavares, F.; Santos, F.; Cunha, M. Unravelling Plant-Pathogen Interactions: Proximal Optical Sensing as an Effective Tool for Early Detect Plant Diseases. Chemistry Proceedings. 2021, 5, 18. https://doi.org/10.3390/CSAC2021-10560

Paper published on 1st July 2021

Classification according to journal: Proceeding Paper

# Unravelling Plant-pathogen Interactions: Proximal Optical Sensing as An Effective Tool for Early Detect plant Diseases †

Mafalda Reis Pereira[1,2,*], Rui C. Martins[3,*], Aníbal Filipe Silva[1,3], Fernando Tavares[1,4], Filipe Neves dos Santos[2], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Centre of Robotics in Industry and Intelligent Systems, INESC TEC, Dr. Roberto Frias, 4200-465, Porto, Portugal

[3] Centre for Applied Photonics, INESC TEC, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[4] Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), Rua Padre Armando Quintas, nº 7, 4485-661, Vairão, Portugal

*** Correspondence:** Mário Cunha mccunha@fc.up.pt

† Presented at the 1st International Electronic Conference on Chemical Sensors and Analytical Chemistry, 01—15 July 2021; Available online: https://csac2021.sciforum.net/.

**Keywords**

Plant disease detection, Plant Pathology, Proximal sensing, Spectroscopy, Precision agriculture, Principal Component Analysis

**Abstract**

This study analyzed the potential of proximal optical sensing as an effective approach for early disease detection. A compact, modular sensing system, combining direct UV-Vis spectroscopy with optical fibers, supported by a Principal Component Analysis (PCA) was applied to evaluate the modifications promoted by the bacteria *Xanthomonas euvesicatoria* in tomato leaves (cv. Cherry). Plant infection was achieved by spraying a bacterial suspension ($10^8$ CFU $mL^{-1}$) until run-off oc-curred, and a similar approach was followed for the control group where only water was applied. A total of 270 spectral measurements were performed on leaves, on five different time instances, in-cluding pre- and post-inoculation measurements. PCA was then applied to the acquired data from both healthy and inoculated leaves, which allowed their distinction and differentiation, three days after inoculation when unhealthy plants were still asymptomatic.

## 1. Introduction

Biotic agents, specifically pests and pathogens, cause significant losses in crop yields from levels that can range between 20% and 40% (Savary, Ficke et al. 2012). Chemical phytosanitary products are usually applied to prevent and combat these organisms. However, their usage can negatively impact the environment, mainly when applied to treat plant diseases that appear suddenly and spread to large scales (Zhang, Yang et al. 2020).

Nowadays, phytopathology methods are considered major challenges because to be implemented they often rely on the presence of indicator visible signs of the infection (disease symptoms), which frequently only manifest themselves at the middle to late stages of the process, compromising the effectiveness of phytosanitary measures (Lowe, Harrison et al. 2017). An example is the scouting technique, which involves inspecting a crop field to detect and identify infected plant through disease symptoms (Parker, Shaw et al. 1995). Despite being extremely useful, this approach requires specialized trained observers (who must be capable of identifying disease symptoms and distinguishing them from those caused by other abiotic stresses), can be labor-intensive, time-consuming, expensive (Sankaran, Mishra et al. 2010, Liaghat, Ehsani et al. 2014, Mahlein 2016, Khaled, Abd Aziz et al. 2018, Ali, Bachik et al. 2019). Moreover, this approach can be an inefficient in the early stages of the infection and on large areas. Other strategies consist of laboratory-based techniques, namely serological and molecular tests, largely used due to their sensitivity, accuracy, and effectiveness. They include enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction (PCR) methods. Their development boosted plant disease diagnosis since they allow the simultaneous processing of several samples and perform a precise pathogen identification. Furthermore, PCR enables the detection of pathogens that have not been cultured. Nevertheless, these procedures present some limitations, especially in the early phase of the infection process, due to the uneven spread of pathogens inside plants, compromising their effectiveness in analyzing asymptomatic samples (Sankaran, Mishra et al. 2010, Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015). Other drawbacks can also be enumerated since they require several hours to be completed, require the realization of detailed sampling procedures, and destructive sample preparation, not allowing a follow-up of the disease progression (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015).

Therefore, arises the necessity of developing fast, accurate, and selective in-vivo techniques for plant disease detection. These innovative approaches must provide

complementary information to the current methods applied in the phytopathology field and combine with them. Several non-invasive methods have been developed in the last decade and proved to be sensitive, consistent, standardize, rapid, cost-effective, and have high throughput (Golhani, Balasundram et al. 2018). Hyperspectral spectroscopy (HS) is one of them and seems to be effective in estimating a wide variety of plant chemical, biophysical, and metabolic traits in living tissue (Thenkabail, Smith et al. 2000, Zhao, Reddy et al. 2003, Delalieux, van Aardt et al. 2007, Jain, Ray et al. 2007, Blackburn and Ferwerda 2008, Abdel-Rahman, Ahmed et al. 2010, Couture, Serbin et al. 2013), namely foliar structure, plant chemical composition, water concentration, and metabolic status (Agrios 2009). Through spectral measurements in the visible (VIS, 400-700 nm), near-infrared (NIR, 700-1100 nm), and shortwave infrared wavelengths (SWIR, 1100–2500 nm), this approach assesses changes in optical properties of leaves, which derive from interactions between light, chemical bonds, and cellular structure (Curran 1989). Briefly, modifications in plants' reflectance in the VIS range are mostly related to pigment concentration and physiological processes such as photosynthesis. In turn, changes in the NIR are correlated with leaf structure and internal scattering processes. The SWIR region is affected by leaf structural and chemical composition (including lignins' and proteins) and water content (Hunt and Rock 1989, Guyot 1990, Jacquemoud and Baret 1990, Jones and Vaughan 2010, Haq and Ijaz 2020).

Since phytopathogens induce physiological, biochemical, and structural changes in host plants, HS seems to be promising in plant disease detection, identification, and quantification (Grisham, Johnson et al. 2010, Mahlein, Steiner et al. 2010, Menesatti, Antonucci et al. 2013, Arens, Backhaus et al. 2016, Couture, Singh et al. 2018, Gold, Townsend et al. 2020, Riefolo, Antelmi et al. 2021). Hyperspectral sensors can be used alone or mounted in different platforms allowing the performance of mapping, monitoring, scouting, and application tasks (Zhang, Yang et al. 2020). Their flexibility allows them to assess leaf, single-plant, canopy (proximal sensing), and even plot and regional scales (remote sensing) (Zhang, Yang et al. 2020). Some examples, sorted by measurement scale, include handheld sensors, rail systems, vehicle, and tractor-mounted systems, drones UAVs, as well as aircrafts and satellites (Thomas, Kuska et al. 2018).

Despite the possibilities provided by these optical devices for simple, rapid, non-destructive disease detection and identification, its application is still very limited due to the scarce of extensive agronomic and phytopathological studies aiming to explore their full potential. Their Technology Readiness Levels (TRLs) is close to TRL3 (analytical and experimental critical function and/or characteristic proof-of-concept) (Hirshorn and Jefferies 2016). Hence, this study aimed to evaluate the potential of UV-Vis spectroscopy

to detect diseased tomato leaves and discriminate between healthy and infected leaves, through a multi-temporal approach. Furthermore, it was also analyzed the capability of this technology in detecting changes in the reflectance spectrum of infected leaves before the first symptoms became visible.

## 2. Materials and methods

### 2.1. Experimental design

Tomato (*Solanum lycopersicum* L.) plants of the cultivar Cherry were grown in 200 mL pots containing a commercial potting substrate, in a walk-in plant growth chamber under controlled conditions (temperature of 25-27 ºC, humidity of approximately 60%, and photoperiod of 12 / 12 h). Plants were divided into two groups, being one of them inoculated with Xanthomonas euvesicatoria LMG 905 (Xeu) bacteria, and the other being treated with sterile distilled water only (Control group). Plants were inoculated in the laboratory, at the growth stage of 5-6 fully expanded leaves, by spraying until they became fully wet, and run-off occurred. The bacterial suspensions used for these inoculation assays consisted of 1 x $10^8$ cells / mL. They were prepared from a 48-h-old culture grown on YDC medium (yeast extract, 10.0g; dextrose, 20.0g; CaCO3, 20.0g; agar, 15.0g; distilled water up to 1.0 liter). The inoculated plants were then covered with transparent polythene bags for 48 h to increase the relative humidity that fosters bacterial entry into plant tissues through natural openings such as stomata (Lamichhane 2015). Plants were daily monitored for symptom development for 7 days.

At the same time, to verify if the bacteria cultures used in these inoculation tests were viable, 20 µL of Xeu solution were culture in different Petri dishes containing YDC media. After 48 h was possible to observe the bacteria growth in both nutrient media, proving that bacteria were viable at inoculation.

### 2.2. Spectral measurements

Hyperspectral data were collected in vivo from the adaxial side of healthy and infected tomato plant leaves by a compact benchtop system consisting of a D2 (deuterium) light source (Ocean Optics model DH-2000-BAL), a spectrometer (Ocean Optics model HR4000), a transmission optical fiber bundle (UV), and a stainless-steel slitted reflection probe for sample measurement. The spectrometer operated in the 195-1100 nm wavelength range with a high spectral response and good optical resolution of 0.025 nm (full width at half maximum - FWHM). The measurements were carried out using an experimental setup in the laboratory. An LED light source was placed beneath the leaf and provided homogeneous illumination to its entire surface. The light signal

from the sample analyzed was guided to the entrance lens of the spectrometer by the fiber-optic cable placed perpendicularly 1 cm above the measured surface. Specialized software was used for data acquisition and processing. Data acquisition was performed with 10 scans for an integration period of 60 ms, in three leaves per plant, on nine locations on each leaf.

### 2.3. Data pre-processing

A spectral pre-processing method was required to reduce the instrumental noise. In this regard, a pretreatment with a Fast Fourier Transform (FFT) was carried out on spectral data to smooth/denoise it. FFT is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. The DFT is obtained by decomposing a sequence of values into components of different frequencies (Heideman, Johnson et al. 1985). Spectral data pre-processing was performed with RStudio software.

### 2.3. Data processing – Analytical Techniques

Spectral data were subjected to a Principal Component Analysis (PCA), a multivariate data analysis technique used to reduce the dimensionality, while preserving its structure, by projecting it into a new coordinate system. It can preserve the total variance of the dataset and minimize the mean square approximate errors. PCA uses eigenvectors and eigenvalues to define the reduced subspace (representing the original coordinate system). It originates principal components (PC) which are linear combinations of interrelated variables. PC1 accounts for the maximum possible proportion of the variance information of the original dataset (explained by the eigenvalue), and subsequent principal components (PC2, PC3, …) account for the maximum proportion of the unexplained residual variance, and so forth (Lee, Alchanatis et al. 2010, Liu, Cheng et al. 2012).

Contigou hyperspectral wavebands present redundant information (Thenkabail, Smith et al. 2002). The application of a PCA allows the transformation of this type of high-dimensional data into a few wavebands that contain most of the information in the original bands. The importance of these hyperspectral bands in each PC is then established based on the magnitude of eigenvectors or factor loadings for crop biophysical and biochemical traits, being that the higher the eigenvector, the higher is the importance of the band. So, PCA allows the selection of the best wavebands to model biophysical and biochemical quantities and the elimination of redundant bands (by highlighting the main bands) (Zhang, Migliavacca et al. 2021).

## 3. Results

The spectral response properties of tomato leaves to the stress caused by *Xanthomonas euvesicatoria* LMG 905 is very important for discriminating bacterial infection levels in precise pest management using hyperspectral proximal sensing data. The averaged raw spectral curves of healthy and diseased tomato leaves were slightly different in some spectral ranges, namely through the visible region of the wavelength spectrum (~ 420-730 nm).

Figure 1 presents the principal components (PCs) Gabriel plot for the healthy (Con) and diseased (Xeu) leaves spectra, three days after inoculation (before the appearance of the first symptoms). The PCA algorithm has obtained two PCs accounting for 99.6% of the total variance. PC1 (94.3%) discriminates the effects on the variance of these two types of tomato leaves, which is more evident in PC2 (5.3%).

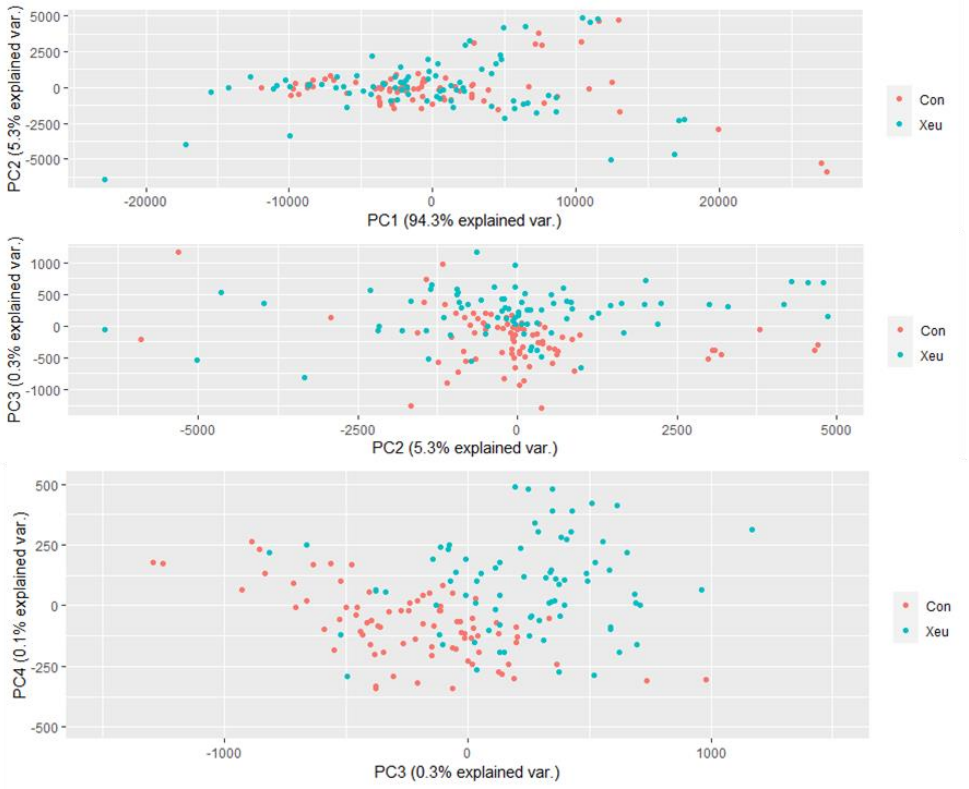



**Figure 1** Gabriel plot of PC1, PC2 and PC3 resulting from the PCA of the dataset three days after inoculation (all leaves were asymptomatic, showing no symptoms of the disease caused by *Xanthomonas euvesicatoria* LMG 905).

The wavelengths that have a higher contribution in these PCs are in the interval of ~454 – 654 nm (visible range of the wavelength spectrum). The ones between ~492 – 510 nm (essentially the blue region of the electromagnetic spectrum) explain 30% of the variance of the PC1, whereas ~454-461 nm (blue region) explain 40% of the variance

of the PC2 and 50% of the PC3. In all the first four dimensions of this analysis, the wavelengths ranging from approximately 445-480 nm (blue) and 580-700 nm (red) were the ones that explain most of the variance of the data.

This evidence can be related to the symptoms caused by Xeu, since these bacteria cause small, brown, angular lesions on leaves (which can be surrounded by a yellow hallow with the time), affecting the levels of photosynthetic pigments (contributing especially to the reduction of the chlorophyll levels, whose absorption features are more evident in the blue and red ranges of the VIS spectral region), cellular content and structural arrangement.

## 4. Discussion

The spectral behavior of tomato plants depends on their biochemical and structural profile. In Brief, plants' spectral signature in the visible spectral region (400–700 nm) depends mainly on the content of photosynthetic pigments. These compounds are good absorbers of red and blue wavelengths. Of the major pigments, Chlorophyll a (Chl a) has maximum absorption in the 410–430 and 600–690 nm regions, whereas Chlorophyll b (Chl b) has maximum absorption in the 450–470 nm range. The green part of the spectrum, on the other hand, is less strongly absorbed resulting in a reflectance peak in the green domain (at about 550 nm) (Jacquemoud and Baret 1990). In the NIR region, plants' spectral response is related to their structure, structural components, and internal scattering processes. Likewise, the SWIR region is also affected by leaf structural and chemical composition (including the action of lignin's and proteins) and water content (Hunt and Rock 1989, Guyot 1990, Jacquemoud and Baret 1990, Jones and Vaughan 2010, Haq and Ijaz 2020).

Since phytopathogens cause changes in plants' biochemical and structural composition, affecting the levels of photosynthetic pigments and structural elements, tracking changes in plants' spectral behavior can allow an indirect analysis of their phytosanitary status. Generally, unhealthy plants have more reflection in the blue and red regions and lower reflectance in the NIR. In fact, stress usually causes a rapid decrease of chlorophylls which exposes the absorption characteristics of other pigments, such as carotenoids (responsible for the yellowing of the leaves) and xanthophylls (responsible for the reddening of the leaves). With continuing stress, leaf structures decompose, resulting in extra intra-leaf scattering and an increased NIR signal. At the same time, concentrations of brown pigments, which absorb radiance in the VIS and at the onset of the NIR, can increase leading to a flattening of the red edge. Also, the absorption in the SWIR decreases due to reduced leaf moisture. With a decay of the leaf

tissue, the absorption features characteristic of healthy plants gradually disappear (Nagler, Daughtry et al. 2000).

Our findings seem to be in accord with the previous information showing evidence that UV-Vis spectroscopy can be suitable for plant disease assessment in laboratory conditions. Data collected in a randomized experimental design combined with a PCA allowed the discrimination of healthy and diseased tomato leaves, even at the third day after bacteria inoculation, when no visual symptoms were observable. Most of the variance of the data can be comprised with the first four PCs. In all of them, the wavelengths that explain most of the variance of the data ranged from approximately 445-480 nm (blue) and 580-700 nm (red), which was expected since *Xanthomonas euvesicatoria* causes tissue lesions, degrading the chlorophylls levels, and affecting their absorption features in these spectral regions.

Therefore, our results can be related to those obtained in different researches where sensor-based approaches proved to be capable of assessing modifications in plants' spectral behavior, allowing the detection, identification, and quantification of different types of plant diseases (Grisham, Johnson et al. 2010, Mahlein, Steiner et al. 2010, Menesatti, Antonucci et al. 2013, Arens, Backhaus et al. 2016, Couture, Singh et al. 2018, Gold, Townsend et al. 2020, Riefolo, Antelmi et al. 2021). They involve the capture and analysis of the optical properties of plants, within different regions of the electromagnetic spectrum, and their relationship with modifications in plant physiology, namely alterations in tissue color, structural composition, and transpiration rate (Blackburn and Ferwerda 2008). These non-invasive methods have been explored in the last decade, presenting the benefits of being sensitive, consistent, standard, high throughput, rapid, and cost-effective (Nagler, Daughtry et al. 2000), surpassing the limitations of the current methods used in plant disease detection.

## 5. Conclusions

The present study suggests that UV-Vis spectroscopy can be a potential tool for the early detection of plant diseases under laboratory conditions, even when unhealthy plants are asymptomatic. Despite these findings, its application is still very limited due to the scarcity of comprehensive agronomic and phytopathological studies aiming to explore their full potential, and to the development of applied advanced statistical approaches for data analysis. More research is necessary, especially in field conditions where more external factors have to surpass, including atmospheric, edaphic, and biotic conditions. Future research should also include more stress levels to discriminate not only healthy leaves from the diseased ones but also different levels of disease severity.

**Acknowledgments**

Mafalda Reis-Pereira and Aníbal Filipe Silva were supported by fellowships from Fundação para a Ciência e a Tecnologia (FCT) with the references SFRH/BD/146564/2019 and DFA/BD/9136/2020, respectively. Rui C. Martins acknowledges Fundação para a Ciência e Tecnologia (FCT) research contract grant (CEEIND/017801/2018). This research was supported by the project 'SpecTOM – Metabolomics Tomography Spectroscopy System', University of Porto, Fundação Amadeus Dias and Santander-Universities Grant.

# Appendix B | Paper II

Reis Pereira, M.; Tosin, R.; Martins, R.; Santos, F.N.d.; Tavares, F.; Cunha, M. Enhancing Kiwi Bacterial Canker Leaf Assessment: Integrating Hyperspectral-based Vegetation Indexes in Predictive Modeling. Engineering proceedings. 2023, 48, 22. https://doi.org/10.3390/CSAC2023-14920

Paper published on 5th October 2023

Classification according to journal: Proceeding Paper

# Enhancing Kiwi Bacterial Canker Leaf Assessment: Integrating Hyperspectral-based Vegetation Indexes in Predictive Modeling

Mafalda Reis Pereira[1,2], Renan Tosin[1,2], Rui C. Martins[2], Filipe Neves dos Santos[2], Fernando Tavares[3,4], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculda-de de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Cam-pus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

*** Correspondence:** Mário Cunha mccunha@fc.up.pt

**Keywords**

Kiwi, Bacterial canker, *Pseudomonas syringae*, Plant pathology, Optical sensing, In-field diagnosis, Vegetation index

**Abstract**

The potential of hyperspectral UV–VIS–NIR reflectance for in-field, non-destructive discrimination of bacterial canker on kiwi leaves caused by *Pseudomonas syringae* pv. *actinidiae* (Psa) was analyzed. Spectral data (325–1075 nm) of twenty kiwi plants were obtained in-vivo, in-situ, with a handheld spectroradiometer in two commercial kiwi orchards in northern Portugal, for 15 weeks, resulting in 504 spectral measurements. The suitability of different vegetation indexes (VIs) and applied predictive models (based on supervised machine learning algorithms) for classifying non-symptomatic and symptomatic kiwi leaves was evaluated. Eight distinct types of VIs were identified as relevant for disease diagnosis, highlighting the relevance of the Green, Red, Red-Edge, and NIR spectral features. The class prediction was achieved with good model metrics, achieving an accuracy of 0.71, kappa of 0.42, sensitivity of 0.67, specificity of 0.75, and F1 of 0.67. Thus, the present findings demonstrated the potential

of hyperspectral UV–VIS–NIR reflectance for non-destructive dis-crimination of bacterial canker on kiwi leaves.

## 1. Introduction

Bacterial Canker of Kiwi (BCK) disease, caused by *Pseudomonas syringae* pv. *actinidiae* (Psa), is accountable for numerous epidemics in Kiwi orchards annually (Balestra, Mazzaglia et al. 2009, Scortichini, Marcelletti et al. 2012). Scouting and laboratory-based techniques (e.g., Polymerase Chain Reaction – PCR –, and Enzyme-linked Immunosorbent assay – ELISA) are currently applied as diagnostic procedures. While insightful, these methods are hindered by labor intensiveness, time requirements, complex sampling, and unsuitability for rapid real-time field decisions, thus limiting their use in disease monitoring and field mapping (Fang and Ramasamy 2015, Martinelli, Scalenghe et al. 2015).

Early diagnosis, especially before symptoms' visible appearance, is of paramount importance in plant disease diagnosis. This proactive approach allows for timely and targeted intervention, reducing the spread of the disease and minimizing crop damage. It also enables more efficient resource allocation and cost-effective management strategies, safeguarding agricultural productivity and food security.

Hyperspectral Spectroscopy (HS) techniques have alternatively been recently applied as an innovative indirect plant disease diagnostic tool capable of retrieving relevant information about host-pathogen interactions related to the host's biochemical and biophysical modifications. Briefly, changes promoted by pathogens related to plants' pigment concentration and physiological processes (e.g., photosynthesis) produce changes in the quantitative and qualitative patterns of plants' spectral behavior, namely in the visible region of the electromagnetic spectrum (VIS, 400–700 nm). In turn, modifications in leaf water levels, chemical composition (i.e., lignin and protein content), structure, and internal scattering processes impact the spectral signatures in infrared wavelengths (IR, 800–2500 nm) (Hunt and Rock 1989, Jones and Vaughan 2010). Hence, HS could be successfully applied in the detection of pests (Herrmann, Berenstein et al. 2017, Zhang, Wang et al. 2017) and fungi (Yu, Anderegg et al. 2018, Skoneczny, Kubiak et al. 2020), bacteria (Bagheri, Mohamadi-Monavar et al. 2018), and viruses (Morellos, Tziotzios et al. 2020) affecting different crops, even at non-symptomatic stages (Gold, Townsend et al. 2020).

Nevertheless, data collected from HS frequently presents redundant information from proximal bands. Hence, only a few spectral wavelengths might help assess plant disease (Caicedo, Verrelst et al. 2014, Rivera, Verrelst et al. 2014). Approaches

including statistical signal processing, mathematical combinations of different bands, and applied predictive models may be computed to extract meaningful information, reduce data dimensionality, and/or select relevant features (Gold, Townsend et al. 2020, Meng, Lv et al. 2020). Vegetation Indices (VIs) exemplify these techniques, as they are numerical measures derived from the parametric formulations of different spectral bands or wavelengths associated with essential plant biophysical parameters like photosynthetic pigments, structural molecules, and water content. These indices are widely employed because of their simplicity and comprehensiveness, users' limited knowledge requirement, fast processing, and computationally inexpensiveness (Verrelst, Camps-Valls et al. 2015). VIs formalizations can be combinations of two-bands (most frequent case), three-bands, and four or more bands (combination of two VIs) (Verrelst, Camps-Valls et al. 2015). Among the most frequently computed VIs are the Normalized Difference Vegetation Index (NDVI) (Thenkabail, Smith et al. 2002, Zarco-Tejada and Sepulcre-CantÃ³ 2007), and the Enhanced Vegetation Index (EVI) (Huete, Didan et al. 2002, Hunt Jr, Daughtry et al. 2011), which are effective in assessing parameters related to the plant's status and structure. VIs developed specifically for parameter estimation (e.g., leaf's photosynthetic pigment and water levels) are frequently employed. Some examples include the Anthocyanin reflectance index (ARI) (Gitelson, Merzlyak et al. 2001), Browing Reflectance Index (BRI) (Merzlyak, Gitelson et al. 2003), Chlorophyll Green (Chlgreen) (Gitelson, Keydan et al. 2006), and Coloration Index (CI) (Escadafal, Belghith et al. 1994), among others. Furthermore, Vegetation Indices (VIs) can undergo band optimization procedures, enhancing their spectral sensitivity to the target parameters and enabling a more comprehensive analysis of the variable under consideration (Verrelst, Malenovský et al. 2019).

The present research aims to compare the suitability of VIs and classification modeling for discriminating non-symptomatic and BCK symptomatic kiwi leaves in-field, using ground-level UV–VIS hyperspectral reflectance assessments.

## 2. Methods

### 2.1. Experimental site

Two commercial orchards cultivated with kiwi plants (*Actinidia deliciosa*) were monitored in 2020, both located at Guimarães, Portugal: one situated in Caldas das Taipas (CT; 41°29′09.8″ N 8°21′54.3″ W), and the other in Briteiros (BT; 41°30′53.3″ N 8°19′20.5′ ′W). Twelve feminine kiwi plants of the variety Bo.Erika® in CT, and eight in BT were chosen, identified with tape, and classified according to the absence or presence of typical BCK visual symptoms (i.e., minor greasy dark lesions which turn

brown to black overtime, and are usually randomly spread on leaves surface). Visual phenotyping was performed on both the adaxial and abaxial sides of the leaves.

### 2.2. Ground-based hyperspectral reflectance acquisition

A portable spectroradiometer (ASD FieldSpec® HandHeld 2, ASD Instruments, Boulder, CO, USA) was used for leaf spectra capturing between May and August 2020 (9 visits), ending when the full development of Psa symptoms was reported in the plants' growing season. More details of the spectra measurement procedure can be found in (2022).

A total of 504 spectral averaged signatures were collected in both test sites, and the dataset was balanced regarding class distribution (Table 1).

**Table 1** Number of test sites, visits, plants, and leaves assessed per location of experimental sites (Reis-Pereira, Tosin et al. 2022).

| Experimental site | Sites | Visits | Plants | Non-symptomatic leaves | Symptomatic leaves | Total measurements |
|---|---|---|---|---|---|---|
| *Briteiros (BT)* | 1 | 9 | 8 | 89 | 127 | 216 |
| *Caldas das Taipas (CT)* | 1 | 8 | 12 | 192 | 96 | 288 |
| *Total* | 2 | 17 | 20 | 281 | 223 | 504 |

### 2.3. Data modeling

Spectral pre-processing was performed by computation of a multiplicative scatter correction (MSC) (Kucheryavskiy 2020). A total of 751 wavelength predictors were considered (325–1075 nm). Due to overlapping nature of hyperspectral data and multi-scale interference, auto-correlated signals may arise across various scales (Martins, Barroso et al. 2022). Thus, techniques capable of identifying the most relevant wavelengths or bands for discrimination and not considering redundant information are essential.

In this regard, reflectance data were processed into 32 spectral VIs, resulting in 41 distinct band combinations (Table A1). To calculate them, the wavelengths considered were: i) the ones enumerated in their original formula (as indicated in Table A1) or ii) default values chosen by the authors, namely 450 nm (representing the Blue region of the electromagnetic spectrum), 550 nm (Green), 680 (Red), Red Edge (700 nm), and 800 nm (NIR).

Applied predictive modeling was then performed using a model with a built-in Feature Selection (FS) method called Flexible Discriminant Analysis (FDA) (Figure 1). Leaf symptomatology was used as a binary variable in the models tested taking the

values ‘No’ (asymptomatic) and ‘Yes’ (symptomatic). The dataset was split into training (70% of random observations) and validation data (the remaining 30% of the observations), following a holdout method. A resampling approach was performed followed by a repeated cross-validation strategy using a repeated 10-fold cross-validation to estimate model evaluation criteria. The confusion matrix (CM), accuracy score, kappa coefficient, and F1-score were considered to determine model performance. A detailed description of these metrics applied, and about the R packages used can be found in (Reis-Pereira, Tosin et al. 2022).

## 3. Results

Model results showed the capacity of classifying the kiwi leaf measurements into ‘Non-symptomatic’ and ‘Symptomatic’ with 0.71 accuracy (proportion of correctly classified instances), 0.42 of Cohen’s kappa (agreement between predicted and actual classes beyond random occurrence), 0.67 of sensitivity (ability to identify diseased measurements), 0.75 of specificity (ability to identify healthy assessments), and 0.67 of F1 score (harmonized measure of precision and recall) for the test set (Table 2). Confusion matrix (CM) results (Table 3) demonstrate that 63 samples were correctly classified as non-symptomatic (True Negatives), and 44 as symptomatic (True Positives). Nevertheless, 21 measurements were wrongly classified as symptomatic (False Positives), and 22 as non-symptomatic (False Negatives). Thus, the model performs better at predicting non-symptomatic assessments than symptomatic measurements. These findings indicate a reasonably effective model performance, with an overall ability to distinguish between classes and make accurate predictions.

The built-in Feature Selection tool highlighted eight distinct VIs for sample discrimination, namely the Chlorophyll Green (Chlgreen), modified Simple Ratio (mSR), Coloration Index (CI), Simple Ratio Greenness Index (GI), Browning Reflectance Index (BRI), Ashburn Vegetation Index (AVI), Hyperspectral perpendicular VI (PVIhyp), and Reflectance at the inflexion point (Rre). These VIs are mostly based in the NIR, Red and Green regions of the electromagnetic spectrum (Table 2).

**Table 2** Classification results of the Flexible Discriminant Analysis (FDA) model computed for the train and test datasets. Legend: Acc. – Accuracy, Kap. – Kappa coefficient, Sen. – Sensitivity, Spe. – Specificity, Pre. – Precision, Rec. Recall, F1 – F1 score.

| Modeling Approach | | Acc. | Kap. | Sen. | Spe. | Pre. | Rec. | F1 |
|---|---|---|---|---|---|---|---|---|
| FDA | Train | 0.76 | 0.48 | 0.68 | 0.80 | 0.73 | 0.68 | 0.70 |
| | Test | 0.71 | 0.42 | 0.67 | 0.75 | 0.68 | 0.67 | 0.67 |

**Table 3** Vegetation Index (VI) importance for class discrimination and Confusion Matrix (CM) results according to Flexible Discriminant Analysis. Legend: Pred – Predicted, 'No' – Non-symptomatic, 'Yes' – Symptomatic.

| VI | Wavelength (nm) | Importance (a.u.) |
|---|---|---|
| Chlgreen | 553, 800 | 100 |
| mSR | 705, 750 | 67.15 |
| CI | 450, 700 | 52.94 |
| GI | 554, 677 | 44.45 |
| BRI | 450, 690 | 40.55 |
| AVI | 400, 994 | 33.71 |
| PVIhyp | 800, 1000 | 24.46 |
| Chlgreen | 530, 730 | 19.65 |
| Rre | 670, 780 | 16.46 |

**CM Train**

| Pred | 'No' | 'Yes' |
|---|---|---|
| 'No' | 157 | 50 |
| 'Yes' | 40 | 107 |

**CM Test**

| Pred | 'No' | 'Yes' |
|---|---|---|
| 'No' | 63 | 22 |
| 'Yes' | 21 | 44 |

Chlgreen– Chlorophyll Green, mSR – Modified Simple Ratio, CI – Coloration Index, GI – Simple Ratio Greenness Index, BRI – Browning Reflectance Index, AVI – Ashburn Vegetation Index, PVIhyp – Hyperspectral perpendicular VI, Rre – Reflectance at the inflexion point

## 4. Discussion

Eight distinct VIs (nine wavelength combinations) were identified as highly relevant for disease discrimination. They mostly consider the NIR, Green, and Red spectral regions. These findings present biological significance since they are coherent with the impact of *Pseudomonas syringae* pv. *actinidiae* (Psa) in kiwi leaves. Briefly, these pathogens cause modifications in pigment concentration and physiological processes (e.g., photosynthesis), resulting in changes in plants' spectral behavior in the VIS wavelengths (Blue, Green, Red). Furthermore, they cause changes in the leaf water levels, chemical composition (namely lignin and protein content), structure, and internal scattering processes which impact the NIR features (Hunt Jr and Rock 1989, Jones and Vaughan 2010). Similar spectral regions were also identified as relevant for late blight, target and bacterial spots detection in tomato leaves (Lu, Ehsani et al. 2018), and for the assessment of Cercospora leaf spot, sugar beet rust and powdery mildew in sugar beet plants (Mahlein, Rumpf et al. 2013). Model evaluation metrics also supported the model ability in discriminating non-symptomatic from symptomatic samples. Model performance may be enhanced by further fine-tuning, particularly in addressing models' sensitivity and minimizing the occurrence of false negatives.

Hyperspectral data may have redundant information in adjacent bands, and only a few wavelength features might be interesting in classifying a diseased plant (Rivera, Verrelst et al. 2014). For that reason, in crop remote sensing (both, ground, aerial and satellite-based solutions) spectral VIs are still the most common approaches studied to identify and manage biotic stresses in different crops (Verrelst, Rivera et al. 2015). Despite its substantial inherent potential, the discernment of the responsiveness of this extensive array of VIs to the target variable remains occasionally ambiguous. Furthermore, concerns related to the susceptibility to disturbances from confounding

elements can arise, mostly encompassing fluctuations in leaf or canopy properties, background soil reflectance, solar illumination, and atmospheric composition. Such a confluence of factors can generate instabilities in the spectral attributes of surfaces (Morcillo-Pallarés, Rivera-Caicedo et al. 2019). Furthermore, VIs were developed when the first applications of broadband sensor occurred, when only a small set of spectral bands were available and the computational power was limited. With the development of narrowband devices (i.e., with a few hundred spectrally narrow bands), this VIs may use the available information within the spectral observation range inefficiently, often relying on only a partial spectral subset. Algorithms for extracting optimized band information were thus created, utilizing well-established index formulations such as simple ratios and normalized differences. These algorithms involved the correlation of all potential band combinations to generate 2D correlation matrices, allowing for the visual identification of the most effective band combinations. Nevertheless, this approach can conduct to optimize indices which are strongly case specific, successfully optimized for local applications but not to generic cases (Verrelst, Camps-Valls et al. 2015). FS non-parametric methods which evaluate all the spectral wavelengths provided by hyperspectral sensors constitute an interesting option for disease assessment, providing more robust and customized information for modeling data class characteristics, and greater model performance (Thenkabail, Lyon et al. 2018, Reis-Pereira, Tosin et al. 2022). Thus, future research is needed to better explore different information extraction (e.g., modeling) approaches suitable to comprehend plant–pathogen interactions and their effect on host spectral behavior.

## 5. Conclusion

The present work aimed to apply hyperspectral reflectance in-field measurements for the diagnosis of bacterial canker of kiwi (BCK) disease, which is caused by the bacteria *Pseudomonas syringae* pv. *actinidiae* (Psa). Different vegetation indices were computed, and later used to classify symptomless and symptomatic kiwi leaves signatures. Chlgreen, mSR, CI, GI, BRI, AVI, PVIhyp, and Rre were signed as the most relevant for disease discrimination, highlighting the Green, Red, Red Edge, and NIR regions of the electromagnetic spectrum. These findings are in line with the metabolic and structural changes promoted by the pathogen in the host tissues. Classification modeling allowed disease discrimination with fair model metrics, showing the suitability of this approach for disease assessment. Nevertheless, further research exploring different Feature Selection methods considering a broader range of wavelengths is advised.

## Acknowledgements

Mafalda Reis-Pereira and Renan Tosin were supported by a fellowship from Fundação para a Ciência e a Tecnologia (FCT) with the references SFRH/BD/146564/2019 and SFRH/BD/145182/2019, respectively. Rui C. Martins acknowledges Fundação para a Ciência e Tecnologia (FCT) research contract grant (CEEIND/017801/2018).

## Supplementary materials

## Appendix A

**Table A1** Spectral Vegetation Indices (VIs) computed in this study.

| Vegetation Indices | Formula | Ref. |
|---|---|---|
| Ashburn Vegetation Index (AVI) | $2.0 \times \mathrm{NIR} - \mathrm{RED}$ | (Ashburn 1979, Bannari, Morin et al. 1995) |
| Anthocyanin reflectance index (ARI) | $\frac{1}{GREEN} - \frac{1}{RED}$ | (Gitelson, Merzlyak et al. 2001) |
| Blue Green Pigment Index (BGI) | $\frac{BLUE}{GREEN}$ | - |
| Browning Reflectance Index (BRI) | $\frac{\frac{1}{GREEN} - \frac{1}{RED}}{NIR}$ | (Merzlyak, Gitelson et al. 2003) |
| Chlorophyll Green (Chlgreen) | $\left(\frac{NIR}{GREEN}\right)^{(-1)}$ | (Gitelson, Keydan et al. 2006) |
| Coloration Index (CI) | $\frac{RED - BLUE}{RED}$ | (Escadafal, Belghith et al. 1994) |
| Chlorophyll Index Green (CIgreen) | $\frac{NIR}{GREEN} - 1$ | (Gitelson, Viña et al. 2003, Ahamed, Tian et al. 2011, Hunt Jr, Daughtry et al. 2011) |
| Chlorophyll Index Red Edge (CIrededge) | $\frac{NIR}{RED\ EDGE} - 1$ | (Gitelson, Viña et al. 2003, Ahamed, Tian et al. 2011, Hunt Jr, Daughtry et al. 2011) |
| Chlorophyll vegetation index (CVI) | $NIR * \frac{RED}{GREEN^2}$ | (Datt, McVicar et al. 2003) |
| Double Difference Index (DD) | $(749nm - 720nm) - (701nm - 672nm)$ | (Le Maire, François et al. 2004, Main, Cho et al. 2011) |
| Enhanced Vegetation Index (EVI) | $2.5 \times \frac{NIR - RED}{(NIR + 6RED - 7.5BLUE) + 1}$ | (Huete, Didan et al. 2002, Hunt Jr, Daughtry et al. 2011) |

| | | |
|---|---|---|
| Green atmospherically resistant vegetation index (GARI) | $\frac{NIR-(GREEN-(BLUE-RED))}{NIR-(GREEN+(BLUE-RED))}$ | (Gitelson, Kaufman et al. 1996, Gitelson, ViÃ±a et al. 2003) |
| Green-Blue NDVI (GBNDVI) | $\frac{NIR-(GREEN+BLUE)}{NIR+(GREEN+BLUE)}$ | (Wang, HUANG et al. 2007) |
| Global Environment Monitoring Index (GEMI) | $\left(n\times(1-0.25n)-\frac{RED-0.125}{1-RED}\right)$ $n=\frac{2\times(NIR^2-RED^2)+1.5\times NIR+0.5\times RED}{NIR+RED+0.5}$ | (Pinty and Verstraete 1992) |
| Simple Ratio Greenness Index (GI) | $\frac{GREEN}{RED}$ | (Zarco-Tejada, Miller et al. 2001, Main, Cho et al. 2011) |
| Green Normalized Difference Vegetation Index (GNDVI) | $\frac{NIR-GREEN}{NIR+GREEN}$ | (Ahamed, Tian et al. 2011, Hunt Jr, Daughtry et al. 2011) |
| Tasselled Cap – vegetation (GVI) | $-0.2848\times Blue-0.2435\times Green-0.5436\times Red+0.7243\times NIR+0.0840\times SWIR-0.1800\times SWIR$ | (Schlerf, Atzberger et al. 2005, Lee, Alchanatis et al. 2010) |
| Infrared percentage vegetation index (IPVI) | $\frac{\frac{NIR}{NIR+RED}}{2}\times(NDVI+1)$ | (Crippen 1990, Kooistra, Leuven et al. 2003) |
| Log Ratio (LogR) | $log\left(\frac{NIR}{RED}\right)$ | - |
| Misra Green Vegetation Index (MGVI) | $-0.386\times GREEN-0.530\times RED+0.535\times REDEDGE+0.532\times NIR$ | (Misra, Wheeler et al. 1977, Bannari, Morin et al. 1995) |
| Modified NDVI (mNDVI) | $\frac{NIR-RED}{NIR+RED-2\times BLUE}$ | (Huete, Liu et al. 1997, Main, Cho et al. 2011) |
| Modified Simple Ratio (mSR) | $\frac{NIR-BLUE}{RED-BLUE}$ | (Kooistra, Leuven et al. 2003, Main, Cho et al. 2011) |
| Modified Simple Ratio 2 (mSR2) | $\left(\frac{NIR}{RED}\right)-\frac{1}{\sqrt{\left(\frac{NIR}{RED}\right)+1}}$ | (Chen 1996) |
| Normalized Difference NIR / Red Normalized Difference Vegetation Index (NDVI) | $\frac{NIR-RED}{NIR+RED}$ | (Thenkabail, Smith et al. 2002, Zarco-Tejada and Sepulcre-CantÃ³ 2007) |
| Normalized Green (NG) | $\frac{GREEN}{NIR+RED+GREEN}$ | (Sripada, Heiniger et al. 2006) |

| | | |
|---|---|---|
| Normalized Near Infrared (NNIR) | $\frac{NIR}{NIR + RED + GREEN}$ | (Sripada, Heiniger et al. 2006) |
| Hyperspectral perpendicular VI (PVIhyp) | $\frac{NIR - a \times 807 - b}{(1 + a^2)^{0.5}}$ $a = 1.17, b = 3.37$ | (Schlerf, Atzberger et al. 2005) |
| Plant Senescence Reflectance Index (PSRI) | $\frac{RED - BLUE}{NIR}$ | (Sims and Gamon 2002, Apan, Held et al. 2003) |
| Reflectance at the inflexion point (Rre) | $\frac{RED + NIR}{2}$ | (Clevers, De Jong et al. 2002) |
| Red-Edge Stress Vegetation Index (RVSI) | $\frac{718 + 748}{2} - 733$ | - |
| Structure Intensive Pigment Index (SIPI) | $\frac{NIR - BLUE}{NIR - RED}$ | (Zarco-Tejada, Miller et al. 2001, le Maire, Francois et al. 2004) |
| Simple Ratio (SR) | $\frac{NIR}{RED}$ | - |

# Appendix C | Oral communication - 4th Annual Conference of the EuroXanth COST Action

## Tracking changes on host physiological traits promoted by Xanthomonas euvesicatoria: proximal optical sensing as an innovative tool for plant disease detection

Mafalda Reis-Pereira[1,2,+], Rui C. Martins[3], Filipe Monteiro-Silva[1,3], Fernando Tavares[1,4,+], Filipe Neves dos Santos[2], Mário Cunha[1,2]

[1] Faculty of Sciences, University of Porto (FCUP), Rua Campo Alegre s/n, 4169-007, Porto, Portugal

[2] Centre of Robotics in Industry and Intelligent Systems, INESC TEC, Dr. Roberto Frias, 4200-465, Porto, Portugal

[3] Centre for Applied Photonics, INESC TEC, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[4] Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), Rua Padre Armando Quintas, nº 7, 4485-661, Vairão, Portugal

[+] Corresponding authors: mafalda.r.pereira@inesctec.pt (Mafalda Reis Pereira); mario.cunha@inesctec.pt (Mário Cunha);

**Abstract**

*Xanthomonas euvesicatoria* (Xeu) is a bacterial pathogen known to cause disease in crops of high economic importance worldwide threatening their yield, quality, and economic value. The current methods used to assess this pathogen often depend on the presence of visible signs of the infection, which frequently manifest themselves only in the late stages of this process, compromising the effectiveness of protection measures. Therefore, complementary methods based on proximal optical sensing (POS) have recently been explored. Based on evidence that plant-pathogen interactions promote changes in the biochemical and internal structures of the host, resulting in modifications to their optical properties, this study evaluated the potential use of a POS as an effective technique for the early detection of pathogen infection. A compact, modular sensing system, combining direct UV-Vis spectroscopy with optical fibers, supported by a robust Self-Learning Artificial Intelligence (SLAI), was used to assess the modifications promoted by Xeu in tomato leaves (cv. Cherry). Plant infection was performed by

spraying a bacterial suspension (1.0 x $10^8$ cells/$mL^{-1}$) until run-off occurred, and a similar approach was followed for the control group where only water was applied. A total of 270 spectral assessments were performed on leaves, on five different dates, which included pre- and post-inoculation measurements. The spectral signatures were then analyzed by principal components analysis coupled with an innovative SLAI algorithm, which allowed the distinction and differentiation of healthy and infected leaves. These findings indicate that this non-destructive, in vivo POS approach may be a promising tool for detecting the changing spectral behavior of diseased plant leaves.

**Keywords**

Plant disease detection, Plant Pathology, Proximal sensing, Spectroscopy, Precision agriculture

**Keywords**

Mafalda Reis-Pereira and Aníbal Filipe Silva were supported by fellowships from Fundação para a Ciência e a Tecnologia (FCT) with the references SFRH/BD/146564/2019 and DFA/BD/9136/2020, respectively. Rui C. Martins acknowledges Fundação para a Ciência e Tecnologia (FCT) research contract grant (CEEIND/017801/2018). This research was supported by the project 'SpecTOM – Metabolomics Tomography Spectroscopy System', University of Porto, Fundação Amadeus Dias and Santander-Universities Grant.

**Supplementary materials**

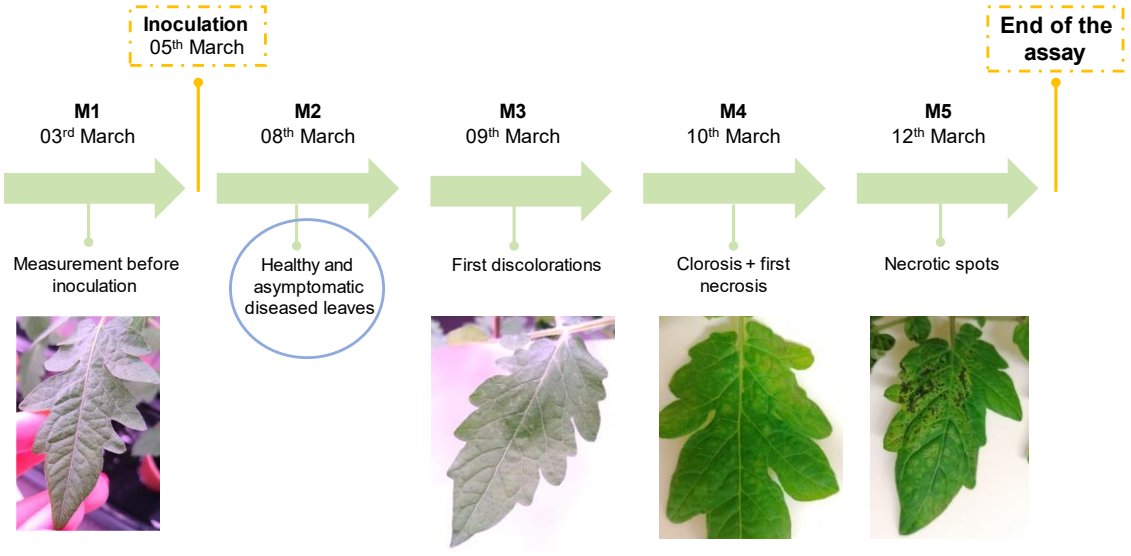



**Figure 1** Diagram showing the moment of the bacterial inoculation assay performed in tomato leaflets, along with the moment of performance of the phenotypical and spectral measurements (M) overtime. Here it is possible to see the appearance and development of the lesions through time.

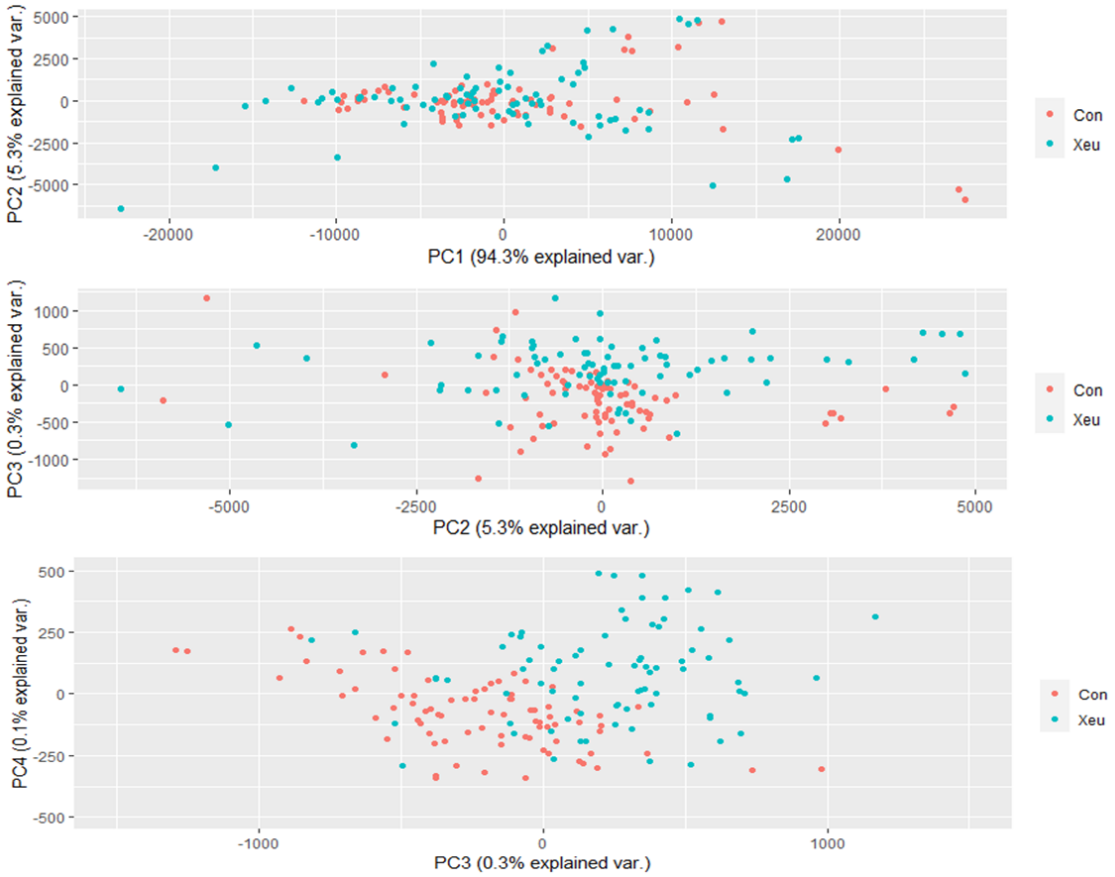



**Figure 2** Principal Component Analysis (PCA) results of the principal component (PC) 1, 2, and 3 resulting from the PCA of the dataset three days after inoculation (all leaves were asymptomatic, showing no symptoms of the disease caused by *Xanthomonas euvesicatoria* LMG 905).

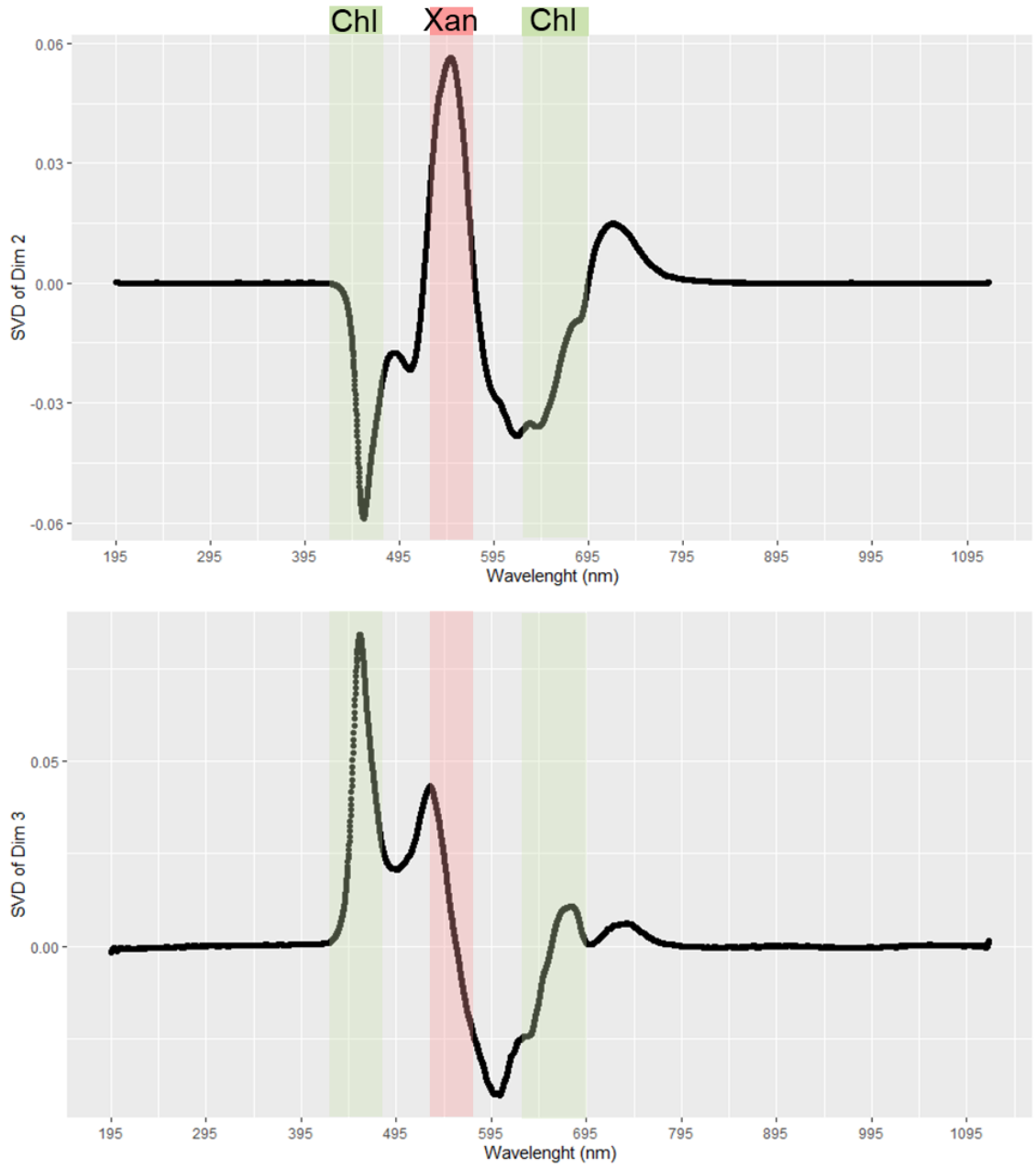



**Figure 3** Principal Component Analysis (PCA) loading results of the principal component (PC) 2, and 3 resulting from the PCA of the dataset three days after inoculation (all leaves were asymptomatic, showing no symptoms of the disease caused by *Xanthomonas euvesicatoria* LMG 905).

# Appendix D | Oral communication - 4º Encontro Biologia Funcional e Biotecnologia de Plantas

## Diagnostics of bacterial plant diseases: proximal optical sensors as new tools for an early detection

Mafalda Reis-Pereira[1,3], Fernando Tavares[1,2], Filipe Neves dos Santos[3], Mário Cunha[1,3]

[1] Faculty of Sciences, University of Porto (FCUP)

[2] Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO)

[3] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) - Centre for Robotics in Industry and Intelligent Systems (CRIIS)

**Abstract**

Early diagnosis of plant diseases, alongside their mapping in the field, are justified by agronomic, environmental, economic, and humanitarian reasons. These practices prevent a crop from being severely affected, as well as allow the targeted application of protection products. A reduction in the use of pesticides and herbicides is expected, which translates into a beneficial impact on the protection of the environment and ecosystem services, on the income of the producer and on the quality of the product that reaches the final consumer. Aligned with this, the European Commission has the aim of applying the set of policy initiatives called *Green Deal*, namely the mechanisms established in the *Farm to Fork* to reduce the use of plant protection products and fertilizers by 50% by 2030.

Currently, detection and identification of phytopathogens are done through direct methods, being the most used the visual assessment of symptoms and the use of molecular and serological techniques. The first approach, despite being very useful, can be demanding and may not be suitable for monitoring all crops. In turn, molecular and serological methods allow the processing of several samples, precise identification of phytopathogens, identification of strains with different virulence, and characterization of the pathogens' diversity. However, these methods can be ineffective for the detection of pathogens in asymptomatic plants, require specialized resources, and do not allow the tracking of all infected plants in cultivated areas.

To respond to these limitations, indirect diagnostic methods have been emerging. They are based on plant-pathogen interactions, which may result in changes in the internal and biochemical structure of leaves. These changes promote modifications in

the optical properties of the host plants which can be detected by optical proximity sensors (multi / hyperspectral, fluorescence or thermal). Thus, arises the hypothesis of analyzing the spatio-temporal pattern of the development of plant diseases through their spectral properties, to allow an early diagnosis in an easy, non-destructive, and specific way.

In this context, a doctoral project was conceived on the theme 'Early detection and identification of plant diseases caused by bacteria based on proximal sensing from a precision agriculture perspective'. This project aims to develop assessment methods, based on the spectral properties of plants, for the early detection and identification of bacteria responsible for diseases in crops. The tests are being carried out under laboratory and field conditions, using different pathovars of *Xanthomonas* spp. and *Pseudomonas syringae*, as well as different crops (kiwi, walnut, and tomato). Its implementation will contribute to enhance the detection of diseases, allowing an early intervention and the development of predictive methods to map diseases, preventing a crop from being severally affected and granting a focused application of protection products.

**Supplementary materials**

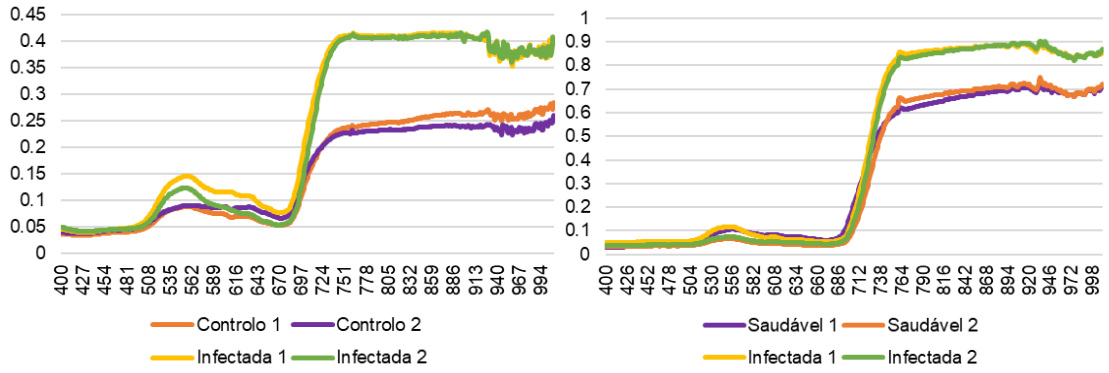



**Figure 1** Examples of spectral signatures of tomato plants (left) and kiwi plants (right) obtained using a spectroradiometer, a portable proximal detection sensor. Legend: Control – 'Controlo', Diseased – 'Infectada'.

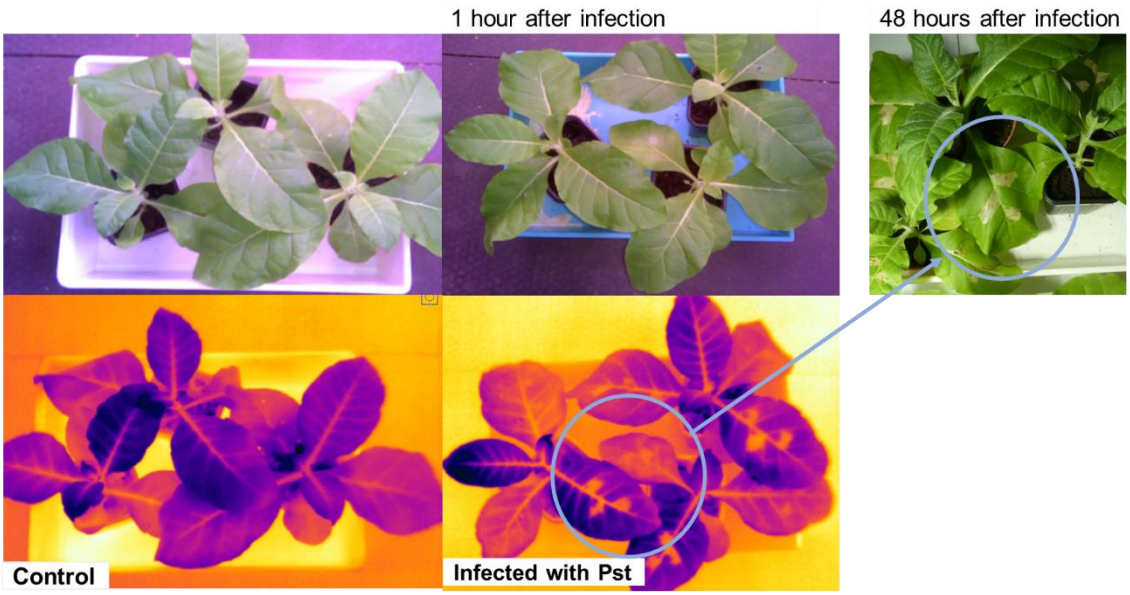



**Figure 2** Images captured using a thermal camera as part of monitoring the infection of tobacco plants in the laboratory. The upper line contains the RGB images collected 1 h (left) and 48 h (right) after infection, and the bottom line shows the corresponding thermal imaging. In the thermal images, it is possible to observe yellow areas on the leaf, corresponding to the areas surrounding the places where the infiltration was carried out even before symptom development. After 48 h, in the inoculated tobacco leaves occurred the full formation of bacterial lesions in the place where here were previously yellow spots in the thermal image.

# Appendix E | Poster presentation - 14th European Conference on Precision Agriculture Unleashing the potential of Precision Agriculture (ECPA2023)

## Early assessment of tomato bacterial spot through proximal hyperspectral sensing: testing data preprocessing approaches and applied modeling in diagnostics of plant diseases

Mafalda Reis Pereira[1,2,+] , Fernando Tavares[1,3], Filipe Neves dos Santos[2], Mário Cunha[1,2]

[1] Faculty of Sciences of the University of Porto (FCUP), Rua Campo Alegre s/n, 4169-007 Porto, Portugal

[2] Centre of Robotics in Industry and Intelligent Systems (INESC TEC), Dr. Roberto Frias, 4200-465 Porto, Portugal

[3] Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), Rua Padre Armando Quintas, nº 7, 4485-661 Vairão, Portugal

[4] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

[+] Corresponding authors: mafalda.r.pereira@inesctec.pt (Mafalda Reis Pereira), mccunha@fc.up.pt (Mário Cunha)

**Abstract**

Predictive modeling based on hyperspectral transmittance data can be a fast and cost-effective approach for improving crop disease assessment. This study investigated the potential of transmittance for in-vivo detection of *Xanthomonas euvesicatoria* on tomato leaves. Infection was accomplished by spraying leaves with a bacterial suspension. Hyperspectral assessments were randomly performed on different leaves for 18 days in a dark room, building a data set of 2430 observations. A supervised machine learning model was tested to discriminate between control and diseased leaves, as well as between healthy, pre-symptomatic, and symptomatic samples. The best leaves' classification accuracy before symptom appearance achieved 85% (healthy vs pre-symptomatic) and 90% (control vs diseased). These findings support the application of in-vivo spectral measurements for disease diagnosis.

**Abstract**

Tomato disease, Xanthomonas spp. diagnosis, Early detection; Proximal Sensing, Hyperspectral data

## Introduction

Plant diseases are responsible for causing major losses in numerous crops worldwide, affecting their yield, and their economic and nutritional value. Early disease detection, promoted by applied predictive classification methods, allows a more immediate and precise intervention, preventing a crop from being severely affected. A reduction in the usage of phytosanitary products is expected, which translates into a beneficial impact on the protection of the environment and ecosystem services, on the producer's income and on the quality of the product that reaches the final consumer. Proximal hyperspectral spectroscopy approaches combined with applied predictive classification models are a helpful solution for assisting producers in early disease diagnosis in vivo tomato plants. In this regard, spectral data must be collected and evaluated to retrieve qualitative and quantitative information, identifying divergences between samples with different health statuses.

## Objectives

The aims of this research were i) to verify if the spectral behaviour of healthy and diseased tomato leaves presented differences; ii) to investigate the capacity of applied predictive models to early detect bacterial tomato diseases (diagnose in pre-symptomatic stages); iii) to develop applied predictive models to classify leaves according to their treatment group (control vs. inoculated plants), and their health status (healthy, pre-symptomatic, and symptomatic).

## Materials and methods

Tomato plants were cultivated in a walk-in plant growth chamber, and divided into two groups, one of them being inoculated with *Xanthomonas euvesicatoria* LMG 905 (Xeu) bacteria and the other being treated with sterile distilled water only (control group, Con) according to [1]. Plants were monitored daily for symptom development for 18 days. Hyperspectral data were randomly collected in vivo from the adaxial side of leaves using an in-house compact benchtop system composed by laptop, mini spectrometer (TM Series C11697MB, Hamamatsu Photonics K.K., Japan), and a transmission optical fibre bundle with a reflection probe. The probe was placed 1 cm above the sample, in a dark room, and a white LED light was used to provide even illumination to the abaxial surface of the leaf. Measurements were taken from 2430 points, belonging both to healthy and diseased leaves.

To assess the predictive modelling of bacterial diseases in tomato leaves, only the spectral region of 400 to 800 nm was used. Raw and normalized spectra were used for data analysis. The normalized spectral signatures were obtained through the division of leaves raw spectral signatures by the spectral signal of the white LED source (according to the time of exposure of the spectral acquisitions). Spectral modelling was then applied to classify tomato leaves pooled according to their health status (HS, independent variable): HS1 (control and Xeu disease plants), and HS2 (healthy, pre-symptomatic and symptomatic). For each approach, class discrimination was performed by date (days after inoculation, DAI). The datasets were randomly divided into training data and validation data (70 / 30%), following a holdout method [2], for each measurement date. To determine which wavelengths predictors were more relevant to diagnose tomato bacterial disease caused by Xeu a Flexible Discriminant Analysis (FDA) was computed (using a repeated 10-fold cross-validation). Different metrics were retrieved to investigate model performance, namely accuracy, Confusion Matrix, Kappa coefficient, and F1-Score according to Reis-Pereira et al. [3].

**Results**

Tomato plants infected with Xeu bacteria showed the first visual typical chlorotic disease symptoms between 12 to 15 DAI, only evolving to the necrotic stage at 17 to 18 DAI. Healthy leaves presented a spectral signature divergent from diseased leaves in raw and normalized data, even before symptom appearance. Spectral divergences were more evident in the ranges of approximately 425-460 nm, 520-585 nm for the raw data, and 425-515 nm, 640-710 nm, 710-770 nm for the normalized set.

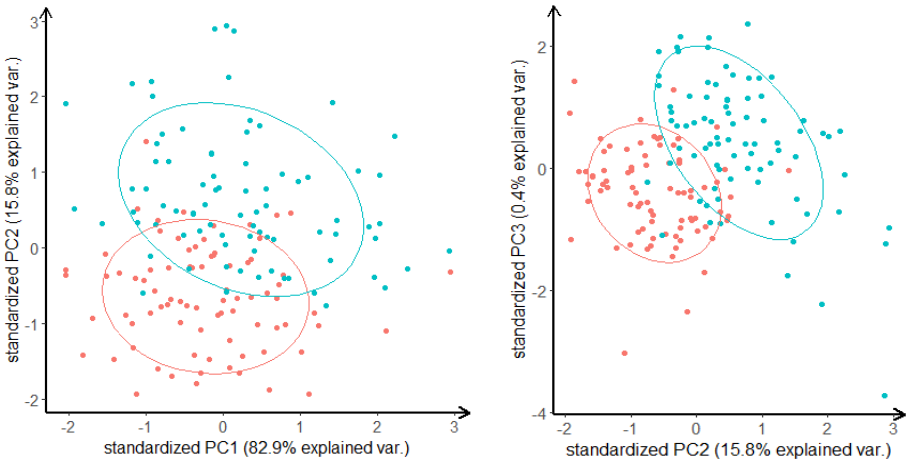



**Figure 1** Biplot of PCA results of raw data at the 8th DAI (before symptom appearance).

The best modelling approach before symptom appearance, for Control and Xeu HS1 classification, was achieved by applying FDA predictive model in both spectral data sets at the 8th DAI, demonstrating an accuracy of 0.90, kappa of 0.79, and f1-measure

of 0.90. For 'healthy' and pre-symptomatic discrimination, the best strategy involving the computation of the same model presented an accuracy of 0.85, kappa of 0.71, and f1-measure of 0.85. After the first symptoms 10 DAI appeared, the best HS1 classification was achieved when normalized data was used. The model registered an accuracy of 0.96, kappa of 0.92, and f1-score of 0.96. In HS2 prediction, is possible to see a NaN value of f1 for the 'symptomatic' class due to the reduced number of symptomatic samples (Table 1).

**Table 1** Model evaluation metrics (accuracy - Ac, kappa score - Kp, and f1-measure - F1) for test sets, when raw and normalized data were used, at 6, 8, and 10 days after infection (DAI).

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Health Status 1** | | | **Health Status 2** | | | | **Health Status 1** | | | **Health Status 2** | | | |
| *DI* | 6 | 8 | 10 | 6 | 8 | 10 | | 6 | 8 | 10 | 6 | 8 | 10 | |
| *Ac* | 0.77 | 0.90 | 0.94 | 0.75 | 0.85 | 0.75 | | 0.77 | 0.90 | 0.96 | 0.75 | 0.85 | 0.75 | |
| *Kp* | 0.54 | 0.79 | 0.75 | 0.50 | 0.71 | 0.55 | **Raw** | 0.54 | 0.79 | 0.92 | 0.50 | 0.71 | 0.56 | **Norm** |
| *F1* | 0.80 | 0.90 | 0.94 | 0.77 | 0.84 | 0.70,0.86,0.29 | | 0.80 | 0.90 | 0.96 | 0.77 | 0.85 | 0.73,0.88,NaN | |

## Discussion and conclusions

In-vivo hyperspectral spectroscopy combined with applied predictive classification was explored to diagnose bacterial tomato leaf disease caused by Xeu bacteria. Even in early infection stages, spectral separability between healthy and diseased leaves was observed, allowing for accurate classification of the HS1 group (90% accuracy) and HS2 discrimination (85% accuracy). These results demonstrate the potential of applied predictive classification modelling using hyperspectral point data to early detect bacterial crop diseases on leaves.

Further research is suggested to better understand the host-pathogen interactions, and their impact on the crop's spectral signature. This can lead to the development of more cost-effective devices, and agricultural practices (e.g. phytosanitary treatments), leading to more efficient and environmentally friendly agricultural practices. Spectroscopic sensors can, withal, be coupled with different measuring platforms, allowing for spectral data studies from the leaf to the canopy scale.

## Acknowledgments

This work was financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, within the project OmicBots, with reference PTDC/ASP-HOR/1338/2021. M. Reis Pereira was supported by an FCT fellowship with the reference SFRH/BD/146564/2019.

## Poster

# Early assessment of tomato bacterial spot through proximal hyperspectral sensing

*Mafalda Reis Pereira*[1,2], *Fernando Tavares*[1,3,4], *Filipe Neves dos Santos*[2], *Mário Cunha*[1,2]

[1] Faculty of Sciences of the University of Porto (FCUP), Portugal; [2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Portugal; [3] Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), Portugal; [4] BIOPOLIS, Portugal. Correspondence: mafalda.r.pereira@inesctec.pt, mccunha@fc.up.pt

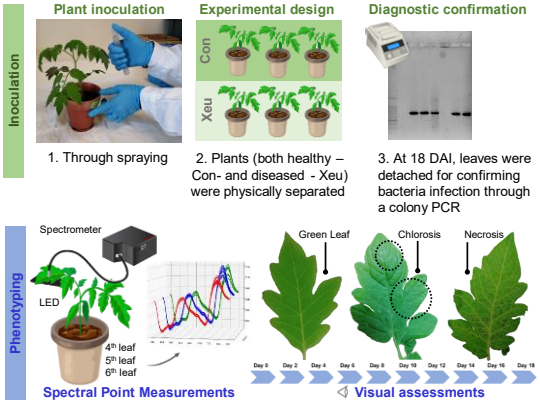
## 1 Purpose

Plant diseases are responsible for causing major losses in numerous crops worldwide, affecting their yield, and their economic and nutritional value. Early disease detection, promoted by applied predictive classification methods, allows a more immediate and precise intervention, preventing a crop from being severely affected. A reduction in the usage of phytosanitary products is expected, which translates into a beneficial impact on the protection of the environment and ecosystem services, on the producer's income and on the quality of the product that reaches the final consumer. Proximal hyperspectral spectroscopy approaches combined with applied predictive classification models are a helpful solution for assisting producers in early disease diagnosis in vivo tomato plants. In this regard, spectral data must be collected and evaluated to retrieve qualitative and quantitative information, identifying divergences between samples with different health statuses.

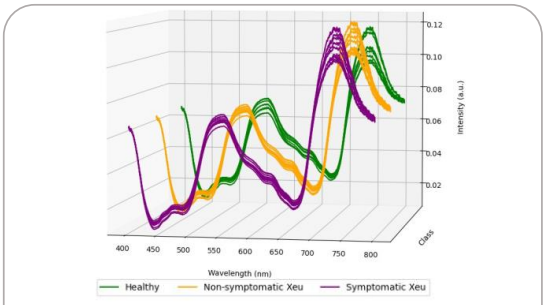
**The aims of the study were to evaluate:**

1) If the spectral behaviour of healthy and diseased tomato leaves presented differences;
2) The capacity of applied predictive models to early detect bacterial tomato diseases (diagnose in pre-symptomatic stages);
3) The development of applied predictive models to classify leaves according to their treatment group (control vs. inoculated plants), and their health status (healthy, pre-symptomatic, and symptomatic).

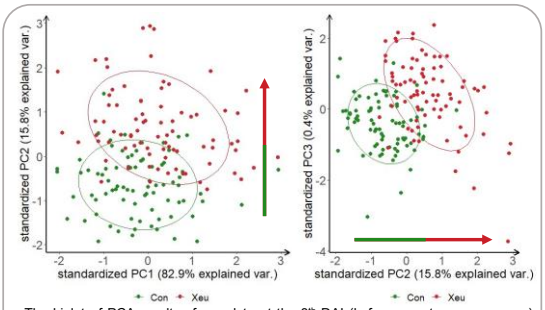
## 2 Methods




## 3 Major Findings




Tomato plants infected with Xeu bacteria showed the first visual typical chlorotic disease symptoms between 12 to 15 DAI, only evolving to the necrotic stage at 17 to 18 DAI. Healthy leaves presented a spectral signature divergent from diseased leaves in raw and normalized data, even before symptom appearance. Spectral divergences were more evident in the ranges of approximately 425-460 nm, 520-585 nm for the raw data, and 425-515 nm, 640-710 nm, 710-770 nm for the normalized set.




The biplot of PCA results of raw data at the 8th DAI (before symptom appearance) showed that healthy (Control) and Xeu diseased leaflets presented a spectral separation gradient through PC1 and PC2.

| | Health Status 1 | | | Health Status 2 | | | | Health Status 1 | | | Health Status 2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DI | 6 | 8 | 10 | 6 | 8 | 10 | | 6 | 8 | 10 | 6 | 8 | 10 | |
| Ac | 0.77 | 0.90 | 0.94 | 0.75 | 0.85 | 0.75 | Raw | 0.77 | 0.90 | 0.96 | 0.75 | 0.85 | 0.75 | Norm |
| Kp | 0.54 | 0.79 | 0.75 | 0.50 | 0.71 | 0.55 | | 0.54 | 0.79 | 0.92 | 0.50 | 0.71 | 0.56 | |
| F1 | 0.80 | 0.90 | 0.94 | 0.77 | 0.84 | 0.70,0.86,0.29 | | 0.80 | 0.90 | 0.96 | 0.77 | 0.85 | 0.73,0.88,NaN | |

The best modelling approach before symptom appearance, for Control and Xeu HS1 classification, was achieved by applying FDA predictive model in both spectral data sets at the 8th DAI, demonstrating an accuracy of 0.90, kappa of 0.79, and f1-measure of 0.90. For 'healthy' and pre-symptomatic discrimination, the best strategy involving the computation of the same model presented an accuracy of 0.85, kappa of 0.71, and f1-measure of 0.85. After the first symptoms 10 DAI appeared, the best HS1 classification was achieved when normalized data was used. The model registered an accuracy of 0.96, kappa of 0.92, and f1-score of 0.96. In HS2 prediction, is possible to see a NaN value of f1 for the 'symptomatic' class due to the reduced number of symptomatic samples.

## 4 Conclusions

In-vivo hyperspectral spectroscopy combined with applied predictive classification was explored to diagnose bacterial tomato leaf disease caused by Xeu bacteria. Even in early infection stages, spectral separability between healthy and diseased leaves was observed, allowing for accurate classification of the HS1 group (90% accuracy) and HS2 discrimination (85% accuracy). These results demonstrate the potential of applied predictive classification modelling using hyperspectral point data to early detect bacterial crop diseases on leaves.

Further research is suggested to better understand the host-pathogen interactions, and their impact on the crop's spectral signature. This can lead to the development of more cost-effective devices, and agricultural practices (e.g., phytosanitary treatments), leading to more efficient and environmentally friendly agricultural practices. Spectroscopic sensors can, withal, be coupled with different measuring platforms, allowing for spectral data studies from the leaf to the canopy scale.

**Acknowledgements**

This work was financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, within the project OmicBots, with reference PTDC/ASP-HOR/1338/2021. M. Reis Pereira was supported by a FCT fellowship with the reference SFRH/BD/146564/2019.

# Appendix G | Poster presentation - II Plant Pests and Diseases

## Forum - Redefining Concepts, Mechanisms & Management Tools

## A review on the main challenges in early diagnostics of plant diseases based on proximal sensing

Mafalda Reis Pereira[1,3,+], Fernando Tavares[1,2], Filipe Neves dos Santos[3], Mário Cunha[1,3,+]

[1] Faculty of Sciences, University of Porto (FCUP), Rua Campo Alegre s/n, 4169-007, Porto, Portugal

[2] Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), Rua Padre Armando Quintas, nº 7, 4485-661, Vairão, Portugal

[3] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) - Centre for Robotics in Industry and Intelligent Systems (CRIIS), Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

[+] Corresponding authors: mccunha@fc.up.pt (Mário Cunha); m.r.pereira@inesctec.pt (Mafalda Reis Pereira)

**Abstract**

Pathogen infections are among the main factors that threaten crop production, being their early detection an important step to efficiently manage plant diseases. Current methods of disease detection often depend on the presence of visible signs of the infection, which often manifest themselves only in the late stages of the process, compromising the effectiveness of protection measures. They are classified as direct methods and include sensitive and accurate molecular and serological laboratory-based techniques. These approaches despite being extremely useful are labor-intensive, time-consuming, and require detailed sample processing. Therefore, alternative indirect methods have recently been explored, introducing new perspectives in the phytopathology field. They assume that plant-pathogen interactions cause changes in the internal and biochemical structure of leaves, resulting in modifications on the optical properties of the host that can be detected by sensors often couple with artificial intelligence. This review presents some of these indirect proximal sensing (PS) techniques, including hyperspectral, thermal, fluorescence, and gas chromatography approaches. A literature search following PRISMA protocols was conducted in the Web of Science database for publications that investigated the suitability of PS for plant

disease assessment. Research shows that data obtained through PS techniques can be specifically analyzed to extract useful information, allowing the distinction between healthy and infected plants, and ultimately the identification of a specific disease. Therefore, PS seems an accurate, fast, and intuitive tool for crop disease diagnosis, although its technology readiness level (TRL) is still low and some technical difficulties must be surpassed, reducing errors associated with the measurement.

**Keywords**

Plant disease detection, Proximal sensing, Precision agriculture, Optical sensing, Gas chromatography

**Funding**

Reis-Pereira, M. was supported by a fellowship from FCT (Fundação para a Ciência e a Tecnologia) with the reference SFRH/BD/146564/2019

**Award**

Award Certificate

This certifies that

**Mafalda Pereira**

has won the award for best e-poster communication at the event **II Plant Pests and Diseases Forum**: **Redefining Concepts, Mechanisms and Management Practices** held online on March 24th 2021. The event was organized by PhD Students from the **Integrative Biology and Biotechnology Laboratory (iB2Lab)** and **PlantStress Lab**, from the Faculty of Sciences of University of Porto.

**Conceição Santos**
Full Professor
Deputy Director of Faculty of Sciences, University of Porto
Principal Researcher of $iB_2$ Lab

**Fernanda Fidalgo**
Associated Professor
Principal Researcher of PlantStress Lab

Organization: Patronage: Sponsors:

## Poster

# A REVIEW ON THE MAIN CHALLENGES IN EARLY DIAGNOSTICS OF PLANT DISEASES BASED ON PROXIMAL SENSING

*Mafalda Reis Pereira, Fernando Tavares, Filipe Santos, Mário Cunha*

### I. BACKGROUND

Plant pathogen infections are among the main factors that affect crop quality and productivity, resulting in economic losses and risk of environmental damages. Therefore, its detection is an important step to successfully manage diseases in greenhouses, orchards, and fields. Plant-pathogen interactions promote changes on a wide variety of plant chemical, biophysical, and metabolic traits.

Current methods of disease detection often depend on the presence of visible signs of the infection, which often manifest themselves only in the late stages of the process, compromising the effectiveness of protection measures. They are classified as direct methods and include sensitive and accurate molecular and serological laboratory-based techniques. These approaches despite being extremely useful are labor-intensive, time-consuming, and require detailed sample processing.

Therefore, alternative indirect Proximal Sensing (PS) methods have recently been explored, introducing new perspectives in the phytopathology field. They assume that plant-pathogen interactions cause changes in the internal and biochemical structure of leaves of the host that can be detected by different types of sensors, often couple with artificial intelligence.

These PS techniques allied with recent technologies, such as global positioning system (GPS) and variable rate spray systems, allow disease mapping and the prospect of Precision Disease Management (PDM). These aim to target phytosanitary products where and when needed and at an appropriate dose. Such practices, in line with the Precision Agriculture concept, make an important contribution in assisting farmers to reduce agricultural inputs, resulting in fewer costs, as well as less residues in crop products and on the environment.

### II. METHODS

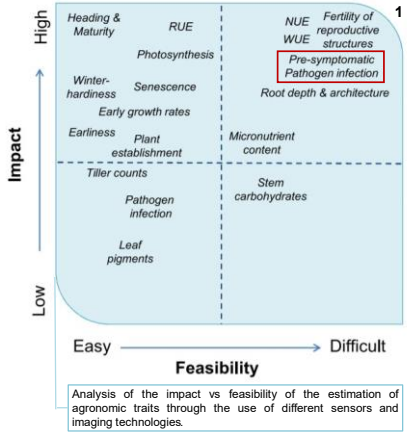
A literature search following PRISMA protocols was conducted in the Web of Science database for publications up to February 2021 that investigated the suitability of Proximal Sensing techniques for plant disease assessment.

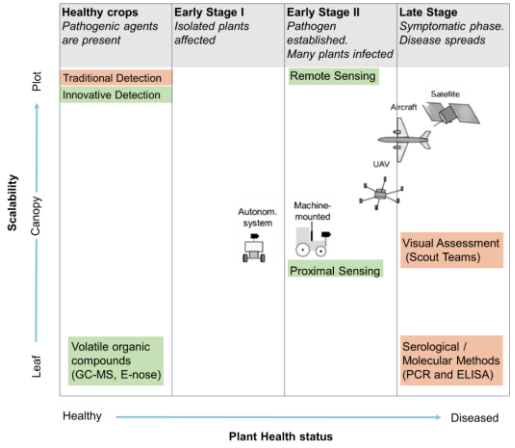

### III. FINDINGS

In recent years, there has been an increasing trend towards using non-destructive, sensor-based methods for detecting plant diseases.

These techniques can be spectroscopy or imaging-based and provide reliable and accurate technical assistance for real-time disease detection and monitoring. The most applied approaches include multi/hiper-spectral, fluorescence, thermal, laser induced breakdown spectroscopy (LIBS), biophoton emission detection, and volatile organic compound assessment techniques. These sensors can be coupled on different platforms to form functioning measurement systems, allowing the realization of mapping, monitoring, scouting, and applying tasks. Their flexibility allows them to assess leaf, single-plant, canopy, and even plot and regional scales.

Nevertheless, the minor differences between healthy and diseased plants at an early stage of the infection process can compromise the feasibility of the disease detection and identification. Thus, early plant disease detection continues to be extremely difficult, particularly when made prior to the appearance of the first symptoms of the infection.




**Proximal Sensing methods**

Most used sensor based on plant–pathogen interactions:

- Multispectral
- Hyperspectral
- Fluorescence
- Thermal
- Biophotons
- LIBS
- GC-MS

### IV. FUTURE WORK

Proximal Sensing techniques must be explored and adjust to allow: (i) the detection of a disease at early points in time, (ii) the differentiation among different diseases, (iii) the separation of diseases caused by biotic and abiotic stresses, and (iv) the quantification of disease severity.

Sensor fusion approaches, as well as sensor combination with molecular and serological methods, could be a new approach which must be tested in future research.

The interpretation of sensor data, especially spectral data, is still a limiting factor for the detection, identification and quantification of diseases in crop production. Its implementation into robust decision support systems by transdisciplinary cooperation will improve the acceptance and realization of precision crop protection.

**REFERENCES**

1. https://doi.org/10.1016/j.plantsci.2019.01.011
2. Adapted from https://doi.org/10.3390/s21010171and https://doi.org/10.1007/s13593014-0246-1
3. Adaptad from https://doi.org/10.1007/978-94-017-9020-8_4

**ACKNOWLEDGES**

Mafalda Reis Pereira is pleased to acknowledge her PhD grant nº. SFRH/BD/146564/2019 awarded by Fundação para a Ciência e Tecnologia (FCT).

# Appendix F | Magazine article I

## Sensores óticos de proximidade para diagnóstico avançado das doenças das plantas

Mafalda Reis Pereira[1,2], Fernando Tavares[1,3], Filipe Neves dos Santos[2], Mário Cunha[1,2,*]

[1] Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, Porto, 4169-007, Portugal

[2] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

Há razões agronómicas, ambientais, económicas e humanitárias que justificam o desenvolvimento de novos métodos de diagnóstico precoce de doenças das plantas, assim como o seu mapeamento no campo, em linha com a agricultura de precisão.

Atualmente, a deteção e identificação de stress e doenças das plantas é feita através de métodos de diagnóstico diretos, sendo os mais utilizados, a técnica de diagnóstico visual e os métodos moleculares e sorológicos. O primeiro método consiste em verificar a cultura em busca de sinais indicadores visíveis (que frequentemente se manifestam nos estados intermédios a tardios da infeção). Esta técnica, demorada e exigente, pode não ser conveniente no acompanhamento de todas as culturas, estando limitada pela sua área de cultivo e pela fenologia das plantas. Por sua vez, os métodos moleculares e sorológicos revolucionaram a deteção de doenças em plantas, pois permitem o processamento de uma grande quantidade de amostras, a identificação precisa de agentes patogénicos, a identificação de estirpes com diferente virulência e a caracterização da diversidade e evolução das populações de fitopatogénicos. Apesar da sua importância para diagnósticos de fitopatologias, estes métodos raramente se revelam eficazes para a deteção de patógenos em plantas assintomáticas, requerem recursos técnicos e humanos especializados, não são imediatos e não permitem o rastreamento de todas as plantas infetadas em áreas cultivadas, o que é particularmente importante para avaliações fitossanitárias completas.

Com o objetivo de dar resposta às limitações destes métodos visuais e laboratoriais, têm vindo a surgir os chamados métodos de diagnóstico indiretos (proxy). Estes baseiam-se na ocorrência de interações entre a planta e o patógeno que a afeta, podendo resultar na criação de estruturas patogénicas e em mudanças na estrutura interna e bioquímica das suas folhas. Estas alterações fisiológicas promovem mudanças nas propriedades óticas das plantas hospedeiras, especificamente na sua refletância e emitância, que podem ser detetadas por sensores óticos de proximidade (por exemplo, através de sensores multi / hiperespectrais, sensores de fluorescência ou termografia). Surge, assim, a hipótese de acompanhar o padrão espácio-temporal de desenvolvimento das doenças das plantas através da sua refletância e emitância, permitindo um diagnóstico precoce de forma rápida, fácil, não destrutiva e específica.

Neste âmbito, foi concebido um projeto de doutoramento na temática da 'Deteção e identificação precoce de doenças das plantas provocadas por bactérias com base na reflectância hiperespectral numa ótica de agricultura de precisão'. Este projeto tem como objetivo desenvolver métodos preditivos, baseados nas propriedades espectrais das plantas, para a deteção e identificação precoce de bactérias responsáveis por doenças em culturas agrícolas. Através da combinação da ciência fundamental (ex. fisiologia vegetal, e bioquímica), de sensores óticos diversos e de técnicas de inteligência artificial, desenvolvem-se metodologias para testes confiáveis e rápidos. Os ensaios estão a ser realizados em condições de laboratório e de campo, utilizando como caso de estudo diferentes patovares do género *Xanthomonas* spp. e da espécie *Pseudomonas syringae*, assim como diferentes culturas agrícolas, nomeadamente kiwi, nogueira, tabaco e tomate. A sua validação justificará a extensão deste tipo de estudos a outras culturas e outros agentes patogénicos, como os fungos que são igualmente responsáveis por danos e perdas nas culturas agrícolas. Após a validação, este sistema será ainda incorporado num braço robótico para efetuar o mapeamento de zonas de risco das doenças ao nível da parcela.

Este projeto identifica a oportunidade de se efetuar uma intervenção precoce, graças à deteção de doenças das plantas num estado inicial. De facto, através da prevenção e controlo da propagação da infeção e do agente patogénico, assim como da modificação de práticas culturais, é possível intervir e evitar que uma cultura seja totalmente afetada e comprometida. Por outro lado, é ainda reconhecida a possibilidade de se utilizarem métodos preditivos na elaboração do mapeamento das doenças ao nível da parcela, permitindo uma aplicação direcionada e precisa de produtos fitofarmacêuticos. Desta forma, prevê-se uma redução do uso de pesticidas e herbicidas, o que se traduz num impacto benéfico na proteção do meio ambiente e dos

serviços ecossistémicos, nos proveitos do produtor e na qualidade do produto que chega ao consumidor final. Este projeto está, assim, alinhado com os desafios que a agricultura europeia enfrenta atualmente no âmbito do *Green Deal*, nomeadamente dos mecanismos de operacionalização estabelecidos no *Farm to Fork* de redução de 50% do uso de produtos fitofarmacêuticos e de fertilizantes até 2030.

O desenvolvimento de métodos de diagnóstico mais automatizados, objetivos e sensíveis é, por todas estas razões, crucial para impulsionar a deteção de doenças em culturas de interesse agronómico.

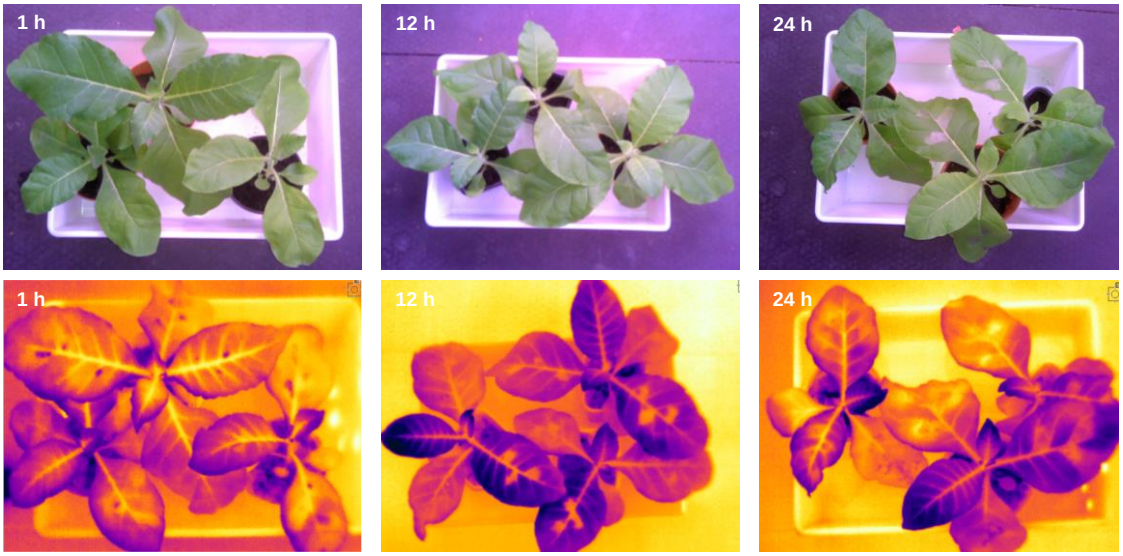



**Figure 1** Imagens capturadas com recurso a uma câmara térmica no âmbito da monitorização da infeção de plantas de tabaco em laboratório. A primeira coluna contém as imagens RGB e térmicas capturadas 1 h após a inoculação. É possível visualizar os locais onde foi realizada a infiltração (manchas azul-escuras). Essas manchas geralmente são circundadas por uma área de temperatura mais elevada (de cor amarela). A segunda coluna contém o mesmo tipo de imagens 24 h após o processo de inoculação. Na imagem térmica é possível observar áreas amarelas na folha, correspondentes às áreas circundantes dos locais onde a infiltração foi realizada. Na terceira coluna, é possível observar que 24 h depois da inoculação ocorreu a formação de lesões visíveis nos locais onde anteriormente já existiam áreas amarelas na imagem térmica.

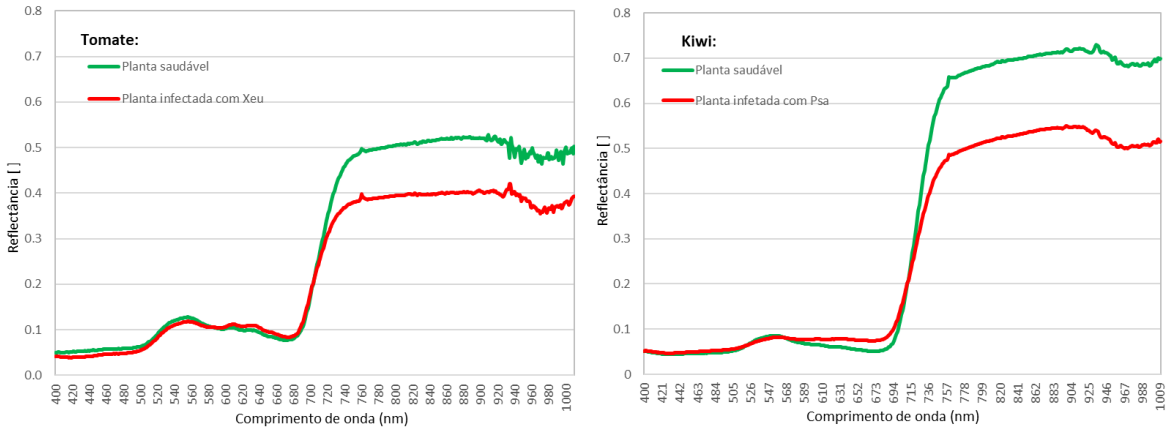



**Figure 2** Espectro de reflectância de plantas de tomate e de kiwi saudáveis e infetadas com diferentes bactérias (*Xanthomonas euvesicatoria* – Xeu – no Tomate e *Pseudomonas syringae* pv. *actinidea* – Psa – no Kiwi). O ensaio da cultura do kiwi foi realizado em campo e o da cultura do tomate em condições controladas (câmara walk-in). Em ambas as culturas foi possível observar que o espectro das plantas saudáveis, quando comparado com o das plantas infetadas, apresenta uma maior reflectância nos comprimentos de onda da zona do Infravermelho próximo e do visível do espectro eletromagnético, nomeadamente na região do vermelho.