



UNIVERSIDAD
NACIONAL
DE COLOMBIA



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Early detection of asymptomatic anthracnose disease in mango by hyperspectral imaging and machine learning

Carlos Alberto Velásquez Hernández

Thesis Supervisor:

Ph.D., Jose Blasco Ivars

Ph.D., Flavio Prieto Ortiz

Ph.D., Nuria Aleixos Borrás

Universidad Nacional de Colombia - Departamento de ingeniería mecánica y mecatrónica

Universitat Politècnica de València - Departamento de Ingeniería Gráfica

Instituto Valenciano de Investigaciones Agrarias - Centro de Agroingeniería

June, 2025

Early detection of asymptomatic anthracnose disease in mango by hyperspectral imaging and machine learning

Carlos Alberto Velásquez Hernández

Co-tutelle Ph.D. thesis to obtain the degree of:

Doctor en Ingeniería Mecánica y Mecatrónica by Universidad Nacional de Colombia

Doctor en Diseño, Fabricación y Gestión de Proyectos Industriales by Universitat Politècnica de València

Thesis Supervisor:

Ph.D., Jose Blasco Ivars

Ph.D., Flavio Prieto Ortiz

Ph.D., Nuria Aleixos Borrás

Line of research:

Machine learning and computer vision for spectral analysis

Advanced optical systems for fruit quality inspection and other applications. Precision agriculture

Research group:

GAUNAL

Universidad Nacional de Colombia - Departamento de ingeniería mecánica y mecatrónica

Universitat Politècnica de València - Departamento de Ingeniería Gráfica

Instituto Valenciano de Investigaciones Agrarias - Centro de Agroingeniería

June, 2025

Dedicado a:

Cenobia y Edison, quienes me dieron el ser.
Alberto, mi mentor, quien le dio forma.
Rosa, mi segunda madre, quien ayudó a criarlo
A mi hermano, tíos y demás familiares.

Ubuntu. Hoy soy porque somos, porque fueron.

Agradecimientos

Desde que emprendí este camino hacia el aprendizaje sobre cómo hacer ciencia, nunca imaginé por todo lo que pasaría. La meta siempre estuvo clara, y más allá de mover la frontera del conocimiento, como se suele indicarnos en el inicio de este camino, está el orgullo de haber sido el primero en mi familia que emprendió este camino de hacer ciencia, de investigar.

El trabajo recopilado en esta memoria doctoral es un resumen de los años que duró esta aventura científica. Percibir nuevas experiencias, nuevas formas de ver la vida, de hacer ciencia, de ver el mundo hicieron parte de este trabajo. No fue un camino fácil, pero sí enriquecedor y lleno de mucho aprendizaje y muchas aventuras que poco quedarán reflejadas en este documento. Hacer ciencia requiere de entereza, dedicación, pasión por descubrir y hacer visible lo invisible, tener el criterio y conocimiento para planificar, ejecutar y analizar lo ideado sin tenerle miedo al fracaso.

No obstante, en este camino recorrido, y entendiendo que es el momento indicado para agradecer, quiero dar las gracias a mis directores, Flavio Prieto, Jose Blasco y Nuria Aleixos. A Nuria, por su guía, comprensión y carisma durante todo este proceso. A Flavio, por su guía, paciencia e invaluable dedicación durante todos estos años. A Jose, por su sabiduría, comprensión y transmisión del conocimiento y del quehacer científico. Hacer ciencia implica rodearse de personas que te aporten en tu quehacer y en tu vida. Y no tengo duda de haberme rodeado con los mejores, mis directores.

A la universidad nacional de Colombia, mi alma mater, gracias por ser mi segunda casa. Al Instituto Valenciano de Investigaciones Agrarias (IVIA) y al Centro de Agroingeniería del IVIA en Moncada (España) por abrirme las puertas y permitir la realización de mi investigación. A Carlos Ruiz, Sergio, Enrique, Andres, Diego, Amanda por su valiosa ayuda y amistad en Valencia. A Brayan, William, Felipe, Kevin, Andrés y tantos de mi querida “nacho” que aportaron con su conocimiento, amistad y aliento a culminar este estudio. A todos mis amigos y seres queridos, gracias.

Finalmente, quiero agradecer y dedicar esta memoria a los verdaderos autores de este trabajo, mi familia. Agradezco a mi madre, mi motor y orgullo; a mi padre, por su luz; a mi hermano, a mis tíos, a mi sobrina, a mis mascotas y a todos mis familiares por su paciencia, amor y cariño conmigo durante todos estos años. Toda gloria y todo el orgullo de lo logrado por mi persona es por y para ustedes.

Carlos Alberto Velásquez Hernández
22 de agosto de 2024

Our senses enable us to perceive only a minute
portion of the outside world.

Nikola Tesla

Resumen

Una de las enfermedades más relevantes en mango es la antracnosis, causada principalmente por *Colletotrichum sp.* La relevancia de la antracnosis se debe a su alta incidencia en los cultivos de mango, causando alrededor del 60% de las pérdidas de mango en todo el mundo, aunque, en regiones con periodos de alta pluviosidad y humedad, la incidencia ha llegado a alcanzar el 100% de la producción de fruta de mango. El control actual de la antracnosis en mango se realiza de forma visual, lo que significa que la detección se realiza en una fase tardía (visible) y está sujeta a la pericia, entrenamiento y agudeza visual de un experto humano. Con el auge de las técnicas de aprendizaje automático, los sistemas de visión hiperespectral han encontrado el complemento para el desarrollo de aplicaciones robustas de detección y clasificación de objetos. Esto ha desencadenado una revolución tecnológica en la inspección y el control de calidad de la fruta, impulsando la investigación y el desarrollo de nuevas soluciones que aborden retos latentes en la agroindustria como la detección temprana de enfermedades en frutas. Así pues, el objetivo de esta investigación ha consistido en desarrollar una metodología para la detección de antracnosis en mango (frutos y hojas) en una fase temprana utilizando imágenes hiperespectrales Vis-NIR y técnicas de aprendizaje automático. Además, esta investigación ha evaluado el potencial de los datos espectrales para caracterizar los síntomas de antracnosis en mango con el fin de establecer un conjunto de longitudes de onda relevantes que representen los principales cambios espectrales inducidos por la enfermedad. Esta investigación también ha abarcado la selección de muestras representativas de frutos y hojas de mango de 3 variedades, la inoculación controlada de *Colletotrichum sp.* y la adquisición de imágenes hiperespectrales en condiciones de laboratorio. Además, ha definido un marco para implementar diferentes filtros de preprocesado espectral y modelos de aprendizaje automático para clasificar y detectar muestras enfermas de antracnosis. También se han implementado varias técnicas de selección de características para seleccionar un conjunto representativo de longitudes de onda (10) que representaran las diferentes etapas de desarrollo de los síntomas de la enfermedad. Como resultado, se obtuvieron modelos de clasificación, basados en redes neuronales y análisis discriminante cuadrático, con métricas superiores al 95%, mostrando ser eficientes en la detección de muestras enfermas de antracnosis 48-72 horas antes de que los síntomas se hicieran distinguibles. Además, se ha encontrado que existen varias regiones espectrales en el rango Vis-NIR relacionadas con los síntomas de antracnosis, permitiendo su caracterización desde los estadios tempranos (no visibles) hasta los tardíos. Los hallazgos reportados en esta investigación han evidenciado el enorme potencial de las imágenes hiperespectrales y los modelos de aprendizaje automático para la detección temprana de la antracnosis en mango con espectro completo o reducido, pudiendo servir de base para el desarrollo de novedosos sistemas de detección en frutos de mango con sistemas de visión multiespectral.

Abstract

One of the most relevant diseases in mango is anthracnose, mainly caused by *Colletotrichum sp.* The relevance of anthracnose is due to its high incidence in mango crops, causing about 60% of mango losses worldwide, although, in regions with periods of high rainfall and humidity, the incidence has reached 100% of mango fruit production. The current control of anthracnose in mango is done visually, which means that detection is done at a late (visible) stage and is subject to the expertise, training and visual acuity of a human expert. With the rise of machine learning techniques, hyperspectral vision systems have found the complement for developing robust object detection and classification applications. This has triggered a technological revolution in fruit inspection and quality control, driving the research and development of new solutions that address latent challenges in agro-industry such as early detection of fruit diseases. Therefore, this research has aimed to develop a methodology for detecting anthracnose in mango (fruits and leaves) at an early stage using Vis-NIR hyperspectral imaging and machine learning techniques. Furthermore, this research has assessed the potential of spectral data to characterise anthracnose symptoms in mango to establish a set of relevant wavelengths representing the main spectral changes induced by the disease. This research has also defined a methodology that involved selecting representative samples of mango fruits and leaves of 3 varieties, the controlled inoculation of *Colletotrichum sp.* and the acquisition of hyperspectral images under laboratory conditions. Additionally, a framework was defined to implement different spectral pre-processing filters and machine learning models for classifying and detecting anthracnose-diseased samples. Also, several feature selection techniques to select a representative set of wavelengths (10) representing the different stages of disease symptom development were implemented. As a result, classification models based on neural networks and quadratic discriminant analysis were obtained with metrics above 95% and were shown to be efficient in detecting anthracnose-diseased samples 48-72 hours before visible symptoms became distinguishable. Furthermore, it was found that there are several spectral regions in the Vis-NIR range related to anthracnose symptoms, allowing their characterisation from early (non-visible) to late stages. The findings reported in this research have evidenced the enormous potential of hyperspectral imaging and machine learning models for early detection of anthracnose disease in mango with full or reduced spectrum and could serve as a basis for the development of novel detection systems in mango fruits with multispectral vision systems.

Resum

Una de les malalties més rellevants del mango és l' antracnosi, principalment causada per *Colletotrichum* sp. La importància de l'antracnosi és deguda a l'alta incidència en el cultiu de mangos, que causa al voltant del 60% de pèrdues a nivell mundial, tot i que en regions amb períodes d'alta pluviositat i humitat la incidència pot arribar al 100% de la producció de fruita. Actualment, el control de l'antracnosi en mango es fa visualment, la qual cosa implica que la detecció es produeix en una etapa (visual) tardana i que està condicionada a l'experiència, entrenament i agudeses visual de la persona experta. Amb l'auge de les tècniques d'aprenentatge automàtic, els sistemes de visió hiperespectral han trobat el complement per al desenvolupament d'aplicacions robustes de detecció i classificació d'objectes. Aquesta circumstància ha desencadenat una revolució tecnològica en la inspecció de fruita i el control de qualitat, i ha impulsat la recerca i el desenvolupament de noves solucions que adrecen reptes latents en la indústria agroalimentària com la detecció primerenca de malalties de la fruita. Així doncs, aquesta recerca ha consistit a desenrotllar una metodologia per a la detecció d'antracnosi en mango (fruits i fulles) en una etapa primerenca emprant imatge hiperespectral Vis-NIR i tècniques d'aprenentatge automàtic. A més, aquesta recerca ha avaluat el potencial de les dades espectrals per a caracteritzar els símptomes d'antracnosi en mango per establir un conjunt rellevant de longituds d'ona que representen els principals canvis a nivell espectral introduïts per la malaltia. Aquesta recerca ha definit una metodologia que ha inclòs la selecció de mostres representatives de fruits i fulles de mango de 3 varietats, la inoculació controlada de *Colletotrichum* i l'adquisició d'imatges hiperespectrals en condicions de laboratori. Addicionalment, s'ha definit un marc de treball per a implementar diferents filtres de preprocessament espectral i models d'aprenentatge automàtic per a la detecció i classificació de mostres amb la malaltia d'antracnosi. També s'han implementat diverses tècniques de selecció de característiques per tal de triar un conjunt representatiu de longituds d'ona (10) que representen les diferents fases del desenvolupament dels símptomes de la malaltia. Com a resultat, s'han obtingut models de classificació basats en xarxes neuronals i anàlisi discriminant quadràtic amb mètriques per damunt del 95% i han mostrat ser eficients en detectar mostres amb antracnosi 48-72 hores abans que els símptomes visibles esdevingueren distingibles. A més, es va trobar que hi ha diverses regions espectrals en el rang Vis-NIR relacionats amb símptomes d'antracnosi, cosa que permet la caracterització des d'etapes primerenques (no visibles) fins a tardanes. Els resultats d'aquesta investigació evidencien l'enorme potencial de la imatge hiperespectral i els models d'aprenentatge automàtic per a la detecció primerenca de l'antracnosi en mango amb tot o part de l'espectre i pot servir de base per al desenvolupament de nous sistemes de detecció en fruit de mango amb sistemes de visió multiespectral.

Content

Agradecimientos	vii
Resumen	xi
Abstract	xii
Resum	xiii
1. Introduction	1
1.1. General introduction	2
1.1.1. Mango crop and production	2
1.1.2. Anthracnose in mango	5
1.2. Justification	7
1.3. Identification and problem statement	10
1.4. Objectives	11
1.4.1. General objective:	11
1.4.2. Specific objectives:	11
1.5. Contributions	11
1.6. Thesis structure	13
2. State of the art	15
2.1. Machine learning models applied to spectral analysis in agriculture	16
2.2. Non-destructive inspection and evaluation of fruits	18
2.3. Non-destructive inspection and evaluation of mango crop	20
2.3.1. Non-destructive estimation of mango fruit properties	20
2.3.2. Beneficial compounds estimation from mango fruits and leaves	23
2.3.3. Other studies related to non-destructive inspection in mango	24
2.4. Anthracnose disease	25
2.4.1. Anthracnose detection in fruit crops	26
2.4.2. Anthracnose detection on mangoes	27
3. Data collection and preprocessing	31
3.1. Gathering representative mango samples to study the anthracnose symptom development in laboratory	31
3.1.1. Fruit selection from 3 mango cultivars	31

3.1.2.	Leaf selection from 3 mango cultivars	33
3.1.3.	Inoculation of <i>Colletotrichum gloeosporioides</i> in a representative group of mango samples	35
3.1.4.	Spectral acquisition system	37
3.1.5.	Hyperspectral imaging conditions	40
3.1.6.	Spectral data pre-processing	41
3.1.6.1.	Background removal filter	42
3.1.6.2.	Spectral smoothing filter	42
3.1.6.3.	Outlier filtering (box plot)	43
3.1.6.4.	Data standardisation	43
3.1.6.5.	ROI selection and data labelling	44
3.2.	Flowchart for hyperspectral image pre-processing	45
4.	Design and train machine learning models	47
4.1.	Machine learning models used in this research	47
4.1.1.	Naive Bayes	48
4.1.2.	Models based on discriminant analysis	48
4.1.3.	Decision tree models	49
4.1.4.	CART	50
4.1.5.	Random Forest	50
4.1.6.	Boosting and variants	51
4.1.7.	Support vector machine	53
4.1.8.	Artificial neural networks	53
4.2.	Hyperparameter optimisation	55
4.3.	Performance metrics applied to ML models	56
4.4.	Flowchart for training and validation of classification models	58
5.	Identification of important wavelengths	59
5.1.	Feature selection techniques applied to spectral characterisation of anthrac- nose in mango	59
5.1.1.	Principal component analysis for feature selection (PCA4FS)	60
5.1.2.	Random forest for feature selection (RF4FS)	61
5.1.3.	Single-layer perceptron for feature selection (SLP4FS)	62
5.1.4.	Genetic algorithm for feature selection (GA4FS)	64
5.1.5.	Forward stepwise selection based on DA	65
5.2.	Flowchart for selection of relevant wavelengths and validation with classifica- tion models	68
6.	Results	69
6.1.	Data collection and preprocessing	69
6.1.1.	Descriptive and visual analysis of the mango samples	69

6.1.2.	Filtered hyperspectral image bank	73
6.1.3.	Geometric and reflectance correction of hyperspectral images	74
6.1.4.	ROI selection and projection	78
6.1.5.	Spectral data pre-processing	82
6.2.	Design and train machine learning models	86
6.2.1.	Training and test dataset definition	86
6.2.2.	Experimental conditions before model implementation and assessment	87
6.2.3.	Best fit dataset for AI-based models in mango fruit assays for anthracnose symptom detection	88
6.2.4.	Biclass models to discriminate the anthracnose diseased classes in mango fruits	91
6.2.5.	Multiclass models for detection of anthracnose symptoms in mango fruit	97
6.2.6.	Best fit dataset for AI-based models in mango leaf assays for anthracnose symptom detection	101
6.2.7.	Biclass models to discriminate the anthracnose diseased classes in mango leaves	104
6.2.8.	Multiclass models for detection of anthracnose symptoms in mango leaves	110
6.2.9.	Validation of early detection models of anthracnose in mango samples at surface level	116
6.3.	Identification of relevant wavelengths	122
6.3.1.	Feature selectors applied to the spectral characterisation of anthracnose on mango fruit	123
6.3.2.	Multiclass models for anthracnose detection on mango fruits	127
6.3.3.	Relevant wavelength selection for spectral characterisation of anthracnose on mango fruit	130
6.3.4.	Feature selectors applied to the spectral characterisation of foliar anthracnose on mango leaves	133
6.3.5.	Multiclass models for foliar anthracnose detection on mango	137
6.3.6.	Relevant wavelength selection for spectral characterisation of anthracnose on mango leaves	142
7.	Discussion	145
8.	Conclusion	157
8.1.	Conclusions	157
8.2.	Planned work	159
A.	Appendix: Results of mango fruit assays	160
A.1.	Biclass results for the mango fruit assays	161
A.1.1.	Performance metrics results per dataset	161

A.1.2. Performance metrics results by classification model	165
A.2. Multiclass results for the mango fruit assays	169
A.3. Multiclass results for the mango fruit assays with selected bands	170
B. Appendix: Results of mango leaf assays	171
B.1. Biclass results for the mango leaf assays	172
B.1.1. Performance metrics results per dataset	172
B.1.2. Performance metrics results by classification model	177
B.2. Multiclass results for the mango leaf assays	182
B.3. Multiclass results for the mango fruit assays with selected bands	183
Bibliography	184

List of Figures

1-1. Production in tonnes of mango-producing countries worldwide in 2023. Taken from FAO [50]	3
1-2. Mango fruits of the 'Kent' (a), 'Keitt' (b) and 'Osteen' (c) cultivars harvested for consumption based on the physiological maturity (size and colour) of each cultivar	4
1-3. Global mango production in tonnes relative to the total cultivated area (ha) during 2013 - 2023. Taken from [50]	4
1-4. Mango production in Colombia in tonnes relative to the total cultivated area (ha) from 2007 - 2022. Taken from [7]	5
1-5. Symptoms of anthracnose in (a) fruits and (b) mango leaves. (c) Acervuli of <i>Colletotrichum sp.</i> on mango leaf (salmon-coloured pimples on lesion) and (d) <i>Colletotrichum sp.</i> colony under microscope (blue tubular structures) . .	6
1-6. Anthracnose cycle in mango. Taken from Uddin et al. [158]	8
3-1. A representative group of selected mango fruits of the (a) 'Keitt' and (b) 'Kent' cultivars for the experimental development of this research	32
3-2. Mango fruit samples of 'Kent', 'Osteen' and 'Keitt' cultivars over 6 days, showing the changes in pigmentation as the fruit ripens when harvested . . .	33
3-3. A representative group of selected mango leaves of the 'Kent' (a) and 'Keitt' (b) cultivars for the experimental phase in this research	34
3-4. Hyperspectral acquisition system (description in section 3.1.4) used for spectral mango imaging	38
3-5. Images acquired of the same samples (fruit and leaf) showing the matching points (linked by a coloured line) that allowed merging both images into one	39
3-6. Motorised lifting assembly for the positioning of mango samples inside the hyperspectral image acquisition dome	40
3-7. Motorised lifting assembly for the positioning of mango samples inside the hyperspectral image acquisition dome	41
3-8. Representation of the motorised lifting assembly for the positioning of mango samples (fruits and leaves) inside the hyperspectral image acquisition dome .	42

3-9. (a) ‘Keitt’ mango fruit sample and (b) ‘Tommy Atkins’ mango leaf sample with selected ROIs (cyan colour). The internal regions in the samples, coloured in blue, represent the induced lesions with punch and surgical needle, for fruit and leaf respectively	44
3-10. Flowchart for pre-processing hyperspectral images acquired on mango fruits and leaves in the laboratory	46
4-1. Representation of a single-layer perceptron neural network	54
4-2. Flowchart for training classification models of spectral data collected in the mango sample essays (fruits and leaves) throughout this research	58
5-1. a) Percentage of variance explained for the first 10 PCs, where the first 6 PCs explain 99 % of the variance of data under analysis; and b) Weight functions of the six selected PCs, where the absolute peaks of each function establish the contribution of each wavelength to the composition of PCs	60
5-2. Schematic representation of an SLP with two inputs, two neurons and one output. Taken from [27]	64
5-3. Flowchart for selecting a representative set of wavelengths for characterising anthracnose symptoms in mango	68
6-1. Mango fruits from the ‘Keitt’, ‘Kent’ and ‘Osteen’ cultivars of the control (sound) group over 9 days of monitoring and acquisition of hyperspectral images	71
6-2. Mango leaves from the control group of ‘Keitt’, ‘Kent’ and ‘Tommy Atkins’ cultivars over 9 days of monitoring and hyperspectral imaging	71
6-3. Mango fruit of ‘Keitt’, ‘Kent’ and ‘Osteen’ cultivars from the diseased group over 9 days of monitoring and hyperspectral image acquisition	72
6-4. Mango leaves from the diseased group of ‘Keitt’, ‘Kent’ and ‘Tommy Atkins’ cultivars over 9 days of monitoring and hyperspectral imaging	72
6-5. Hyperspectral image of a ‘Keitt’ mango fruit without correction (a) vs. reflectance corrected image (b) in the 760 nm band	74
6-6. Hyperspectral image of a ‘Tommy Atkins’ mango leaf without correction (a) vs. reflectance corrected image (b) in the 760 nm band	75
6-7. Matching algorithm and RANSAC algorithm for the computation of the transformation matrix between hyperspectral images of two sample fruits (one control and one diseased) of the <i>cv.</i> ‘Keitt’	75
6-8. Matching algorithm and RANSAC algorithm for the computation of the transformation matrix between hyperspectral images of two sample leaves (one control and one diseased) of the <i>cv.</i> ‘Tommy Atkins’	76

6-9. Hyperspectral image resulting from a ‘Kent’ mango fruit after geometric and reflectance corrections. Image (a) shows the 760 nm band, while image (b) shows an RGB reconstruction of the hyperspectral image at 680, 550 and 480 nm	76
6-10. Hyperspectral image resulting from a ‘Tommy Atkins’ mango leaf after geometric and reflectance corrections. Image (a) shows the 760 nm band, while image (b) shows an RGB reconstruction of the hyperspectral image at 680, 550 and 480 nm	77
6-11. ROI selection in two mango fruits of the <i>cv.</i> ‘Keitt’ (control and inoculated) on day 6 of acquisition. The ROIs correspond to the areas delimited in cyan colour, while the blue ones the regions removed and the yellow asterisks the points selected for perspective correction	79
6-12. ROI selection in two mango leaves (control and inoculated) of the <i>cv.</i> ‘Tommy Atkins’ on day 7. The ROIs correspond to the areas delimited in cyan colour, while the blue ones the regions removed and the yellow asterisks the points selected for perspective correction	79
6-13. ROI projection on samples (control and inoculated) of ‘Keitt’ mango fruit during all days of acquisition. The projection for this assay was made on the 6 lesions caused by a surgical punch from day 6 (last day of acquisition) to days 2-5	80
6-14. ROI projection on samples (control and inoculated) of ‘Tommy Atkins’ mango leaf during all days of acquisition. The projection for this experiment was made on the 5 surgical needle lesions from day 7 (last day of acquisition) to days 3-6	80
6-15. Group of spectral signatures of a control sample (control) and a diseased sample (days 2 and 6) of the ‘Keitt’ mango fruit experiment (mean spectra in red)	82
6-16. Group of spectral signatures of a control sample (Control) and a diseased sample (days 2 and 7) of the ‘Tommy Atkins’ mango leaf experiment (mean spectra in red)	83
6-17. Comparison of the mean spectral signature of each defined class (Raw), the implemented pre-processing filters (the Savitzky-Golay and the outlier filter) and the SNV method for the ‘Keitt’ fruit mango data	85
6-18. Comparison of the mean spectral signature of each defined class (Raw), the implemented pre-processing filters (the Savitzky-Golay and the outlier filter) and the SNV method for the ‘Tommy Atkins’ leaf mango data	85
6-19. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the <i>cv.</i> ‘Keitt’ (e1-kt assay)	89
6-20. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the <i>cv.</i> ‘Keitt’ (e4-kt assay)	89

6-21. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the e2-kn assay (<i>cv.</i> ‘Kent’)	90
6-22. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the e3-os assay (<i>cv.</i> ‘Osteen’)	91
6-23. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the <i>cv.</i> ‘Keitt’ (e1-kt assay) and the Fil+noSG dataset	92
6-24. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e4-kt assay (<i>cv.</i> ‘Keitt’) and the Fil+noSG dataset	92
6-25. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e2-kn test of the ‘Kent’ mango cultivar and the Fil+noSG dataset	94
6-26. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e3-os assay (<i>cv.</i> ‘Osteen’) and the Fil+noSG dataset	95
6-27. Confusion matrices of the MLP and SLP classification models for the e2-kn test	99
6-28. Confusion matrices of the SVM and QDA classification models for the e3-os test	101
6-29. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the ‘Tommy Atkins’ e5-to experiment . .	102
6-30. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the ‘Tommy Atkins’ e6-to experiment . .	102
6-31. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the <i>cv.</i> ‘Keitt’ e7-kt experiment	103
6-32. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the <i>cv.</i> ‘Keitt’ e8-kt experiment	103
6-33. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the <i>cv.</i> ‘Kent’ e9-kn experiment	104
6-34. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the ‘Tommy Atkins’ e5-to experiment and the Fil+noSG dataset	105
6-35. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the ‘Tommy Atkins’ e6-to experiment and the Fil+noSG dataset	106
6-36. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e7-kt ‘Keitt’ assay and the Fil+noSG dataset	107

6-37. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e8-kt ‘Keitt’ assay and the Fil+noSG dataset	107
6-38. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e9-kn experiment of the <i>cv.</i> ‘Kent’ and the Fil+noSG dataset	108
6-39. Confusion matrices of the QDA and SVM classification models for the e5-to and e6-to assays	112
6-40. Confusion matrices of the QDA and SVM classification models for the e7-kt and e8-kt experiments	114
6-41. Confusion matrices of the SVM and SLP classification models for the e9-kn assay	116
6-42. Model prediction result of the control samples for the a) ‘Keitt’, b) ‘Kent’, c) ‘Osteen’ and d) ‘Keitt’ fruit assays	118
6-43. Model prediction result of the control samples for the ‘Keitt’ (a and b), ‘Tommy Atkins’ (c and d) and ‘Kent’ (d) leaf assays	119
6-44. Model prediction result of the inoculated samples for the a) ‘Keitt’ and b) ‘Kent’ fruit assays	120
6-45. Model prediction result of the inoculated samples for the a) ‘Osteen’ and b) ‘Keitt’ fruit assays	120
6-46. Model prediction result of the inoculated samples for the ‘Tommy Atkins’ and ‘Kent’ leaf assays: a) a5-to, b) a6-to and c) a9-kn	121
6-47. Model prediction result of the inoculated samples for the ‘Keitt’ leaf assays: a) a7-kt and b) a8-kt	122
6-48. Group of wavelengths selected by the feature selectors, per defined dataset, for the e1-kt assay of the ‘Keitt’ mango fruit cultivar	124
6-49. Group of wavelengths selected by the feature selectors, per defined dataset, for the e4-kt assay of the ‘Keitt’ mango fruit cultivar	125
6-50. Groups of wavelengths selected by the feature selectors, per defined dataset, for the e2-kn assay of the ‘Kent’ mango fruit cultivar	126
6-51. Group of wavelengths selected by the feature selectors, per defined dataset, for the e3-os assay of the ‘Osteen’ mango fruit cultivar	127
6-52. Group of wavelengths selected by the feature selectors, per defined dataset, for the ‘Tommy Atkins’ e5-to assay on mango leaves	134
6-53. Group of wavelengths selected by the feature selectors, per defined dataset, for the ‘Tommy Atkins’ e6-to assay on mango leaves	135
6-54. Group of wavelengths selected by the feature selectors, per defined dataset, for the e7-kt ‘Keitt’ mango leaf assay	136
6-55. Group of wavelengths selected by the feature selectors, per defined dataset, for the e8-kt ‘Keitt’ mango leaf assay	137

6-56. Group of wavelengths selected by the feature selectors, per defined dataset, for the e9-kn test of the ‘Kent’ mango leaf cultivar	138
7-1. Temporal evolution (from day 2 to day 4) of ‘Osteen’ mango fruits in colour image (RGB) and the 3 most relevant wavelengths (770, 790 and 800 nm) . .	146
7-2. Temporal evolution (from day 2 to day 4) of ‘Keitt’ mango fruits in colour image (RGB) and the 3 most relevant wavelengths (450, 760 and 950 nm) . .	147
7-3. Temporal evolution (from day 2 to day 4) of ‘Tommy Atkins’ mango leaves in colour image (RGB) and the 3 most relevant wavelengths (950, 460 and 470 nm)	148
7-4. Temporal evolution (from day 2 to day 4) of ‘Kent’ mango leaves in colour image (RGB) and the 3 most relevant wavelengths (950, 720 and 460 nm) . .	148
7-5. Mean spectra of a) ‘Osteen’ and b) ‘Keitt’ mango classes and the 10 most relevant wavelengths for the anthracnose spectral characterisation	151
7-6. Mean spectra of a) ‘Kent’ and b) ‘Tommy Atkins’ mango classes and the 10 most relevant wavelengths for the foliar anthracnose spectral characterisation	151
7-7. Mean spectral signature of the classes defined for mango fruit assays with a) mechanical damage (taken from [162]); b) <i>Colletotrichum gloeosporioides</i> inoculation on the ‘Keitt’ cultivar; and c) <i>Colletotrichum gloeosporioides</i> inoculation on the ‘Kent’ cultivar	153
7-8. Mean spectral signature of the control and inoculated classes per day of acquisition for the ‘Keitt’ (a), ‘Kent’ (b) and ‘Osteen’ (c) assays with mango fruits	154
A-1. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e1-kt assay)	161
A-2. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Kent variety (e2-kn assay)	162
A-3. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Osteen variety (e3-os assay)	163
A-4. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e4-kt assay)	164
A-5. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e1-kt assay)	165

A-6. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Kent variety (e2-kn assay)	166
A-7. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Osteen variety (e3-os assay)	167
A-8. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e4-kt assay)	168
B-1. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Tommy Atkins variety (e5-to assay)	172
B-2. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Tommy Atkins variety (e6-to assay)	173
B-3. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e7-kt assay)	174
B-4. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e8-kt assay)	175
B-5. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Kent variety (e9-kn assay)	176
B-6. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Tommy Atkins variety (e5-to assay)	177
B-7. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Tommy Atkins variety (e6-to assay)	178
B-8. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e7-kt assay)	179
B-9. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e8-kt assay)	180
B-10. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Kent variety (e9-kn assay)	181

List of Tables

3-1. Mango fruit sets for the experimental development of this research work . . .	33
3-2. Mango leaf sets for the experimental phase of this research	35
3-3. Results associated with pre-assays with mango samples and inoculation strategies for <i>Colletotrichum gloeosporioides</i>	36
6-1. Filtered sample group of mango assays reported in Tables 3-1 and 3-2 after the experiments under laboratory conditions. DPM: dried plant material (senescence), NS: no symptoms	73
6-2. Total hyperspectral images acquired and corrected per assay, cultivar and acquisition days	77
6-3. Number of spectral signatures per class obtained from the ROIs selected in each assay and mango cultivar	81
6-4. Spectral signatures obtained using the outlier filter (W/O) versus the raw data	84
6-5. Area under the ROC curve (AUC) of the QDA, SVM and SLP models for the e1-kt, e2-kn, e3-os and e4-kt fruit assays, and by day of study	96
6-6. Performance metrics of the classification models for the <i>cv.</i> ‘Keitt’ e1-kt assay	97
6-7. Performance metrics of the classification models for the <i>cv.</i> ‘Keitt’ e4-kt assay	98
6-8. Performance metrics of the <i>cv.</i> ‘Kent’ e2-kn assay detection models	99
6-9. Performance metrics of the ‘Osteen’ e3-os assay	100
6-10. Area under the ROC curve (AUC) of the QDA and SLP models for the e5-to, e6-to, e7-kt, e8-kt and e9-kn leaf assays, and by day of study	109
6-11. Performance metrics of the classification models for the ‘Tommy Atkins’ e5-to experiment	111
6-12. Performance metrics of the classification models for the ‘Tommy Atkins’ e6-to experiment	111
6-13. Performance metrics of the classification models for the <i>cv.</i> ‘Keitt’ e7-kt assay	113
6-14. Performance metrics of the <i>cv.</i> ‘Keitt’ e8-kt assay classification models . . .	113
6-15. Performance metrics of the <i>cv.</i> ‘Kent’ e9-kn assay classification models . . .	115
6-16. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e1-kt assay of the <i>cv.</i> ‘Keitt’ on mango fruit	128
6-17. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e2-kn assay of the <i>cv.</i> ‘Kent’ on mango fruit	128

6-18. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e3-os assay of the <i>cv.</i> ‘Osteen’ on mango fruit	129
6-19. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e4-kt assay of the <i>cv.</i> ‘Keitt’ on mango fruit	129
6-20. Sets of the 10 most representative wavelengths by best-performing feature selector in each assay conducted for the three mango fruit cultivars	130
6-21. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e5-to <i>cv.</i> ‘Tommy Atkins’ experiment on mango leaves	138
6-22. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e6-to <i>cv.</i> ‘Tommy Atkins’ experiment on mango leaves	139
6-23. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e7-kt assay of the <i>cv.</i> ‘Keitt’ on mango leaves	140
6-24. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e8-kt assay of the <i>cv.</i> ‘Keitt’ on mango leaves	140
6-25. Performance metrics of the top 10 classification models with the best fitting feature selector and dataset for the e9-kn test of the ‘Kent’ mango leaf cultivar	141
6-26. Sets of the 10 most representative wavelengths by best-performing feature selector in each assay conducted for the 3 mango leaf cultivars	142
A-1. Performance metrics of the dataset classification models for the Keitt variety e1-kt assay.	169
A-2. Performance metrics of the dataset classification models for the Keitt variety e4-kt assay.	169
A-3. Performance metrics of the detection models for the Kent variety e2-kn assay.	169
A-4. Performance metrics of the classification models for the Osteen e3-os assay. .	170
B-1. Performance metrics of the classification models for the Tommy Atkins e5-to experiment.	182
B-2. Performance metrics of the classification models for the Tommy Atkins e6-to experiment.	182
B-3. Performance metrics of the classification models for the Keitt variety e7-kt assay.	182
B-4. Performance metrics of the Keitt variety e8-kt assay classification models. . .	183
B-5. Performance metrics of the Kent variety e9-kn assay classification models. . .	183

List of Algorithms

1.	PCA for feature selection algorithm	61
2.	Random forest for feature selection algorithm	62
3.	Single-layer perceptron for feature selection algorithm	64
4.	Genetic algorithm for feature selection	66
5.	Forward stepwise feature selection algorithm	67

List of Abbreviations

Abbreviation	Meaning
AI	Artificial intelligence
ANN	Artificial neural network
CART	Classification and regression tree
CV	Computer vision
CVS	Computer vision system
<i>cv.</i>	cultivar
DL	Deep learning
DM	Dry matter
F	Firmness
GA	Genetic algorithm
GBM	Gradient boosting machine
HSI	Hyperspectral images
KNN	k-nearest neighbours
LDA	Linear discriminant analysis
LCTF	Liquid crystal tunable filter
ML	Machine learning
MLP	Multi-layer perceptron
NIR	Near-infrared
PA	Precision agriculture
PCA	Principal component analysis
QDA	Quadratic discriminant analysis
RF	Random Forest
RGB	Red-Green-Blue
RPI	Ripening index
SLP	Single-layer perceptron
SSC	Soluble solid content
sLDA	stepwise Linear discriminant analysis
sQDA	stepwise Quadratic discriminant analysis
SVM	Support vector machine
TA	Titrateable acidity
TSS	Total soluble solids

Abbreviation	Meaning
UV	Ultraviolet
UAV	Unmanned aerial vehicle
VI	Vegetation index
Vis	Visible
XGB	eXtreme gradient boosting

1. Introduction

Using non-invasive tools in inspecting and monitoring fruit crops in pre-harvest and postharvest helps producers improve their management and risk control practices by providing information to diagnose problems affecting the harvest and fruit marketing. One pathogen of great interest due to the large damages caused in agricultural products is *Colletotrichum sp.* [93, 99, 164], responsible for anthracnose, a severe disease that causes large losses in the harvest and postharvest of fruits such as [98, 167], mango [19, 120], olive [48], tea [175], and avocado [141], among others [33, 82, 168]. Mango is a fruit appreciated worldwide [47, 52] but susceptible to anthracnose, which hinders their exportation as this disease can spread to healthy fruits during storage and transit. Hence, there is a large interest in developing methods to control this damaging disease, including early detection, using non-invasive solutions such as those provided by machine vision and artificial intelligence systems.

The current method for anthracnose diagnosis in mango is visual inspection, which implies that the detection is only possible at advanced disease stages. Moreover, it is subject to the human expert's quality, training and visual acuity. Therefore, the infection goes undetected during quality controls of the harvested fruits, developing the spores that can spread the disease without control to other fruits throughout the storage, transportation, and marketing processes.

That is why there is an urgent need to develop non-invasive diagnostic tools for producers which allow the detection of anthracnose in its early stages and reduce the incidence of the human factor in its detection. For this purpose, the use of sensors that capture specific information from the electromagnetic spectrum, such as cameras sensitive to ultraviolet (UV), visible (Vis) and near-infrared (NIR), emerges as a potential solution for the early detection of the disease. Hyperspectral cameras, capable of capturing spatial information in regions of the spectrum such as Vis and NIR (Vis-NIR), have been used in recent years to develop automated non-destructive fruit inspection systems [60, 136], the correlation of abnormal temperature rises with specific diseases [67, 177] and early detection of diseases on crops [166, 169]. Particularly, studies such as [169] demonstrated that detecting diseases in previsual stages is possible by setting up protocols for acquiring, calibrating and analysing spectral data.

Therefore, this research addresses the study and use of hyperspectral images (HSI) collected

on three (3) mango cultivars under laboratory conditions to detect the symptoms associated with anthracnose in both mango fruits and leaves before they are visible. The study focuses on three cultivars on mango and was performed using controlled inoculations of *Colletotrichum gloeosporioides* in selected samples to guarantee 1) the development of lesions compatible with anthracnose; 2) an incidence greater than 90 % of symptomatic samples; and 3) the repeatability of the procedures applied to allow the spectral study of the disease over time. Additionally, exploratory analysis and preprocessing of the spectral data were carried out to correct the hyperspectral images, obtain relative reflectances, detect outliers, and improve the quality of the spectral signatures extracted from the selected regions of interest in the hyperspectral images.

A main goal was to test the ability of machine learning (ML) techniques to correlate the spectral information in the HSI with the relevant anthracnose symptoms in mango. Feature selection techniques were used to obtain a group of relevant wavelengths to characterise the anthracnose symptoms in mango samples spectrally. Both the design of the ML models and the selection of relevant wavelengths were assessed through a descriptive, statistical and physiological analysis that allowed to determine the best model for detecting symptoms associated with anthracnose in mango.

Results showed that detecting anthracnose symptoms in the early stage, in fruits and leaves, in the three cultivars of mango studied was feasible. The disease could be detected between 48 and 72 hours after the pathogen was inoculated and between 24 and 48 hours before the symptoms were visible. Additionally, a group of relevant wavelengths was found, allowing the spectral characterisation and detection of disease symptoms in both leaves (10 wavelengths) and fruits (10 wavelengths).

1.1. General introduction

1.1.1. Mango crop and production

Mango, *Mangifera Indica L.*, also known as the “king of fruits”, is a perennial crop with a tropical and, recently introduced, subtropical climate, native to India but widely cultivated in Asia, Latin America, northern and southern Africa and Europe (Spain and Italy), see Figure 1-1. It is commercially grown at 0 to 1,600 meters above sea level. It has been determined that for every 120 m of height or 1 degree of temperature variation in northern or southern latitude, relative to the equator, there is a delay in flowering of 4 days [39], allowing countries with different thermal floors or territorial extension over the hemispheres to have extended mango production cycles per year.

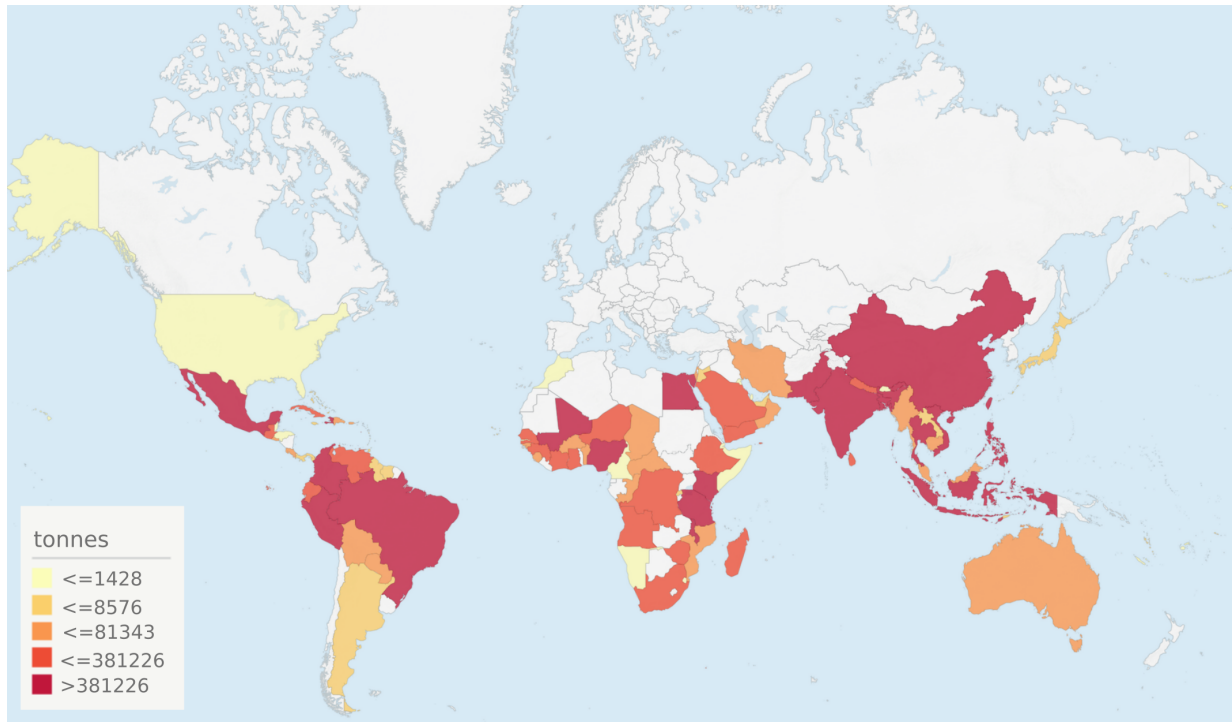


Figure 1-1. Production in tonnes of mango-producing countries worldwide in 2023. Taken from FAO [50]

The mango crop develops and produces in a range of 24 °C to 32 °C, although excessive temperatures cause a decrease in production because of a higher flower drop at the blossom stage. The ideal rainfall for this crop ranges between 1,500 – 2,000 millilitres throughout the year [39]. Above this range, the tree has greater foliage and lower production, while the fruit-filling stage below this threshold (drought) can reduce its production and quality.

Once the fruits are set, they develop quickly, reaching a commercial size at 7 weeks (depending on the cultivar) and being ripened after 12 - 13 weeks, the time at which the fruit can be harvested and consumed for a higher nutritional contribution (Figure 1-2). The mango fruit is a source of vitamins A, B and C, although it is poor in calcium, phosphorus and iron [139]. The content of macro- and micronutrients largely depends on the cultivar, of which more than 1,600 have been listed worldwide [25]. Despite this important number of cultivars, only 7 have commercial importance due to their nutritional contribution and high demand in international markets: *cv.* ‘Kent’, *cv.* ‘Keitt’, *cv.* ‘Tommy Atkins’, *cv.* ‘Haden’, *cv.* ‘Ataulfo’, *cv.* ‘Osteen’ and *cv.* ‘Palmer’ [25,113].

Mango is one of the most consumed tropical fruits in predominantly local markets [52]. In the last decade, the cultivated area has increased by 12.46 %, reaching more than 6 million ha in 2022 with approximately 59.1 million tonnes, see Figure 1-3. In Colombia, mango is the 3rd fruit crop with the highest production after citrus and avocado, reaching an annual production of 346 thousand tonnes in a harvested area of 33.6 thousand ha, see Figure 1-4. The

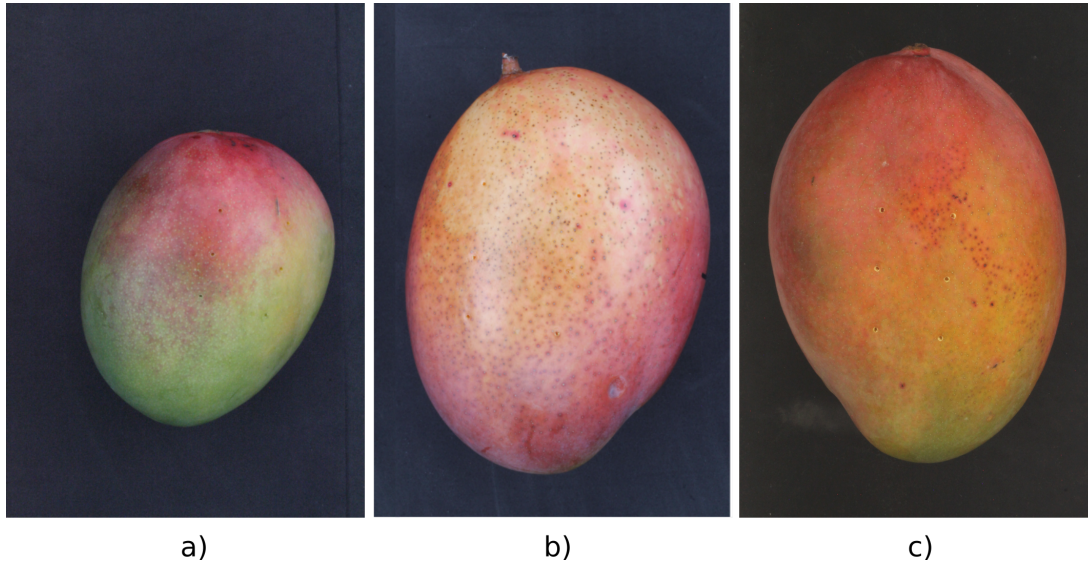


Figure 1-2. Mango fruits of the 'Kent' (a), 'Keitt' (b) and 'Osteen' (c) cultivars harvested for consumption based on the physiological maturity (size and colour) of each cultivar

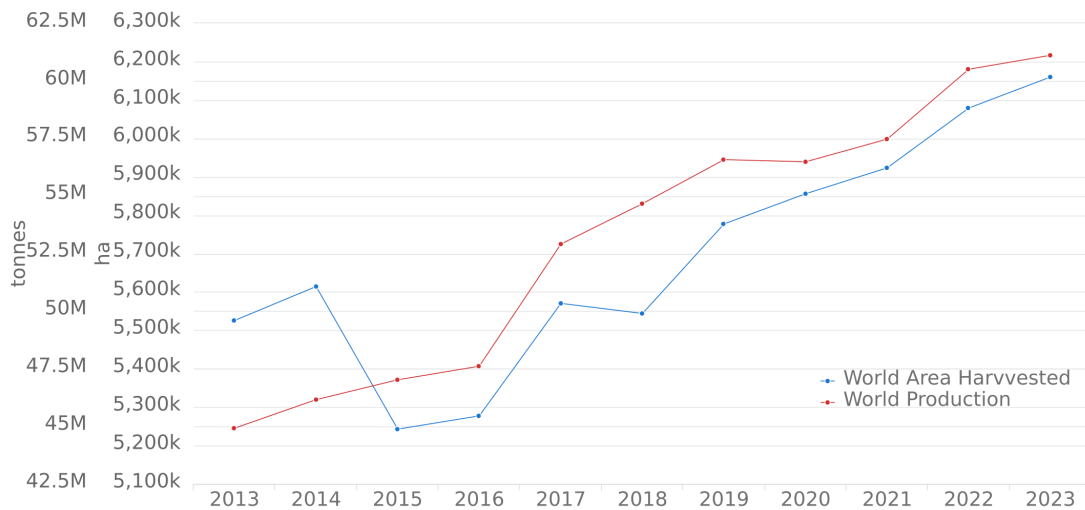


Figure 1-3. Global mango production in tonnes relative to the total cultivated area (ha) during 2013 - 2023. Taken from [50]

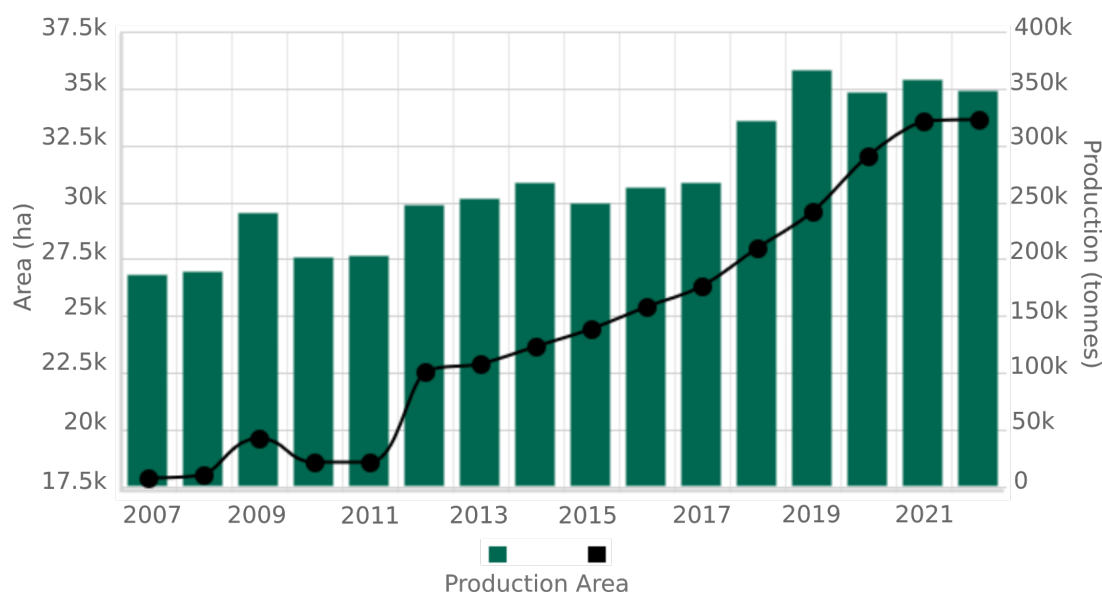


Figure 1-4. Mango production in Colombia in tonnes relative to the total cultivated area (ha) from 2007 - 2022. Taken from [7]

main export markets for Colombia are concentrated in North America (USA and Canada) and Europe (Netherlands and Germany). On the other hand, in Spain, mango cultivation has experienced a large expansion in recent years due to the climate conditions of some regions (mainly Andalusia), its geographical proximity to the European market and the export potential to supply the demand of countries such as France, Germany and Portugal. By 2022, mango production in Spain reached 40,000 tonnes [35] in an area greater than 4,800 ha [160].

Despite having a growing international market, which demands more and better fruits, its export value has been limited to 3.89 % of the world production in 2022 (approximately 2.3 million tonnes) [49]. This is attributed mainly to the high incidence of diseases and pests, followed by climate changes, production costs, insufficient supply chains and the lack of technological assistance to support the production. This situation is evident in the yield (ton/ha) both at world level (Figure 1-3) and in Colombia (Figure 1-4), where it can be observed that the yield at world level went from 9.91 to 9.84 in the period 2019-2022, while in Colombia this decrease, for the same period, went from 12.38 to 10.31. Therefore, to achieve sustainable and efficient production, research and development in mango is mandatory, given the interest in improving the crop's yield and quality and increasing its export market.

1.1.2. Anthracnose in mango

The incidence of this disease in tropical and subtropical regions is one of the main reasons for the export market of mango to be reduced to small production lots, since anthracnose is easy

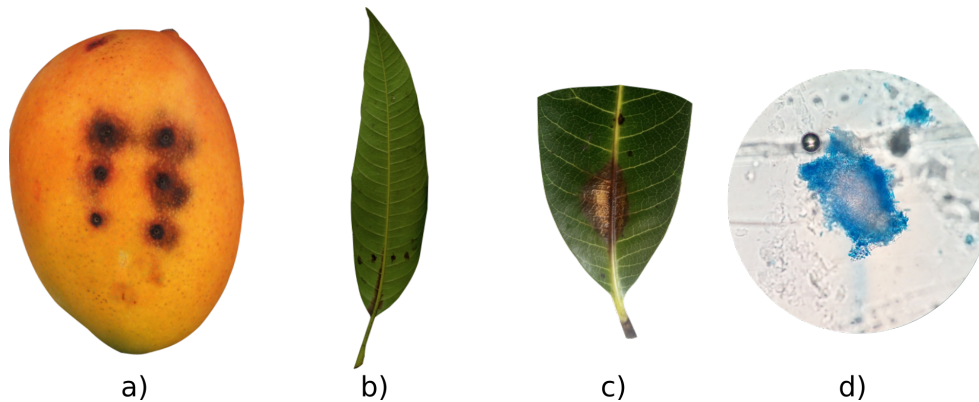


Figure 1-5. Symptoms of anthracnose in (a) fruits and (b) mango leaves. (c) Acervuli of *Colletotrichum* sp. on mango leaf (salmon-coloured pimples on lesion) and (d) *Colletotrichum* sp. colony under microscope (blue tubular structures)

to spread and difficult to detect and control. In Southeast Asia, this disease is the leading cause of the non-export of mango worldwide, and in producing countries such as India, it has been reported that 90 % of annual mango production is affected by anthracnose [74]. In Costa Rica, where periods of high rainfall and humidity have been reported, the incidence has reached 100 % of the mango fruit production [158]. In other producing countries such as Thailand, its incidence has been estimated at 62.8 % [40], in China around 30-60 % [90], in Indonesia 50 % of mango losses [83], in Bangladesh around 25-30 % [158], in Colombia has up to 40 % of the production [53, 54] and globally around 60 % of mango losses [34, 158].

In mango, anthracnose affects leaves, twigs and fruits [40]. It can also be found in the blossoms [158]. The most common species of *Colletotrichum* sp. are *Colletotrichum gloeosporioides* and *Colletotrichum Acutanum* [93, 105, 164], although more than 37 species have been genetically identified and associated with anthracnose disease in mango [24]. Common symptoms show up as oval or irregular brown to deep brown sunken point-sized spots of varying sizes spread all over the leaf surface, most prominently in young leaves but also older leaves [40]. As the leaves age, these spots become larger lesions that are either circular or irregular and surrounded by a red halo [128]. In fruits, this disease produces sunken necrotic spot lesions in the mango peel that coalesce until they cover the whole surface, see Figure 1-5. These visible symptoms affect the visual aspect of fruits, causing rejection by consumers and, therefore, limiting their commercialisation to local markets.

Figure 1-6 shows the anthracnose infection cycle in mangoes. Firstly, the dissemination of *Colletotrichum* sp. conidia reaches a host (tree) through rain or irrigation of water to the crop. Subsequently, pathogenic spores establish themselves in infection areas such as leaves, branches, and panicles. After spore germination, the appressorium penetrates the cuticle and epidermis of immature fruits or young tissue. The infection starts after penetrating

these structures and remains quiescent (latent infection) until the climacteric fruits begin to ripen. This means the pathogen remains latent until the host conditions allow its development. After the beginning of fruit ripening, characteristic anthracnose lesions appear on mangoes that quickly expand and join with other anthracnose lesions. On these visible lesions, conidia masses appear on the symptomatic tissue, especially during rainy or humid weather. Orange or salmon-coloured pimples characterise these structures, see Figure 1-5.

Currently, anthracnose control methods are based on applying fungicides on the crop throughout the blossom stage to prevent infection of the growing fruit. This application method is often extended to postharvest, combined with periods of fruit immersion in hot water, expensive cold storage chains and visual inspection controls. Likewise, anthracnose detection methods are focused on visual inspection by experts, which implies that the detection is carried out at the visible (or late) stage. This method is demanding in terms of time, costs, and labour (it can cause visual fatigue in the human eye) and is subject to the particular expertise of the workers, as latent infections are difficult to detect in the early stages. This disease has aroused the scientific community's interest, given the challenges involved in its correct detection and control in postharvest. Studies reported in the literature have been focused on (1) determining its incidence and severity in different crops [46, 126, 156], (2) identifying the factors that stimulate its growth [115], (3) developing new treatments to prevent/control the infection [146, 170] or (4) identifying early symptoms associated with this disease [30, 48].

As reference works on anthracnose in mango, Uddin et al. [158] and Dofour et al. [40] presented a review of the state-of-the-art treatments and trends in anthracnose in mango. The reproductive cycle of this pathogen is also detailed in depth, and recommendations are given for the promotion of new research areas in this field.

1.2. Justification

Currently, the productivity and quality of the mango crop are affected by the lack of technology for disease control in the farming sector. Current mango inspection methods are carried out visually, resulting in late detection of crop problems and relying on an expert's visual acuity and training. This is primarily due to the limited access to and availability of practical producer assistance tools in the field. In addition, pressure from international markets, such as the USA, the EU and China, to improve the quality of the fruit has led to a stagnation of mango exports: in 2022, only 3.89 % of global mango production was destined for export [50]. The growth potential of this market is high. However, improving fruit control and production practices remains imperative. Colombia, a privileged tropical country producing mangoes 10 out of the 12 months of the year, and Spain, a European country that can supply the European mango market, can benefit economically from this

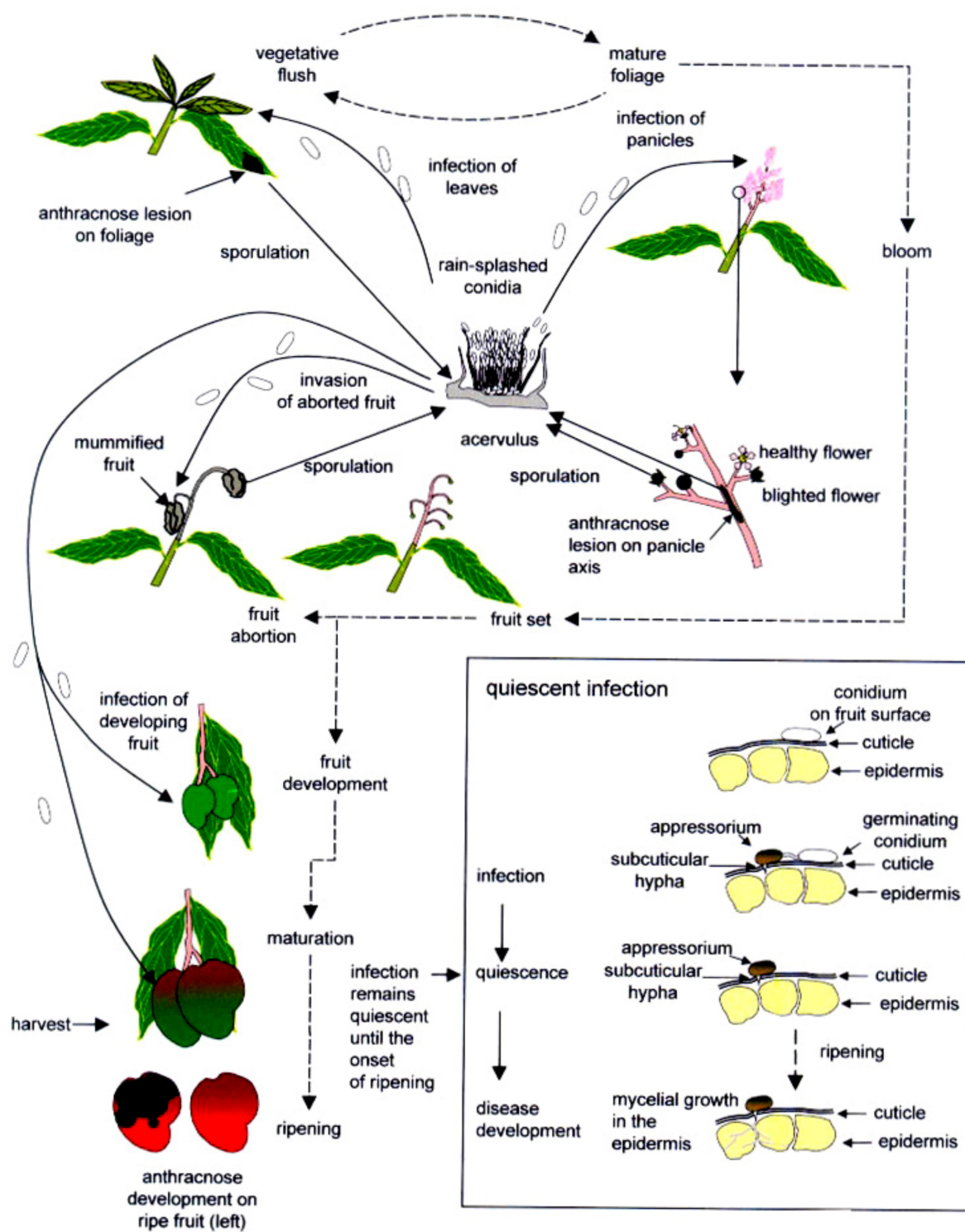


Figure 1-6. Anthracnose cycle in mango. Taken from Uddin et al. [158]

agricultural product if their productive capacity is enhanced.

Therefore, there is a need to develop tools for monitoring and assessing mango crops at the pre-harvest and postharvest stages that help to improve their production technologically. One of the phytosanitary problems affecting mango crops throughout their production (pre- and postharvest stages), and possibly one of the reasons for the stagnation of mango exports, is anthracnose. This disease is among those of most significant interest and impact in mango, as it affects the crop in pre-harvest and postharvest and is responsible, according to estimates presented in [34, 158], for 60 % of fruit production losses worldwide.

In the literature, few studies have focused on anthracnose detection in mangoes. Studies such as those presented in [100, 155] proposed a model for fruit disease detection when the first symptoms became visible. A study associated with the fruit resistance mechanisms to anthracnose is presented in [73]. Recently, studies have evaluated the potential of hyperspectral imaging and machine learning-based detection models to detect anthracnose symptoms on *cv.* ‘Nam Dok Mai’ and sugar mango fruit were conducted in [149, 150] and [30, 132], respectively. All these studies contributed to the understanding and characterisation of anthracnose symptoms in mango fruits from spectral analysis and have revealed the feasibility of using ML-based detection models to assist in inspecting and controlling this disease from the early stages of development. As recommendations, the authors of these studies also reported the need to extend research in this field to other mango cultivars and foliar anthracnose detection. In addition, implementing and comparing different ML techniques to obtain the most robust and efficient detection model for this phytosanitary problem is also a latent need.

Likewise, no studies related to the feature selection to spectrally represent the different stages of anthracnose symptom development in mango have been reported in the literature. Feature selection is important in spectral analysis and disease detection as it removes non-relevant information for the detection models and enhances the model training/learning in the spectral ranges of higher contribution. This approach reduces the number of wavelengths needed to detect those symptoms and, hence, the computational cost of data processing.

Consequently, developing this research is vital to advancing the development of robust and effective methods for the early diagnosis of anthracnose in mango. Furthermore, it is essential to develop detection models applied to phytosanitary control, as they could help farmers monitor and prevent the pathogen outbreak in mango production. This thesis advances to determine non-destructive methods for early anthracnose detection in mango using hyperspectral image analysis, ML model training and feature selection techniques.

1.3. Identification and problem statement

Given the advantages of non-destructive inspection systems such as spectroscopy and hyperspectral imaging, spectral data is a possible solution to detecting early (pre-symptomatic) anthracnose symptoms. Studies such as those reported in [30, 149] showed the potential of this technology and AI-based models to help with post-harvest inspection and control of this disease. In [30], for example, there was evidence of spectral regions that could be useful for spectrally characterising anthracnose symptoms, while in [149, 150] there was proof that hyperspectral imaging could contribute to the development of non-destructive spectral vision solutions. These works also demonstrated the effectiveness of developing experiments with mango fruits under controlled conditions, given the challenges in monitoring and temporal assessment of anthracnose-associated symptoms from the initial inoculation stage.

Undoubtedly, these reference works contributed to the understanding of the symptomatological development of the disease for its later detection, but the limitation to a few AI models and the non-application of spectral-level feature selectors to correlate this information with the physiology of the infection remain latent challenges. Hence, to provide a deep and proper response to this need, further research in this field is still required as the development of affordable spectral vision systems, given the current costs of hyperspectral technology, continues to be a technological barrier for the agro-industrial sector.

Similarly, a study concerning detecting early symptoms of anthracnose in leaves (an organ of great relevance for disseminating the pathogen) has not been reported to date in the state-of-the-art. Only studies on detecting leaf anthracnose at the visible (symptomatic) stage have been reported [74, 126]. Hence, further research in this area continues to be needed to move the frontier of knowledge that leads to the development of future solutions that help the diagnosis of anthracnose in the field, i.e. in preharvest stages.

In addition, when the scientific databases Science Direct and Google Scholar were consulted, with the formula "*anthracnose detection AND hyperspectral imaging AND mango*", it became even more evident that there is a lack of related studies that contribute to providing a solution to the problem of assisting producers in the early detection of anthracnose in mango. While in the Science Direct database, only three articles are directly related to the mentioned formula, the number of articles in Google Scholar increased to seven. From the articles reported in Science Direct, one corresponded to a publication based on the results of this research. In contrast, in Google Scholar, three publications come from the results of this research (see section 1.5 - Contributions). The rest of the studies found [30, 132, 149, 150] were referenced in the state-of-the-art in section 2.4.2.

Based on the above exposition, the study developed in this doctoral thesis revolved around a

question concerning the use of hyperspectral information and machine learning-based models for the early detection of anthracnose in mango. The research question posed for this thesis was:

- What are the correlations between spectral information and anthracnose-associated symptomatology in mango?
- Is this correlation(s) strong enough to detect anthracnose at the pre-symptomatic stage?

1.4. Objectives

1.4.1. General objective:

Develop predictive models for early detection of anthracnose in fruit and leaves of three mango varieties using hyperspectral imaging and machine learning techniques, and obtaining a reduced set of important wavelengths.

1.4.2. Specific objectives:

1. **Data collection and preparation:** Gather representative samples of fruits and leaves from the three mango cultivars under study from infected and healthy plants. Preprocess hyperspectral data to remove noise and normalise it for subsequent analysis.
2. **Design and train machine learning models:** Develop predictive models using machine learning techniques such as supervised and unsupervised classification to detect the presence of asymptomatic anthracnose in mango samples.
3. **Evaluation of model accuracy:** Assess the accuracy and robustness of the developed models. Compare the performance of different models and select those with the best performance.
4. **Identification of important wavelengths:** Employ feature selection techniques to identify a reduced set of wavelengths most relevant for anthracnose detection in mangoes.

1.5. Contributions

With the study carried out in this doctoral thesis, several contributions were achieved and reported to each of the objectives, covering a variety of products ranging from student training (undergraduate and master students), published scientific articles, posters, and attendance

at international conferences to the funding of research projects.

For the sake of clarity, the main products generated by this doctoral thesis are listed below:

Published articles

- Velásquez, C., Aleixos, N., Gomez-Sanchis, J., Cubero, S., Prieto, F., J. Blasco, 2024. Enhancing anthracnose detection in mango at early stages using hyperspectral imaging and machine learning. *Postharvest Biology and Technology*, vol. 209, p. 112732. DOI: 10.1016/j.postharvbio.2023.112732.
- Velásquez, C., Prieto, F., Palou, L., Cubero, S., Blasco, J., Aleixos, N. 2023. New model for the automatic detection of anthracnose in mango fruits based on vis/nir hyperspectral imaging and discriminant analysis. *Journal of Food Measurement and Characterization*, pp. 1-11, 2023. DOI: 10.1007/s11694-023-02173-3.
- Rincón-Fonseca, R. I., Velásquez-Hernández, C. A., Prieto-Ortiz, F. A. 2021. Spectral noise reduction in hyperspectral images using the discrete wavelet transform. *Journal of Research, Development and Innovation*, vol. 11(3), pp. 601-616. DOI: 10.19053/20278306.v11.n3.2021.13359.

Attendance at congresses

- Velásquez, C. (2022, March 30 - April 1). New spectral model for anthracnose detection in mango fruits based on hyperspectral analysis and discriminant analysis techniques [Virtual]. Second Latin American Congress on Precision Agriculture (CLAP 2022), Manfredi, Córdoba, Argentina.
- Espinel, O. F., Velasquez, C. A., Prieto, F. A. (2021, 15-17 September). Detection of anthracnose in mango leaves by hyperspectral image analysis. 2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1-6, Popayán Colombia. DOI: 10.1109/STSIVA53688.2021.9592005.

Posters

- Velásquez, C. (2022, June 29 - July 1). Detection of anthracnose in mango cv ‘Keitt’ using hyperspectral imaging and artificial neural networks. FRUTIC 14th International Symposium, Valencia, Spain.
- Velásquez, C. (2022, June 29 - July 1). New model for the automatic detection of anthracnose in mango fruits based on vis/nir hyperspectral imaging and discriminant analysis. FRUTIC 14th International Symposium, Valencia, Spain.

R&D projects

- Velasco, A., Velasquez, C., Prieto, F., Argüello, H. 2021. Development of an optical-computational system to estimate the COS of agricultural soils through spectral images and artificial intelligence in citrus crops in Santander [Proyecto inversión R&D, sistema general de regalías - Colombia].
- Velasquez, C., Prieto, F. 2020. Field detection of anthracnose at the foliar level in mango trees using spectral image analysis and ML techniques [R&D project, Universidad Nacional de Colombia].
- Velasquez, C., Lizarazo, I., Blasco, J., Aleixos, N., Prieto, F. 2020. Early in situ detection of leaf-level diseases using spectral imaging. Case studies: anthracnose in mango and verticillium in potato [R&D project, Universidad Nacional de Colombia].

Student training

- Montenegro, A., Prieto, F. (Tutor). 2021. Detection and classification of anthracnose in mango using hyperspectral imaging and deep learning techniques [Master thesis, Universidad Nacional de Colombia].
- Garzón, L., Prieto, F. (Tutor). 2022. Detection of anthracnose in mango leaves using spectral image analysis and machine learning techniques [Tesis pregrado, universidad nacional de Colombia].
- Peraza, D., Prieto, F. (Tutor). 2021. Implementation of deep learning models for the analysis of hyperspectral signatures of agricultural products [Tesis pregrado, universidad nacional de Colombia].

1.6. Thesis structure

This doctoral thesis is divided into eight chapters. The first chapter presents an introduction, the research problem addressed, the objectives developed and the contributions achieved by the research development. Chapter 2 exposes the state-of-the-art related to the research work to point out the current state-of-the-art in the disease detection field with spectral information, emphasising the studies related to anthracnose in mango. Chapters 3, 4 and 5 develop each of the objectives, showing the materials, methods, and methodology used in this research; while Chapter 6 presents the results obtained and related to these chapters. Likewise, Chapter 7 presents a discussion on the central problem addressed in this doctoral thesis, giving a frame of reference with the related state-of-the-art studies and answering the research question with the results obtained from this research. Chapter 8 presents the main conclusions of the development of this study, as well as future work and trends related to the results obtained to extend further the scope of the findings and results achieved in this

research.

At the end of this manuscript, the bibliography referenced throughout this doctoral thesis is presented, and the appendices previously mentioned throughout the eight chapters of this doctoral thesis are attached.

2. State of the art

Agriculture is a field of great interest for researchers, given the low level of technology associated with the production processes and selection of agricultural products, and the impact of this activity on climate change and the world population. Fields of knowledge such as computer vision (CV) [10, 57, 62, 130], robotics [16, 28, 42] and artificial intelligence (AI) [13, 125] are some of the areas, transversal to agriculture, that have served as tools for the development and invention of new processes that improve the quality of agricultural products and reduce the impact of this activity on the environment.

With the emergence and use of modern and more specialised sensors, such as spectral sensors, considerable progress has been made in the non-destructive inspection and evaluation of crops and fruits. Such sensors can be grouped depending on the dimensionality of the spectral data they collect (spectrometers) or the spatial-spectral data they acquire (hyperspectral cameras -this term encompasses multispectral cameras-, monochromatic or thermal cameras). The spectrum these spectral sensors can acquire ranges from the UV, the Vis, and the NIR to the mid-infrared (MIR). The combination and use of these sensors has allowed, for example, rapid estimations of fruit properties without destroying them [43, 110, 136], the detection of temperature increases associated with specific disease symptoms [67, 177] and the early detection of crop diseases [31, 114, 144, 169]. Reference studies on the use of hyperspectral images (HSI) are presented in [6, 75], the concepts related to hyperspectral cameras, types of sensors and their differences, their use in unmanned aerial vehicles (UAV), HSI processing techniques, software available to date and reference machine learning (ML) models are referenced in these works. Additionally, trends and applications in various fields such as agriculture, mining and forestry are presented.

Within agriculture, a field of interest for researchers and producers, given its impact on the production supply chain, is the inspection and control of diseases. This field has recently gained interest as phytosanitary control of products is time-consuming, is mainly done by human visual inspection, restricts the marketing of agricultural products locally and directly affects the productivity of the crop if proper in-field phytopathological control is not applied. Specifically, mango crop production has been affected by the lack of technological development supporting disease diagnosis and control. Consequently, mango efforts have focused on the early detection of diseases with a higher incidence in production and greater severity in the fruit. The disease with the greatest impact on mango, also reported in other tropical

fruit crops, is anthracnose, caused mainly by the pathogen *Colletotrichum gloeosporioides*. Its impact on mango production is such that it is estimated to be responsible for the loss of up to 60 % of worldwide mango production [34, 158] and has even been reported that up to 100 % of the production has been affected in countries like Costa Rica [158].

Given the breadth of ML techniques and HSI analysis applied to disease detection and quality control in crops, studies in mango have mainly focused on assessing the internal quality of the fruit and estimating physico-chemical properties of the fruit, relying on chemometric techniques. Likewise, work related to detecting relevant diseases in the pre- and post-harvest stages has been reported. Therefore, to provide a comprehensive state-of-the-art related to mango crops and anthracnose, the following sections detail the work associated to non-destructive fruit inspection, emphasising mango and the techniques of ML, HSI and chemometrics. Then, the state-of-the-art concerning detecting crop diseases by HSI techniques is presented, with the studies related to anthracnose and those focused on mango the most interesting ones.

2.1. Machine learning models applied to spectral analysis in agriculture

Machine learning, an AI area, has been of interest in spectral analysis and agriculture due to the ability of these models to process large volumes of data and make predictions from new data or observations that were not used in the model training. This model training is performed iteratively and recursively, which implies that at each iteration, the model assesses its performance and redirects its learning towards valuable datasets while discarding observations considered as outliers. In this way, an ML model learns to abstract the most relevant features or patterns to characterise the training data and to predict the behaviour or response of new data whose outcome (the variable that typically is being estimated or predicted) is unknown.

The relevance of ML models lies in the enormous variety of approaches proposed and implemented in spectral analysis. Some machine learning models are artificial neural networks (ANN), decision trees (DT), support vector machines (SVM), Bayesian models, and genetic algorithms (GA). In agriculture, these models have been used in crop disease detection, as evidenced in [77, 124, 169] where they extracted features from RGB images for detection, segmentation and classification of field scenes used for aerial crop monitoring and tracking [21, 64, 76, 143], or crop reconstruction with RGB-D (red, green, blue - depth) data to estimate plant foliage and volume through temporal changes [41, 58]. Other applications include the design of crop yield estimation models [26, 29, 163], fruit counting on the tree [131], pesticide bioactivity estimation on maize leaves [112], and fruit damage detection [32, 123, 173],

among other applications. The variety of applications and studies reported in the literature demonstrates the usefulness and versatility of ML models for developing solutions to assist in agricultural production activities.

Nevertheless, the use of information in the visible spectrum constrains the potential of these models, which can process large amounts of data, as it restricts their predictive capacity in applications such as the estimation of internal properties in fruits or the early detection of phytosanitary problems. For this reason, ML models have been used in spectral analysis to propose new approaches in the development of disruptive technological solutions: detecting early disease symptoms in strawberry [69], mango [150], olive [48], rice [172], walnut [15], cotton [9], among others; estimating internal quality parameters such as firmness (F), titratable acidity (TA), total soluble solids (TSS), dry matter (DM) [104, 110, 135, 153], among others; and predicting nutritional status [95, 116] in crops are just among the applications reported in the literature that demonstrate the potential of ML and spectral analysis.

Recent work in this area demonstrates the usefulness and validity of these techniques for applied research in agriculture. For example, in [174], the detection of chilling injuries in banana using a multiple linear regression model (MLR) and a flexible in-situ Vis-NIR sensing system is studied; in [91], the potential of Vis-NIR hyperspectral imaging for early detection of bruised lesions in peaches using a partial least squares-discriminant analysis (PLS-DA) classification model is assessed. Regarding fruit quality estimation in [88], the soluble solids content (SSC) in Tributi citru, an important citrus fruit in China, using hyperspectral imaging (in the ranges 400-1000 nm and 900-1700 nm), the partial least squares regression (PLSR) and support vector regression (SVR) models and the successive projections algorithm (SPA) and correlation coefficient analysis (CCA) selectors are studied. Similar to this work, [96] evaluates the potential of ML models (XGBoost, LightGBM, CatBoost, and Random Forest -RF-) to estimate SSC and TA in peach, apricot, and cherry. These recent studies showed the continued usefulness and relevance of ML models for non-destructive solutions for the estimation/detection of quality parameters or common fruit damage, for example. Although considered standard AI techniques in other areas, these methods are still considered relevant in spectral analysis, inspection, and monitoring of crops.

ML, however, has not been the only AI approach successfully implemented in spectral analysis. Deep learning (DL), an ML variant, has recently emerged as an area of interest because of the rise of more powerful computational systems and the properties that these techniques have brought to the estimation and classification tasks of large volumes of data. Instead of traditional ML techniques, DL can extract high-level features (e.g., evaluating surface textures), process data in parallel, and transfer learning from one model to another to enhance and obtain generalisable and robust learning models [85]. Thus, DL techniques have shown promising results in feature extraction and detection of diseases, faces and vehi-

cles [22, 48, 72, 85], as they allow unsupervised learning of robust features compared to ad-hoc features defined in traditional methods. Among the most widely used structures are convolutional neural networks (CNNs), feed-forward neural networks (FFNs), radial-base function networks (RBFNs), recurrent neural networks (RNNs), or autoencoders, among others [72].

Due to the growth in computing capabilities, it is possible to design deeper networks with larger parameters, improving detection and classification results. This kind of technique has recently been used in agriculture. For example, in [97, 106] they used a single-layer perceptron (SLP) to detect vegetation diseases, while in [108] a multi-layer model to detect yellow rust in the wheat crop is presented; and in [84] a CNN to detect arctic vegetation was implemented and described. In [32], 3 CNNs were implemented and compared to detect surface bruises at an invisible (early) stage in plums using the Vis-NIR hyperspectral data and wavelength selection approach; while in [48] a study related to anthracnose disease detection in olives was carried out using Vis-NIR hyperspectral images acquired in laboratory conditions and the PCA compression strategy to adapt hyperspectral cubes to standard DL-input data. Likewise, in [1], the authors used an RBFN to detect avocado laurel wilt disease in the early and late stages of infection. These reference studies showed remarkable performance in combining spectral data with DL models and envisaged the potential of AI solutions in the agroindustry, as more reliable and robust DL models continue to be presented and published in the literature. As DL is an active and recent line of research in spectral analysis, in [12] an overview of the challenges to be addressed in this area is exposed, an introduction to wavelength selection in hyperspectral data with DL techniques is discussed in [176], while in [79] a comparison between ML and DL methods and an introduction to the challenges in disease detection in agriculture with DL is done.

Finally, given the breadth and variety of techniques and applications reported in the literature and related to ML and DL, in [20, 81, 138, 154] compiled papers are presented as a framework for ML and DL techniques to deepen their use in agriculture employing spectral data: [138] focused on ML techniques, [20, 81] on DL models, and [154] on wavelength selection methods. Similarly, in sections 2.2 and 2.4 of this document, reference works are presented in the areas of fruit quality parameter estimation and disease detection, respectively, with the studies related to mango crops and anthracnose disease being the ones with more focus and exposition.

2.2. Non-destructive inspection and evaluation of fruits

The relevance of non-destructive fruit inspection and quality control techniques lies in the potential use of sensors, such as spectral cameras, to make rapid and reliable estimates of internal or external parameters of the fruit. In agriculture, numerous studies demonstrate the viability and potential use of these sensors. In [17], a review of non-destructive inspection

techniques on thick-skinned fruits (including hyperspectral techniques) is presented. This work exposes the problem inherent to the low precision achieved in the models regarding thin-skinned fruits: the inability of current sensors to penetrate the fruit peel without affecting it, and make reliable estimates. However, the authors also advocate a more extensive analysis to determine new correlations or indices that allow the inspection of this kind of fruit in a non-destructive manner.

On the other hand, in [63] a review related to the determination of fruit maturity and ripening indices is exposed, the authors also established some of the advantages of the use of HSI over spectroscopy systems (point-to-point systems): 1) HSI provide spatial resolution combined with spectral information; 2) HSI allow the entire surface of the fruit to be inspected with a single data capture, which is a fundamental advantage, as fruits do not have a homogeneous composition and surface.

Likewise, some studies related to predicting and classifying fruits according to their maturity and ripening are [5, 111, 171]. In [111], a study is conducted with persimmon to classify ripeness, detect the level of astringency and predict the F of the fruit. Using a PLSR model, the authors achieved an accuracy of 97.9 % in the classification of the fruit according to its ripeness. Meanwhile, in [5], a study to classify banana ripening after harvesting is carried out. As a result, the authors determined high correlations between the banana ripening index and specific wavelengths (532, 660 and 785 nm). Based on a PLSR model, the classification model achieved an accuracy of 97.53 %. On the other hand, in [171], a work related to the prediction of the ripening state of strawberries is presented. Through a hyperspectral image analysis and predictive models, the authors determined the most representative wavelengths for the prediction of fruit ripening and, through a model based on an SVM, they achieved an accuracy higher than 85 % in predicting the ripening state of strawberries.

Nevertheless, not only have maturity and ripening indices been established with hyperspectral techniques. In the literature, studies focused on the detection of other properties of fruits have been reported. This is the case of melon, where in [153] an evaluation of the hardness and sweetness indices was made based on the information obtained from hyperspectral images. Likewise, studies on the spectral estimation of micro and macronutrients in Persimmon leaves *cv.* Rojo Brillante [3, 4] and citrus [2] were carried out. As relevant findings, the authors estimated with a PLSR model nutrients such as nitrogen and phosphorus in persimmon, and potassium and boron in citrus fruits.

All these works reported in the literature have demonstrated the feasibility and potential of using spectral technology to estimate or discriminate internal and external properties of fruits of interest for human consumption. Consequently, to validate the usefulness of this technology in this research, a review of studies related to the mango crop, both in fruits and

leaves, was carried out.

2.3. Non-destructive inspection and evaluation of mango crop

Mango, being a product of great international demand due to its nutritional quality and organoleptic properties of great value for the human diet [71, 103], has generated great interest in the scientific community that seeks to improve the processes of planting and production in the field and to guarantee the quality of the fruit in postharvest. The demand for fruit in good condition, with good flavour and visual appearance, is one of the characteristics demanded by the international market and which rewards the entire production chain with great economic benefits. This is why the non-destructive study of this fruit has been developed in the last decade using spectral technology and AI-based estimation models.

For this purpose, the rapid and robust estimation of the internal properties of the fruit and the inspection and control of diseases are seen as fields of interest in mango. In the literature, most of the studies focus on the study and quality control of the fruit, with the estimation of properties (titratable acidity, total amount of soluble solids, firmness, among others) and compounds (flavonoids, mangiferin, among others) being the areas of further development. In addition, studies related to mangoes include the study of specific compounds in leaves, given that the intake of these compounds and their use in medical treatments has been shown to have a positive impact on the treatment of diseases such as cancer, diabetes, obesity and even hepatocellular protection [80]. Also noteworthy in the literature is the recent interest in the study of diseases of high prevalence in mango crops, with anthracnose being one of the most important to the scientific community due to its direct (and widely reported) impact on the loss of mango production.

Considering the above, this section presents the most relevant studies in these two areas (estimation of properties and compounds) associated with studying mango by using AI techniques and spectral analysis (this term includes multi- and hyperspectral analysis). The idea with this state-of-the-art is to establish the potential of spectral data and AI techniques for the non-invasive estimation of properties in mango and their potential to be used in disease detection.

2.3.1. Non-destructive estimation of mango fruit properties

As mango fruits provide great benefits for the diet and human health due to the nutrients and minerals it contain, the non-invasive estimation of the main fruit properties has been a constantly evolving field. Internal properties such as TA, DM, F, and TSS, among others, are

the main parameters studied and reported in the literature, since they allow for establishing fruit quality indices and knowing the ideal ripeness. By estimating these parameters, the final consumers can have a fruit with the highest standards of nutritional and taste quality.

Reference studies on the prediction of internal properties (or quality attributes) of mango fruit to estimate fruit quality are presented in [43, 135, 147, 159], all these studies propose estimation methods, spectral ranges of higher potential or models capable of predicting one or more properties, representing an important advance in the study of these new technologies at the service of the mango industry. [159] evaluated, for example, the potential of NIR spectroscopy to estimate pH and TSS by calibrating regression models such as kernel partial least square regression (KPLSR), SVMR, and multi-predictor local polynomial regression (MLPR) and preprocessing filters. As a result, they determined the feasibility of integrating in a single model, MLPR, with different preprocessing filters the estimation of both properties: while for pH they found that the MLPR model with Gaussian smoothing filter obtained the best performance, for TSS prediction they found that the model fitted best with the Savitzky-Golay filter (SGF). A similar study on TSS prediction was carried out by [147], where in addition to studying this fruit attribute, the feasibility of its prediction in a real inspection line was evaluated. The study consisted of the evaluation of 182 *cv.* ‘Nam Dok Mai Sithong’ mangoes arranged on a conveyor and an online diffuse reflectance scanning system using a diode array short-wave Vis-NIR spectrometer. As a result of this work, the authors were able to obtain a PLSR model combined with moving average smoothing (MAS) with a correlation coefficient of the calibration set and a prediction set of 0.80 and 0.74, respectively, showing the real potential of bringing these spectral solutions to practical applications in industry.

Likewise, an extensive study is presented in [135], which assessed for 10 days the attributes F, TSS, TA, pH and β -carotene content to compute a ripening index using 188 mangoes *cv.* ‘Nam Dokmai’ sub*cv.* ‘Si Thong’ and analysing the spectral data acquired with a NIR spectrometer (800-2500 nm) and Vis-NIR HSI system (450-998 nm). The base model employed was a PLSR using original and pre-treated spectra from both devices. As results highlight, the authors found that non-significant differences in all fruit quality attributes were found between positions at the initial ripening stage, as well as that the NIR spectroscopy provided only slightly better prediction performances than the HSI system. However, they also urged that further studies employing new calibration models on other mango cultivars would be an important advance in this field.

Other studies, such as those reported by [11] and [109], showed the potential of artificial intelligence-based models for estimating internal properties in mango. [11] demonstrated that the support vector machine (SVM) obtains better results for estimating Brix degrees (indicating the sweetness of a fruit) with NIR spectroscopy than a traditional PLSR model:

the coefficient of determination (R^2) obtained from SVM for training/calibration and testing dataset were 0.96 and 0.95, respectively; while for the PLSR model were 0.89 and 0.86 respectively. This study was conducted on 80 mango samples (60 for training and 20 for testing) of the 'Chokanan', 'Rainbow', and 'Kai Te' cultivars. A similar study was reported in [109]. In this study, a comparison of PLSR, SVMR and ANN models for estimating TA in mango with NIR spectroscopy was carried out. As a novel result, it was found that the neural networks outperformed the other models, ANN>SVMR>PLSR, with the first 4 PCs (PCA was applied before calibrating the models): the R^2 of calibration (R_{cal}^2) and prediction (R_{pred}^2), the root-mean-square error of calibration (RMSEC) and prediction (RMSEP), and the ratio of prediction to deviation (RPD) were 0.97, 0.89, 25.29 mg*100g⁻¹, 28.42 mg*100g⁻¹ and 4.02, respectively.

Concerning ripening indices (RPI), the works presented by [37, 136, 161] stand out. The main objective of these studies was to establish a metric that expresses the incidence of the internal and external parameters that influence the ripening stage of the fruit and thus offer, in a fast, reliable and non-destructive way, information on fruit quality. In [102] an initial study was proposed to determine SSC, a parameter closely related to the RPI, of Irwin mango fruit using a hyperspectral imaging system. The prediction model used was an ANN. The mango SSC was predictable with a correlation coefficient (R_{cv}^2) of 0.79 and RMSEcv (root-mean-square cross-validation error) of 0.069. According to the authors, these results can be improved by considering the quality of the sensor to be used, and the optimal distance and geometry of the object, as brightness conditions could vary from one measurement to another and thus bias the data collected. In [161], a similar work was presented, but it extended the measured internal parameters (F, TSS and TA) and fruit pigmentation to integrate them into a Manila mango ripening classification model. This proposed classification model consisted of a combination of PCA and MDA (multivariate discriminant analysis) techniques. As a result, the authors obtained a classification rate of 97.44 % thanks to the combination of PCA and MDA, which allowed the reduction of the dimensionality of the spectral data and the calibration of the mathematical discriminant model.

An internal quality index (IQI) and an *cv.* 'Osteen' mango RPI, based on NIR and Vis spectroscopy, were proposed in [37]. The physical and chemical properties addressed in this work were F, flesh and peel colour, TA and TSS. As a result, the authors found that predicting RPI and IQI indices from Vis and NIR spectral analysis was possible. In addition, this work established a strong correlation between F and the other analysed mango properties. Results of this work were confirmed in [137], where a study of the F, TA and TSS properties was made with hyperspectral images. As a result, the authors found a strong relationship of F with specific wavelengths of the captured images, as well as a strong correlation of F with TA. Furthermore, the authors state that this kind of image has the potential to be used in non-destructive applications for the detection of mango properties (F, TA and ripeness).

The previously referenced works emerged with great relevance given that they established the relationship between the content of internal properties of the fruit with an RPI that indicates to both producers and consumers the current organoleptic ripeness of the fruit rapidly and reliably, without the need to destroy the fruit. Additionally, these studies stood out since the results and findings have allowed, for example, the spectral characterisation of mango fruit, reducing the number of wavelengths needed to calculate the RPI and, therefore, reducing the economic costs associated with the use of spectral technology: the establishment and validation of these indices with groups of wavelengths enables the use of multispectral sensors, which are more affordable when designing machine vision systems in agricultural production lines.

Based on the above literature review in the mentioned areas, the enormous potential of spectral sensors in non-destructive estimation and quality control of fruit is evident. Likewise, the design and development of estimation or classification models using AI techniques add value to the growth of knowledge in this field and the technological development of the agricultural sector, particularly in the mango-producing sector. However, it is also clear that there is a need to develop further AI spectral solutions for other mango cultivars, new producing countries/markets and new approaches that improve the performance of the systems proposed to date and allow their applicability in real production chains (by being affordable and reliable).

2.3.2. Beneficial compounds estimation from mango fruits and leaves

Another active line of research related to mango crops is the study and evaluation of the potential of several bioactive compounds in mango fruits and leaves [68, 71, 133]. Such compounds, as reported in the review presented in [80, 122], offer benefits for human health given the benefits of the biological activity in diseases such as diabetes, obesity, cancer, and liver damage, among others. In addition, the study of these compounds has also been complemented by the development of predictive models that provide rapid and reliable estimates of their presence (or absence) and the identification of mango cultivars with the highest potential to be used in in-depth laboratory studies for the extraction, disposition and use of these compounds in innovative products.

A comprehensive study on the nutritional composition, pigmentation and arrangement of phytochemical compounds in mango fruit is presented in [103], while similar work on mango leaves is reported in [80]. Both studies demonstrate the benefits of consuming mango fruit and the potential of extracts from mango leaves. A common conclusion and finding from these studies is that the mango crop is rich in flavonoids and phenolic acids, which are compounds of great interest to the pharmaceutical and cosmetic industry [119] given their

correlation with the anti-inflammatory, antifungal and, recently, anticancer activity.

In [178], a comprehensive survey of the techniques for estimating the content of flavonoids is presented, with spectral analysis through UV-Vis and NIR spectroscopy and hyperspectral cameras being the most outstanding. The authors demonstrated the feasibility and potential use of these sensors to determine the presence of these compounds, given that they have characteristic spectral signatures that are easy to differentiate from other phytochemical compounds. Furthermore, they stated that spectral technology-based systems had a technological advantage over standard methods as they allowed non-destructive and reliable estimations. However, they also called for further research to refine the presented estimation models and exploit the potential of current hyperspectral systems, given the spatial component these sensors can capture.

One of the main flavonoids that has recently gained interest is anthocyanin [86, 103, 151]. This pigment, responsible for the reddish pigmentation in fruits and leaves of trees, is widely present in mango crops [151]. Its relevance lies in the protection against cell damage [86] and the antifungal activity that serves as a defence mechanism of the tree against infections such as those caused by *Colletotrichum sp.*, the pathogen responsible for anthracnose in mangoes [151]. As a main result of the studies related to this pigment, an absorption peak that determines the presence and accumulation of this pigment was found in the range 520-540 nm [118]. However, neither the presence of this pigment nor the spectral response in the mentioned range has been correlated with the ripening process of the fruit, since ripening in mango fruits responds to other physiological changes and processes.

Undoubtedly, the versatility of chemometric techniques, AI and spectral analysis have provided new opportunities for science to develop new extracts beneficial to human health. The enhancement of these studies and the validation with new and improved mango cultivars is also considered a future work to be developed, as new approaches and medicines will be needed shortly to preserve the good condition (health) of human beings.

2.3.3. Other studies related to non-destructive inspection in mango

However, in mango crop production, research has not only focused on assessing fruit quality or estimating beneficial compounds. Studies are also reported in the literature evaluating the potential of spectroscopy, spectral analysis and AI models to estimate, for example, the quality of the juice produced from mango pulp [55], the internal damage in fruits such as internal browning [51] and spongy tissue [129], and the feasibility to discriminate the mango cultivar spectrally [157].

In [162], a classification model for Manila mango fruit based on the induced mechanical damage and hyperspectral image analysis was proposed. This work studied the selection of

the most relevant spectral bands (wavelengths) for detecting mechanical damage induced by a pendulum. Classification models were trained with the selection of wavelengths, and their results were compared against the full collected spectrum. As a result, the early mechanical damage detection model was able to classify fruits with an accuracy of 97.72 % when using the full spectrum and 91.14 % when using specific wavelengths, showing the potential to spectrally characterise this type of damage in mango fruit, as the metrics obtained (full versus selected wavelengths) achieved remarkable performances (above 91 % in accuracy) and did not vary considerably from each other (the difference was less than 7 %).

On the other hand, [65] presented a vegetation classification scheme with UAV over a mango crop using a hyperspectral system based on a liquid crystal tunable filter (LCTF). The authors presented an SVM classification model that achieved an accuracy of 94.5 %. Among the results and conclusions presented, the authors expressed the importance for the classifier of using the shaded areas of the mango trees, as these regions provide valuable information: 81 % was the accuracy achieved with only the illuminated areas, while 94.5 % was the accuracy achieved when the shaded areas were included. In addition to this result, the LCTF system was presented as an alternative for acquiring hyperspectral images with a UAV, despite requiring considerable acquisition time.

All the studies reported in this section undoubtedly allowed to demonstrate the enormous potential in the non-destructive estimation of internal parameters in mango fruits, in determining beneficial compounds for human health and non-invasive detection of internal damage in fruits through spectral sensors, spectral analysis and AI-based models. Hence the importance of this research in approaching the study of anthracnose detection through these spectral and AI methods as they emerge as the most promising and potential approach to detect this disease in mango.

2.4. Anthracnose disease

Anthracnose is a disease characterised by symptoms that mainly cause necrosis of the material infected by the pathogen. Its lesions are identified as circular, brownish-black, sunken, and have a superficial growth that allows them to coalesce with other anthracnose lesions or other diseases. Benefiting from the climatic conditions of tropical climates (high relative humidity and temperature), this disease affects many fruit crops such as mango, strawberry, banana, cassava, avocado, papaya, etc. Additionally, the pathogen has developed an infection mechanism that allows it to remain latent and inactive (quiescent infection) until the host conditions are ideal for its development (typically when sugars and starches in fruits are available). For this reason, to develop effective methods for detecting anthracnose in fruit crops, approaches have been proposed to seek its early detection through machine vision, spectral analysis and AI techniques. All studies reported in these fields and related to

anthracnose have improved the understanding of the anthracnose infectious process (stages) from the early to the visible stages of the disease.

2.4.1. Anthracnose detection in fruit crops

In this section, works related to anthracnose detection in fruit crops (different from mango) will be detailed, since their review provides several perspectives to address the problem of assessing the feasibility of early detection of symptoms associated with anthracnose in mango.

Works related to the detection of anthracnose symptoms with vision systems are reported in crops such as strawberry [70, 98, 166], aonla [78], cotton [9], walnut [15], tea [38, 175], pear [89, 94], cucumber [92], among others. From these studies, those related to strawberry crops stood out, as they were developed in a timeline seeking first to evaluate spectroscopy to detect anthracnose under controlled conditions [166], then to extend the work to field detection using spectral vegetation indices and aerial equipment [98] and finally to compare different AI models to detect foliar anthracnose at an early stage [70]. The importance of these works lies in the thoroughness and completeness of assessing this disease, given the challenges in terms of detection and control. On the other hand, the works presented in [38, 92, 94] addressed the challenge of detecting anthracnose at the foliar level, since it is in this organ of the plant where the greatest problems of crop affectation and damage occur. To do so, the authors implemented several machine learning techniques such as SVM, KNN, RF, ANN, K-means and discriminant models such as LDA, successive projection algorithm LDA (SPA-LDA), among others. As a main and common result among these studies, SVM-based models obtained the best performance metrics: for tea, pear and cucumber crops, an accuracy of over 93 % [38], 93.25 % [94] and 94.7 % [92], respectively, was obtained.

Similarly, novel studies related to anthracnose detection in olive and tea crops are reported in [48, 175]. A model for early detection of olive anthracnose based on deep neural networks (ResNet101 architecture) with hyperspectral images in the Vis-NIR range (450-1050 nm), acquired in the laboratory on 250 olives, is presented in [48]. Each hyperspectral image consisted of 61 spectral bands of 1392×1040 pixels. These images were spectrally compressed using a trainable linear transformation and a spatial reduction to allow this type of data to be coupled with the convolutional networks. As a result, the authors reported that the ResNet101 network achieved an accuracy = 91.8 %, recall = 96.88 % and specificity = 73.08 %, which represented a promising advance in developing detection systems for anthracnose in olive trees. On the other hand, [175] exposed a study on the detection and discrimination of three common phytosanitary problems in tea plants: anthracnose (AH), green leafhopper (*Empoasca spp.*) (GL) and sunburn (BU). The work aimed to evaluate the possibility of detecting these three problems and discriminating them through hyperspectral analysis. As a result, the authors obtained a KNN model with an accuracy of over 95 % to discriminate

these phytosanitary problems, while doing an individual detection they achieved average accuracies of 98 % for GL, 86 % for AH and 83 % for BU; however, the authors recommended further validation or adaptation of their system to ensure adequate performance.

Thanks to these works referenced in the literature, the scientific community's interest in the study of this disease was confirmed, given the challenges it represents for its control and proper detection. It was also clear that research in this field was ongoing, as there was no universal consensus on the approach to detect this disease at the early stages of development. Additionally, it was also evident the potential of AI techniques and hyperspectral imaging for the characterisation of the symptoms associated with this disease, however, the development of robust diagnostic solutions requires a thorough study of the crop and the predominant *Colletotrichum sp.* strain, as specific studies and technological adaptations (sensors, methodologies, mathematical models) must be carried out to assure the success of the proposed approach to the problem of inspection and control of anthracnose.

2.4.2. Anthracnose detection on mangoes

Given the economic interest and the serious effect of anthracnose on the mango crop, affecting up to 100 % of the production [158], over the last years, research works have been carried out to support the detection of this disease in mango using the technological tools available to date. One of the first studies on anthracnose in mango was presented in [36]. They implemented a computer vision system (CVS) to quantify anthracnose lesions on *cv.* 'Manila' mango using a spinning image capture system. Among the features of the presented CVS were the characterisation of the size, number of points and their distribution on the fruit. Similar to this work, in [155] a methodology for detecting anthracnose in mango using colour imaging was presented. The authors focused their study on estimating the fruit damage percentage based on the area of the detected lesion, its perimeter and greyscale intensity. In the same line of evaluating the potential of colour vision systems to detect anthracnose, in [100], a study was carried out to detect the disease on mango fruits, using colour cameras. As a result, the authors obtained a classification model of affected fruits based on SVM with an accuracy of 83.34 %. However, a point in common with these studies is that the study and analysis was carried out with visible anthracnose lesions, which implied a late detection and, therefore, a late control of the disease.

Likewise, developing a CNN model (based on the AlexNet architecture) for detecting anthracnose in *cv.* 'Nam Dok Mai' mango fruits with RGB images was reported in [165]. As a result, the CNN model achieved an accuracy of 81.36 % for anthracnose-free (healthy) mangoes and 79.55 % for diseased fruits, indicating the potential of these deep networks for the detection of a difficult-to-control disease such as anthracnose. Similar to this work, studies evaluating the development of anthracnose in mango fruits through AI-based models

trained with RGB images have been reported [74,100,126,155]. These works have illustrated the potential of deep neural networks in anthracnose detection; however, their analyses have focused on detection at a late (visible) stage of the disease.

Few studies have addressed the foliar anthracnose in mango using AI-based models to characterise and detect symptoms [18,74,134,140,148]. One of the first studies on foliar anthracnose in mango was presented in [74]. This study assessed anthracnose to establish the disease severity by capturing images from a DSLR camera over 30 mango leaf samples. Preprocessing techniques to enhance image quality were also applied, and an SVM model was trained to detect anthracnose symptoms and to correlate with leaf pigment. As a result, in this work was established a potential correlation of the visible spectrum with the degradation of pigments such as chlorophyll *a* and *b*, and carotenoids.

On the other hand, in [148], a multilayer convolutional neural network (MCNN) was proposed for classifying mango samples infected by anthracnose. A total of 2200 RGB images were acquired in this study from different locations and conditions and separated into training (80 % of the images) and testing (remaining 20 %) groups. To deal with the size of the images used in CNN, images were first preprocessed for contrast enhancement and then rescaled to a 128×128 pixel. The trained MCCN model obtained an accuracy = 97.13 %, missing report rate = 2.87 % and false report rate = 0 %, denoting the potential of the DL methods to detect late-stage anthracnose symptoms. Similar results were obtained by [18] and [140], which obtained 96.6 % and 95.5 % accurate outcomes with their models. The difference with these studies lies in the use of a CNN with no preprocessing or feature extraction in [18], while in [140] the authors presented a vein pattern segmentator with an SVM classifier to extract and detect symptoms on the main leaf area. Likewise, in [126], a CNN architecture was used to train a DL model to identify and classify anthracnose, bacterial black spot, and sooty mould in mango. Using 380 images representing the healthy and diseased classes, different data augmentation techniques were applied to improve generalisation and a pre-trained MobileNetV2 model was used for the learning stage. The disease classification was done via an SVM at the final stage of the MobileNetV2 model. Recently, in [134], a CNN-based approach, called LeafNet, to detect seven of the most common diseases of mango using images of the leaves was presented. This model was trained using data from mango diseases in Bangladesh, where anthracnose, die back, powdery mildew, sooty mould, among others, were included in their database. As a result, the LeafNet model reached an average accuracy, F-score, and specificity of 98.55 %, 99.47 %, and 99.88 %, respectively, allowing them to state that their modelling approach was at the top of the most advanced classification models.

Although the studies mentioned above have contributed to the characterisation of the anthracnose symptoms in mango, the use of information in the visible spectrum implies that

the models are specialised in detecting symptoms in the late (visible) stage, so their contribution in the inspection and control of this disease is done in a corrective way (not a predictive form). A novel work merging RGB information with spectra in the UV range is presented in [132]. In this study, an LDA model to detect anthracnose in sugar mango fruits was presented. This study was carried out with 108 fruits, half of the samples (54) were inoculated with *Colletotrichum sp.* and stored for 8 days to collect daily images and characterise the disease development through a severity scale from 0 to 4, where 0 was a healthy fruit, 1 corresponded to early anthracnose lesions, 2 to moderate lesions and 3-4 to severe lesions. As a result, the UV-LDA-based model achieved an average accuracy of 97 % when using the severity scale, showing, according to the authors, the feasibility of detecting anthracnose in early stages with the use of UV-A illumination.

In the same line as [132], the literature reports the study carried out by [30] to evaluate the potential for predictive detection of anthracnose in mango fruits. This work reported a study to detect anthracnose on sugar mangoes with spectral signatures collected with a portable spectroradiometer in the Vis-NIR range (350-1900 nm). The fruit bank comprised 200 fruits (100 control units and 100 units inoculated with *Colletotrichum sp.*) stored for 11 days. During the experiment, 8,400 spectral signatures were collected. The models implemented in this work were RF and SVM, with the SVM model being the best one: 100 % accuracy when using the full spectrum and 92 % accuracy when using 29 selected wavelengths. Certainly, this work represented a reference in the spectral study of anthracnose in mango fruits; however, the authors proposed the need to extend its study to more robust models (such as ANN) and the use of hyperspectral images since the characteristic lesions of anthracnose grow and expand as the infection develops.

On the other hand, a preliminary study merging Vis-NIR hyperspectral imaging, chemometrics and pseudo-colour image processing to evaluate the feasibility of detecting anthracnose in early-stage *cv.* ‘Nam Dok Mai’ mango fruits was carried out in [149]. 40 mango fruits inoculated with *Colletotrichum gloeosporioides* in the equatorial region of each fruit were used, and hyperspectral images were acquired every 2 days from day 0 to 10 days after inoculation, resulting in 6 classes to be evaluated: 1 healthy and 5 related to the development of the anthracnose lesion. Using a linear discriminant factor analysis (DFA) model, fused with a PCA method to choose the PCs of the largest weight for the discrimination of the six classes, the feasibility of using fused hyperspectral imaging, pseudo-colour image processing and chemometrics to develop reliable solutions for disease control could be demonstrated. Similar to this work, in [150], a study that evaluated the potential of Vis-NIR hyperspectral imaging with the PCA method and SVM model for detecting anthracnose symptoms was presented. Experimentally, in this study, 60 fruits of the same cultivar and with the same periodicity described in [149] were used: Nam Dok Mai mango fruits were monitored on days 0, 2, 4, 4, 6 and 8 after pathogen inoculation. As a result of this work, they achieved an SVM

model capable of discriminating with 98 % of the defined classes (per day of acquisition), which implies that from the first days of disease development, symptoms could be detectable through spectral information. As a relevant fact, it should be noted that on day 4 after inoculation with the pathogen, there was a slight decrease in the accuracy of the model, a decrease that was corrected from day 6 after inoculation. However, as a common point of the studies mentioned above, an exhaustive analysis of other classification models, feature selection or dimensionality reduction methods, and further study with different mango cultivars was not evident.

Undoubtedly, all these works marked a relevant milestone in the study and understanding of anthracnose in mango since they proposed systems capable of detecting the presence of anthracnose lesions in mango (both fruit and leaf). The relevance and interest of the initial efforts that addressed the challenge of characterising the symptoms associated with anthracnose at the spectral level to reduce the number of wavelengths needed to detect the disease can also be highlighted. However, as reported in those papers, the need to establish a set of relevant wavelengths physiologically related to the early stages of anthracnose is still latent. In addition, answering the essential research question on the ability of current AI-based vision systems to detect early symptoms associated with anthracnose is an urgent matter, as robust solutions are needed to support the task of phytosanitary control throughout the harvesting and post-harvesting stages of this prized fruit crop.

3. Data collection and preprocessing

Anthracnose produces a localised infection, which implies that the disease develops in the organ (or part of the plant) where the pathogen is inoculated without systematically affecting the individual, i.e. the symptoms associated with anthracnose do not cause a generalised expression of the symptoms (in the whole individual), but a punctual one where the pathogen was inoculated. This disease behaviour has allowed the symptomatology of this disease to be studied in laboratory experiments with samples taken from the tree. Therefore, since fruits and leaves are the critical points for anthracnose-induced infection, their study under laboratory conditions was selected since the experimental conditions (inoculation, sample monitoring and data acquisition) could be controlled. The mango cultivars used for the study were 'Keitt', 'Kent' and 'Osteen' for fruits and 'Keitt', 'Kent' and 'Tommy Atkins' for leaves. The cultivars selected in fruit represent those of most significant interest for the international mango trade, as well as cultivars of great interest for the growing and recent mango-producing market in Colombia and Spain. On the other hand, the cultivars selected in leaves corresponded to cultivars with different levels of susceptibility to anthracnose, being 'Kent' moderately susceptible, 'Keitt' moderately resistant and 'Tommy Atkins' low susceptible [23, 45, 127].

3.1. Gathering representative mango samples to study the anthracnose symptom development in laboratory

For this research, mango fruit and leaf samples were collected from producing farms dedicated to research and commercialisation in Spain during the 2020 - 2021 seasons. The selection and collection of plant material was carried out following the protocol established in the "Guía para la recolección y preservación de muestras botánicas en campo" document [61] and the technical assistance from an expert researcher in the crop to choose the best samples of both fruits and leaves. The sample sets used for the experiments with (and without) inoculum of the pathogen causing anthracnose on mango are described below.

3.1.1. Fruit selection from 3 mango cultivars

For the mango fruits, a mango-producing farm for local consumption and export located in the region of Motril (Spain) and the mango germplasm bank of the experimental station



Figure 3-1. A representative group of selected mango fruits of the (a) ‘Keitt’ and (b) ‘Kent’ cultivars for the experimental development of this research

La Mayora from the Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM) in Málaga (Spain) were chosen. The mango fruit selection of the ‘Keitt’, ‘Kent’ and ‘Osteen’ cultivars was carried out according to the external visual appearance of the fruits: samples without visible lesions, with proper physiological maturity (fruit size suitable for harvesting) and similar external pigmentation (dark green colour with reddish regions), see Figure 3-1. This selection was performed with the guidance of an expert researcher from the IHSM, who provided the indications to obtain sets of fruits with a similar physiological maturity stage for each cultivar. In addition, the selected fruits were harvested and sent to the Centro de Agroingeniería of the Instituto Valenciano de Investigaciones Agrarias (IVIA) in Moncada (Spain) in less than 48 hours for later inoculation and laboratory monitoring.

Visually, the ‘Kent’ and ‘Osteen’ cultivars show similarities during their ripening phase, with a change from dark green to yellow-orange in the skin of the mango. In contrast, the ‘Keitt’ cultivar starts its ripening phase with a yellowish pigmentation that turns orange-red as the fruit ripens. Figure 3-2 shows three samples of the cultivars mentioned above, where the changes in pigmentation can be perceived as a result of the ripening of the fruits once they have been harvested.

Table 3-1 describes each set of fruits selected, per cultivar, for the laboratory experiments. The number of samples per cultivar and set was established based on the fruit availability with the selection criteria described above and the maximum possible number of fruit that could be acquired daily with the hyperspectral vision system (see section 3.1.4). On the other hand, the pre-assay reported in Table 3-1 was used exclusively to verify the appropriate inoculation method to obtain the development of anthracnose lesions on mango fruits; therefore, no spectral analysis was performed because its purpose was purely to test the inoculation methodology. In addition, these experiments were carried out in two different mango

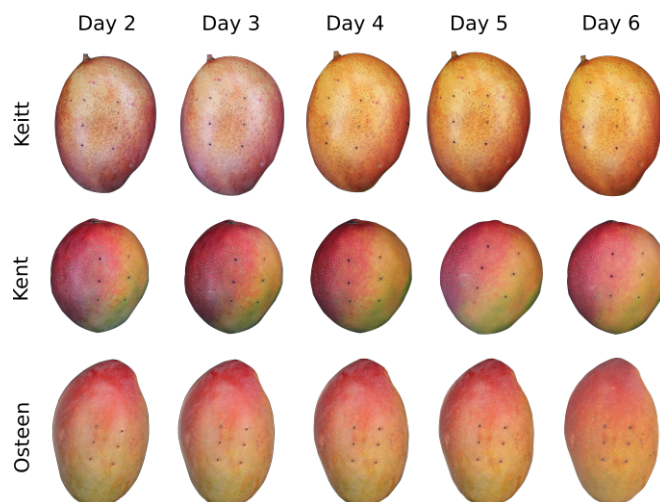


Figure 3-2. Mango fruit samples of 'Kent', 'Osteen' and 'Keitt' cultivars over 6 days, showing the changes in pigmentation as the fruit ripens when harvested

Table 3-1. Mango fruit sets for the experimental development of this research work

Type	cultivar	Samples	Weigh (g)				Width (mm)			
			Min	Max	Mean	Dev std	Min	Max	Mean	Dev std
Pre-assay	'Kent'	10	326	491	397	52	65	70	67	3
Assay	'Kent'	60	350	690	492.4	76	72	93	82	4
Assay	'Keitt'	55	603	988	822.7	85.7	82	100	91.1	4
Assay	'Osteen'	100	380	732	529	67	76	93	85	3
Assay	'Keitt'	100	651	983	849	77	77	93	85	3

seasons: pre-assay, 'Kent' (60 samples) and 'Keitt' (55 samples) assays corresponded to the 2020 season, while 'Osteen' (100 samples) and 'Keitt' (100 samples) assays corresponded to the 2021 season.

The data reported in Table 3-1 shows that the cultivar with the highest average weight was 'Keitt', corresponding to the cultivar with the widest width, measured over the fruit's equatorial area. Due to space limitations and the maximum load of the automatic lifting system, which raises the fruit inside the inspection chamber for image acquisition, mango fruit weighing more than 1,000 g was discarded for this cultivar. On the other hand, the 'Kent' and 'Osteen' cultivars, as previously mentioned, presented similar weight and width values.

3.1.2. Leaf selection from 3 mango cultivars

The experimental station *La Mayora* from the IHSM was chosen for the mango leaf samples, as it has the most complete, controlled and proper mango germplasm bank, and therefore the most suitable plant material in Spain and Europe. The mango leaf selection of



Figure 3-3. A representative group of selected mango leaves of the ‘Kent’ (a) and ‘Keitt’ (b) cultivars for the experimental phase in this research

the ‘Keitt’, ‘Kent’ and ‘Tommy Atkins’ cultivars was performed according to their location and external visual appearance: leaves without visible lesions, with dark, green-dark or green pigmentation, of medium age and coming from the second-third of the tree. This leaf selection was also performed with the guidance of an expert researcher from the IHSM to ensure the relevance of the sample sets. They were packed, transported and processed to the laboratory according to the “Guía para la recolección y preservación de muestras botánicas en campo” [61]. Like the fruit sets, the mango leaves were received at the Centro de Agroingeniería of the IVIA in Moncada (Spain) within 48 hours of being cut from the tree.

Figure 3-3. shows different sets of samples selected for the experimental development of this research under laboratory conditions. Even though the selected cultivars, according to [23,45,127], present different susceptibility to the development of anthracnose symptoms, at a visual level the appearance of the samples does not differ from one cultivar to another, except for the average size of the selected samples, where the ‘Keitt’ cultivar presented the largest leaves and the ‘Tommy Atkins’ cultivar the smallest leaves in length and width.

Table 3-2 shows the sample list selected for the pre-assays and experiments fulfilled throughout this work. As with the fruits, the number of samples was determined by the availability of suitable plant material on the tree and the technical limitations at the laboratory level (hyperspectral and colour image acquisition times). Also, experiments were carried out in the 2020 and 2021 mango seasons: the pre-assays and the first assays of each cultivar corresponded to the 2020 season, while the second assays (and the third one for the ‘Keitt’ cultivar) corresponded to the 2021 season. The pre-assays were developed to find the best method of inoculation of the pathogen on the leaves, as it was mandatory to ensure the plant material (leaf) quality throughout its storage and the anthracnose development study. More details on the usefulness of the pre-assays are given in section 3.1.3.

Table 3-2. Mango leaf sets for the experimental phase of this research

Type	cultivar	Samples	Height (mm)				Width (mm)			
			Min	Max	Mean	Dev std	Min	Max	Mean	Dev std
Pre-assay	'Kent'	20	152	173	160	2	63	76	70	2
Pre-assay	Irwin	20	85	131	106	3	55	78	65	3
Pre-assay	'Keitt'	30	144	170	155	4	53	61	57	2
Pre-assay	Tommy A.	30	86	105	96	3	44	70	56	3
Assay	Tommy A.	30	75	132	101	5	50	78	62	4
Assay	Tommy A.	20	70	101	77	4	52	75	62	3
Assay	'Keitt'	30	139	165	149	4	46	58	62	2
Assay	'Keitt'	30	126	170	152	5	49	58	53	2
Assay	'Kent'	40	146	185	172	3	56	81	69	3
Assay	'Kent'	40	133	171	153	3	49	75	65	3
Assay	'Keitt'	40	130	161	142	4	48	60	55	2

The table above shows that the leaves of the 'Kent' and 'Keitt' cultivars have the greatest average height (measured from the tip to the petiole of the leaf) of the studied cultivars, being larger than 142 mm. On the other hand, in the width of the leaves, the cultivars presented a similar average width, varying between 55 and 70 mm. Additionally, from the three mango cultivars used in the leaf assays, the 'Tommy Atkins' cultivar had limited availability of leaves, given that the IHSM germplasm bank, for preservation of this particular cultivar, did not have enough plant material for the second assay (40 samples were required for each experiment).

3.1.3. Inoculation of *Colletotrichum gloeosporioides* in a representative group of mango samples

Before the conduct of the assays with the samples mentioned in Tables 3-1 and 3-2, a solution (inoculum) was prepared with *Colletotrichum gloeosporioides* in a concentration of 1×10^6 spores/ml. The inoculum for the assays was prepared by researchers from the Centro de Poscosecha at IVIA, who guaranteed its concentration and quality for mango samples. In addition, as a protocol for disinfection of the samples before inoculation, each sample was immersed in a 0.5 % hypochlorite solution for 30 seconds, then in distilled water for 30 seconds, and finally dried with paper towels, as reported in various related studies [99,128,156].

With the inoculum prepared and the samples disinfected, pre-assays were carried out to establish the best inoculation method that would allow the adequate development of anthracnose lesions on mango without significant loss of quality of the plant material under study. Since the leaves, once detached from the tree, started a senescent process (contrary to what happens with the fruits), pre-assays of inoculation were carried out to establish if the disease developed visibly before the death (drying) of the plant material. In addition,

Table 3-3. Results associated with pre-assays with mango samples and inoculation strategies for *Colletotrichum gloeosporioides*

Type	cultivar	Samples	Inoculation method	Inoculum points	Monitoring days	Monitored surface	Incidence (%)	
							Abaxial	Adaxial
Fruit	'Kent'	10	Injury	6	9	1 surface	100	
Leaf	'Kent'	10	Injury	5	9	Abaxial Adaxial	100	96
		10	Deposition	5	13	Abaxial Adaxial	100	66
Leaf	Irwin	10	Injury	5	9	Abaxial Adaxial	100	94
		10	Deposition	5	13	Abaxial Adaxial	88	28
Leaf	'Keitt'	10	Injury	5	10	Abaxial Adaxial	96	90
		10	Adaxial deposition	12	20	Abaxial Adaxial	3.75	7.5
		10	Abaxial deposition	5	13	Abaxial Adaxial	88	22
Leaf	Tommy A.	10	Injury	5	11	Abaxial Adaxial	94	90
		10	Adaxial deposition	12	20	Abaxial Adaxial	1.25	2.5
		10	Abaxial deposition	5	13	Abaxial Adaxial	84	0

these assays established monitoring and image acquisition times on the samples inoculated with the pathogen. The following inoculation strategies were tested: 1) adaxial localised deposition, 2) abaxial localised deposition, and 3) localised deposition on the induced lesion.

Table **3-3** reports the monitoring times of each pre-assay, the inoculation strategy implemented and the final incidence of each pre-assay. As a result of this experimental phase, it was established that the most rapid (and at the same time effective) method to allow the development of anthracnose lesions is the localised deposition of the solution on an induced lesion. In the case of fruits, the induced lesion was made with a surgical punch of 1 mm diameter and 2 mm deep, while a surgical needle was used for leaves. This strategy allowed the development of lesions with a diameter greater than 20 mm in at least 90 % of the samples under study and in an average time of 9 days after inoculation. During the days of continuous monitoring of the samples, no significant incidence of senescence in the leaves and overripening in the fruits was reported, since the preservation of the samples under ideal conditions (relative humidity > 85 % and a temperature of $23^{\circ}\text{C} \pm 1^{\circ}\text{C}$) allowed the plant material conservation of more than 15 days in fruits and 20 days in leaves.

On the other hand, from the results obtained from the pre-assays reported in Table **3-3**, it was evident that the strategy of the abaxial localised inoculum deposition on leaves performed better than the method used by the adaxial localised deposition. This finding agrees with the results reported in [107, 156], which also evaluated the best method of inoculation on mango leaves and explained that the defence constituent substances in leaves are found in greater proportion on the adaxial side, as this side is exposed to solar radiation and weathering. Therefore, this defence mechanism on the adaxial side makes infection by external pathogens more difficult, while on the abaxial side, not having a significant accumulation of these substances, infection by external agents is easier. However, the times reported for the abaxial localised deposition were superior to the induced lesion method, making this

latter method more suitable for developing assays with *Colletotrichum gloeosporioides* in both mango fruits and leaves.

As relevant aspects of this inoculation method in induced lesions, and considering that anthracnose develops symptoms locally, it was decided to make six lesions on one side of the mango fruit and five on one side of the leaf. To compare the development of anthracnose lesions in inoculated samples (hereafter diseased samples) with non-inoculated samples (hereafter sound samples), in the group of sound samples, an equal number of lesions were made and covered with distilled water. The strategy aimed to replicate the same experimental conditions between sound and diseased samples, with the only difference being the use of inoculum in one group and distilled water in the other. This strategy also increases the number of anthracnose lesions per sample (and thus per assay) as these can be isolated and analysed individually since the pathogen does not induce a systemic but a localised infection.

Finally, based on the pre-assays, the monitoring time of the experiments was set to a maximum of 9 days, since above this threshold time, visible anthracnose lesions were achieved in all pre-assays with the selected inoculation method. However, it was also established that an assay could be finished earlier than this threshold time if an incidence of more than 90 % of anthracnose lesions was achieved in the samples.

3.1.4. Spectral acquisition system

The hyperspectral acquisition system used consisted of a monochromatic camera (CoolSNAP ES, Roper Scientific, Inc, AZ, USA) sensitive in the 320 - 1030 nm range and two liquid crystal tunable filters (LCTF) sensitive in the visible (400 - 720 nm) and near-infrared (650 - 1100 nm) ranges (Varispec VIS-07 and NIR-07, Cambridge Research & Instrumentation, Inc., MA, USA). The image acquired by this vision system corresponded to an image of 1392×1040 pixels. In addition, the system was integrated into a dome-shaped inspection chamber, illuminated by 12 x 25 W halogen lamps equally distributed around the scene. Figure 3-4 shows the spectral acquisition system used for this research.

For the LCTF working range, the extremes of the filters were delimited to avoid a low signal-to-noise ratio (SNR) at the sensitive edges of the filters (450 - 720 nm for the Vis filter and 680 - 980 nm for the NIR filter), and the spectral resolution was set to 10 nm. This spectral resolution was determined based on the absorbance and reflectance peaks in plant material, which are typically wider than the chosen spectral resolution. Similarly, including lower resolutions increases the number of spectral images acquired (increasing data acquisition, processing and analysis times) and the redundancy of these images, given that the spectral variation between one wavelength and another is usually minimal when studying plant ma-

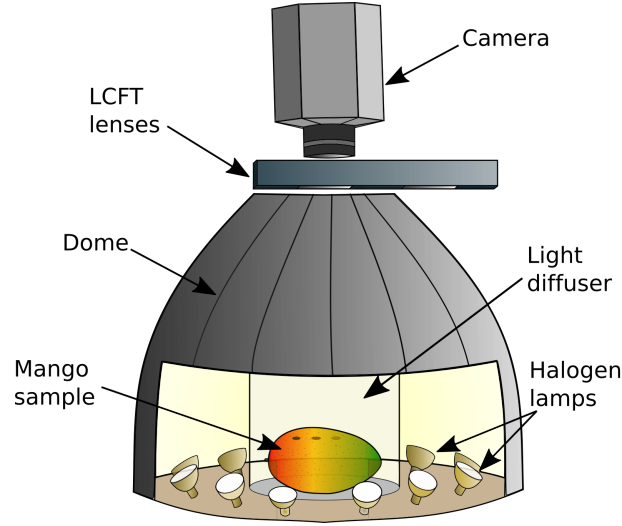


Figure 3-4. Hyperspectral acquisition system (description in section 3.1.4) used for spectral mango imaging

terials.

During the acquisition of the hyperspectral images, white (Spectralon 99 %) and black (LCTF filters covered and light sources turned off) image references were captured with each filter to correct the images of the samples and obtain the relative reflectance. Equation 3-1 was used to perform this spectral correction, the image of each sample (I_{ref}) results from the normalisation of the original image (I_{raw}) concerning the range established by the white reference (W_{ref}) minus the black reference (B_{ref}). Additionally, as a filtering strategy for the corrected images, it was established that reflectance values higher than 0.99 or lower than 0 were discarded given the reflectance scale (from 0 to 1) and the fact that the used reflectance target has a certified reflectivity of 0.99 (nothing could be whiter than this reference panel).

$$I_{ref} = \frac{I_{raw} - B_{ref}}{W_{ref} - B_{ref}} \quad (3-1)$$

Since each filter captures an independent hyperspectral image and has different efficiency curves and positions (within the acquisition system), a merging procedure was performed to obtain a single hyperspectral image in the 450 - 980 nm range. A point extraction and matching algorithm was applied in the 700 nm band (of both filters). Figure 3-5 shows two examples (a fruit and a leaf) of this phase, showing the matched points of interest between the images (linked by straight coloured lines) and the unmatched points of interest. Following the interest point matching algorithm, a RANSAC algorithm was applied to find both images' transformation matrix and correct the LCTF filters' spatial offset. Additionally, since the hyperspectral images of the LCTF filters have overlapping spectral ranges (680 - 720 nm),

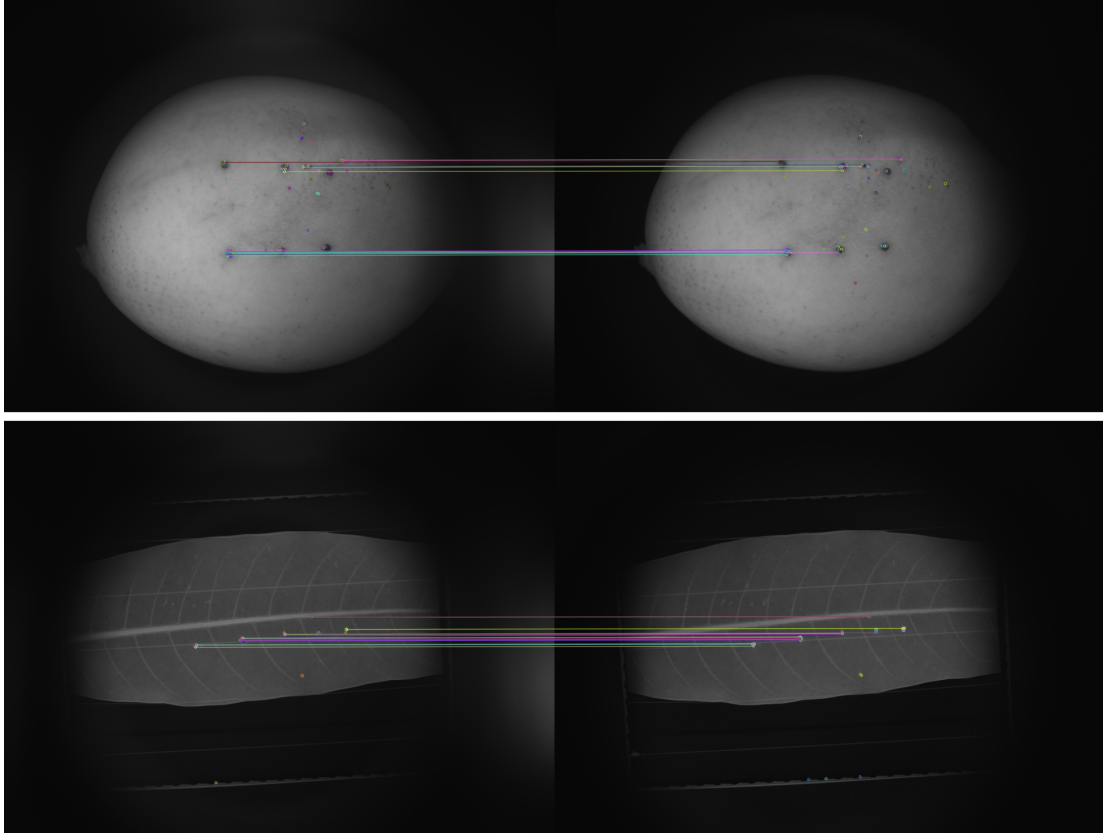


Figure 3-5. Images acquired of the same samples (fruit and leaf) showing the matching points (linked by a coloured line) that allowed merging both images into one

a median filter, Equation 3-2, was applied between the original bands of this overlapping range to obtain the resulting band of the hyperspectral Vis-NIR image.

$$I_{Vis-NIR} = \frac{I_{Vis} + I_{NIR}}{2} \quad (3-2)$$

As a final result of the hyperspectral acquisition system, hyperspectral images were obtained in the range 450 - 980 nm (Vis-NIR range) with a resolution of 10 nm, i.e., images of 54 wavelengths.

As a complement to this hyperspectral image acquisition activity, daily RGB images were taken with a Canon EOS 500D camera (F32, f 35 mm, ISO 800) in a light-controlled chamber to have a visual reference of the mango samples and the surface and visible changes they showed. Figure 3-6 shows a) the setup of this RGB image acquisition chamber with a mango fruit and b) the scene captured with the previously mentioned setup.



Figure 3-6.: Motorised lifting assembly for the positioning of mango samples inside the hyperspectral image acquisition dome

3.1.5. Hyperspectral imaging conditions

Since the proposed study involved daily monitoring of the same sample for 9 days (or less), a controlled lifting system was implemented to the spectral acquisition system to ensure that the acquisition distance (distance from the camera to the sample) was repeatable. This motorised lifting system consisted of a servomotor that positioned all samples at a distance of $140 \text{ mm} \times 2 \text{ mm}$, obtaining a spatial resolution of 0.115 mm/pixel , i.e. an image covering an area of $160 \times 120 \text{ mm}$. In this way, the acquisition conditions of the same sample were respected throughout the experiment since the acquisition plane (at a distance of 140 mm) remained the same. This solution validated the temporal analysis of a sample and its comparison with the other samples of the same test, since the recorded reflectance was constrained to the same plane or scene. Considering the above, for the hyperspectral image acquisition on mango fruits, the length of the fruits was measured before inoculation to compensate for the distance the motorised system had to travel to maintain the established acquisition plane. Figure 3-7 shows the motorised lift for positioning mango samples (representation) and the real motorised lifting assembly used in this work.

For the hyperspectral image acquisition on leaves, the length of the leaves was not measured because the thickness of the leaves was less than the resolution of the motorised system (2 mm). However, the leaves had natural folds that made the image acquisition difficult as the days passed. For this reason, a leaf sample holder was designed to insert leaves and ensure they remained flat in the acquisition plane. Due to the difference in size between the mango leaf cultivars, this holder was magnetised and included nylon strings that allowed positioning of the leaf area to be acquired regardless of the leaf width. Figure 3-8 shows the leaf sample holder and its assembly (representation) inside the hyperspectral acquisition system. In addition, given the thickness of the mango leaves, the bottom of this holder was

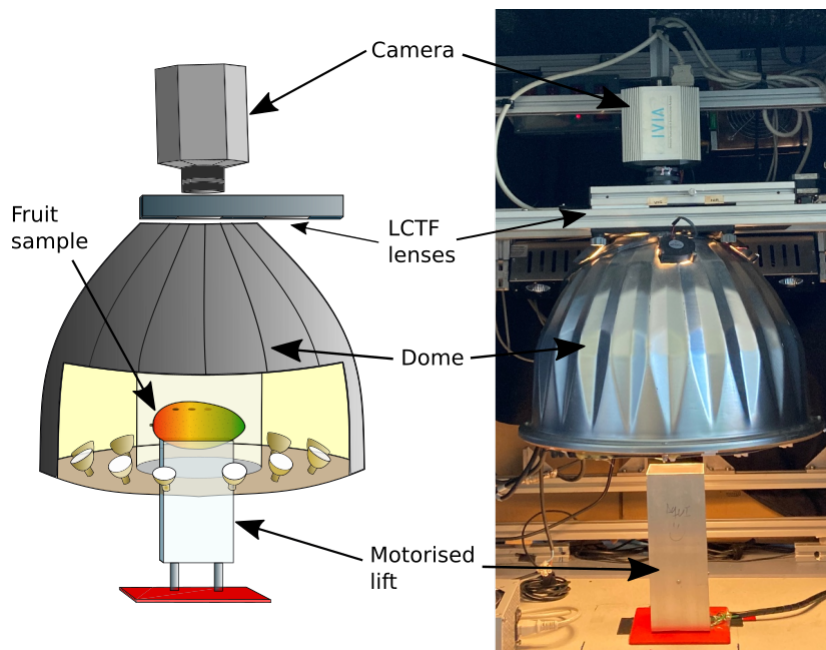


Figure 3-7. Motorised lifting assembly for the positioning of mango samples inside the hyperspectral image acquisition dome

matte black to avoid transmittance effects that would affect the diffuse reflectance acquired by the hyperspectral camera. This phenomenon is similar to the acquisition process in the field, where the tree leaves do not have a reflective surface on the back of the leaf, which affects the recorded data.

3.1.6. Spectral data pre-processing

An essential part of spectral and hyperspectral analysis is data pre-processing. This procedure aims to improve data quality to reduce the incidence of noise (both internal and external) in the data, to improve the design and training of AI-based models, and to represent digitally and without bias the phenomenon or object under study.

In this research, since a spectral characterisation of symptoms associated with anthracnose was proposed, several filters were applied to remove non-relevant data (scene background in the images), to reduce spectral noise and eliminate outliers, and to standardise variables (wavelengths) to avoid bias in the spectral feature selectors. Additionally, the strategy for selecting the regions of interest (ROIs) and the class definition for the model training was also established. As an important point, an anthracnose lesion was considered an individual and temporally differentiated ROI, as each visible lesion was localised and showed changes over time.

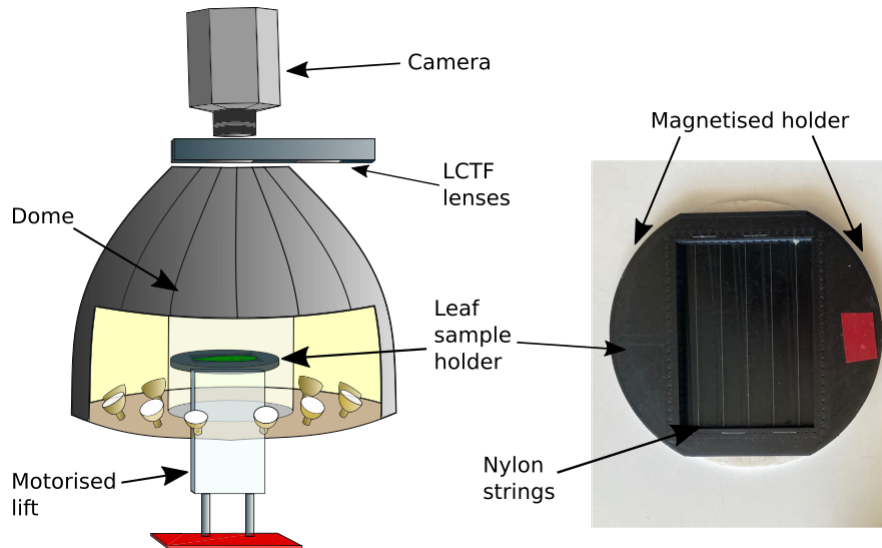


Figure 3-8. Representation of the motorised lifting assembly for the positioning of mango samples (fruits and leaves) inside the hyperspectral image acquisition dome

The methods used and assessed in this hyperspectral image pre-processing step are listed below.

3.1.6.1. Background removal filter

The main function of this technique is to remove data, sections or objects that were not of interest in an image but had been recorded by the image acquisition system. For this research, the object of interest was defined as the mango fruits or leaves contained in the hyperspectral images; therefore, any section or area outside these objects was considered as background and was removed.

A thresholding filter was used to remove scene background, and the 700 nm wavelength was selected as image thresholding since this wavelength has a distinctive reflectance peak associated with plant materials: fruits and leaves increase their reflectance in this region, which allows them to be easily differentiated and segmented from the background (which is not related to plant materials).

3.1.6.2. Spectral smoothing filter

The Savitzky-Golay (SG) filter is one of the most widely used smoothing methods in spectroscopy and signal analysis, given its features to remove the incidence of noise that causes deformations (variations) in a portion or a complete one-dimensional signal (as in spectral signatures). In spectroscopy, this filter minimises spectral noise at specific wavelengths or portions of the spectrum that do not truly represent the phenomenon or object under study, thus preserving the most essential spectral features for further analysis. Additionally, this

method may configure the degree of the polynomial to smooth the signal and combine it with other methods, such as the 1st or 2nd derivative, to enhance the relevant features in the spectra further.

To set up an SG filter, the size of the smoothing window (number of variables affecting the estimate, w), the degree of the polynomial (degree of filter fitting to the signal, m) and the derivative order, d , need to be established. When $d=0$, there is no derivative filter, while $d = 1, 2, \dots$ means the derivative order to be used. To determine the filter parameters, tests were carried out with different configurations, resulting in the SG filter ($w=11$, $m=3$, $d=0$) as the best fit without loss or relevant deformation of the appropriate spectral features in the collected data.

3.1.6.3. Outlier filtering (box plot)

Outlier detection is a strategy to improve the quality of the data or the sets (classes) established to ensure the maximum performance of calibrated models. Given the number of wavelengths that compose a spectral signature, in spectral analysis, techniques for detecting these outliers are usually applied on a global basis, since the individual treatment of a wavelength does not allow the outlieriness of a value to be established: depending on the phenomenon under study, a value or variation at a specific wavelength may be typical.

Therefore, the spectral signature outlier filter applied in this research was based on the extended box plot. Since the box plot allows the detection of extreme values (outside the limits of the plot) of a variable by defined class, for this study, the presence of 6 wavelengths in the same spectral signature with these extreme values was considered an outlier and thus removed from the data set. The threshold of 6 wavelengths corresponded to the (rounded) 10 % of the 54 wavelengths that compose a spectral signature. Experimentally, a threshold higher than the defined threshold did not preserve particular spectral features in the signatures (intra-class spectral diversity was lost), while a lower threshold did not impact the outlier detection.

3.1.6.4. Data standardisation

Data standardisation, also known as the standard normal variate (SNV) method, is a technique of standardising variables applied to a dataset to remove external effects and represent only those changes due to the variation of the measured variable [145]. In spectral analysis, this technique is often applied to correct the light scattering to avoid bias in the feature (wavelengths) selection and to enhance the learning (and hence the performance) of AI-based models.

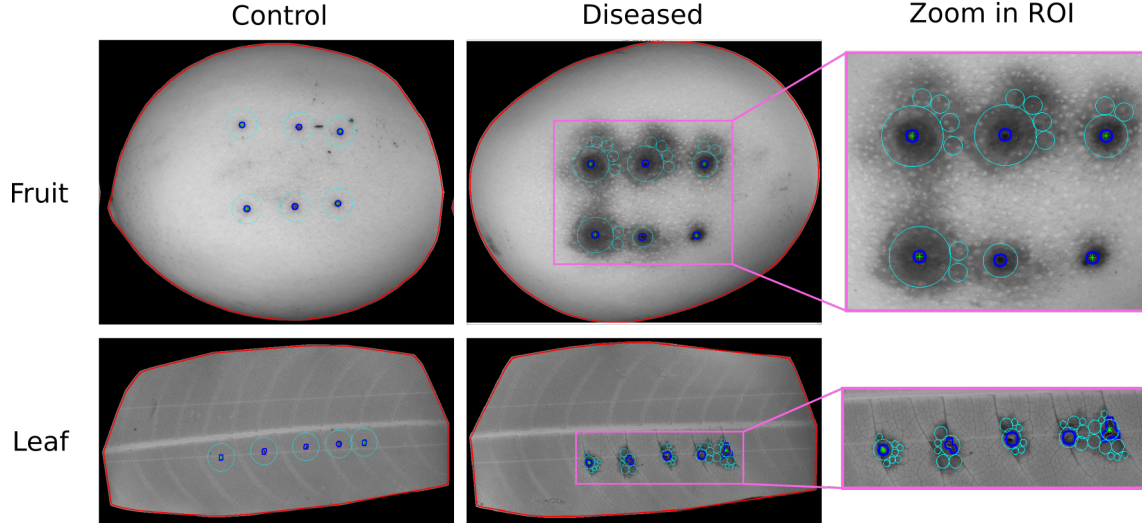


Figure 3-9. (a) ‘Keitt’ mango fruit sample and (b) ‘Tommy Atkins’ mango leaf sample with selected ROIs (cyan colour). The internal regions in the samples, coloured in blue, represent the induced lesions with punch and surgical needle, for fruit and leaf respectively

This method focuses on the mean of each variable, removing the magnitude of one variable over the others and constraining the data dispersion of each variable to 1 standard deviation. Thus, it allows the representation of the spectral changes due to the reflectance of the phenomenon under study (anthracnose symptoms) and not to external factors or noises.

3.1.6.5. ROI selection and data labelling

The ROI selection was made according to the lesions on the samples: six lesions on mango fruits and five on leaves. Starting from the last day of acquisition of each assay, where necrotic anthracnose lesions were visible, circular regions with a radius between 1 and 3.45 mm around each lesion were extracted from the fruits, while for the leaves, individual ROI up to 3 mm were selected. The size of the ROI was chosen to cover the anthracnose lesion consistently. Additionally, to avoid bias in the spectral analysis of the punch-induced lesions on fruits and the surgical needle on leaves, regions over these areas were determined to remove them from the hyperspectral image and thus from each selected ROI. Figure 3-9 shows two mango samples, a fruit of the ‘Keitt’ cultivar and a ‘Tommy Atkins’ leaf on which ROIs (cyan colour) were extracted on the last day of acquisition, day 6 for the fruit and day 7 for the leaf. In this figure, the blue-coloured regions represent the induced lesions in the samples removed from the ROIs as they do not belong to the infection lesions.

After the ROIs were selected on the last day of the assay, they were projected onto the rest of the images of each sample (previous days of acquisition) to ensure that each ROI represented the same area of the sample during each acquisition day. In this way, an ROI represents the temporal development of anthracnose symptoms over a specific sample region. To do the

ROI projection, a perspective correction algorithm was applied to calculate the transformation matrix between images to identify and correct the view offset of one image to another. This procedure was necessary because the position of the same sample photographed on different days varied, so to project a ROI from one day to another properly, this projection had to be included in the ROI selection.

On the other hand, since each ROI was composed of at least 500 spectral signatures (pixels of the hyperspectral image), a subsampling step was performed by calculating the average spectral signature of a 3×3 pixel window. This window size was selected considering that the minimum detectable size of a visible anthracnose lesion was 1 mm in diameter. Also, this subsampling was applied taking into account the benefits of the spectral signature smoothing in reducing noise and decreasing spectral data (in an order of 9) without losing the quality and representativeness of the extracted ROIs.

After the ROI selection in the mango samples of each assay, the data (spectral signatures) were labelled for further analysis and use in the training and testing of the classification models. For this purpose, the acquisition day was used as the class label for the data. This class definition was established as this research sought to establish the period of symptom detection as early as possible after pathogen inoculation, so the daily data analysis was considered the most appropriate strategy to address this issue. Furthermore, foreseeing that symptom development from one sample to another may vary, a clustering technique (K-means) was experimentally tested for class definition, resulting in several clusters (classes) similar to the number of classes obtained with the days of acquisition. Therefore, including a clustering step in the anthracnose symptom detection and classification model was unnecessary.

Likewise, for the control class, corresponding to the sound samples, a single class was established because this class represented an anthracnose-free mango sample regardless of the study day. The above definition in a single control class also described the fact that this class implicitly included the natural process of ripening (in fruits) and senescence (in leaves) in mango, therefore, the models or feature selectors that were implemented had to differentiate the spectral variations produced by these natural processes (intra-class variance) from the anthracnose infection (inter-class variance).

3.2. Flowchart for hyperspectral image pre-processing

For pre-processing the hyperspectral images acquired in the experiments, the following flowchart was proposed to obtain a bank of filtered and labelled spectral signatures to analyse and design a detection model of anthracnose symptoms in mango. Figure 3-10 shows the defined flowchart.

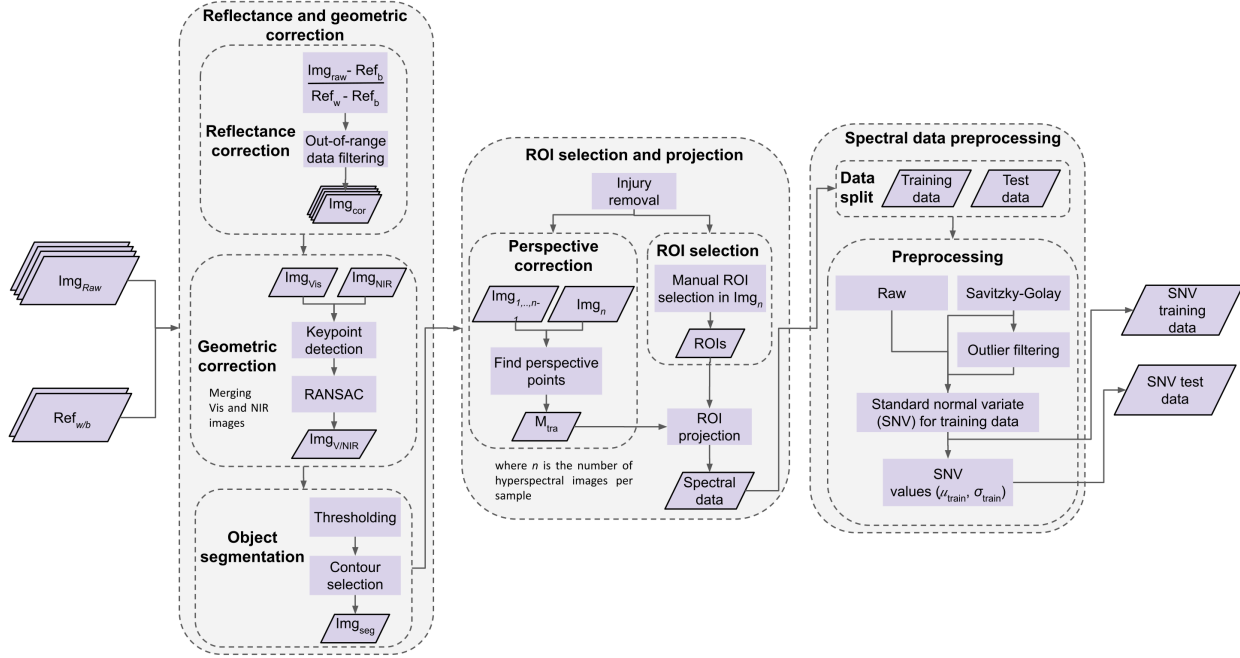


Figure 3-10. Flowchart for pre-processing hyperspectral images acquired on mango fruits and leaves in the laboratory

This diagram was complemented by developing a C++ application to automate several tasks in image pre-processing, such as reflectance correction with black and white references, the merge of Vis and NIR images, background removal, and ROI selection and projection. Given the special approach to the ROI selection, this task was performed manually, as automating the anthracnose lesion detection was indeed the aim of this research. The results associated with the assays performed in this research, pre-processing filters and the ROI selection strategy, are presented and discussed in section 6.1.

4. Design and train machine learning models

This chapter presents the implementation, assessment and discussion of several machine learning (ML) models to establish the feasibility of early anthracnose detection in mango fruits and leaves using the hyperspectral image analysis in the Vis-NIR range (450-980 nm). To do this, this chapter includes a section where the ML models implemented in this work are described, as well as the relevant hyperparameters in the design and calibration of each ML model. This section is complemented by describing the hyperparameter optimisation strategy to obtain the best version of each model and the performance metrics used to objectively assess and compare the models in the conducted mango assays (described in Chapter 3). Likewise, a flow chart to outline the methodology used for the correct training and validation of the techniques, as well as the metrics used for the selection of the best detection models for the studied mango cultivars, is included in this chapter.

On the other hand, the results associated with this chapter are presented in section 6.2. The results of the optimised ML models, both multiclass and biclass versions, are also analysed and compared to establish the feasibility of the early detection of anthracnose symptoms in mango. Given that the mango samples correspond to fruits and leaves, the results of the multiclass and biclass models are presented differently for each sample type. Finally, throughout the development of the section mentioned above, a discussion about the results and the best models for each mango cultivar is addressed.

4.1. Machine learning models used in this research

Given the wide variety of models proposed over the years, ML techniques can be divided into two main areas: standard (or traditional) ML techniques and artificial neural network (ANN) techniques. Standard ML techniques comprise models widely implemented and used as benchmarks in the scientific community due to the proven performance of these techniques in analysing, abstracting and learning features from data. Within these techniques, the models used in this study were: based on decision trees (classification and regression tree -CART-, random forest -RF-), Naive Bayes, based on discriminant analysis (LDA, QDA, PLS-DA), support vector machine (SVM), boosting (BST), gradient boosting machine (GBM) and extreme gradient boosting (XGB). On the other hand, ML models based on ANN opened

a new approach in spectral analysis due to the great variety of techniques proposed and the positive results obtained to date. The ML-ANN techniques used in this study are the single-layer perceptron (SLP) and the multi-layer perceptron (MLP).

4.1.1. Naive Bayes

Model based on Bayes' theorem (Eq. 3.1), calculates the conditional probability that an observation belongs to a class or group based on evidence (predictor values). This model assumes that all predictors are independent, so the likelihood of data belonging to a class is not biased by the dependence between predictors. In addition, Naive Bayes has been used as a classification model given the performance shown in works related to disease detection and multiclass discrimination, among others. Additionally, being a mathematical model, it is not affected by training and recursive validation steps like other AI models. Therefore, it has become a reference model in classification problems with AI models.

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} \quad (4-1)$$

In spectral analysis, the premise of independence among predictors impacts model performance since contiguous wavelengths have a high correlation (redundant), thus biasing the model's ability to separate the defined classes. Additionally, the correlation among predictors resulting from the phenomenon under study may be affected by the model's performance since the redundancy between wavelengths, in some instances, tends to mask the changes/correlations produced by the phenomenon under study. For example, in the spectral research of crop diseases, disease symptoms may make changes in pigmentation or plant material degradation that could be detected in several ranges of the Vis-NIR spectrum, however, redundancy between predictors could cause the model to concentrate its learning in a single region of the spectrum, thus reducing the relevance of the studied phenomena (symptoms associated with a disease). Therefore, the researcher or developer should consider this weakness of the Naive Bayes model in their analyses to ensure the reliability and relevance of their study.

4.1.2. Models based on discriminant analysis

Discriminant analysis (DA) is a supervised classification technique widely used in spectral analysis due to its robustness to highly variable population data. DA-based methods classify an observation (new data) according to its features (weights) and how these relate (linearly or non-linearly) to groups of data whose characteristics are known a priori. DA contrasts the assumption of class mean equality to determine their potential to be distinguishable (or not). To do so, DA-based techniques look for separate groups by considering the inter- and intra-group variance of the classes: the higher the interclass variance and the lower the

intraclass variance, the better the model fits the data.

As DA models, LDA, QDA and PLSDA techniques were implemented and compared in this study. LDA produces linear decision boundaries that maximise the inter-group variance and minimise the intra-group variance. This method, based on Bayes' theorem (Eq. 4-1), estimates the probability that an observation, given a certain value of the predictors, belongs to each class of the response variable, $P(Y = k|X = x)$. Finally, the observation is assigned to the class k for which the predicted probability was the highest.

LDA is an alternative method to logistic regression when the qualitative variable (response variable) has more than two levels, as in the assays defined in this research, where there are five or more categories (classes) per experiment. Also, PLSDA, a variant of LDA, is a robust algorithm that has proven to be very successful in modelling high-dimensional datasets for various purposes, such as product validation in food analysis and disease classification in medical diagnosis [87].

In contrast, when the data have non-linear relationships and classification cannot be done robustly, due to the lack of flexibility of the LDA model to deal with this drawback, the DA variant called quadratic discriminant analysis (QDA) produces quadratic (curved) boundaries that fit better the data, allowing to separate the groups and overcoming the linear stiffness of the LDA model. QDA considers that each class k has its covariance matrix (Σ_k) and, as a consequence, the discriminant function takes a quadratic form:

$$\log(P(Y = k|X = x)) = -\frac{1}{2}\log |\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log(\pi_k) \quad (4-2)$$

to calculate the posterior probability from equation 4-2, it is necessary to estimate, for each class, the covariance matrix (Σ_k), the mean (μ_k) and the prior probability (π_k) from the sample. Each new observation (sample) is classified in the class for which the posterior probability value is larger. Since the data shown in Table 6-4 had no balanced class distribution, the models' prior class probabilities (π_k) were obtained by dividing the number of samples in each class by the total number of samples.

4.1.3. Decision tree models

The decision tree (DT) is based on hierarchical segmentation to build logical diagrams functional in AI. This class of models does not perform statistical estimation since it does not construct a formal mathematical model or calculate the parameters or weights associated with that mathematical model. However, through iterative and sequential evaluation of the predictors, it seeks to determine the decision boundaries (thresholds) to order the predictors

(build the decision tree) and thus perform a classification or estimation of a new observation. According to Parra (2019) [121], the DT *“is an explanatory and decompositional technique that uses a sequential, iterative and descending division process that, starting from a dependent variable, forms homogeneous groups specifically defined through combinations of independent variables in which all the cases collected in the sample are included”*.

A DT is essentially composed of nodes, branches and leaves. The nodes represent the predictors (input variables) of greatest relevance to the tree, the branches represent the possible values of the predictors, and the leaves are the possible values of the response variable (output variable). Every DT is constructed from a root node that represents the predictor of the highest relevance in the classification or estimation process. The selection of subsequent nodes (next level in the tree) is developed according to the relevance of the predictor variable, thus recursively building the decision tree.

However, positioning two or more nodes at the same level may cause inconsistencies in the tree construction since the same node may belong to different levels of the tree, thus affecting the tree’s predictive ability. This phenomenon is known as overfitting, and to solve this issue, the tree’s growth is usually limited by the number of nodes or levels. This process of optimising the DT is known as pruning and is frequently used in training DT-based models with many predictors.

4.1.4. CART

The standard DT, which employs the abovementioned methodology, is known as Classification and Regression Tree (CART). This model generates binary decision trees, meaning each node is always split into two branches until its final assembly (leaves) is achieved. The advantages of the CART model include 1) its robustness to outliers, 2) the invariance in the structure of its classification or regression trees to monotonic transformations of the independent variables [44], and its interpretability since it allows for obtaining a representation of the defined tree. This enables the establishment of the order of importance of the predictors (nodes) and the defined decision thresholds (branches) to obtain a single prediction (leaves).

4.1.5. Random Forest

This model comprises an ensemble of single decision trees, each trained with a random sample extracted from the original training data by bootstrapping. Bootstrapping is a resampling method, widely used in statistics and ML, to extract datasets with replacement (i.e. data can be selected multiple times) from the original data. This method aims to estimate a population parameter and train models under different datasets (different from the original, but with a normal distribution).

Using bootstrapping in random forest (RF) implies that each tree is trained on slightly different data. Each tree is usually constructed following the rules defined for decision trees or, in the case of this research, for CART-based trees. For the prediction of a new observation, RF averages the predictions of each tree (in the case of regression trees) or selects the most repeated categorical response (class, category or discrete value) among the trees (in the case of classification).

The selection and creation process of the trees in RF is similar to that used by the bagging technique, which improves the results as it decorrelates the trees generated in the process. The fundamental difference between RF and bagging lies in the predictor selection: while in RF this selection is random, in the bagging process, the construction of the trees can be influenced by a small set of dominant predictors and thus be biased. As general recommendations for the random selection of the m predictors in RF from the original p predictors, according to [14], there are:

- The root square of the total number of predictors for classification problems, $m \approx \sqrt{p}$.
- One-third of the number of predictors for regression studies, $m \approx \frac{p}{3}$.
- If predictors are highly correlated, smaller values of m achieve better results.
- The best way to find the optimal value of m is to evaluate the out-of-bag error or to use cross-validation.

By increasing the number of trees created, RF does not face overfitting issues like bagging. Once a certain number of trees is reached, the reduction in test error stabilises.

4.1.6. Boosting and variants

Boosting is a strategy used in several ML methods, including decision trees. Boosting, unlike the bagging technique used in RF, bases its operation on the sequential adjustment of simple decision trees to improve the overall predictive power of the model, i.e., it generates simple decision trees (with less than 10 nodes) which are connected sequentially in such a way that the input of one tree is the output of another. Then, it can test and create multiple tree ensembles, which reduce the error associated with the final prediction of new observations as the error of one tree is slightly adjusted (reduced) in the next one.

As opposed to the bagging method (random forest), boosting does not use repeated sampling (bootstrapping), the difference between the trees that form the ensemble arises because the importance (weight) of the observations changes with each iteration. However, this nested

strategy is sensitive to overfitting the data as the sequential tree creation process may not have a defined stopping criterion, leading to the creation of even one decision tree per observation. Therefore, procedures such as hyperparameter optimisation (see section 4.2) are important in this class of models and cross-validation strategies are recommended not to degrade the predictive potential of these techniques.

In this research work, three boosting algorithms were employed, which are the most widely used in ML: AdaBoost (BST), Gradient Boosting Machine (GBM) and Xtreme Gradient Boosting (XGB). Adaboost, an acronym for adaptive boosting, was one of the first algorithms to become widespread and widely used in ML, given the significant improvement in performance in prediction and classification (binary classification) tasks. Its functioning is based on constructing small decision trees capable of correctly predicting several observations from the training set. In contrast, the prediction errors (misclassified or estimated observations) serve as input for designing a subsequent decision tree. In this way, the error associated with mispredicting observations is reduced with each iteration; however, the lack of control of this iterative process may trigger high overfitting of the model.

Meanwhile, GBM is a generalisation of the AdaBoost algorithm with the advantage of calculating a sequentially differentiable cost function (usually the difference between the actual value and the predicted value), so the cost function of one model becomes the updated cost function of the next one. In this way, the adjustment (always towards reduction) of the residuals or errors of previous models is performed. However, given the recursive nature of this process, to control the overfitting caused by this model, the influence of each cost function on the residual decrease is limited by the λ parameter, also called the learning rate. With this parameter, and by controlling the number of iterations to train the model, the overfitting is controlled, yielding a balance of the bias and variance of the model to make accurate predictions with new data.

Similar to GBM, a boosting-based method is XGB, an enhanced variant of GBM, as it allows for parallelizable computation that directly impacts the training time of this kind of model. Additionally, it changed the method of node splitting of a tree based on the similarity and similarity gain metrics of the new nodes: the node with the highest gain and similarity to its parent node is the one that continues with the tree construction process. By introducing these new metrics and parallelising the mathematical computations, it has become possible to establish the potential of this technique for estimation tasks (regression problems) compared to standard ML techniques.

4.1.7. Support vector machine

A support vector machine (SVM) is a set of ML algorithms developed for classification and regression applications. Although initially designed as a binary classification method, its use has been extended to multiple classification and regression problems. SVMs have proved to be one of the best classifiers for a wide range of situations, which is why it is considered one of the benchmarks in the field of statistical learning and ML.

SVM is based on the computation of a hyperplane capable of separating a dataset of samples corresponding to different categories or classes (clusters). This is why SVM is often called a maximum margin classifier [14]. The intuitive idea of SVM is to ensure that the resulting hyperplane has an optimal separation of the samples closest to it, thus dividing the (training) dataset into the target classes or clusters. Moreover, after constructing this hyperplane, SVM expects that in the evaluation of new samples (test cases), it will be possible to classify them correctly according to their proximity (separation) to this boundary: the greater the distance from this hyperplane, the greater the certainty of a correct prediction.

The simplest hyperplane for separating sets is a straight line (for two dimensions), a plane (for three dimensions) or an n -dimensional hyperplane. However, the data distribution is not often separated, as most classification problems have n predictors whose effective class separation is done with non-linear curves or hyperplanes. To overcome these restrictions, SVM uses kernels to project the information in a larger dimensional space, where the data is adequately separated. The computation of this new space (hyperplane) is done with the dot product of the support vectors, allowing it to change from the original space to another easily. In this way, SVM increases its computational capacity and thus its learning. Among the kernels implemented to perform the hyperplane computation are the linear kernel, radial kernel, polynomial kernel and Gaussian kernel, with the linear kernel being the one with the best performance for classification problems in terms of computational efficiency (model training time) and learning ability. Therefore, this kernel was chosen to be implemented in this research.

4.1.8. Artificial neural networks

Artificial neural networks (ANNs) are mathematical models inspired by the biological behaviour of neurons and how they are organised to form the structure of the brain [121]. Through recursive trials, these networks learn how to organise their internal structure to abstract the model to map the inputs to the outputs (response variables) of the model and thus achieve prediction with maximum performance.

An ANN comprises layers and neurons, with the neurons performing simple mathematical operations to propagate the learning abstracted from the training data between layers and

neurons. The simplest representation of an ANN is known as a single-layer perceptron (SLP), which is composed of a hidden layer, an input layer (input variables) and an output (or response) layer, see Figure 4-1. These layers are composed of neurons (input, hidden and output neurons), which are interconnected through weights that indicate the relevance of a neuron in learning a feature or property of the dataset under analysis. The variation in the number of neurons in the hidden layer, the number of hidden layers and activation functions (mathematical operations performed by neurons in the hidden layer to learn) makes an ANN model unique for each problem.

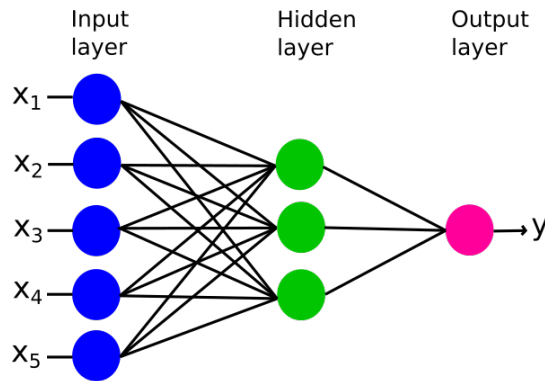


Figure 4-1. Representation of a single-layer perceptron neural network

Given the diversity of existing neural networks, and taking into account their proven usefulness in biclass and multi-class classification tasks and signal analysis (one-dimensional data such as spectral signatures), the single-layer perceptron (SLP) and the multilayer perceptron (MLP) were chosen in this research as the ANN models for analysis and comparison with other reference models in ML. SLP has been a benchmark in this field since it performs well in two-class discrimination and obtains outstanding results in regression applications without increasing the computational cost of training and designing more complex neural networks. However, it is only capable of learning simple patterns and can be limited for applications where the relationships between inputs and outputs are not linear and the weighting of an input compared to another is not controlled. The learning of the hidden layer could be biased. To overcome these limitations, it was found that by combining multiple hidden layers, the network can learn much more complex relationships between predictors and response variables. This structure is known as a multilayer perceptron (MLP) and can be considered the first deep learning (DL) model. Thanks to this class of models, which combines multiple hidden layers and non-linear activation functions, ANN models can learn virtually any pattern. It has been shown that, with enough neurons, an MLP model is a universal approximator for any function [121].

On the other hand, deep learning (DL) reference models are not included in this study as their use is primarily oriented to image analysis (classification), they present significant chal-

lenges in terms of training with hyperspectral images due to the high dimensionality of these images and their adaptability is limited to the processing of one-dimensional signals (such as spectral signatures). DL techniques base their learning on texture analysis, the application of filters, encoders, and descriptors (and their combination in multiple layers of the network), requiring mainly two-dimensional data on which they can operate. In addition, for these techniques to converge to reliable models, they require a large amount of data (in the order of thousands), which makes the collection and storage of hyperspectral data the main problem to be addressed, as the collection, storage and real-time processing of these hypercubes becomes expensive or limited.

In recent years, this topic has begun to gain relevance with technological advances, and the literature has reported works such as Fazari et al. [48], which proposed a compressive model of hyperspectral images to adapt them to standard RGB images and to be able to classify them with DL models. On the other hand, and as a recommendation to the reader, [8, 12, 81] present the challenges, research lines and problems that DL researchers should address to improve the integration of these techniques in hyperspectral analysis (this term includes spectral analysis). However, since this work's main objective was to establish the most relevant wavelengths to detect and characterise anthracnose symptoms in mango, the spectral signatures (one-dimensional data) analysis was identified as the research area. In contrast, this work did not consider the spatial component (primary focus of the DL models). However, it is presented as a future, extensive work of the results and analysis presented in this manuscript.

4.2. Hyperparameter optimisation

Certain ML models have internal parameters that can be adjusted for a given dataset. A hyperparameter optimisation procedure was employed using the grid search technique to obtain the best parameters (and thus the best ML models) for the data and classes defined in this study. Grid search performs an exhaustive search by evaluating all combinations in the defined ranges for the hyperparameters. This strategy has the disadvantage that it can be time-consuming to achieve the best combination of parameters; however, it guarantees the best model for the training data, especially when evaluating the potential of feature selection models. Therefore, considering the need to evaluate the feasibility of characterising and spectrally detecting anthracnose symptoms in mango, this tuning strategy fits the needs of this research.

On the other hand, since the discriminant analysis models (LDA, QDA and PLSDA) and Naive Bayes are mathematical models per se and do not require hyperparameter adjustment, the grid search procedure was not applied to these models. For the rest of the models used in this research, the hyperparameter optimisation was performed after the calibration (tuning)

of the following hyperparameters:

- **CART:** the parameters adjusted were: $minsplit = 10$, $maxdepth = 12$, $xval = (1:10)$.
- **RF:** the tuned parameters for the random forest technique were: $ntree = (10, 100)$ and $mtry = 7$ (derived from $\sqrt{54}$, 54 wavelengths).
- **SVM:** for this technique, a linear kernel was used and $cost = (0.001, 0.01, 0.1, 0.5, 1, 10)$. The parameter $cost$, or C , defines the penalty applied when the obtained model performs a misclassification.
- **GBM:** the parameters to be tuned were: number of iterations, $n.trees = (1:100)$ and complexity of the tree, called $interaction.depth = (1, 2, 3)$; while the parameters set were: learning rate, called $shrinkage = 0.1$, the minimum number of training set samples in a node to commence, called $n.minobsinnode = 10$, and the cross-validation method, being $nfolds = 5$.
- **XGB:** the tuned parameters were: $eta = 0.01$, $maxdepth = 8$, $gamma = 4$ and $subsample = 0.75$.
- **Boosting:** only the size of the cross-validation method, $nfolds < 5$, and the maximum tree size, $mtree < 100$, were adjusted for this model.
- **SLP:** the parameters tuned for this ANN were: $epochs = (1, 5, 10, 20, 20, 50, 100, 200, 500)$, $neurons = (1, 5, 10, 20, 50, 100, 200, 500)$, $nfolds = 3$.
- **MLP:** the parameters tuned for this ANN were: $layers = (2, 3)$, $neurons = (10, 20, 50, 100, 200, 500)$, $epochs = 100$, $nfolds = 3$.

For models based on ANN and boosting, whose learning rate per cycle is defined by the hyperparameter learning rate, the *learningrate* was set at 0.01 as this hyperparameter optimisation has a direct impact on model training time and it has been shown that close to this value the best results are obtained because it handles the over-learning of the models in a training cycle (when this value is very high) and the slowness in updating the weights (parameters) when this value is minimal.

4.3. Performance metrics applied to ML models

To establish the performance of the ML models, it is necessary to measure it through standard metrics or metrics commonly used in the scientific community. The metrics help the researcher to objectively determine whether the obtained model fits the data analysed and whether it is possible to generalise the statistical behaviour of the data to make predictions (or classifications) about new data or observations. Thus, the metrics used for this study were:

- **Accuracy:** represents the percentage of items correctly classified in each class. It corresponds to the quotient between the sum of the elements of the main diagonal of the confusion matrix and the total classified data.
- **Recall:** also called sensitivity or TPR (true positive rate), represents the number of data correctly classified within the positive class out of the total number of true positives. As a metric used for binary classification, the positive class corresponds to the control class for this work. In contrast, the negative class corresponds to all classes defined for the inoculated group.
- **Specificity:** also called TNR (true negative rate), represents the number of data correctly classified within the negative class out of the total number of true negatives. Like the recall metric, in this study, the negative class represents the classes of the inoculated group. Its relevance lies in correctly classifying a true negative, which implies detecting when data does not correspond to the positive class (control class).
- **F1-score:** it performs an overall weighting of the model's performance, implicitly considering recall, specificity and accuracy metrics. Therefore, in the evaluation of the performance of a classification model, calculating this metric allows a more global assessment of the model to be established. Using this metric, a model performs well if the F1-score is close to 1 as well as the recall and specificity metrics.
- **MCC:** Mathews correlation coefficient, or phi coefficient, is a performance metric whose measure is similar to the F1-score with the advantage of being invariant to class balance. Its use is often recommended for problems with unbalanced classes since F1-score, recall, and specificity are sensitive to class balancing. Like the F1-score, a model performs well if MCC is close to 1.

Other metrics that were used in this research to be applied to specific ML models were:

- **Classification Error Rate:** it is defined as the proportion of observations that do not belong to the most common class in the node.

$$E_m = 1 - \max_k(p_{mk}) \quad (4-3)$$

where p_{mk} represents the proportion of observations at node m that belong to class k . Despite the simplicity of this measure, it is not sensitive enough to create good trees, so in practice, other measures are often used.

- **Gini Index:** measures the total variance in the set of K classes of node m . It is considered a measure of node purity.

$$G_m = \sum_{k=1}^K p_{mk}(1 - p_{mk}) \quad (4-4)$$

When p_{mk} is close to 0 or 1 (the node contains mostly observations of one class), the term $p_{mk}(1 - p_{mk})$ is very small. As a consequence, the higher the purity of the node, the lower the value of the Gini index (I_{Gini}).

The CART algorithm uses the Gini index as a splitting criterion, while random forest (RF) and the genetic algorithm (GA) for feature selection use it as a ranking criterion.

4.4. Flowchart for training and validation of classification models

For the training, validation and selection of the best model for classifying spectral samples associated with the ROIs extracted from anthracnose lesions in mango, the following flowchart was defined, see Figure 4-2.

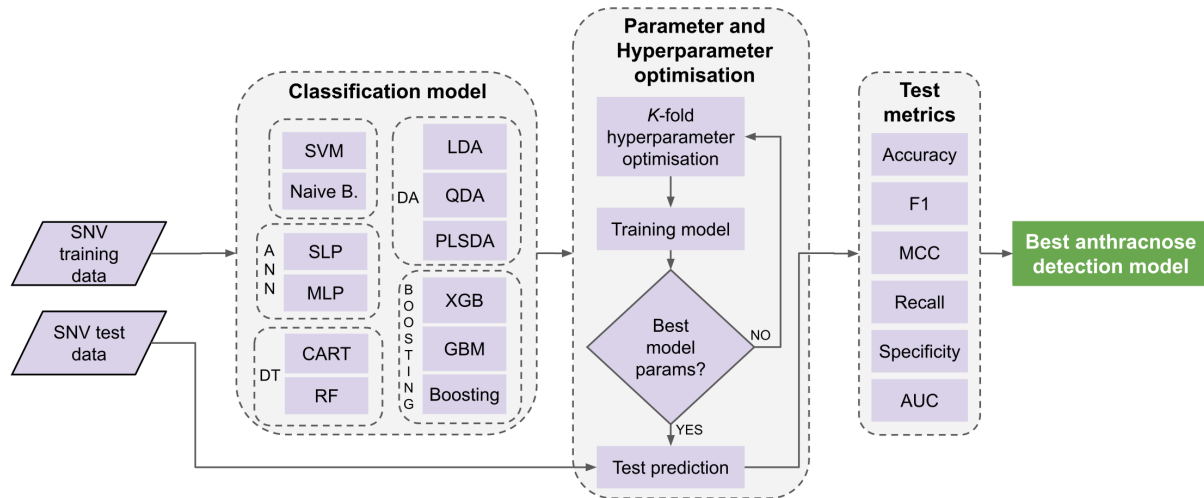


Figure 4-2. Flowchart for training classification models of spectral data collected in the mango sample essays (fruits and leaves) throughout this research

This scheme was coded in R, a statistical language used in research and development. Each ML model was programmed following the authors' guidelines and the recommendations of the researchers who support each R library to avoid biasing the model performance. Finally, the results associated with the training and validation of these ML models are presented and discussed in section 6.2.

5. Identification of important wavelengths

The main objective of this chapter is to show the implementation and assessment of different feature selection techniques that allowed the establishment of a set of relevant wavelengths for anthracnose symptoms in mango. The feasibility of using this set of wavelengths in designing an AI model for detecting the early anthracnose symptoms in mango was also established. For this purpose, a section describing the properties of the feature selection techniques used in this work was included. A flowchart that depicts the stages employed to assess each feature selection technique with the best classification model obtained in Chapter 4 is also presented. In addition, a result section where the performance metrics of the classification models trained with the wavelengths selected by the feature selection techniques is presented. As presented in Chapter 4, this section was divided into the trained models for mango fruits and leaves.

Finally, an interpretation and discussion based on plant physiology concepts are presented to establish the selected wavelengths' relation to anthracnose in mango. This analysis aimed to support and validate the detection models designed in this research.

5.1. Feature selection techniques applied to spectral characterisation of anthracnose in mango

Six well-known techniques in data analysis were implemented as feature (variable) selectors: principal component analysis (PCA), RF, SLP, genetic algorithm (GA), stepwise LDA (SLDA) and stepwise QDA (SQDA). These selectors aimed to find a set of wavelengths that allows for accurately discriminating the classes defined in this study and, therefore, to characterise the different stages of development of the anthracnose symptoms in mango fruits and leaves.

As each feature selector technique processes the data differently and has its own parameters, the configuration of each technique was carried out as follows:

5.1.1. Principal component analysis for feature selection (PCA4FS)

PCA is one of the most widely used unsupervised learning techniques in AI and data science, given the performance demonstrated in various data analysis applications. It allows the observations used in the model training to be expressed in a new dimensional space (principal components) to understand their properties better and make their analysis easier. PCA manages to reduce the initial data dimensionality, thus decreasing the computational cost of operating on the data. To do this, each principal component (PC) is a combination of the input variables, ensuring that these (the PCs) explain as much of the variance of the data as possible. This estimate also means that the latter components contain less predictive value as they do not considerably represent a portion of the variance, being considered as noisy components of the model and helping to reduce the incidence of this issue in the new dimensional space of the data.

However, using PCA as a feature selector (PCA4FS) has been a novel and useful approach, even though it is not possible to directly estimate the weight or importance of a variable in the model, given that it may impact one or more PCs. For this purpose, different approaches have been proposed in the literature. The PCA4FS used in this research was based on the weight analysis of the most relevant variables for the PCs (the PCs that explain at least 99 % of the variance of the data), see Figure 5-1.a. Then, a variable is considered relevant for a PC if the weight associated with it corresponds to a peak or valley in the weight function of the PCs, see Figure 5-1.b, since this variable represents an important portion (in absolute value) in the composition of the PC to which it belongs to. However, by having several PCs, there is a set of relevant variables per PC, therefore a variable ranking based on the frequency and cumulative absolute weight of each variable for all PCs was applied.

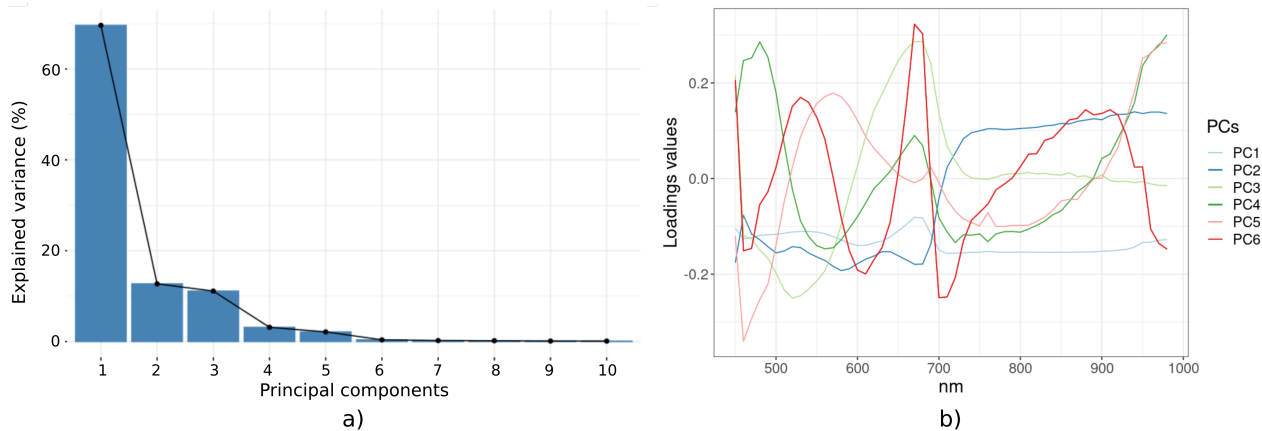


Figure 5-1. a) Percentage of variance explained for the first 10 PCs, where the first 6 PCs explain 99 % of the variance of data under analysis; and b) Weight functions of the six selected PCs, where the absolute peaks of each function establish the contribution of each wavelength to the composition of PCs

For a better understanding of the PCA for feature selection applied in this research, the pseudo-algorithm used is presented below:

Algorithm 1 PCA for feature selection algorithm

- 1: Compute $n - 1$ PCs, where $n = 54$, the number of wavelengths of the spectrum
 - 2: Organise the PCs according to the percentage of explained variance
 - 3: Select the number of PCs that explain 99 % of the data variance
 - 4: From the weight functions, extract the peaks and valleys with the highest absolute weights in each PC. These weights are directly associated with the wavelengths and correspond to their contribution to each PC they belong to
 - 5: Organise all wavelengths according to the prevalence in different PCs, as the first criterion, and the weight in each PC
 - 6: Select the wavelengths based on the prevalence criterion (a wavelength must be on at least half of the PCs). The maximum number of variables was set at 27, half the wavelengths
-

5.1.2. Random forest for feature selection (RF4FS)

Since random forest estimates from several trees, the direct calculation of the predictor importance is difficult to measure as it may be part of several trees at different nodes (tree depth). However, strategies have emerged to accurately measure the relevance of a predictor in the overall performance of RF models. Two of these strategies are the importance of permutation and the increase in node purity.

The strategy used in this research was the increase in node purity, which is based on estimating the purity improvement due to splits involving each predictor of all trees. To understand the node purity related to predictors in RF, it is necessary to comprehend the usefulness of the Gini index (splitting criterion used in the nodes). The Gini index (I_{Gini}), explained in section 4.3, indicates that a node is pure ($I_{Gini} = 0$) if the groups or classes it generates (by the divisions) are homogeneous, i.e. they are correctly classified; otherwise, if a node is impure ($I_{Gini} = 1$) it indicates that the generated classes are not homogeneous and their splitting is due to chance and not to predictors. Thus, predictors whose nodes are considered pure ($I_{Gini} \approx 0$) will be regarded as essential and those whose nodes are impure will be considered as irrelevant ($I_{Gini} \approx 1$).

However, since a predictor is part of different trees at different levels, the purity estimation must be measured in each tree and then averaged to obtain a global metric to quantify the global relevance of a predictor for the RF model. To do this, the I_{Gini} mean decrease ($MeanDecreaseI_{Gini}$) is used for each predictor. This is measured by calculating in each T

tree the change in the relative decrease in I_{Gini} , Δi , produced by the i predictor in all the m node where it is located, see Equation 5-1, thus eliminating the bias of diminishing the importance of a predictor that is found at different levels of the trees. Hence, predictors with a higher $MeanDecreaseI_{Gini}$ are expected to be the most relevant for the model since their contribution allows for an increase in the purity of the nodes ($I_{Gini} \approx 0$) and, therefore, to improve the classification of the data.

$$MeanDecreaseI_{Gini.i} = \sum_T \sum_m \Delta i(m, T) \quad (5-1)$$

As a final step, the average $MeanDecreaseI_{Gini}$ of all predictors ($\overline{MeanDecreaseI_{Gini}}$) was set as the selection threshold since the $MeanDecreaseI_{Gini}$ is computed for all predictors and the final result does not evidence the number of selectors to be chosen. Therefore, a predictor is considered relevant if its metric $MeanDecreaseI_{Gini.i} > \overline{MeanDecreaseI_{Gini}}$.

For a better understanding of the selection procedure of relevant predictors in RF, the following pseudo-algorithm was applied:

Algorithm 2 Random forest for feature selection algorithm

- 1: Create an RF model (optional parametric tuning)
 - 2: Calculate I_{Gini} at the nodes of the model
 - 3: For each i predictors, compute the $MeanDecreaseI_{Gini.i}$ using Equation 5-1
 - 4: Sort the $MeanDecreaseI_{Gini}$ vector in descending order
 - 5: Calculate $\overline{MeanDecreaseI_{Gini}}$
 - 6: Extract predictors such that $MeanDecreaseI_{Gini.i} > \overline{MeanDecreaseI_{Gini}}$
-

5.1.3. Single-layer perceptron for feature selection (SLP4FS)

ANNs have also been studied to understand the weight assignment to each predictor (variable) in the input layer and how this contributes to the final prediction of the model. As mentioned in section 4.1.8, the single-layer perceptron is one of the simplest and most powerful representations of ANNs. This feature has also allowed SLP to be of interest in the feature selection field since their interpretability, although not explicit, is simpler than deep neural networks (MLP or DL) as the analysis focuses on the relation between the hidden layer weights, the input variables, the activation function and the output layer.

Given that a neuron consists of an activation function, a bias (b) and weights (w) associated with the previous layer of neurons, the feature selection strategy is based on objectively

and invariantly pondering the contribution of these weights that link a variable in the input layer to a neuron in the output layer. Figure 5-2 shows a schematic representation of an SLP, where the weights w_{11} , w_{21} , w_{31} and w_{32} interconnect the input x_1 with the output y of the network. This representation, although simple, denotes the challenge of weighting the contribution of these four weights in the response generated by the inputs: ideally, they should only contribute to input x_1 to create a reaction while rejecting any signal from input x_2 , however, the interconnectedness of neurons in the hidden layer makes this task difficult. Furthermore, when multiple inputs, multiple outputs and different numbers of neurons are used, the analysis of the weight matrix becomes challenging, and so does the ranking of variables by relevance.

To overcome these challenges, Gideon (1997) [56] presented one of the most widely used metrics, Q_{ik} , in ANN to quantify the importance of an input variable i regarding the output k of the network and the number j of hidden neurons. This metric is based on the estimation of the matrices P_{ij} and P_{jk} , which quantify the contribution of the input neurons to the hidden layer and the contribution of the hidden layer neurons to the output, respectively. Thus, by the cross product of these matrices, the input variables can be related to the output variables and the importance of these (the input variables) in the correct classification of the data (output variables) can be measured. Equation 5-2 mathematically represents the Gedeon metric,

$$Q_{ik} = \sum_{r=1}^{nh} (P_{ir} \times P_{rk}), \quad (5-2)$$

where i is the input variable, j is the number of hidden neurons, k is the output variable, and P_{ij} and P_{jk} are calculated according to equations Eq. 5-3 and Eq. 5-4.

$$P_{ij} = \frac{|w_{ij}|}{\sum_{p=1}^{ni} |w_{pj}|}, \quad (5-3)$$

$$P_{jk} = \frac{|w_{jk}|}{\sum_{r=1}^{nh} |w_{rk}|}. \quad (5-4)$$

As an advantage of this metric, the magnitude of the total contribution of an input is decoupled from the sign of each weight (thanks to the absolute value in Eq. 5-3 and Eq. 5-4). Likewise, it avoids the recursive evaluation of the input variables to weight the importance of the input variables in the network (brute force evaluation method), improving analysis and variable selection times since they are calculated considering the matrices resulting from Eq. 5-3 and Eq. 5-4.

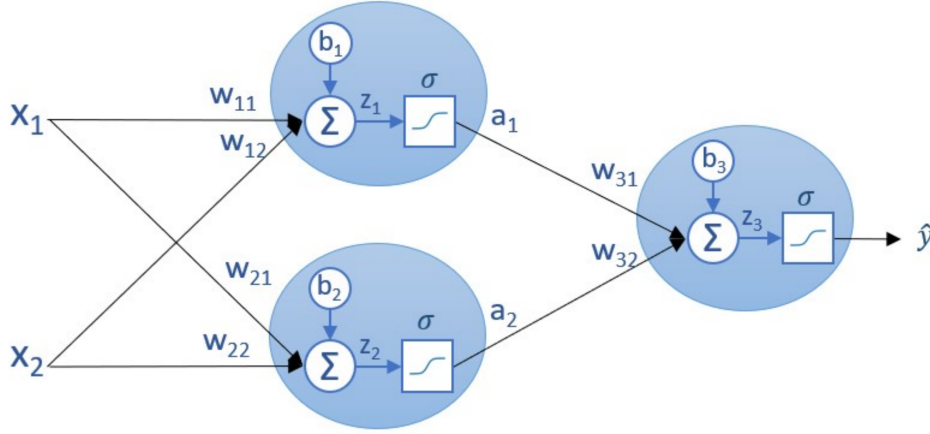


Figure 5-2. Schematic representation of an SLP with two inputs, two neurons and one output. Taken from [27]

Finally, computing the Gedeon metric requires training a neural network to obtain the weight matrix. To do this, the SLP feature selector implemented in this research starts with training an optimised neural network (hidden layer neurons from 1 to 500) to obtain the best neural model that describes the defined classes. Then, from the resulting network weight matrix, the Gedeon metric is calculated following equations Eq. 5-2 - 5-4. After weighing the input variables, the average threshold of this metric, $\overline{Q}$, is calculated to select variables with a metric above this average threshold. The pseudo-algorithm representing the selection procedure of essential features in the SLP model is presented below:

Algorithm 3 Single-layer perceptron for feature selection algorithm

- 1: Train an SLP model (parametric tuning of neurons in the hidden layer)
 - 2: Extract the weight matrix, W , from the trained model
 - 3: Calculate Gedeon's Q metric, according to equations Eq. 5-2 - 5-4, for all input variables i
 - 4: Sort the Q_i vector in descending order
 - 5: Calculate $\overline{Q}$
 - 6: Extract predictors such that $Q_i > \overline{Q}$
-

5.1.4. Genetic algorithm for feature selection (GA4FS)

Genetic algorithms are heuristic optimisation methods used to find the combination of variables that maximise the predictive capacity of a model, among other applications. Their operation is inspired by Darwin's theory of natural selection, indicating that only the best individuals of a generation are capable of combining to generate new individuals, who have

the features of their progenitors and are capable of adapting and recombining with other individuals to improve, generation after generation, the features and properties that allow them to adapt to the environment. In this context, the features of the individuals are the input variables (predictors) that are combined and then evaluated (fitness function) to measure the performance of these variables with the objective of the defined application or objective function (data classification, prediction of the response variable, among others).

Specifically, classification or regression models are used to evaluate the performance of an individual (combination of input variables), with the RF model being the most widely used given its predictive capacity in classification or regression tasks and being the learning model that implicitly allows the importance of the input variables to be measured. Similarly, once the genetic algorithm gets the individual with the best adaptation (fitness) to the training data, a ranking of the input variables is done with the I_{Gini} calculated implicitly and following the selection process described for Random Forest (section 5.1.2): with the $MeanDecreaseI_{Gini}$ metric, a selection threshold is calculated ($\overline{MeanDecreaseI_{Gini}}$) and the variables with a $MeanDecreaseI_{Gini}$ greater than this threshold are selected as variables of relevance for the genetic algorithm.

Genetic algorithm implementations are available in different platforms, languages and approaches. However, in [14], the approach chosen in this study to be implemented and evaluated as a feature selector is presented and discussed. A pseudo-algorithm describing the operation of the genetic algorithm implemented in this research is presented in Algorithm 4.

5.1.5. Forward stepwise selection based on DA

Forward stepwise selection (FSS) is a technique that allows, in a computationally efficient way, the selection of a set of predictors capable of fitting the training data without overfitting. It differs from the brute-force selection method (which determines the best set of variables) because it does not assess all possible combinations of predictors to converge to a solution (which raises the computational cost). Additionally, it is not over-fitted to a given training data set, as the FSS does not obtain the best set of variables. Still, it does obtain the best set of variables to balance out the possible impact of outliers or noise in the data. On the other hand, it differs from backward stepwise selection (BSS) in the iteration process to find the set of variables since it does not start by training the model with all the predictors and then discarding, one by one, the predictors that are less relevant for the model.

FSS starts with the random selection of 1 predictor until the one that provides the best performance for the data classification (or estimation) is determined. After selecting this predictor, the iterative process of including predictors that improve the model performance begins. This process is nested, which means that the predictor comprises the set of variables

with the previously selected ones. Different metrics can be established as a stopping criterion to indicate to the technique that the chosen set has achieved its objective. For this research, as a stopping criterion, it was chosen that the remaining predictors do not increase by 1 % the model performance. Accuracy was selected as a performance metric as it quantifies the overall model's performance to discriminate and correctly separate the defined classes.

Algorithm 4 Genetic algorithm for feature selection

- 1: Create an initial random population of P individuals ($P = 10$) with $n - 1$ characteristics (n is the number of input variables, 54)
 - 2: $iteration \leftarrow 0$
 - 3: $improve_{accuracy} \leftarrow 10\%$
 - 4: **while** $iteration < 100 \wedge improve_{accuracy} > 1\%$ **do**
 - 5: Calculate the fitness function for each individual in the population P , based on an RF ($ntree = 100$)
 - 6: Create a new empty population, P_{new}
 - 7: **while** P_{new} has empty individuals **do**
 - 8: Select two individuals from the existing population, where the probability of selection is proportional to the fitness of the individuals
 - 9: Cross the two selected individuals to generate a new descendant ($P_{crossover} = 0.8$)
 - 10: Apply a random mutation process on the new individual ($P_{mutation} = 0.1$)
 - 11: Add the new individual to P_{new}
 - 12: **end while**
 - 13: $P \leftarrow P_{new}$
 - 14: **end while**
 - 15: Compute the $MeanDecreaseI_{Gini}$ of the variables that compose the individual of the last generation with the best fitness
 - 16: Sort the $MeanDecreaseI_{Gini}$ vector in descending order
 - 17: Calculate $\overline{MeanDecreaseI_{Gini}}$
 - 18: Extract predictors such that $MeanDecreaseI_{Gini} > \overline{MeanDecreaseI_{Gini}}$
-

However, since this process requires a classification model that allows the evaluation of the performance of the selected predictors, the LDA and QDA models were integrated into this algorithm since they do not have hyperparameters that affect the overall performance of the model, eliminating the bias of determining whether the performance of a model is due to the predictors or to the hyperparameters of the models. Additionally, by not having hyperparameters, the computational cost of evaluating predictor sets is more efficient than models that require hyperparameter optimisation.

Consequently, thanks to integrating the FSS technique, the LDA and QDA models and the accuracy performance metric, the sLDA (stepwise LDA with accuracy metric) and sQDA

(stepwise QDA with accuracy metric) feature selectors were defined. A pseudo-algorithm describing the operation of this feature selector for the two versions established (sLDA and sQDA) in this research is presented in Algorithm 5.

Algorithm 5 Forward stepwise feature selection algorithm

```

1: From the  $n$  predictors, select the best-performing predictor for the model
2:  $improve_{accuracy} \leftarrow 0\%$ 
3: while  $(n - 1) > 0$  do
4:   Randomly select a predictor and calibrate the model
5:   Compute the  $improve_{accuracy}$  for the selected predictor
6:   if Is  $improve_{accuracy}$  the best value from all  $n-1$  predictors? then
7:     if  $improve_{accuracy} > 1\%$  then
8:       Include the selected predictor in the predictor set
9:        $n \leftarrow n - 1$ 
10:      Go to step 4
11:     else
12:        $n \leftarrow 0$  ▷ End the predictor selection
13:     end if
14:   else
15:     Go to step 4
16:   end if
17: end while
18: Extract the set of relevant variables

```

Finally, once the selectors had established the relevant wavelength sets for each experiment, the sets were evaluated with the best classification models for each cultivar in Chapter 4 to verify the best set of wavelengths with the best classification model. This involves including a further hyperparameter calibration step as the model has to be adjusted to the input information (the spectral data with the selected wavelengths).

In addition, as each selector is not limited to a defined number of variables, the performance of the classification model is affected by the number of variables used in the model training. Therefore, to make an objective comparison among the selectors, the 10 most important variables for each technique were selected to measure the influence of these selectors on the performance of one or more classification models. The results related to this selection and comparison are exposed and explained in section 6.3.

5.2. Flowchart for selection of relevant wavelengths and validation with classification models

For the estimation of the feature selectors and their subsequent validation with the classification models presented in section 4.1, to define and validate a representative set of wavelengths characterising anthracnose symptoms in mango fruits and leaves, the following flowchart was defined (see Figure 5-3).

The feature selectors were coded in R version 4.3.1, given the robustness and reliability of this statistical language in research and development. Additionally, each feature selector used in this research was programmed following the authors' guidelines and the recommendations of the researchers who support each of the associated R libraries. The results related to the feature selection and ML model training are presented and discussed in section 6.3.

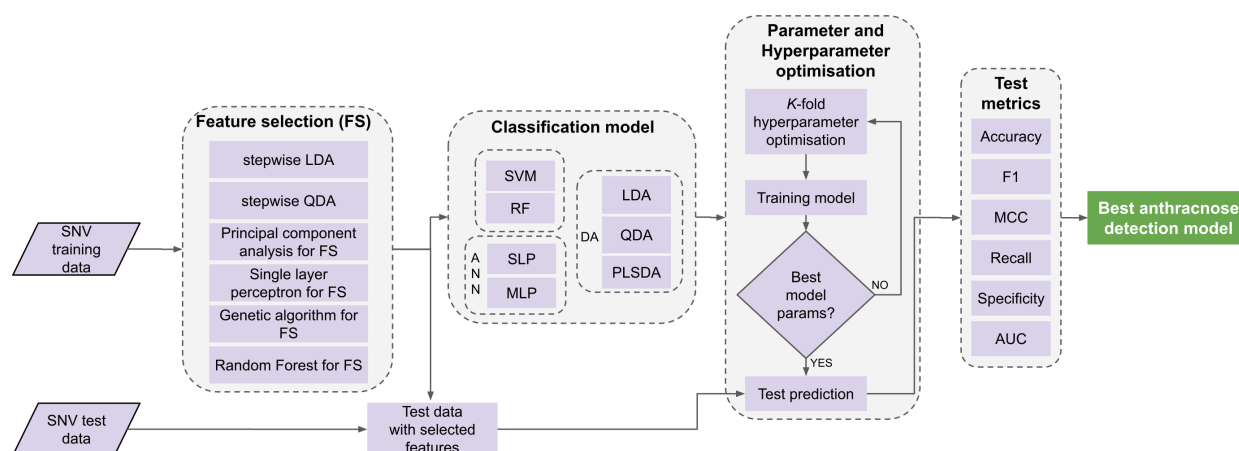


Figure 5-3. Flowchart for selecting a representative set of wavelengths for characterising anthracnose symptoms in mango

6. Results

This chapter reports the results associated with chapters 3, 4 and 5, evidencing the execution and achievement of the objectives listed in section 1.4. The first section of this chapter focuses on showing the results of the methodology used for the selection and inoculation of mango samples, the hyperspectral image collection, and the image pre-processing to obtain quality spectral data that guarantees the adequate training of the classification models for the detection of the symptoms associated with anthracnose in mango.

Then, the results of the design and training of the ML models implemented in this work for the classification (and hence detection) of the classes defined from the data collected in the laboratory experiments are presented. This section starts with obtaining the training and validation groups, passing through the obtaining of the ML models in the two-class and multi-class configurations for both fruit and leaf samples and the selection of the best preprocessing filter and the best ML models for the detection of anthracnose in mango fruits and leaves.

Subsequently, the results associated with selecting the most relevant wavelengths for the spectral characterisation and detection of the anthracnose symptoms in mango from the early (pre-visual) to the late (visible) stage are presented. Finally, to conclude this section, the ML models with their relevant parameters and the best-fit wavelength sets for the early anthracnose detection in each mango cultivar, both in fruits and leaves, are formally described.

6.1. Data collection and preprocessing

6.1.1. Descriptive and visual analysis of the mango samples

Before the cleaning and pre-processing of the hyperspectral images, a characterisation of the developed lesions in each cultivar and assay was carried out. This assessment supported as evidence to establish whether the lesions reported in the diseased groups (inoculated with *Colletotrichum gloeosporioides*) were compatible with anthracnose on mango and whether in the control groups, as a result of the induced lesion, anthracnose symptoms due to cross-contamination occurred during the hyperspectral imaging phase.

Figures **6-1** to **6-4** present a compilation of the three cultivars used for fruits and leaves in the control and diseased groups. Firstly, from the control samples, Figures **6-1** and **6-2**, it was observed that there were no symptoms on the fruits, except for the dried edge of the induced lesion. Similarly, on the leaves of the control groups, necrosis and drying of the edges of the induced lesions were observed, with no considerable alterations (changes in size, colouring or shape) during the assays.

On the other hand, from the results obtained with the diseased groups, Figures **6-3** and **6-4**, it was observed that the cultivar ‘Kent’, in fruits and leaves, was the most susceptible cultivar to anthracnose, since the symptoms were expressed with greater size and incidence than the rest of the cultivars. Similarly, for the Osteen and ‘Keitt’ cultivars in mango fruits, it was observed that the development of the reported lesions was compatible with anthracnose, with the difference that in the *cv.* ‘Osteen’, the incidence at day 6 of the study was already over 90% in all the samples. On the other hand, from the leaf assays, the ‘Keitt’ cultivar was found to be the most resistant cultivar (by the size of the lesions reported) to anthracnose, followed by ‘Tommy Atkins’. This finding agreed with the state-of-the-art, which indicated that the cultivars ‘Tommy Atkins’ and ‘Keitt’ had a medium resistance level to anthracnose. Additionally, it was observed that in the *cv.* ‘Tommy Atkins’, the recurrence of senescence symptoms on leaves was higher than in the rest of the cultivars, with symptoms being reported on the tip and edges of the leaf from day 7 of the study (this finding was also reported in leaf samples of the control group).

As a common point among the symptoms reported in the assays, both in fruits and leaves, it was evidenced that these become distinguishable from day 5 and so on, while in the ‘Kent’ leaf cultivar, this threshold was reported from day 4. This indicated that, at first sight, between days 2, 3 and 4, for most of the tests carried out, the anthracnose symptoms were not recognisable as they had not reached the size and symptoms whereby this disease is distinguishable. Except for a microscopic or close inspection of the sample, which does not apply to the current inspection methodology, it could indicate the presence of the disease between days 3 and 4. Therefore, and considering the findings made from the recording logs in the assays, day 5 was established as the threshold of visible disease detection for all the experiments and cultivars studied in this research.

Finally, it is important to mention that throughout the laboratory experiments (years 2020 - 2021) the number of samples, both fruits and leaves, varied due to the availability of quality plant material in the germplasm bank ‘La Mayora’ that provided the samples for this research, in Table **6-1** the initial samples per cultivar and assay are listed. Particularly, for the assay with the *cv.* ‘Tommy Atkins’ in leaves, there was a limited availability of leaves, given that the germplasm bank had 2 trees in good condition; therefore, having a greater number of leaves limited the production and preservation capacity of this mango cultivar in

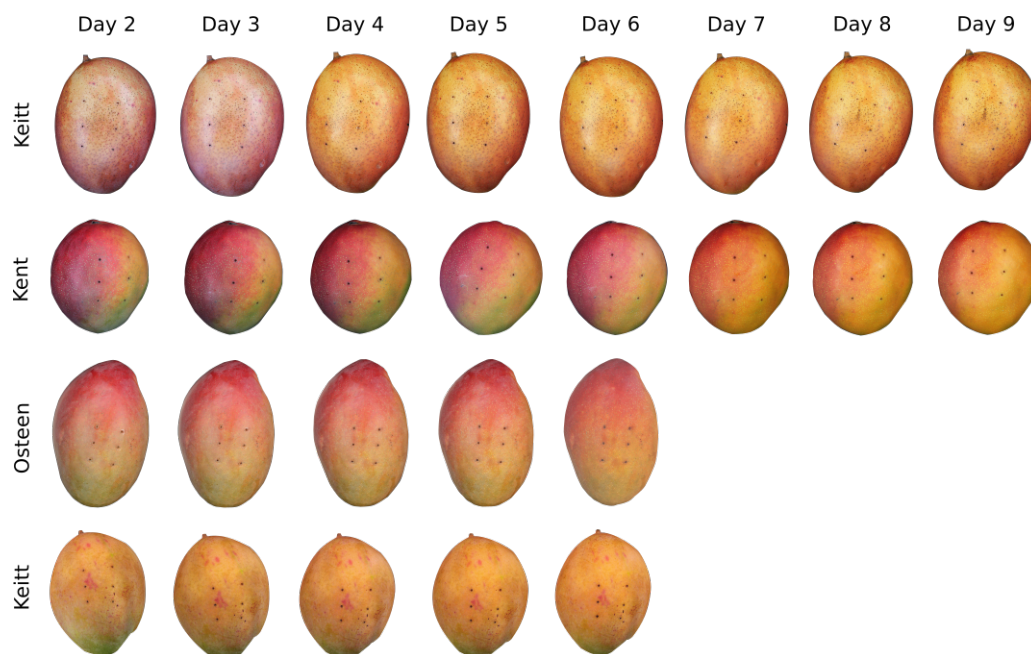


Figure 6-1. Mango fruits from the ‘Keitt’, ‘Kent’ and ‘Osteen’ cultivars of the control (sound) group over 9 days of monitoring and acquisition of hyperspectral images

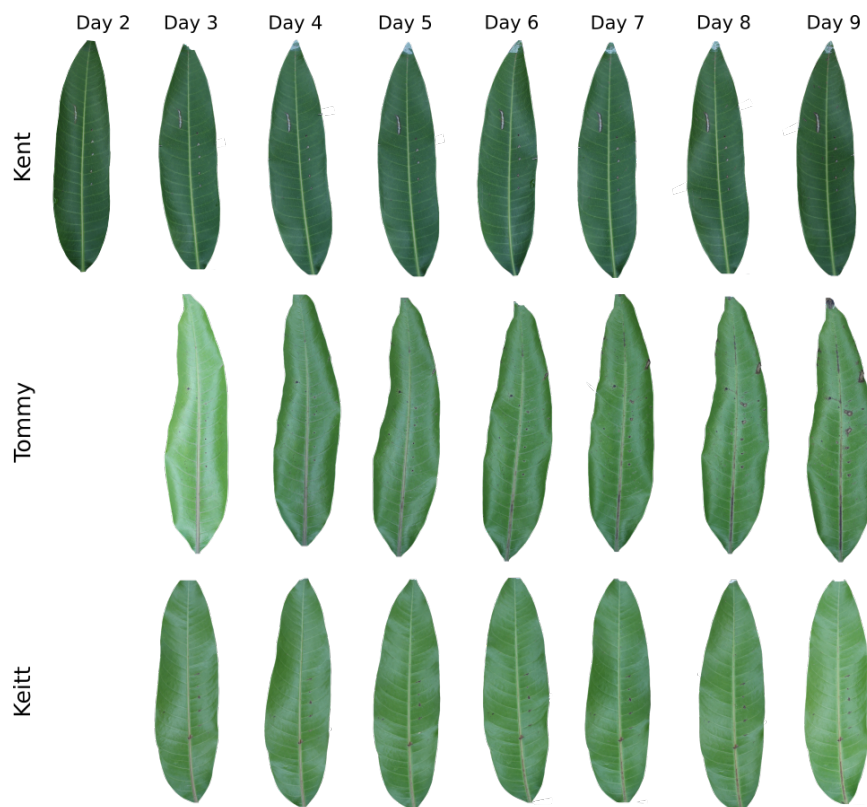


Figure 6-2. Mango leaves from the control group of ‘Keitt’, ‘Kent’ and ‘Tommy Atkins’ cultivars over 9 days of monitoring and hyperspectral imaging

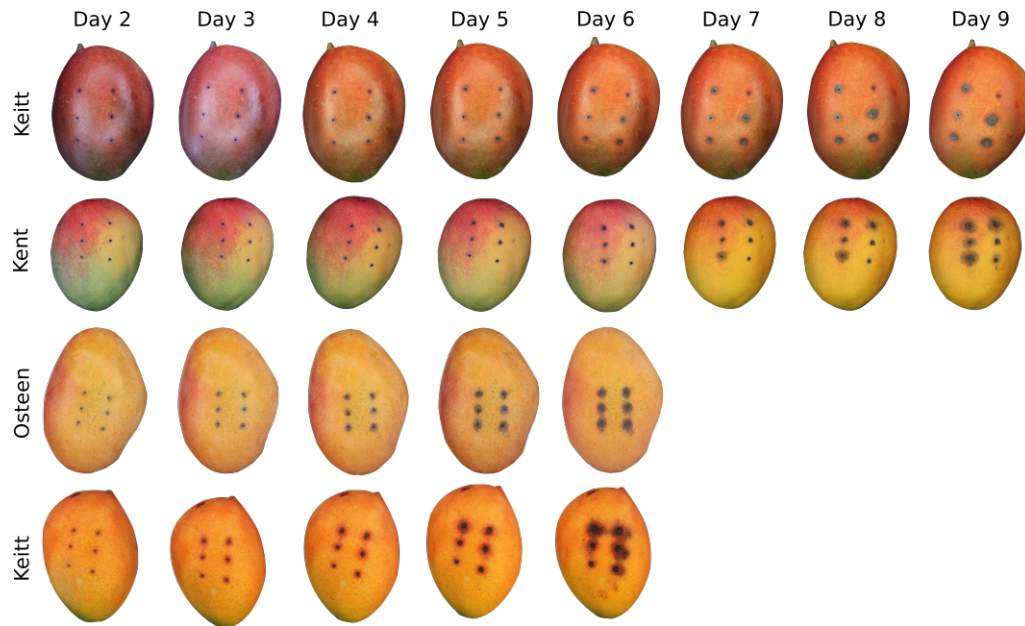


Figure 6-3. Mango fruit of ‘Keitt’, ‘Kent’ and ‘Osteen’ cultivars from the diseased group over 9 days of monitoring and hyperspectral image acquisition

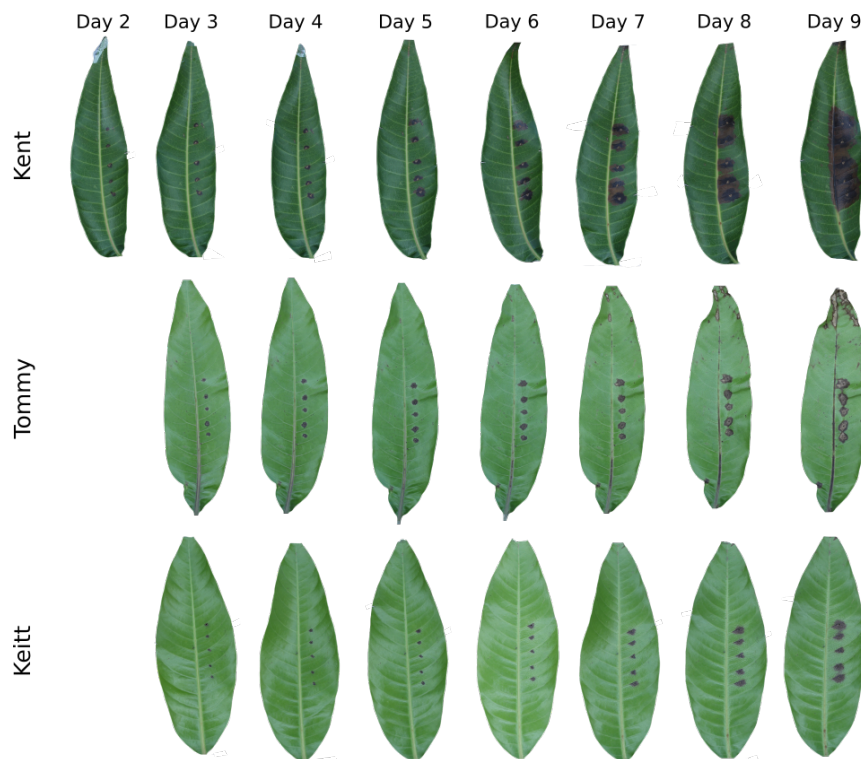


Figure 6-4. Mango leaves from the diseased group of ‘Keitt’, ‘Kent’ and ‘Tommy Atkins’ cultivars over 9 days of monitoring and hyperspectral imaging

the farm. Likewise, the initial number reported in the assays, see Table **6-1**, corresponds to the number of samples that fulfilled the requirements to be considered within the experiments, which implicitly indicates that there were samples discarded, since they presented superficial alterations that could affect during the inoculation and image acquisition phase.

6.1.2. Filtered hyperspectral image bank

Before the pre-processing of the hyperspectral images acquired from the experiments referenced in Tables **3-1** and **3-2**, a filtering of the samples that, after the monitoring days for each assay, did not develop visible anthracnose symptoms was carried out. Table **6-1** shows the information on the cleaned samples per assay and the cause associated with the sample discarding in each assay.

Table 6-1. Filtered sample group of mango assays reported in Tables **3-1** and **3-2** after the experiments under laboratory conditions. **DPM:** dried plant material (senescence), **NS:** no symptoms

Type	cultivar	Samples							Comments
		Monitoring days	Initial		Filtered		Loss rate (%)		
			Control	Diseased	Control	Diseased	Control	Diseased	
Fruit	‘Keitt’	9	10	45	10	45	0.00	0.00	No comments
Fruit	‘Kent’	9	10	50	10	50	0.00	0.00	No comments
Fruit	‘Osteen’	6	10	90	10	90	0.00	0.00	No comments
Fruit	‘Keitt’	6	10	90	10	90	0.00	0.00	No comments
Leaf	Tommy A.	7	10	20	10	20	0.00	0.00	No comments
Leaf	Tommy A.	7	6	14	6	12	0.00	14.29	DPM, NS
Leaf	‘Keitt’	9	10	20	9	16	10.00	20.00	DPM, NS
Leaf	‘Keitt’	9	10	20	10	18	0.00	10.00	No symptoms
Leaf	‘Kent’	9	10	30	10	29	0.00	3.33	No symptoms
Leaf	‘Kent’	9	10	30	7	21	30.00	30.00	No symptoms
Leaf	‘Keitt’	9	10	30	8	15	20.00	50.00	DPM, NS

From this table, it could be noted that the highest proportion of discarded samples was in leaf assays. In these experiments, for the established monitoring time, there were samples (from ‘Keitt’ and ‘Kent’ cultivars) that did not develop visible symptoms of anthracnose in the diseased group, therefore, they were not included in the subsequent analysis as these samples were classified as outliers and had to be removed. Similarly, senescent samples were reported in two trials (‘Tommy Atkins’ and ‘Keitt’), which led to the discarding of these samples because the visible symptomatology in these samples corresponded to the drying (death) of the plant material and not strictly to anthracnose symptoms. Therefore, with the leaf assays, the incidence of anthracnose symptoms above 90 % (established as a metric for assay stop) was biased by removing the symptomless or senescent samples.

However, it is noteworthy that none of the samples used in the study were removed in the fruit experiments, indicating that the methodology used for this research on fruit was

the right one. Similarly, it was found that the inoculation method used in this research allowed the satisfactory development of anthracnose lesions in all cultivars. Also, the sample selection, acquisition and monitoring methodology employed during the development of the assays proved to be an adequate approach for the development of fruit and leaf experiments on mango as no cross-contamination between the control and diseased groups was reported: the sample quality was guaranteed in the assays (except the senescent samples) and the employed storage conditions allowed the proper sample preservation, making anthracnose lesion development possible.

6.1.3. Geometric and reflectance correction of hyperspectral images

Following the flowchart presented in Figure 3-10, reflectance and geometric correction of the hyperspectral images were carried out. Figures 6-5 and 6-6 depict the results of the reflectance correction, showing a circular region with pixels outside the permissible spectral range (white in the figures), so this region (and spectral signatures with this defect) were removed from the reflectance-calibrated hyperspectral image.

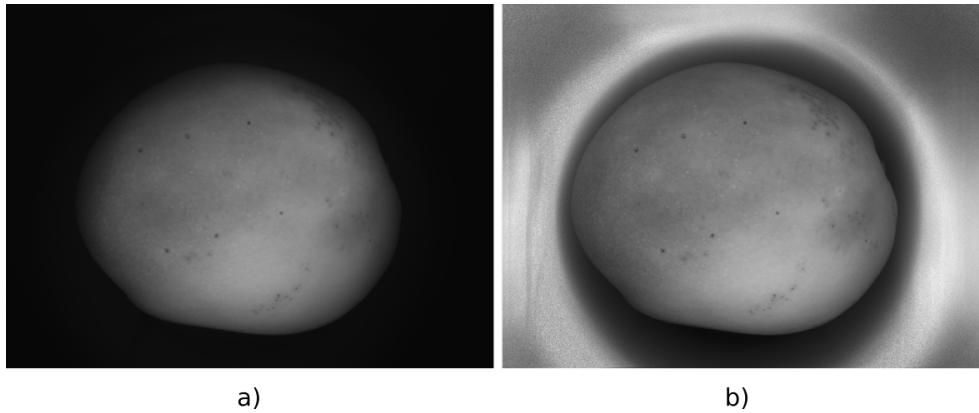


Figure 6-5. Hyperspectral image of a ‘Keitt’ mango fruit without correction (a) vs. reflectance corrected image (b) in the 760 nm band

Figures 6-7 and 6-8 show a schematic of applying the RANSAC algorithm for the geometric correction of the images. The figures show the extracted and matched interest points between the Vis and NIR images (coloured line), which allowed the transformation matrix to be obtained that corrected the spatial offset (horizontal displacement) found in the image acquisition system. Similarly, the incidence of cases where the fruit moved between filters was reported. However, with the application of the RANSAC algorithm, this movement was automatically corrected, thus ensuring the quality of the resulting hyperspectral image.

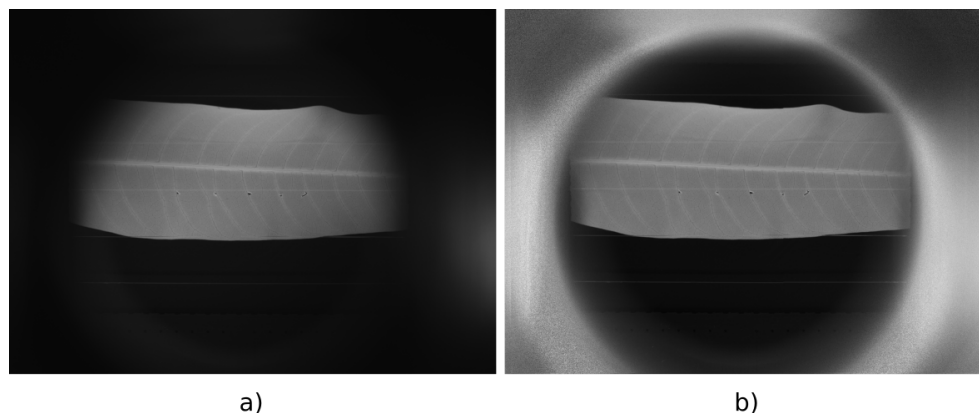


Figure 6-6. Hyperspectral image of a ‘Tommy Atkins’ mango leaf without correction (a) vs. reflectance corrected image (b) in the 760 nm band

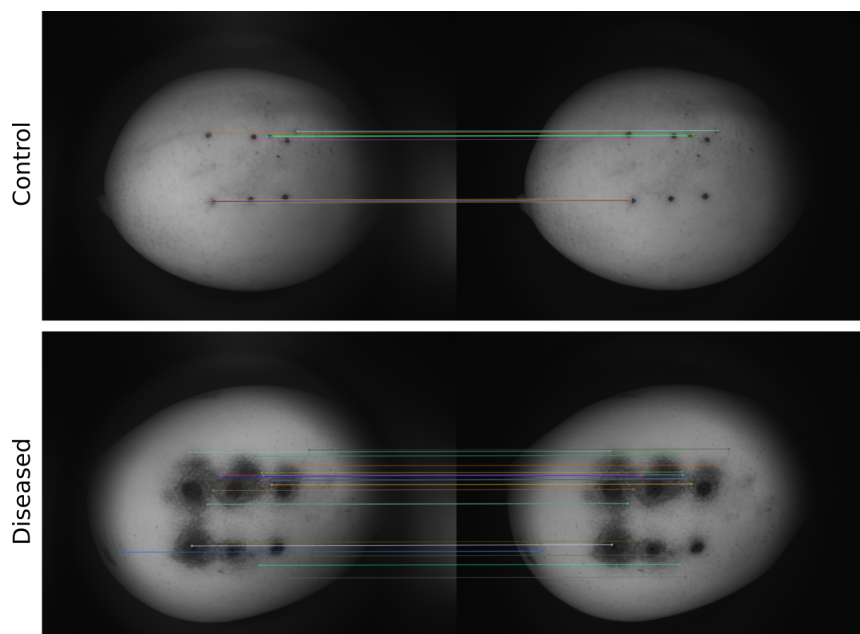


Figure 6-7. Matching algorithm and RANSAC algorithm for the computation of the transformation matrix between hyperspectral images of two sample fruits (one control and one diseased) of the *cv.* ‘Keitt’

After corrections were applied to all hyperspectral images from the assays, the scene’s background was removed by segmenting the object of interest (fruit and leaves) at the 700 nm wavelength. Figures 6-9 and 6-10 depict examples of control group samples of both a fruit and a leaf from the ‘Kent’ and ‘Tommy Atkins’ experiments, respectively, in the 760 nm spectral band and the RGB reconstruction of the hyperspectral image at 680, 550 and 480 nm.

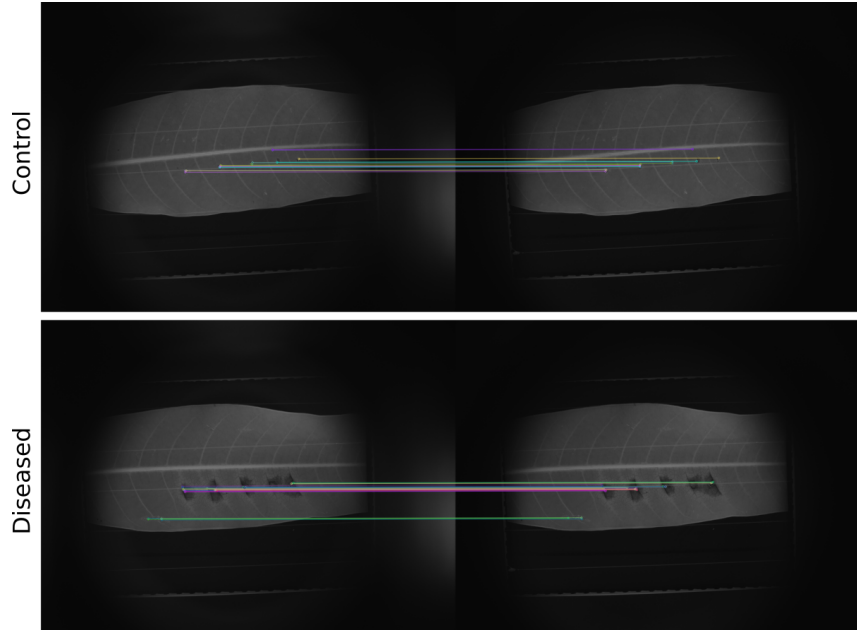


Figure 6-8. Matching algorithm and RANSAC algorithm for the computation of the transformation matrix between hyperspectral images of two sample leaves (one control and one diseased) of the *cv.* ‘Tommy Atkins’

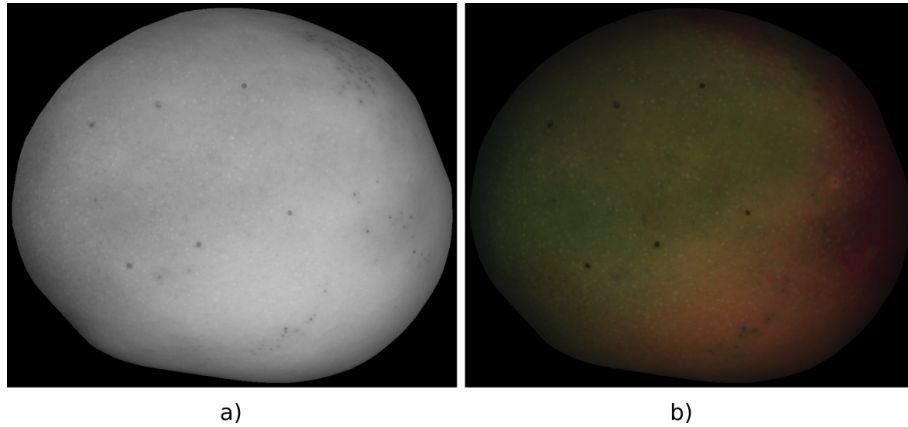


Figure 6-9. Hyperspectral image resulting from a ‘Kent’ mango fruit after geometric and reflectance corrections. Image (a) shows the 760 nm band, while image (b) shows an RGB reconstruction of the hyperspectral image at 680, 550 and 480 nm

Figure 6-9 shows that the resulting image for the fruits matches their characteristic shape (ellipsoid shape), while for the leaves (Figure 6-10), due to the magnetic holder, their shape is affected in the resulting hyperspectral image. However, for the study, the spectral analysis on the induced lesions and the segmentation in the leaf assays were satisfactory and could be performed automatically. As a result, a reduction in the size of the segmented hyperspectral image was achieved, allowing a reduction in the computational demand related to storage, memory loading and processing.

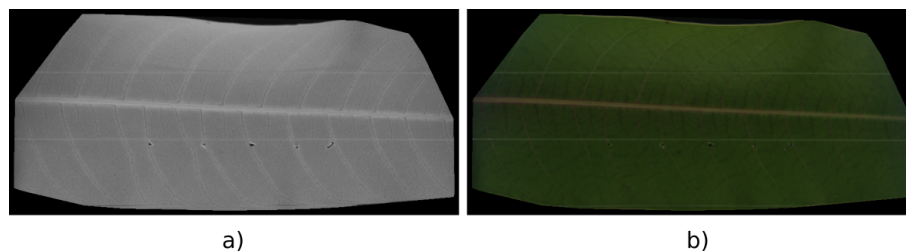


Figure 6-10. Hyperspectral image resulting from a ‘Tommy Atkins’ mango leaf after geometric and reflectance corrections. Image (a) shows the 760 nm band, while image (b) shows an RGB reconstruction of the hyperspectral image at 680, 550 and 480 nm

Table **6-2** details the number of hyperspectral images obtained after the acquisition and correction phase. The table shows the number of days for image acquisition (different from the number of days for sample monitoring) for each experiment. It also presents the number of hyperspectral images acquired by each LCTF filter (Vis and NIR filter), which, according to the geometric correction procedure, resulted in the total number of hyperspectral images (total HSI).

Table 6-2. Total hyperspectral images acquired and corrected per assay, cultivar and acquisition days

Assay ID	Type	cultivar	Hyperspectral image acquisition						
			Image days	Samples		LCTF Images		Total HSI	
				Control	Diseased	Control	Diseased	Control	Diseased
e1-kt	Fruit	‘Keitt’	8	10	45	160	720	80	360
e2-kn	Fruit	‘Kent’	8	10	50	160	800	80	400
e3-os	Fruit	‘Osteen’	5	10	90	100	900	50	450
e4-kt	Fruit	‘Keitt’	5	10	90	100	900	50	450
e5-to	Leaf	Tommy A.	5	10	20	100	200	50	100
e6-to	Leaf	Tommy A.	5	6	12	60	120	30	60
e7-kt	Leaf	‘Keitt’	7	9	16	126	224	63	112
e8-kt	Leaf	‘Keitt’	7	10	18	140	252	70	126
e9-kn	Leaf	‘Kent’	8	10	29	160	464	80	232
e10-kn	Leaf	‘Kent’	8	7	21	112	336	56	168
e11-kt	Leaf	‘Keitt’	8	8	15	128	240	64	120

Finally, based on the total number of HSI per assay reported in Table **6-2**, the ROI selection and projection procedure presented in section 3.1.6.5 was performed.

6.1.4. ROI selection and projection

After the hyperspectral image segmentation, the ROIs on the lesions were manually selected and marked with a punch and surgical needle on mango samples. This selection, as mentioned in section 3.1.6.5, was always performed on the last day of acquisition under the assumption that on this day the anthracnose lesion was in a visible stage (necrotic lesion) and had the largest size.

Figures 6-11 and 6-12 show the ROIs selected in two mango samples, one for a fruit of the *cv.* ‘Keitt’ and the other for a ‘Tommy Atkins’ mango leaf, respectively. As previously mentioned, 6 ROIs per fruit and 5 ROIs per leaf were obtained. These regions are denoted in cyan colour in both images and correspond to areas that comprise the highest percentage of necrotic lesions (grey-black) produced by anthracnose. Additionally, in this image, the area of the induced lesions (removed from the ROIs) is shown in blue, while the perspective points chosen for the ROI projection on all images of the same sample are shown as yellow asterisks. For fruits, at least 4 perspective points were selected, while for leaves at least 2 points were necessary.

Figures 6-13 and 6-14 depict the projections of the mango fruit and leaf ROIs shown in Figures 6-11 and 6-12, respectively. These figures also show the correct projection of ROIs on all images associated with a single sample, validating the approach chosen to automate extracting the same ROI in different images (corresponding to different days of acquisition). However, despite having reported a displacement error in the projection of the smaller ROIs on the leaves, since a smaller number of perspective points were used for these, this error did not affect the quality of the obtained signatures since the ROIs were always chosen inside the visible anthracnose lesion and never on its border or boundary.

After the consolidation and validation of the ROI projection on all samples from all experiments, the data were labelled based on the day of image acquisition for the inoculated groups, while for the control groups, only 1 class was defined as control samples remained sound during the entire experiment. Table 6-3 presents a compilation of the ROIs, defined classes and number of spectral signatures per class obtained after selecting, projecting and labelling the regions of interest in all the assays. An identifier was included to ensure the traceability of the data and avoid confusion among assays: experiments e1 to e4 correspond to mango fruits, while experiments e5 to e11 correspond to leaves. It is also important to mention that these data were used to develop the ML models, the wavelength selection and the spectral analysis, i.e. the study performed in this thesis was applied to the circular ROIs surrounding the lesions, as they represented (and guaranteed) the proper infectious process of anthracnose on the mango samples surface.

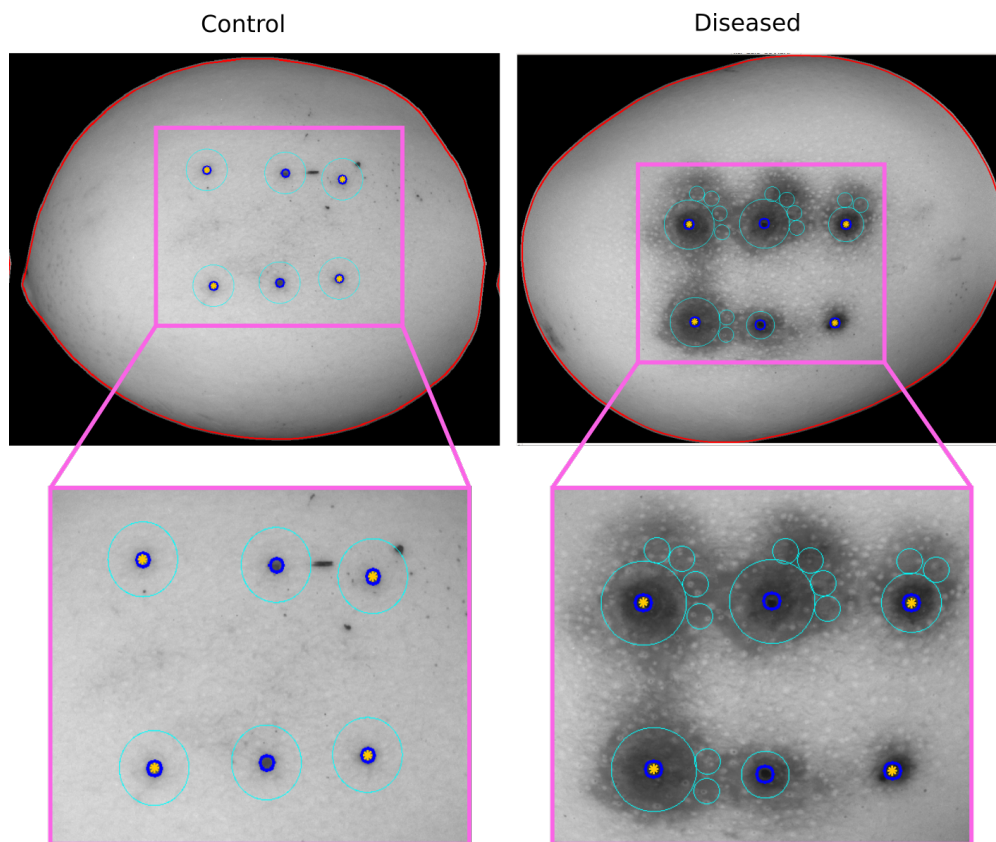


Figure 6-11. ROI selection in two mango fruits of the *cv.* 'Keitt' (control and inoculated) on day 6 of acquisition. The ROIs correspond to the areas delimited in cyan colour, while the blue ones the regions removed and the yellow asterisks the points selected for perspective correction

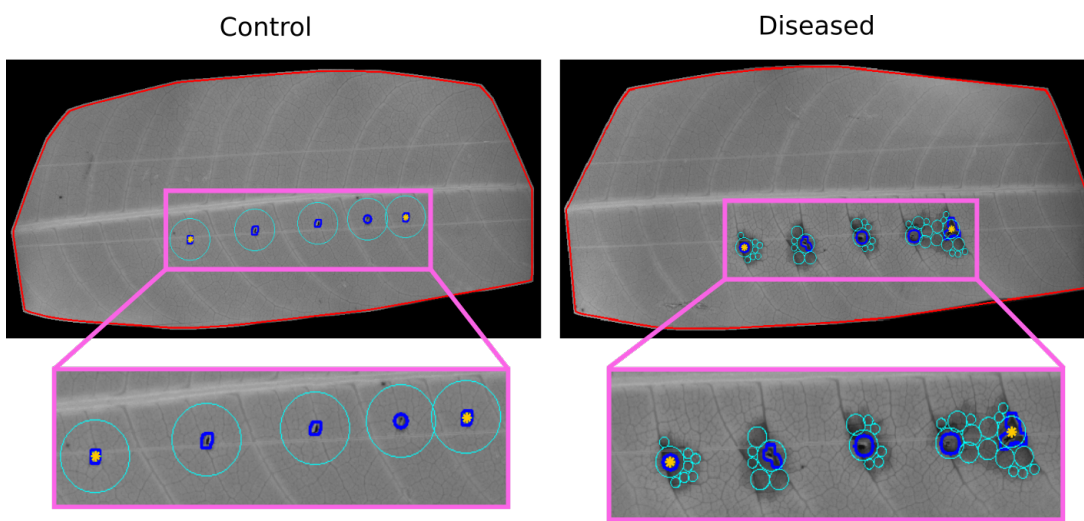


Figure 6-12. ROI selection in two mango leaves (control and inoculated) of the *cv.* 'Tommy Atkins' on day 7. The ROIs correspond to the areas delimited in cyan colour, while the blue ones the regions removed and the yellow asterisks the points selected for perspective correction

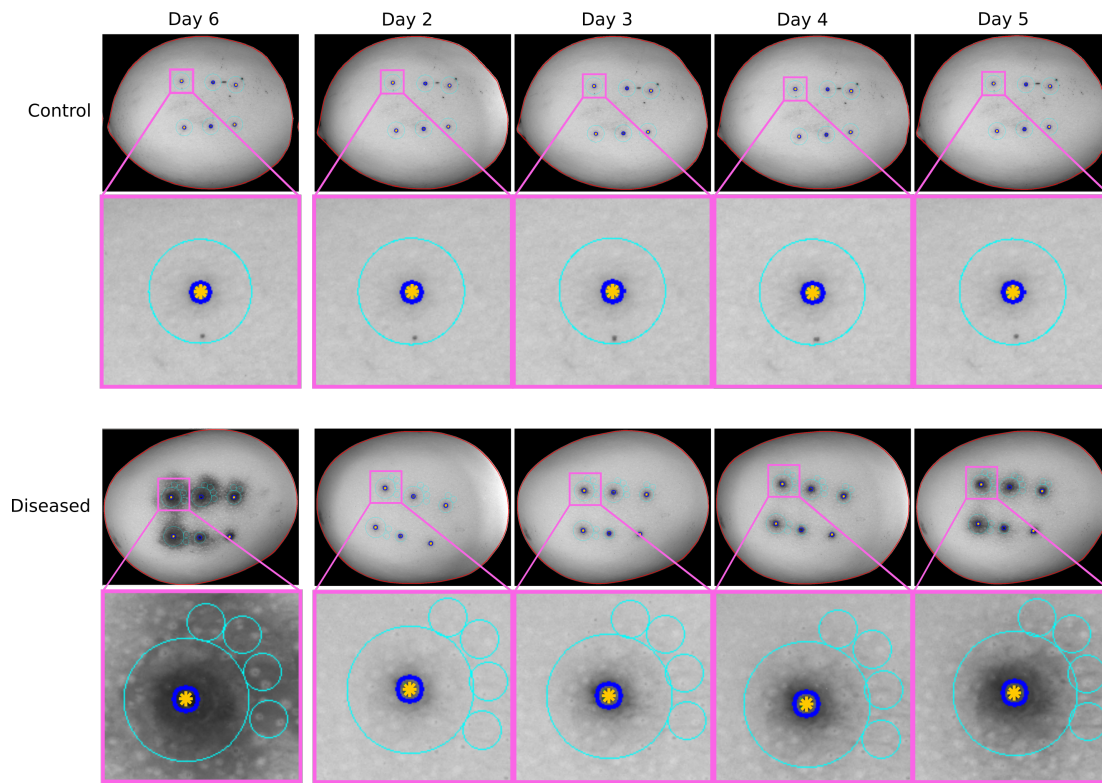


Figure 6-13. ROI projection on samples (control and inoculated) of 'Keitt' mango fruit during all days of acquisition. The projection for this assay was made on the 6 lesions caused by a surgical punch from day 6 (last day of acquisition) to days 2-5

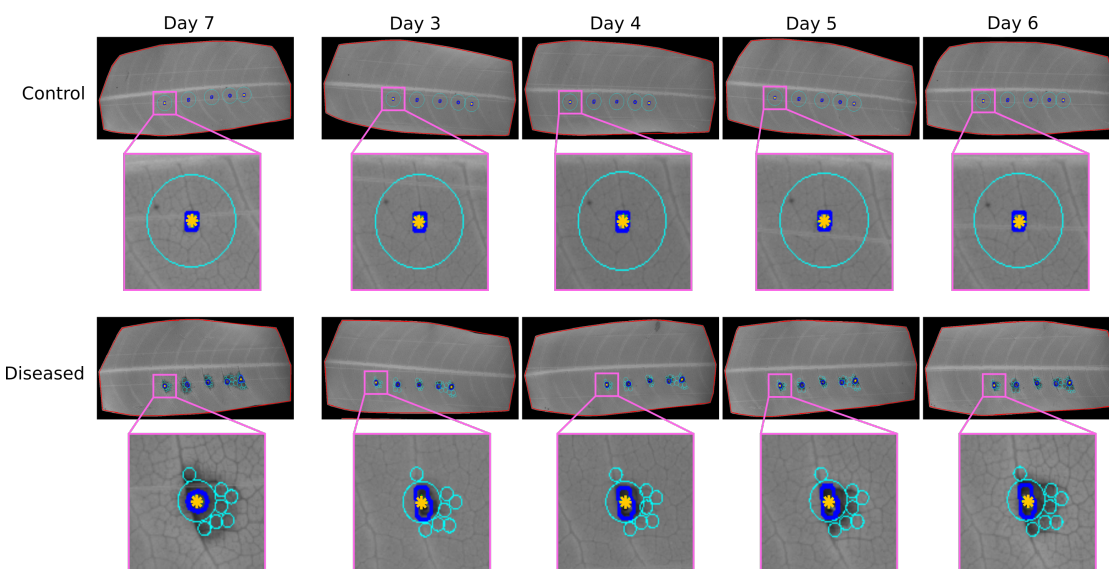


Figure 6-14. ROI projection on samples (control and inoculated) of 'Tommy Atkins' mango leaf during all days of acquisition. The projection for this experiment was made on the 5 surgical needle lesions from day 7 (last day of acquisition) to days 3-6

Table 6-3. Number of spectral signatures per class obtained from the ROIs selected in each assay and mango cultivar

Assay ID	Type	ROI		Classes	
		Control	Diseased	Labels	Spectra per class
e1-kt	Fruit	480	2160	ktc, kti d2, kti d3, kti d4, kti d5, kti d6, kti d7, kti d8, kti d9	15598, 15070, 15100, 15145, 14967, 15093, 15101, 14974, 15046
e2-kn	Fruit	480	2400	knc, kni d2, kni d3, kni d4, kni d5, kni d6, kni d7, kni d8, kni d9	14351, 14035, 13974, 13965, 13963, 13899, 13320, 14030, 13977
e3-os	Fruit	300	2700	osc, osi d2, osi d3, osi d4, osi d5, osi d6	23405, 24061, 23698, 24169, 24248, 24391
e4-kt	Fruit	300	2700	ktc, kti d2, kti d3, kti d4, kti d5, kti d6	24757, 22796, 22618, 22824, 22457, 22812
e5-to	Leaf	250	1875	toc, toi d3, toi d4, toi d5, toi d6, toi d7	10376, 10328, 10334, 10297, 10331, 10369
e6-to	Leaf	150	1440	toc, toi d3, toi d4, toi d5, toi d6, toi d7	17274, 18545, 18545, 18545, 18546, 18546
e7-kt	Leaf	315	3381	ktc, kti d3, kti d4, kti d5, kti d6, kti d7, kti d8, kti d9	1848, 1805, 1778, 1771, 1776, 1773, 1769, 1799
e8-kt	Leaf	350	3388	ktc, kti d3, kti d4, kti d5, kti d6, kti d7, kti d8, kti d9	14370, 15099, 10420, 15088, 15093, 15093, 15095, 15036
e9-kn	Leaf	400	6864	knc, kni d2, kni d3, kni d4, kni d5, kni d6, kni d7, kni d8, kni d9	24868, 25017, 25044, 24871, 24929, 24481, 24874, 25139, 24990
e10-kn	Leaf	280	1680	knc, kni d2, kni d3, kni d4, kni d5, kni d6, kni d7, kni d8, kni d9	2545, 2488, 2490, 2491, 2500, 2498, 2495, 2341, 2502
e11-kt	Leaf	320	2400	ktc, kti d2, kti d3, kti d4, kti d5, kti d6, kti d7, kti d8, kti d9	569, 536, 541, 543, 541, 555, 553, 447, 572

Table 6-3 shows that the e7-kn, e10-kn and e11-kn assays, despite having many ROIs, did not obtain many spectral signatures compared to the other leaf assays. Despite reaching its visible phase, this was because the lesion size development did not generate anthracnose lesions with a large enough area to extract sufficient spectral signatures. In addition, for the e7-kt and e11-kt assays, this situation was also because this cultivar is moderately resistant to anthracnose development. However, comparing the results of these two assays with the e6-to, it can be observed a considerable difference in the number of spectra per class despite having less ROIs: for the *cv.* ‘Tommy Atkins’, the development of anthracnose lesions on leaves was higher than on the *cv.* ‘Keitt’. Therefore, the e10-kn and e11-kt experiments were not used to train classification models, since the number of samples obtained was significantly

lower than the rest of the assays.

6.1.5. Spectral data pre-processing

Once the classes for each assay were defined, the Savitsky-Golay (SG), outlier and SNV filters were applied. These methods aimed to improve the quality of the data (with SG and outlier filters) and to avoid the bias of the feature selectors in the variable (wavelengths) selection and the bias of the classification models in the internal calibration of their weights.

Figures 6-15 and 6-16 show the spectral signatures resulting from the application of the outlier filter for a control mango sample (1 class) and a diseased mango sample (2 classes): Figure 6-15 shows the results for the ‘Keitt’ assay fruit, while Figure 6-16 shows the results for the ‘Tommy Atkins’ leaf sample. As can be noted, this filter managed to remove spectral signatures that did not correspond to the trend or variation of the class, which could improve (reduce) the intra-class variance by eliminating spectral signatures that, *a priori*, corresponded or fitted to another class. Furthermore, it can be noted the difference in the spectra variability within the Vis range of the control class of a fruit sample and a leaf: as a result of pigmentation changes in the fruits due to ripening, there was greater variability in this region, while the leaf pigmentation remained stable, meaning that spectral variability was reduced and conserved over time.

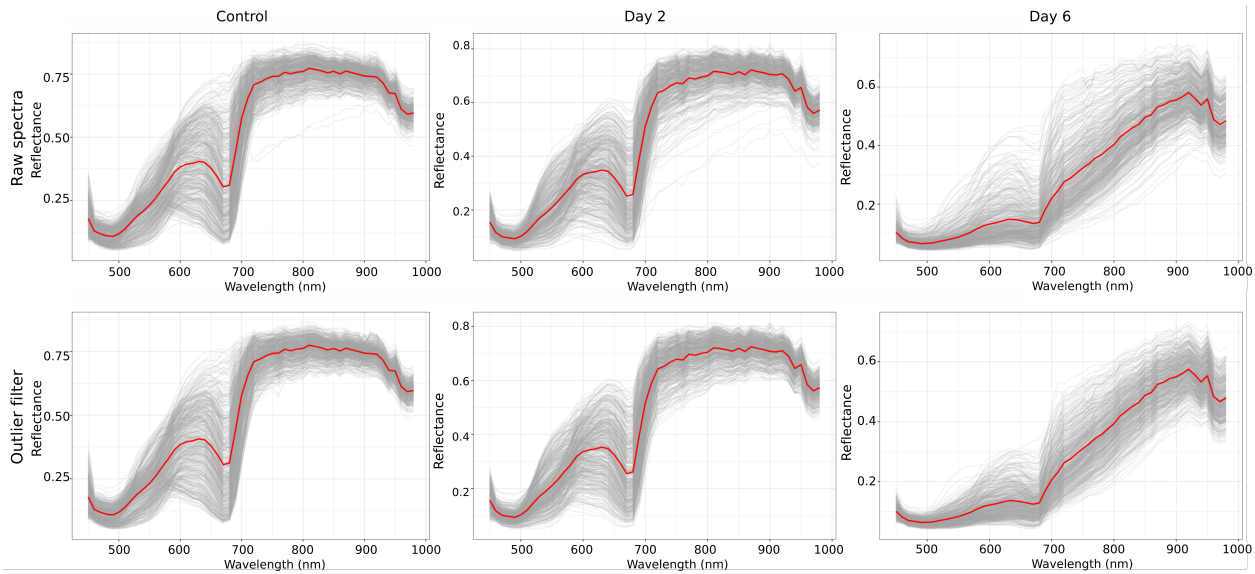


Figure 6-15. Group of spectral signatures of a control sample (control) and a diseased sample (days 2 and 6) of the ‘Keitt’ mango fruit experiment (mean spectra in red)

However, once the results of applying this filter for all experiments were obtained, it was found that the incidence of outliers did not exceed 10.25 % of the data. Table 6-4 shows the

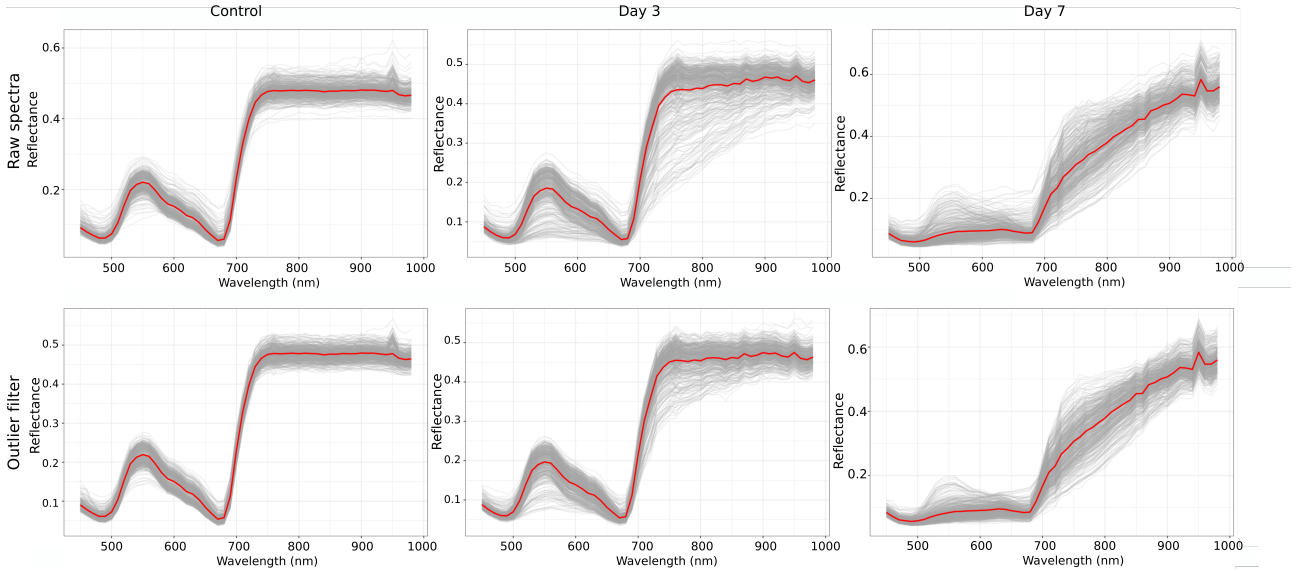


Figure 6-16. Group of spectral signatures of a control sample (Control) and a diseased sample (days 2 and 7) of the ‘Tommy Atkins’ mango leaf experiment (mean spectra in red)

number of outliers per class for all the analysed assays. It can be observed that, globally, the fruit assays had a lower outlier incidence (the highest reported in these assays was 5.09 %) compared to the leaf ones, which had an incidence between 7.34 % and 10.25 %. This finding can be explained by considering the changes induced by the pathogen according to the sample type (fruit or leaf) while it develops and the ROI selection strategy. The first element (changes due to the pathogen) is explained by the development of symptomatic lesions in fruits, which is gradual and homogeneous in the infected peel, while in leaves, given the low availability of nutrients, necrotic lesions develop from the first days of infection. Therefore, using this outlier detection strategy fixes this drawback, which, a priori, allowed for obtaining classes with less intraclass variability and greater inter-class variability. Nevertheless, the importance of this filter was compared against the raw dataset (without the application of filters) to establish whether this outlier filter, like the SG filter, improved the performance metrics of the AI-based detection models.

On the other hand, Figures 6-17 and 6-18 present a comparison between the mean signatures of the classes presented in Figures 6-15 and 6-16, respectively, the SG filter and the SNV method. As can be observed, the SG filter (Figures 6-17.b and 6-18.b) achieved the objective of smoothing the spectral signatures of each class, thus allowing the reduction of spectral level variations that could be associated with spectral noise. However, it was also considered that the cause of these small spectral variations could be related to the phenomenon under study (anthracnose infection). Therefore, the performance of the classification models with this SG filter was also compared against the raw data (without any filter) (Figures 6-17.a and 6-18.a), since the raw data were taken as a baseline reference to measure and quantify

the usefulness (and impact) of the implemented filters.

Table 6-4. Spectral signatures obtained using the outlier filter (W/O) versus the raw data

Assay ID	Classes									Resume	
e1-kt	knc	kni d2	kni d3	kni d4	kni d5	kni d6	kni d7	kni d8	kni d9	# spectra	Data
	15598	15070	15100	15145	14967	15093	15101	14974	15046	136094	Raw
	15021	14602	14578	14498	14401	14251	14731	14614	14481	131177	W/O
e2-kn	ktc	kni d2	kni d3	kni d4	kni d5	kni d6	kni d7	kni d8	kni d9		
	14351	14035	13974	13965	13963	13899	13320	14030	13977	125514	Raw
	13403	13305	13321	13324	13383	13419	12702	13151	13117	119125	W/O
e3-os	osc	osi d2	osi d3	osi d4	osi d5	osi d6					
	23405	24061	23698	24169	24248	24391				143972	Raw
	22717	23354	23331	23965	23093	23390				139850	W/O
e4-kt	ktc	kni d2	kni d3	kni d4	kni d5	kni d6					
	24757	22796	22618	22824	22457	22812				138264	Raw
	24122	21820	22141	22150	21606	21522				133361	W/O
e5-to	toc	toi d3	toi d4	toi d5	toi d6	toi d7					
	10376	10328	10334	10297	10331	10369				62035	Raw
	9620	9055	9399	9875	9960	9407				57316	W/O
e6-to	toc	toi d3	toi d4	toi d5	toi d6	toi d7					
	17274	18545	18545	18545	18546	18546				110001	Raw
	16356	16214	15830	18166	17942	17421				101929	W/O
e7-kt	ktc	kni d3	kni d4	kni d5	kni d6	kni d7	kni d8	kni d9			
	1848	1805	1778	1771	1776	1773	1769	1799		14319	Raw
	1687	1541	1581	1673	1709	1682	1612	1567		13052	W/O
e8-kt	ktc	kni d3	kni d4	kni d5	kni d6	kni d7	kni d8	kni d9			
	14370	15099	10420	15088	15093	15093	15095	15036		115294	Raw
	13021	13438	9231	13170	14383	14438	13902	13010		104593	W/O
e9-kn	knc	kni d2	kni d3	kni d4	kni d5	kni d6	kni d7	kni d8	kni d9		
	24868	25017	25044	24871	24929	24481	24874	25139	24990	224213	Raw
	22339	22224	21981	21176	19706	22058	23722	23991	24034	201231	W/O

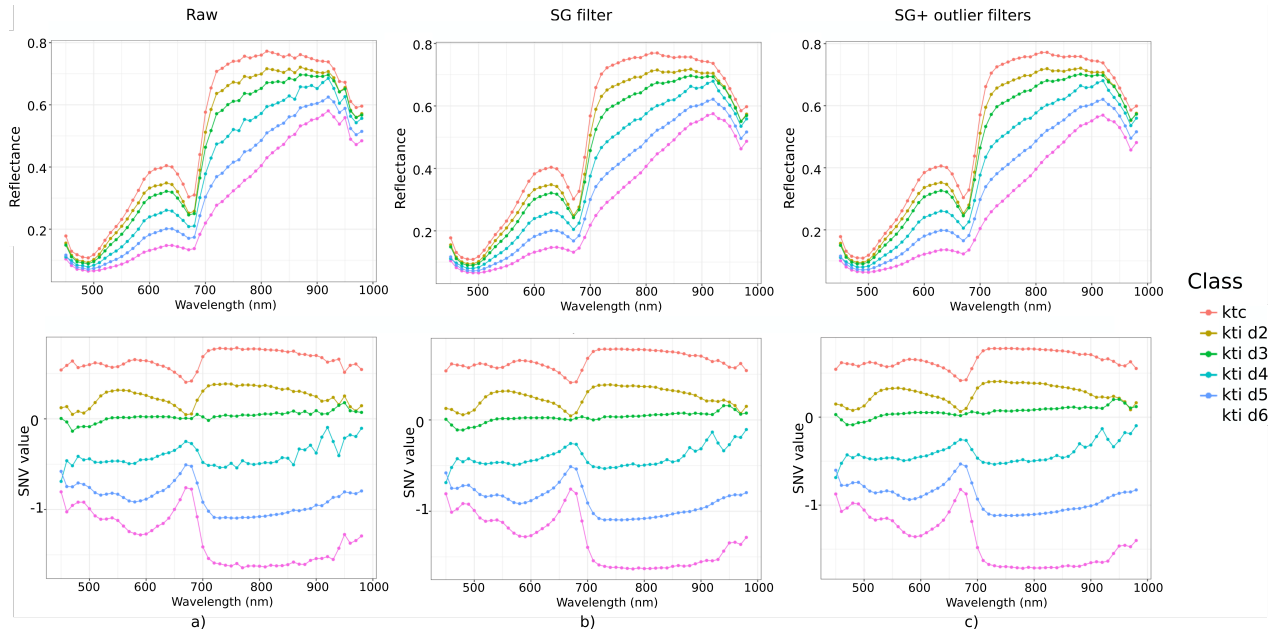


Figure 6-17. Comparison of the mean spectral signature of each defined class (Raw), the implemented pre-processing filters (the Savitzky-Golay and the outlier filter) and the SNV method for the ‘Keitt’ fruit mango data

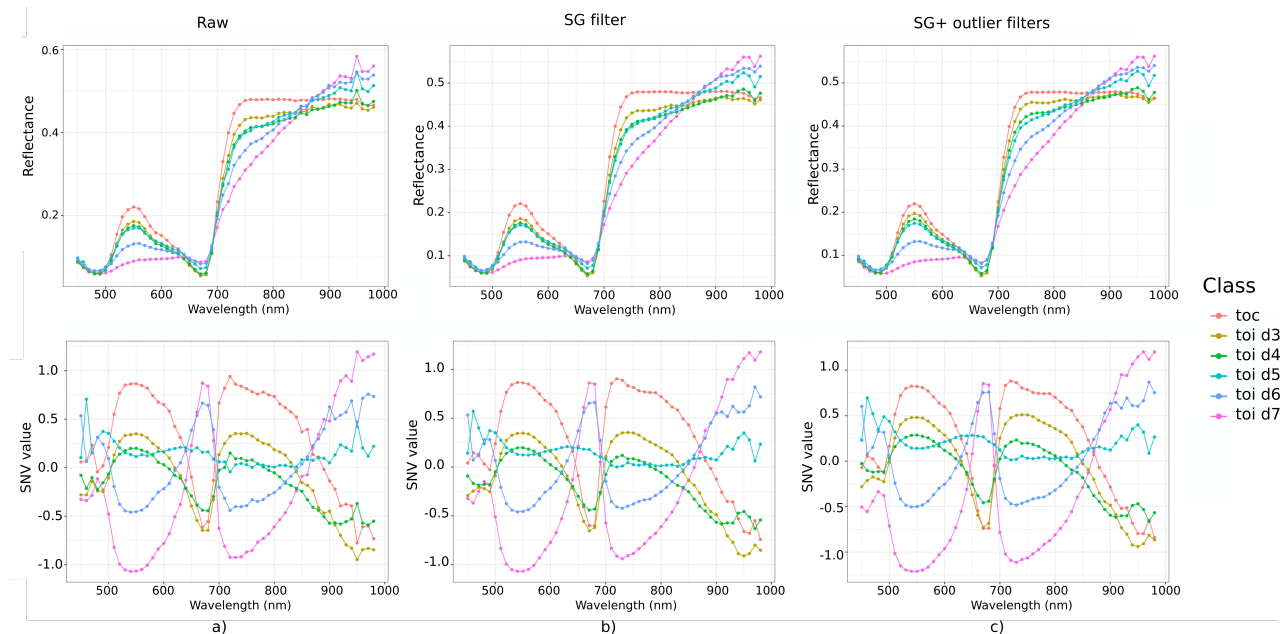


Figure 6-18. Comparison of the mean spectral signature of each defined class (Raw), the implemented pre-processing filters (the Savitzky-Golay and the outlier filter) and the SNV method for the ‘Tommy Atkins’ leaf mango data

Similarly, when applying the data standardisation to the spectral signatures (bottom line

graphs of Figures 6-17 and 6-18, it was noticeable that the characteristic shape of the signatures changed and the features that were not noticeable in the original scale (reflectance) were enhanced. These features are indeed the ones that contribute to the spectral characterisation of the classes and allow the unbiased calibration of AI-based models or feature selectors, since the information represented on this scale is mainly associated with differences induced by the object or material under study (anthracnose lesions on mango in this research). As can also be observed in these figures, the spectral signatures associated with the fruits presented a clear separation among the defined classes, which could allow the models to learn from the data and generalise the behaviour of the classes under study. In contrast, for the leaf assay classes, these differences were not appreciable (except in specific regions of the spectrum), so it was expected to be more difficult for the models to learn to differentiate them at the spectral level.

Finally, sections 3.3 and 4.3 present the performance metrics associated with applying SG and outlier filters in more detail, as their usefulness concerning their contribution to the performance of the AI-based classification models was quantified. In the sections mentioned above, the usefulness of these filters and their combinations is analysed and compared in depth.

6.2. Design and train machine learning models

6.2.1. Training and test dataset definition

Before the design and training of the AI models presented in section 4.1, a selection of the training (model calibration) and test (external validation) groups was made taking into account 1) the class balance, as this greatly affects the bias of the models; 2) the 70 % - 30 % ratio for group selection, where 70 % corresponds to training and 30 % to test; and 3) the selection is done on mango samples and not on spectral signatures.

The 70 % - 30 % ratio applied to the data collected in the tests was based on a standard, commonly applied in AI. It has been found that this ratio allows for the generalisability and validity of the results for the data used in the AI model training. Commonly, an 80 % - 20 % ratio is also applied in this field, being more used to perform internal validations of the models, i.e. the training group, with 80 % of the data, is divided into training and validation (internal) groups with a ratio of 70 % training - 10 % internal validation. This subdivision performs an internal measurement of the trained model before testing (or external validation). However, since in this research, it was chosen to perform a 3-fold cross-validation in the training group, the internal validation of the models was carried out.

On the other hand, sample selection was done on mango samples and not on the spectral

signatures, since in this research it was assumed that the spectral signatures of the single mango sample had the same experimental conditions and were subject to the same pathogen development conditions even when inoculation was carried out in 6 different locations. This consideration removed any eventual bias that might be presumed in the data, as the crossover of spectral signatures in the training and test groups was avoided. However, since the ROIs of a mango sample were not the same, as the development of lesions was unique to each sample, the number of spectral signatures per sample varied, causing a class imbalance that also affected the performance of the models.

Class balance is an impact factor in the training of AI models since a higher number of one class compared to the others biases the learning of the model to discriminate better between the large class over the rest. This can lead to overfitting to one class (whichever has more data) and reduce the discriminatory power of the other classes. The class imbalance was more intense in the leaf assays than in the fruit ones: the anthracnose lesions in leaves corresponded to ROIs scattered on the leaf, while in fruits it was focused on a single circular region, making the number of signatures per class more disparate in leaves. Therefore, a random sub-sampling of the control class (corresponding to the average of the inoculated classes) was carried out to improve the balance of classes, especially the control class for the inoculated classes. This sub-sampling was not considered for the inoculated classes, as the main interest was to assess as much data as possible related to anthracnose symptoms at different phases.

Finally, it is also important to mention that the selection of the training and test groups was randomised. Once these groups were obtained, they were digitally stored to guarantee that all the models in all their training, test or validation phases were subjected to the same data, allowing their comparison according to the obtained performance metrics.

6.2.2. Experimental conditions before model implementation and assessment

To establish the contribution of the outlier and the SG filter, 4 datasets using the training and test groups defined in section 6.2.1 were established per assay:

- **Raw+noSG:** meaning the data has no filters (nor SG filter nor outlier filter) and corresponds to the Raw data in Table 6-4
- **Raw+SG:** meaning the data has only been preprocessed with the SG filter
- **Fil+noSG:** meaning the data has only been preprocessed with the outlier filter, corresponding to the without outliers data in Table 6-4
- **Fil+SG:** meaning the data has been preprocessed with the outlier and SG filters

Also, models were evaluated using the biclass and multiclass classification approaches. The biclass approach (first approach) was intended to evaluate class by class the discriminatory power of the classification models, allowing for understanding and establishing more reliably the stage of development of anthracnose symptoms that can be detected. In contrast, the multiclass approach was included in this research to measure the discriminatory ability of a single model to discriminate all defined classes, i.e. the different stages of development of anthracnose symptoms in mango, ranging from the sound to the symptomatic stage of the disease. Both methods have advantages and disadvantages; however, their implementation in this study was based on the need to establish the day on which it is feasible to begin detecting anthracnose symptoms in mango and if it is possible to do so in a single detection model.

On the other hand, the recall, specificity and F1 metrics, which are calculated for 2 classes, in the multiclass approach express the average result obtained from all the possible two-class combinations. Likewise, as the ROC metric is more related to the discrimination between sound and disease data, this metric is only reported and discussed in the biclass classification approach.

Finally, each model was implemented in R, using the IDE RStudio v. 2022.02.3+492 on Ubuntu Bionic 20.04.5 LTS x64 with AMD® Ryzen 7 5800h processor, 64 Gb RAM + 64 Gb swap RAM and NVIDIA GeForce RTX 3050i 4Gb GPU, and following the official distributions and guidelines of each author.

6.2.3. Best fit dataset for AI-based models in mango fruit assays for anthracnose symptom detection

This study compared the control class (sound, healthy or anthracnose-free) against each disease class of an assay, leading to a classification model for each diseased class that was assessed. The models presented in section 4.1 were trained with the data presented in Table 6-4 for the mango fruit assays, i.e. assays e1-kt, e2-kn, e3-os and e4-kt. The model training followed the flowchart defined in Figure 4-2. Then, the results were consolidated for individual and global analysis to determine the potential of the AI models and the filters implemented for the early anthracnose detection in mango fruits. Accordingly, Figures 6-19 to 6-22 present the results of the accuracy, F1 and MCC metrics per dataset of the classification models implemented for the e1-kt, e4-kt, e2-kn and e3-os assays, respectively.

From the first *cv.* ‘Keitt’ assay results (Figure 6-19), it can be noted that the models achieved better performance when they were trained with diseased classes with more distinguishable anthracnose symptoms: from day 4 onwards, the models incrementally improved their performance in class discrimination, which meant the models could detect anthrac-

nose samples. However, on day 3, an outstanding performance of the models was observed, achieving similar performance metrics as on day 7. This finding was also reported in the e4-kt assay, where similar results were obtained. Figure 6-20 depicts the results of the e4-kt assay for the established datasets. Despite this experiment having fewer study days, it was reported that from day 4 onwards, the models managed to obtain remarkable performance metrics.

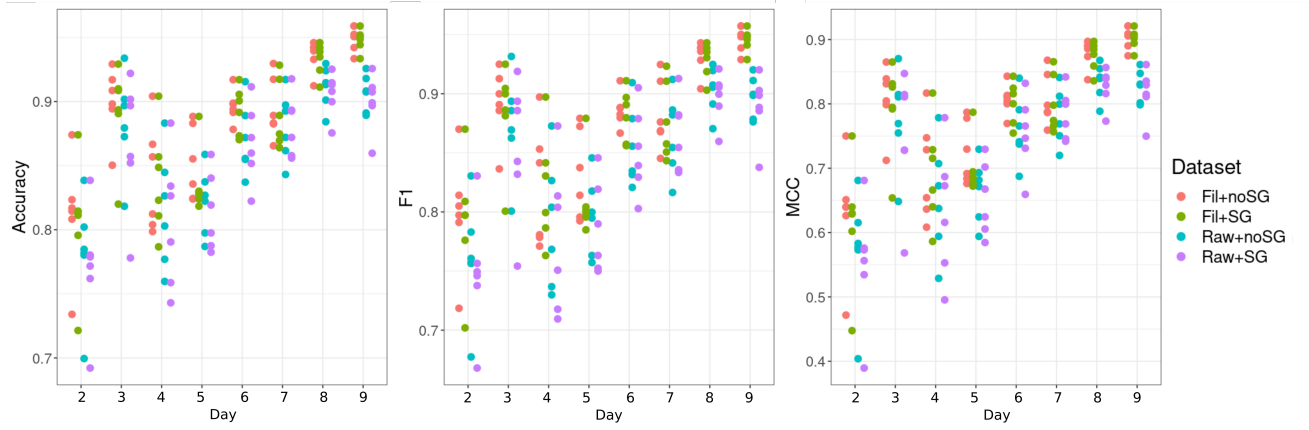


Figure 6-19. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the *cv.* 'Keitt' (e1-kt assay)

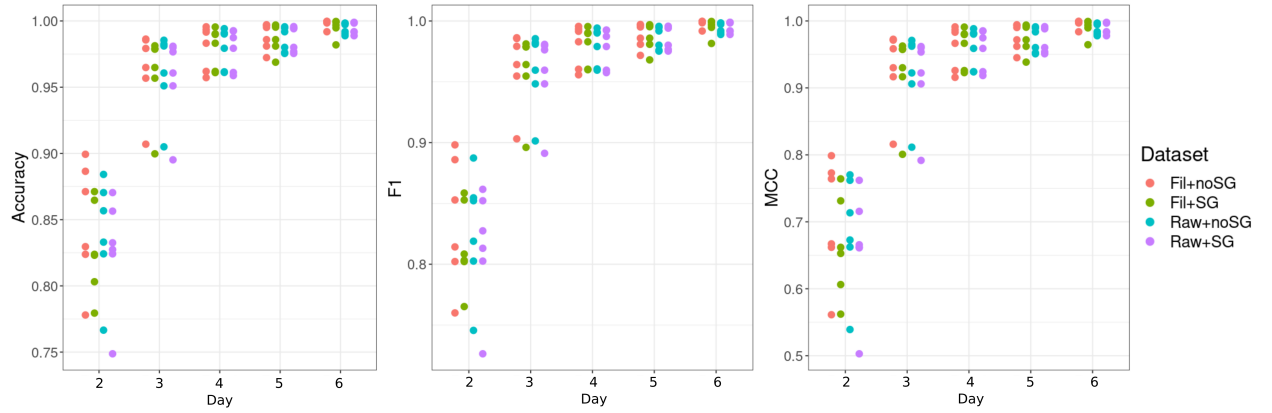


Figure 6-20. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the *cv.* 'Keitt' (e4-kt assay)

Focusing on the datasets, models trained with the outlier filter (Fil+noSG and Fil+SG) obtained better performance from initial days than the raw datasets (Raw+noSG and Raw+SG), see Appendix A.1.1 with full model metrics. Moreover, these datasets (Fil+noSG and Fil+SG) reduced their variance among the trained models, indicating a trend of better fit of the models to the data when outliers were removed. Likewise, when comparing the

Fil+noSG and Fil+SG datasets, there were no relevant variations in model performance when using or not the SG filter, so, from a practical point of view, there was no need to use the SG filter to improve model predictions.

As a difference between these assays (e1-kt and e4-kt), it was observed the magnitude of the scores obtained by the models within the first days of the study (days 3 to 4): while for the e1-kt assay, the minimum metrics were weighted above 0.8 for the outlier filter datasets, for the e4-kt experiment this minimum weighting was above 0.9.

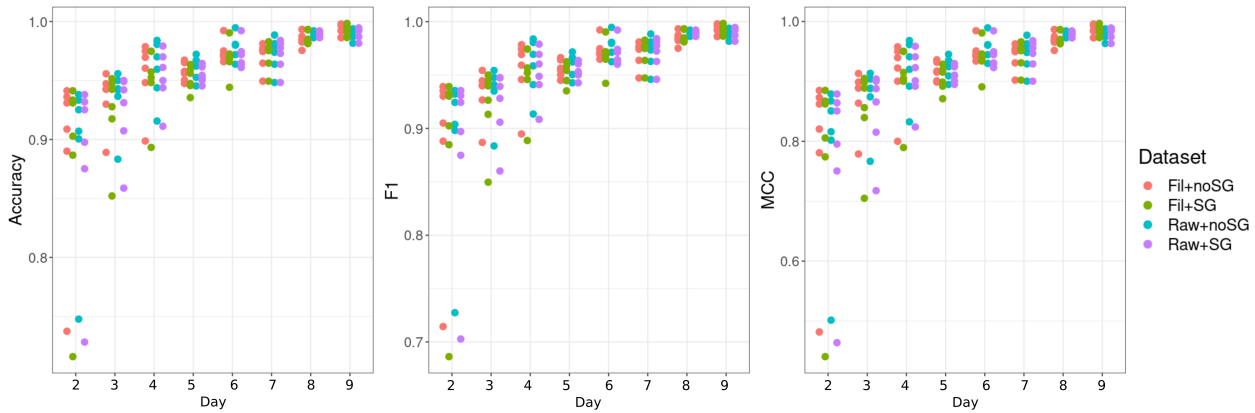


Figure 6-21. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the e2-kn assay (*cv.* ‘Kent’)

On the other hand, Figures 6-21 and 6-22 present the results associated with the e2-kn and e3-os assays, respectively. From the e2-kn experiment (*cv.* ‘Kent’), it was observed that in the first 2 days of the study, the datasets without the SG filter (Raw+noSG and Fil+noSG) showed the best scores. During days 4 to 7, the Raw+noSG dataset reported the best performance, followed by the Fil+noSG dataset, marking a difference with the results reported for the *cv.* ‘Keitt’.

Opposed to the e2-kn assay, the best datasets for the e3-os experiment (*cv.* ‘Osteen’) were Fil+noSG and Fil+SG, although the difference between the scores with the raw datasets was minimal. Additionally, the incremental behaviour in the model performance was not observed in this assay since the scores obtained from the first day of the study were higher than 0.95 and similar to the other days of the study. This result indicates that it is possible to discriminate between sound and diseased samples from the first days of infection, allowing the establishment of an early detection model of anthracnose in ‘Osteen’ mango fruits.

Finally, as the best-fit dataset for mango fruit cultivars, the Fil+noSG dataset was selected based on its regularity in 3 of the 4 assays (e1-kn, e3-os and e4-kn) and its outstanding performance in the e2-kn experiment (despite not being the best in this cultivar). Therefore,

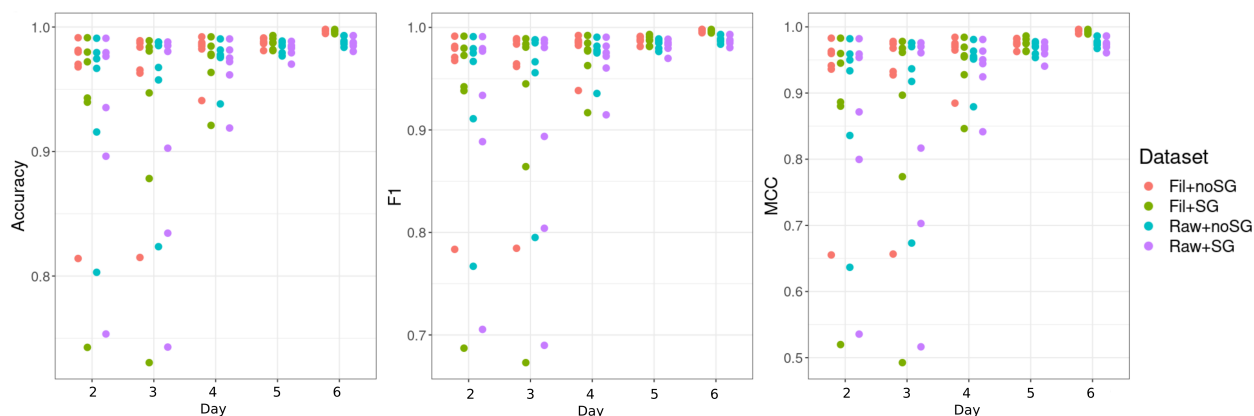


Figure 6-22. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the e3-os assay (*cv.* ‘Osteen’)

the following sections detail the assessment and selection of the best classification models for the mango experiments according to the results obtained with this dataset.

6.2.4. Biclass models to discriminate the anthracnose diseased classes in mango fruits

This section presents the results of the classification models trained with the Fil+noSG dataset. Given the number of models trained per class and assay, the figures and tables listed in this section present the results associated with the 6 best classification models (QDA, LDA, SLP, MLP, SVM and RF), while the remaining models (PLSDA, CART, Naive Bayes, GBM, XGB and BST) are presented in Appendix A.1.2. In addition, the results related to these 6 models, given the high volume of information, are presented per cultivar (e1-kt, e4-kt, e2-kn and e3-os). As a conclusion and discussion of this section, the best classification model for all the studied cultivars was selected.

AI-based biclass models applied to ‘Keitt’ assays

Figures 6-23 and 6-24 present the performance metrics of the biclass models (QDA, LDA, SLP, MLP, SVM and RF) for the e1-kt and e4-kt assays, respectively. From these figures, it is noted that on the first day of the study (day 2) the best models in both assays barely reached values close to 0.9 in the accuracy, F1 and MCC metrics. However, from day 3 onwards, the classification models started their upward trend of improving the class discrimination performance.

From the models trained for the e1-kt assay, the best model throughout this study was QDA. This model stood out with the best in accuracy, F1, MCC and recall metrics, indicating that its overall performance in discriminating between sound and diseased samples

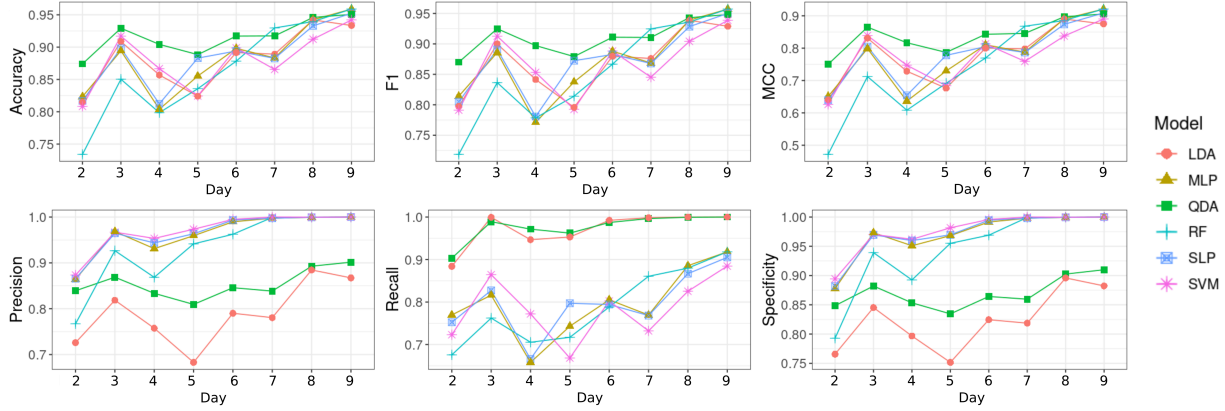


Figure 6-23. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the *cv.* 'Keitt' (e1-kt assay) and the Fil+noSG dataset

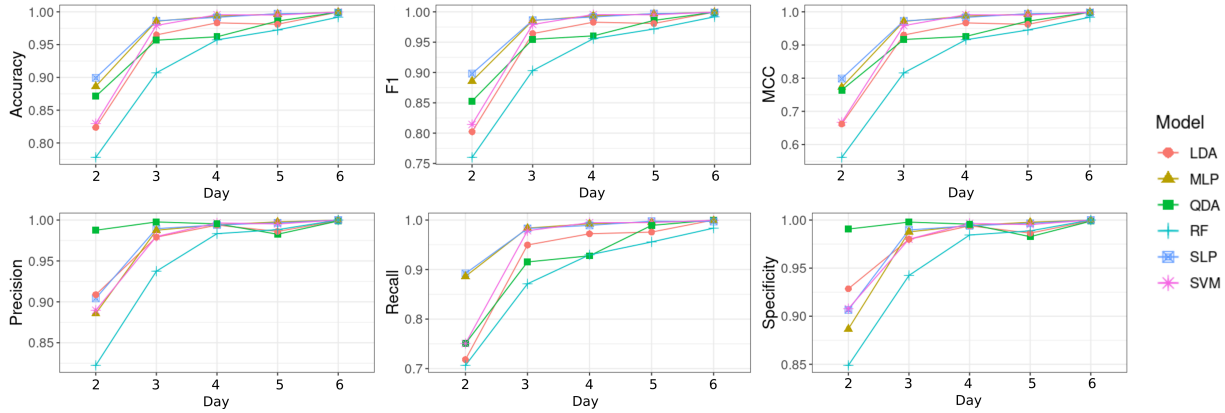


Figure 6-24. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the *cv.* 'Keitt' (e4-kt assay) and the Fil+noSG dataset

(regardless of the day of symptom development) was preserved. This model (with the LDA model) stood out as the one with the best recall metric, indicating its ability to detect when a mango fruit sample was anthracnose-free (positive class). Conversely, for the specificity metric, which measures the model's ability to detect the negative class (diseased class), the QDA model reduced its classification performance compared to the neural models (SLP and MLP) and SVM. However, the recall metric of these last three models during the first 6 days of the study was above 0.8, indicating a weakness in maintaining their discriminatory ability against mango fruit samples that temporarily evolve.

On the other hand, from the models trained with the e4-kt assay dataset, the neural techniques (SLP and MLP) stood out, followed by the QDA and SVM models. Specifically, the SLP model stood out from the first day of study in most metrics. This model obtained scores above 0.9 in accuracy, F1, precision, recall and specificity, while in the MCC metric, it obtained a score of 0.8 on day 2. However, from day 3 onwards, the SLP metrics remained

above 0.95, indicating that from this point onwards the model could discriminate between sound and diseased samples with high reliability. Following this model, with slightly lower metrics, was MLP, a model that showed the same behaviour as the SLP model, indicating that for this experiment the neural models had a good fit to the data.

The third-best model in the e4-kt assay was QDA. This model stood out in the specificity and precision metrics, surpassing the neural models and showing stability in the diseased class (negative class) detection during all days of this study. Conversely, the recall and MCC metrics on day 2 showed a slight issue in the correct classification of the positive (sound) class, as reported in the neural models. This slight issue, present in both assays, is explained by the fact that in the early stages of anthracnose development, the spectral variations are not considerable enough to be generalised (learned) by the models, which is why they had difficulties in discriminating between healthy and diseased samples. However, in the following days of the study (from day 3 onwards), as pathogen-induced physicochemical changes cause large spectral variations, the detection models could learn to detect these changes.

Therefore, and taking into account the behaviour reported in these assays, the QDA and SLP models are considered to be the best classification techniques applied to the problem of spectral detection of anthracnose diseased samples from the first days of disease development on fruits of the ‘Keitt’ mango cultivar. The hyperparameters selected for the SLP in the e4-kt assay were: *neurons* = 200, *epochs* = 502 on day 2, *neurons* = 100, *epochs* = 182 on day 3, *neurons* = 500, *epochs* = 176 on day 4, *neurons* = 50, *epochs* = 256 on day 5 and *neurons* = 5, *epochs* = 501 on day 6.

AI-based biclass models applied to ‘Kent’ assay

Figure 6-25 presents the performance metrics of the biclass models (QDA, LDA, SLP, MLP, SVM and RF) for the e2-kn experiment. The results obtained with this cultivar demonstrated the feasibility of discriminating against anthracnose diseased samples from the first days of infection, i.e., from day 2 after the pathogen inoculation. The models obtained, except RF, managed to obtain scores above 0.9 in the accuracy, F1, precision and specificity metrics, while for the recall and MCC metrics, this minimum threshold was lowered to 0.85 and 0.78, respectively. Despite the results obtained by the RF model in the first days, it showed a progressive improvement in its metrics, indicating that its predictive ability enhanced as anthracnose symptoms developed.

From the models with the best performance and stability during this experiment, the SVM model was the best performing, followed by the SLP and QDA models, see Appendix A.1.2. Particularly, SVM and SLP stood out for their stable and high performance in the precision and specificity metrics, which suggests that they have a greater ability to detect the presence of diseased samples regardless of the stage (day) of development. In contrast, for the recall

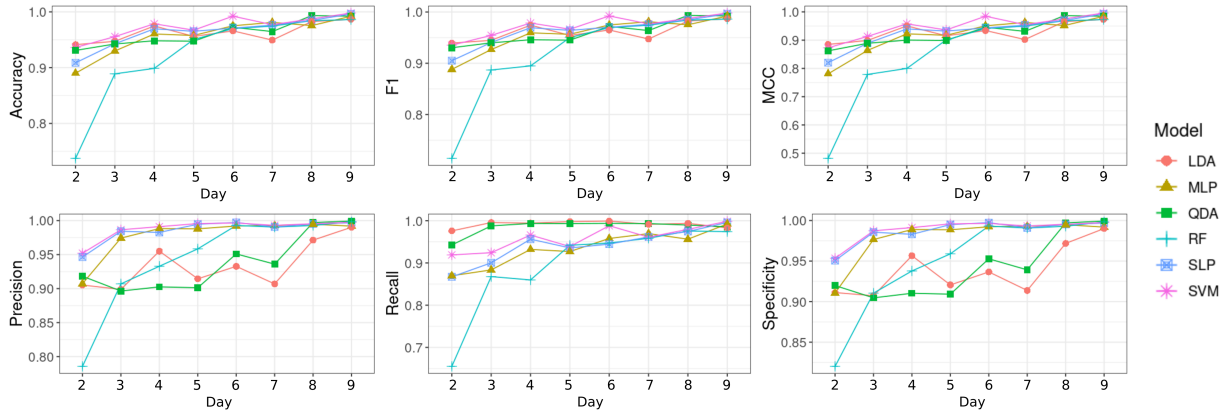


Figure 6-25. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e2-kn test of the ‘Kent’ mango cultivar and the Fil+noSG dataset

metric, the QDA model (with LDA) obtained the best score during the study, followed by the SLP and SVM models. Nevertheless, all the above-mentioned models achieved metrics above 0.9, demonstrating a high ability to discriminate between sound and diseased fruit samples.

Therefore, and considering that this study focused on the design of a classification model capable of detecting anthracnose-diseased samples, the SVM model was selected as the best classification model for this particular assay and cultivar. Finally, as a particularity of this model, it was reported that the hyperparameters reported per study day did not vary considerably. Therefore, the hyperparameter $cost = 0.1$ was chosen for this model in this cultivar.

AI-based biclass models applied to ‘Osteen’ assay

Figure 6-26 presents the performance metrics of the biclass models (QDA, LDA, SLP, MLP, SVM and RF) for the e3-os assay. From this figure, it was observed that for this cultivar, the classification models, except RF, achieved outstanding performance metrics during all days of study: for all 6 performance metrics, scores above 0.92 were achieved. This result demonstrated that for the *cv.* ‘Osteen’, the spectral changes induced by the infection from the first days were distinguishable by most of the classification models. Moreover, it also showed stability in the models, with variations in metrics less than 8 % (the range of the metrics varied between 0.92 and 1).

The best biclass classification models obtained for this cultivar were QDA, MLP and SLP, see Appendix A.1.2. In the accuracy, F1 and MCC metrics, which give an idea of the global behaviour of the model, scores higher than 0.95 were obtained, which is a sign of the discriminatory potential of these three models to distinguish healthy and diseased samples for this cultivar. For the accuracy, recall and specificity metrics, these models reported a slight

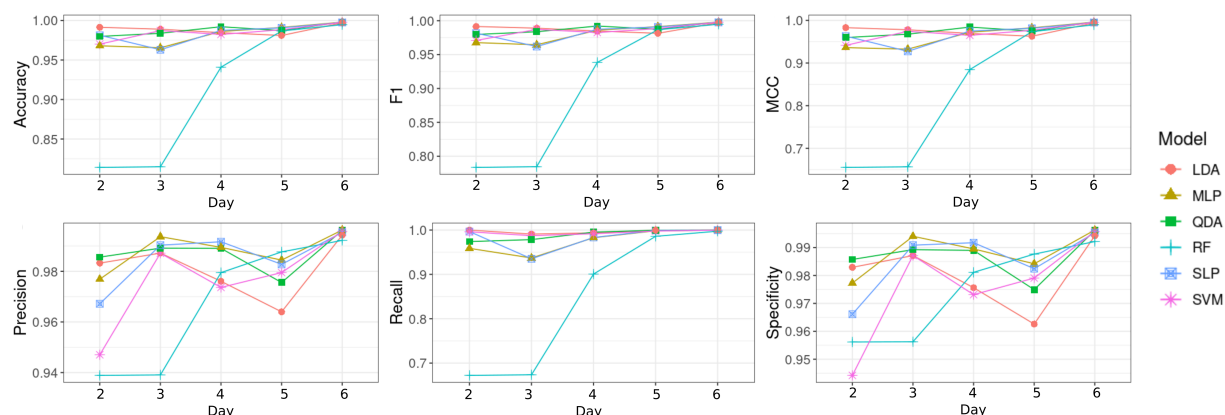


Figure 6-26. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e3-os assay (*cv.* ‘Osteen’) and the Fil+noSG dataset

decrease on some specific days which, in general, did not affect the overall performance of the models: on day 5, the values of these models showed a slight decrease of less than 2 %; while on day 3, the neural models showed a decline of 4 %.

Nonetheless, based on the results obtained and, particularly, the performance obtained by the QDA model, the best classification model, and therefore the best detection model, for the ‘Osteen’ mango cultivar in fruits was QDA. This model showed, one more time, its enormous potential and ability to differentiate sound samples from anthracnose-diseased samples from the first days of symptom development. The scores obtained on the 6 performance metrics showed that this model was close to the ideal classification model for this cultivar.

Best biclass model to detect anthracnose symptoms in mango fruit assays

In the previous sections, the analysis by cultivar of the best classification models was presented, with the QDA, SVM, QDA and SLP models being the most outstanding in the e1-kt, e2-kn, e3-os and e4-kt experiments, respectively. However, in this section, the area under the curve (AUC) metric, calculated from the ROC curve, is presented to establish the model with the best discriminatory power for the 3 cultivars and to evaluate the feasibility of having a unique model for the anthracnose detection in mango fruits.

The AUC metric allows a more concise but effective interpretation of the predictive ability of a model as it can readjust the decision threshold to improve the balance between recall and specificity metrics, i.e. to enhance the detection of sound and diseased samples. Table 6-5 reports the AUC metric scores for the QDA, SVM and SLP models in the mango fruit assays of the three cultivars. . From these results, the positive performance of the selected models was evidenced by the scores for this metric being above 0.8 for all experiments and days of study. In addition, it was validated that the QDA model presented the best performance in the e3-os assay, making it the best classification model for the *cv.* ‘Osteen’. Similarly, the

SLP model in the e4-kt assay was the best-fitting model, although the QDA model reports scores similar to those of the SLP model from day 3 of the study (the difference was less than 0.001).

On the other hand, for the e2-kn assay, where the SVM model had reported the best performance, for the AUC metric it was found that the QDA and SLP models obtained better results, indicating that if the decision threshold of these models were adjusted, the performance could be improved more than the SVM model. The differences between the QDA/SLP and SVM models varied on average by 0.014, while the differences between the QDA and SLP models varied by 0.001. Thus, the SVM model in this metric ranked substantially below the QDA and SLP models.

Similarly, for the e1-kt experiment, where the QDA model was reported to be the best in terms of performance, it was observed that between days 2 and 4 (considered the pre-symptomatic phase) the QDA model obtained the best AUC result, while on the remaining days (days 5 to 9) it was the SLP model. However, it was also observed that the difference in the AUC metric of these two models was greater on the first days (days 2 to 4) than on the rest (days 5 to 9). Therefore, for the objective of finding a model for early detection of anthracnose symptoms in mango fruits, the QDA model fits better for this purpose.

Table 6-5. Area under the ROC curve (AUC) of the QDA, SVM and SLP models for the e1-kt, e2-kn, e3-os and e4-kt fruit assays, and by day of study

Diseased class	Assay e1-kt			Assay e2-kn			Assay e3-os			Assay e4-kt		
	QDA	SVM	SLP	QDA	SVM	SLP	QDA	SVM	SLP	QDA	SVM	SLP
Day 2	0.907	0.809	0.869	0.956	0.936	0.955	0.998	0.970	0.995	0.951	0.829	0.966
Day 3	0.951	0.917	0.944	0.975	0.956	0.979	0.999	0.987	0.998	0.998	0.979	0.999
Day 4	0.931	0.867	0.867	0.990	0.979	0.996	1.000	0.982	0.999	0.999	0.996	1.000
Day 5	0.909	0.825	0.936	0.990	0.967	0.990	1.000	0.988	0.999	0.998	0.995	1.000
Day 6	0.933	0.899	0.951	0.998	0.992	0.998	1.000	0.998	1.000	1.000	1.000	1.000
Day 7	0.940	0.866	0.979	0.995	0.977	0.999						
Day 8	0.952	0.913	0.993	1.000	0.988	0.999						
Day 9	0.964	0.942	0.999	1.000	0.994	1.000						

Based on the above, it was determined that there were 2 classification models with high performance for the anthracnose symptom detection at different stages (days) for the mango cultivars ‘Kent’, ‘Keitt’ and ‘Osteen’: the QDA and SLP models combined with the outlier filter. As a special feature of these classification models, QDA has the computational advantage of being a technique that can be calibrated and readjusted in a shorter time than the SLP model, which allows its use and deployment in real-time applications that require constant updating of the training and calibration database. Consequently, this discriminant model, QDA, is the best fitting anthracnose detection model for the three mango cultivars.

6.2.5. Multiclass models for detection of anthracnose symptoms in mango fruit

The models presented in section 4.1 were trained with the Fil+noSG dataset selected in section 6.2.3 for the e1-kt, e2-kn, e3-os, and e4-kt assays. Following the flowchart shown in Figure 4-2, model training was carried out and test results were analysed to determine the potential of multi-class AI models for the early detection of anthracnose symptoms in mango fruit.

In addition, the models reported in this section corresponded to the neural models (SLP and MLP), SVM and QDA since they achieved the best performance metrics in all fruit experiments, metrics above 0.9. The other models can be consulted in Appendix A.2.

AI-based multiclass models applied to ‘Keitt’ assays

Tables 6-6 and 6-7 report the results associated with the calibrated and trained classification models for the e1-kt and e4-kt experiments of the *cv.* ‘Keitt’ in mango fruit. From these tables, it can be observed that the reported models stood out in this cultivar. Among the metrics of these models, specificity was the most remarkable, with scores above 0.98. This metric, which expresses the ability of the model to detect a negative class, indicated the correct detection of anthracnose-diseased samples regardless of their stage of development. In contrast, the recall metric, which measures the ability of the model to detect the positive class, differed from the specificity scores, as this metric was around 0.91. However, having both metrics above 0.9, it can be stated that the four models could adequately discriminate between both sound and diseased samples (regardless of the stage of the infection).

Table 6-6. Performance metrics of the classification models for the *cv.* ‘Keitt’ e1-kt assay

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
MLP	0.966	0.967	0.996	0.966	0.963	0.968
SLP	0.952	0.952	0.994	0.952	0.947	0.957
SVM	0.940	0.941	0.993	0.940	0.933	0.940
QDA	0.926	0.926	0.991	0.923	0.916	0.927

Specifically, the best models obtained for the e1-kt assay were MLP with *neurons* = 200, *layers* = 2, and *epochs* = 100, and SLP with *neurons* = 200 and *epochs* = 382, while in the e4-kt were MLP with *neurons* = 200, *layers* = 2 and *epochs* = 100, and SVM with *cost* = 0.1. However, it is noteworthy that the difference between the SLP and SVM models in the e4-kt assay was smaller than the same models for the e1-kt one: for the first assay, the difference in performance was close to 2 %, while for the e4-kt assay was less than 0.7

%. Therefore, the SLP model with $neurons = 200$ and $epochs = 382$ was selected as the second-best classification model for this cultivar.

Table 6-7. Performance metrics of the classification models for the *cv.* ‘Keitt’ e4-kt assay

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
MLP	0.952	0.951	0.990	0.951	0.943	0.955
SVM	0.924	0.925	0.985	0.925	0.910	0.925
SLP	0.921	0.919	0.984	0.918	0.909	0.935
QDA	0.910	0.910	0.982	0.910	0.892	0.911

The results obtained with the neural models demonstrate the stability and great potential of the ANNs (especially the MLP) to spectrally characterise the classes obtained from this mango cultivar, which similarly indicates its capacity to detect anthracnose symptoms from stage 0 (anthracnose-free symptoms) to a late stage (visible anthracnose symptoms). Likewise, the results obtained with the SVM model showed that it fulfilled the features to discriminate the classes established in these experiments, showing this model as an alternative detection model.

The fourth model in the experiments was the QDA model, which achieved performance metrics above 0.91 in most of them. This discriminant model was proposed as an alternative to neural and SVM models due to its fast implementation in real-time applications. This result differs from the results obtained in the biclass implementation where QDA was the best model, however, given the initial objective of this research to design and calibrate a classification model with the best performance for detecting anthracnose symptoms in mango, the model that outperformed the rest of the techniques for its outstanding performance in both assays for the *cv.* ‘Keitt’ was MLP with $neurons = 200$, $layers = 2$, and $epochs = 100$, obtaining performance metrics higher than 0.943 in both experiments.

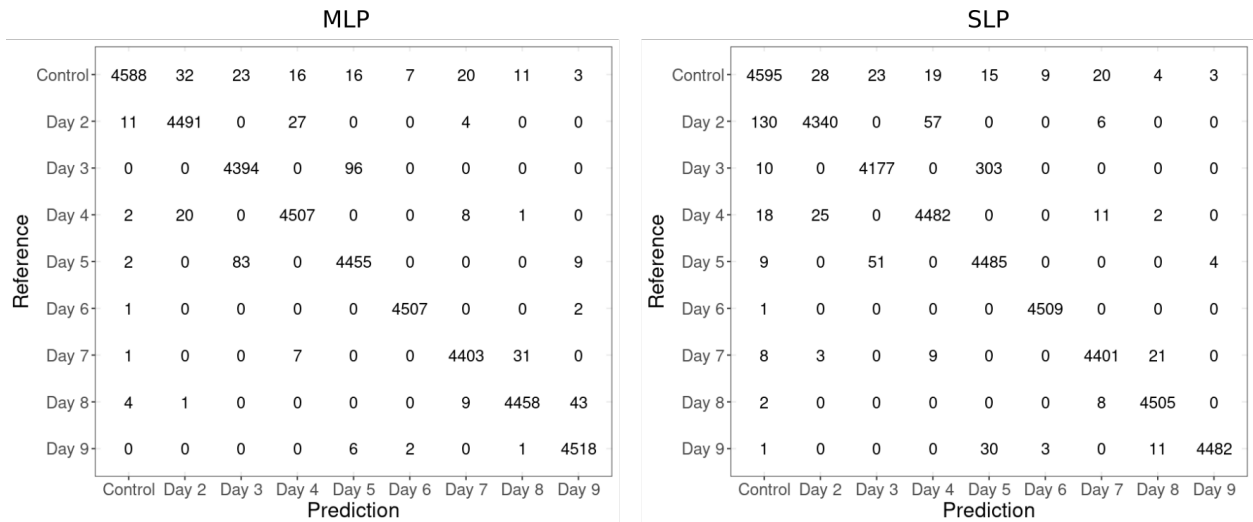
AI-based multiclass model applied to ‘Kent’ assay

Table 6-8 reports the results associated with the classification models for the e2-kn experiment, while Figure 6-27 presents the confusion matrices of the 2 best models. When selecting the best models for this cultivar, the neural models were once again the best classifier, with the MLP model being the best performing, although the SLP model was very close. As a particularity, for this assay, these models obtained different hyperparameters from those reported for the *cv.* ‘Keitt’: for the MLP model, the hyperparameters were $neurons = 200$, $layers = 3$ and $epochs = 100$, and SLP with $neurons = 100$ and $epochs = 477$.

Table 6-8. Performance metrics of the *cv.* ‘Kent’ e2-kn assay detection models

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
MLP	0.982	0.982	0.998	0.982	0.980	0.982
SLP	0.971	0.972	0.996	0.971	0.968	0.972
SVM	0.951	0.951	0.994	0.951	0.945	0.951
QDA	0.935	0.936	0.992	0.934	0.928	0.937

On the other hand, when comparing the confusion matrices of these two models (Figure 6-27), the SLP model accumulated a higher number of misclassified samples in the first row (multi-class false negatives -MFN-) and first column (multi-class false positives -MFP-), even though the performance metrics reported in Table 6-8 indicate that its performance is similar to the MLP model. Adding the MFN and MFP values of the confusion matrix from day 2 to day 9, MLP accumulated 149 misclassified samples compared to the 300 reported by SLP, which notably differentiates the performance between these neural models. The above finding indicated that the SLP model confused healthy with diseased samples (or MFP) and diseased with healthy samples (or MFN) with greater recurrence than the MLP model, being the MFN error the most critical as it implies the non-detection of the disease in any of its infection stages.

**Figure 6-27.** Confusion matrices of the MLP and SLP classification models for the e2-kn test

Consequently, as the best multiclass detection model for the *cv.* ‘Kent’ (e2-kn assay), the MLP model with *neurons* = 200, *layers* = 3, and *epochs* = 100 and the outlier filter was chosen as the best classification and fitting model for the data collected for this cultivar. This result, similar to the *cv.* ‘Keitt’, differs in the best model selection from the biclass im-

plementation, where the QDA model showed better performance in the biclass classification of sound and diseased samples by day of acquisition.

AI-based multiclass model applied to ‘Osteen’ assay

Table 6-9 reports the results associated with the classification models for the e3-os experiment. From this table, it was observed that the neural models did not manage to place themselves in the top positions since, particularly, SVM and the discriminant models (QDA and LDA) obtained better results than the neural models: QDA outperformed the SLP model, LDA outperformed MLP and even PLSDA obtained performance metrics similar to MLP (differences in their metrics were less than 0.5 %). The best classifier was SVM, obtaining in all its metrics scores above 0.96 with the hyperparameters $cost = 10$, $gamma = 0.18$.

Table 6-9. Performance metrics of the ‘Osteen’ e3-os assay

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
SVM	0.969	0.969	0.994	0.969	0.963	0.969
QDA	0.955	0.955	0.991	0.955	0.946	0.955
SLP	0.948	0.947	0.990	0.947	0.940	0.955
MLP	0.931	0.929	0.986	0.928	0.921	0.946

Comparing the confusion matrices of the SVM and QDA models, see Figure 6-28, the MFP and MFN for both models were 122 vs 142 and 129 vs 209, respectively. This result confirmed the remarkable multiclass performance of the SVM model compared to the QDA model, since in the MFN measure, which quantifies the misclassification of diseased samples as healthy, the difference between the models was more than 35 %. On the MFP measure, which measures the misclassification of healthy samples as diseased, this difference was reduced (122 vs 142 for SVM and QDA, respectively); however, the SVM model demonstrated its effectiveness in correctly classifying the data from this assay.

Therefore, as the best multiclass detection model for the *cv.* ‘Osteen’ (e3-os assay), SVM with the hyperparameters $cost = 10$, and $gamma = 0.18$ was chosen as the best classifier. However, as a disadvantage of this model, its training time was longer than that required by neural and discriminant models. This issue could be overcome with the implementation of QDA, which, despite its limitations compared to SVM, also had remarkable metrics and could be quickly deployed and calibrated in real-time applications.

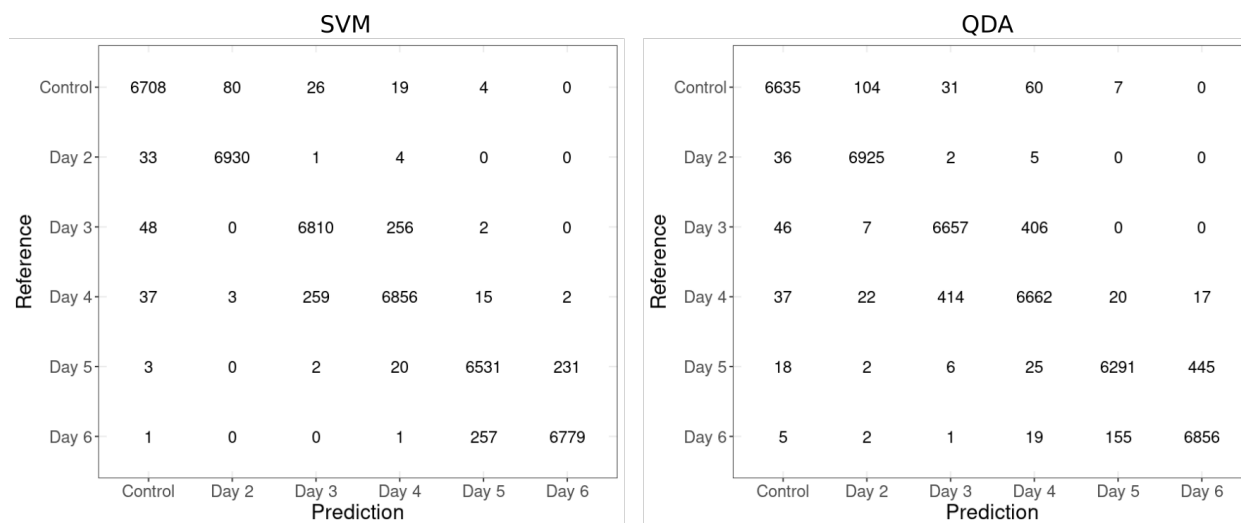


Figure 6-28. Confusion matrices of the SVM and QDA classification models for the e3-os test

6.2.6. Best fit dataset for AI-based models in mango leaf assays for anthracnose symptom detection

As presented in section 6.2.3, this study compared the control class (sound, healthy or anthracnose-free) against each diseased class of the leaf assays, obtaining a classification model for each assessed diseased class. The models presented in section 4.1 were trained with the data presented in Table 6-4 for the mango leaf experiments, i.e. the e5-to, e6-to, e7-kt, e8-kt, e9-kn, e10-kn assays. Following the flowchart defined in Figure 4-2, the training of the models was performed, and then the results were consolidated for individual and global analysis to determine the potential of AI models for early detection of anthracnose symptoms in leaves of the three mango cultivars. Figures 6-29 to 6-33 present the results of the accuracy, F1 and MCC metrics per dataset of the classification models implemented for the tests e5-to, e6-to, e7-kt, e8-kt and e9-kn.

From the *cv.* ‘Tommy Atkins’ results (Figures 6-29 and 6-30), it can be noted that the models achieved better performance when they were trained with diseased classes with more distinguishable anthracnose symptoms, i.e. metrics reduced their variance among the models (metrics were more concentrated) while increasing their scores when late classes were used. Moreover, from day 3, models achieved a remarkable performance in discriminating between diseased and sound samples.

Focusing on the datasets, models trained with the outlier filter (Fil+noSG and Fil+SG) obtained the best performance from the initial days (days 3 and 4) in both assays, although the Raw datasets (Raw+noSG and Raw+SG) achieved good metrics. In addition, as the days passed, the variance among model metrics decreased, indicating the usefulness of this strat-

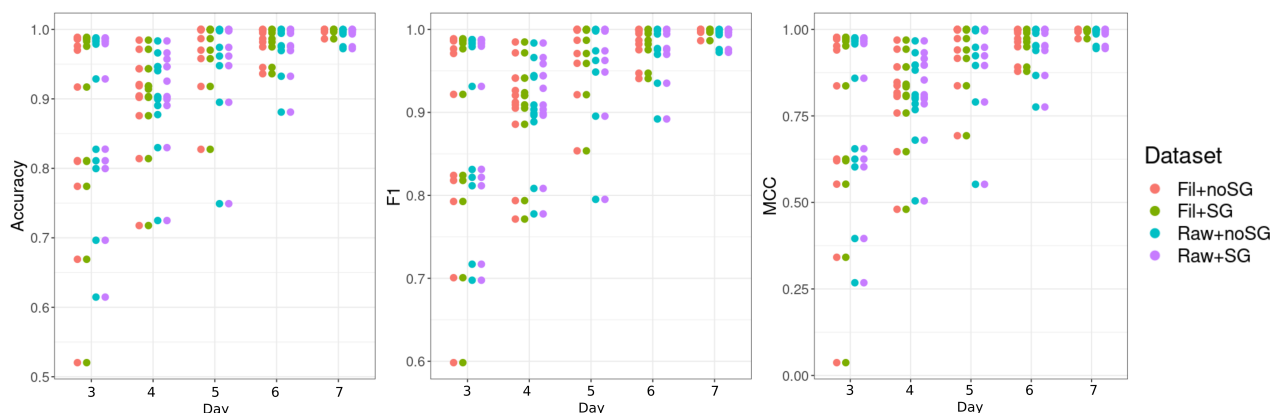


Figure 6-29. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the ‘Tommy Atkins’ e5-to experiment

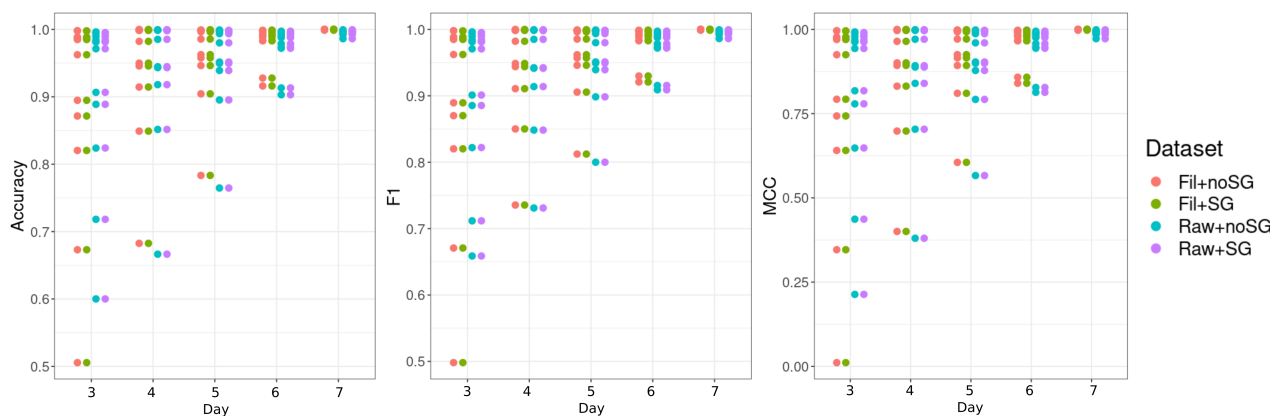


Figure 6-30. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the ‘Tommy Atkins’ e6-to experiment

egy in allowing classification models to fit the data. However, when comparing the Fil+noSG and Fil+SG datasets, there were no relevant variations in model performance when using or not using the SG filter, so there was no need to use the SG filter to improve model predictions.

Likewise, for both ‘Keitt’ assays (e7-kt and e8-kt), Figures **6-31** and **6-32**, from the first day of study (day 3) models achieved remarkable metrics, with outcomes greater than 0.95. Also, it can be noted that the models achieved a better performance when they were trained with diseased classes with more distinguishable anthracnose symptoms (as was reported in the ‘Tommy Atkins’ experiments -e5-to and e6-to-). However, in the e7-kt assay models decreased their overall performance on day 4, being the most notable difference among the leaf assays carried out in this research (this behaviour was not reported in the e8-kt assay, for instance).

Contrary to the ‘Tommy Atkins’ experiments, in the ‘Keitt’ assays, the usefulness of the out-

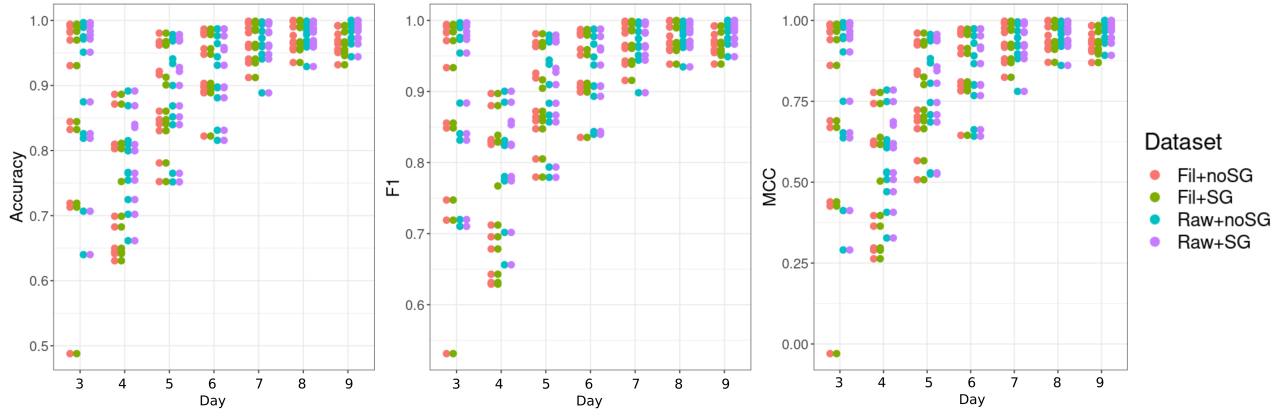


Figure 6-31. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the *cv.* ‘Keitt’ e7-kt experiment

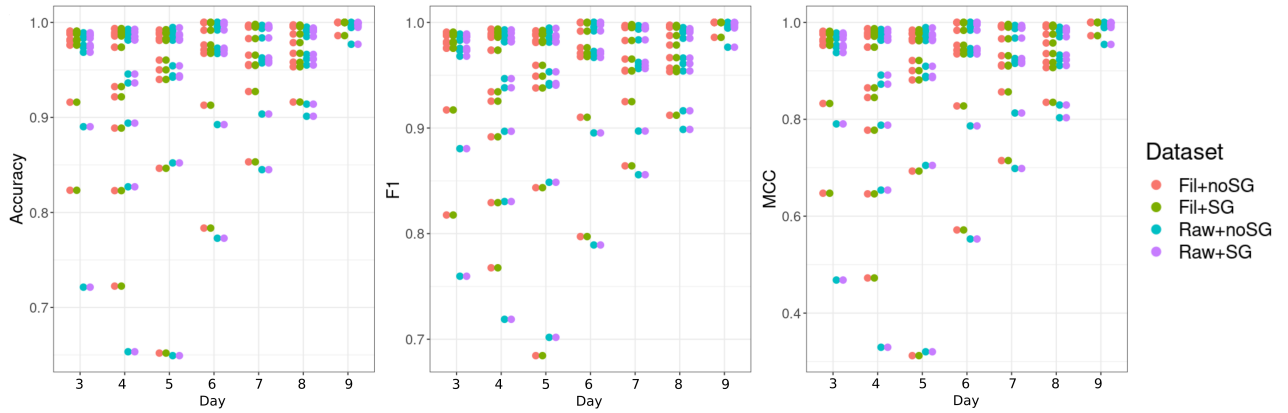


Figure 6-32. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the *cv.* ‘Keitt’ e8-kt experiment

lier filter (Fil+noSG and Fil+SG) was not overwhelming: while in the e7-kt assay datasets without this filter (Raw+noSG and Raw+SG) stood out in the initial days, in the e8-kt experiment the outlier datasets showed a slight score improvement when comparing against the raw datasets, see Appendix B.1.1 with full model metrics.

On the other hand, in the ‘Kent’ assay (e9-kn), Figure **6-33**, remarkable performance metrics were achieved since day 2 (several models reported metrics above 0.95 in accuracy, F1 and MCC), showing the existence of classification models that fitted well the data from the first initial symptoms of anthracnose. However, the increasing trend in performance (reported in the ‘Tommy Atkins’ and ‘Keitt’ cultivars) when evaluating late diseased classes against the sound class was not observed since on days 4 and 5 it was reported a decrease in the metric scores.

Also, in this assay, a unique dataset that showed the best behaviour in performance during

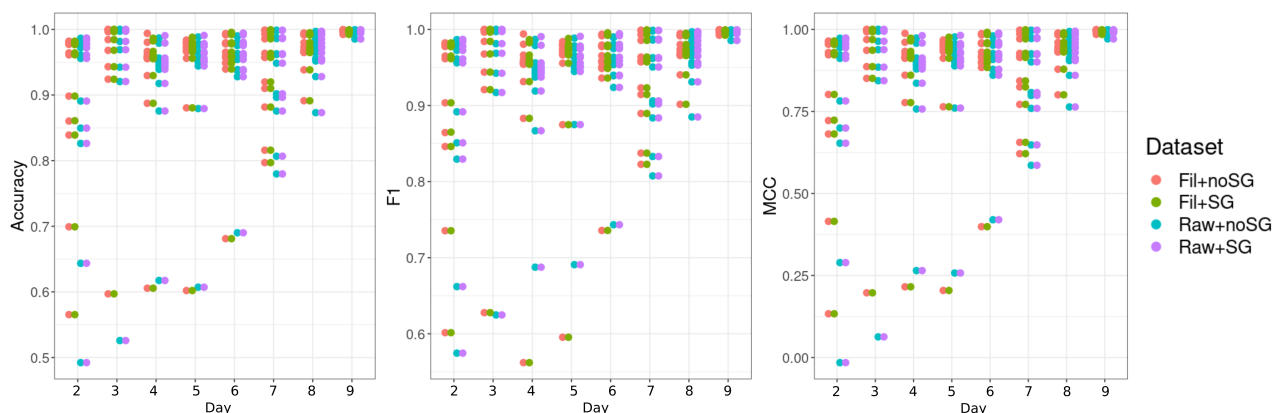


Figure 6-33. Performance metrics (a. Accuracy, b. F1, c. MCC) of the classification models implemented by dataset for the *cv.* ‘Kent’ e9-kn experiment

all days was not found since all datasets showed similar results on top of the best models, see Appendix B.1.1 for details. However, the only notable difference was found in the variance of the metrics of the models, where the datasets with the outliers filter obtained a lower variance, which meant that all models fitted the data better, as the difference between the best and worst model (in performance) was smaller in these datasets. Hence, the Fil+noSG dataset could be selected for this cultivar, considering that the use of the SG filter (Fil+noSG) did not substantially improve the overall performance compared to Fil+noSG.

Despite the Savitzky-Golay filter (SG) being a good method to improve the spectral quality of the samples, it did not report the best performance when combined with the outlier filter. Only when comparing the datasets Raw+noSG and Raw+SG, their contribution to developing better classification models was noted. Consequently, the Fil+noSG dataset was selected to continue with the analysis of the classification model metrics to detect foliar anthracnose symptoms in the three mango cultivars (‘Tommy Atkins’, ‘Keitt’ and ‘Kent’) based on the results reported in all the experiments.

6.2.7. Biclass models to discriminate the anthracnose diseased classes in mango leaves

Considering the best fitting dataset for the assays conducted with leaves of the three mango cultivars was Fil+noSG, this section presents the results associated with the classification models trained with this dataset. Similarly, given the number of models trained per class and assay, the figures and tables listed in this section present the results associated with the 6 best classification models (QDA, LDA, SLP, MLP, SVM and BST), while the remaining model results (RF, CART, Naive Bayes, GBM, XGB and PLSDA) are presented in Appendix B.1.2.

During the analysis of the performance metrics, the positive class will be referred to as the

sound class and the negative class as the diseased class. Furthermore, it is clarified that the recall metric directly measures the ability of the model to correctly classify the positive class, while the specificity metric measures the negative class. Hence, this section will omit these concepts to improve readability, as they were previously introduced.

AI-based models applied to ‘Tommy Atkins’ assays

Figures 6-34 and 6-35 present the performance metrics of the QDA, LDA, SLP, MLP, SVM and BST biclass models, per day, for the e5-to and e6-to experiments, respectively. From these figures, it can be noted that from day 5, the classification models presented outstanding performance metrics, allowing for the verification of their efficacy in the spectral differentiation of the classes: metrics were higher than 0.975. Furthermore, on day 3, the models also reported remarkable performance in discriminating the sound and diseased classes in both assays. However, on day 4, a 5 % decrease in the scores of the SLP, MLP and LDA models was evidenced in the e5-to assay. This decrease was noticeable in the specificity and precision metrics of the SLP and MLP models and the recall metric of the LDA model.

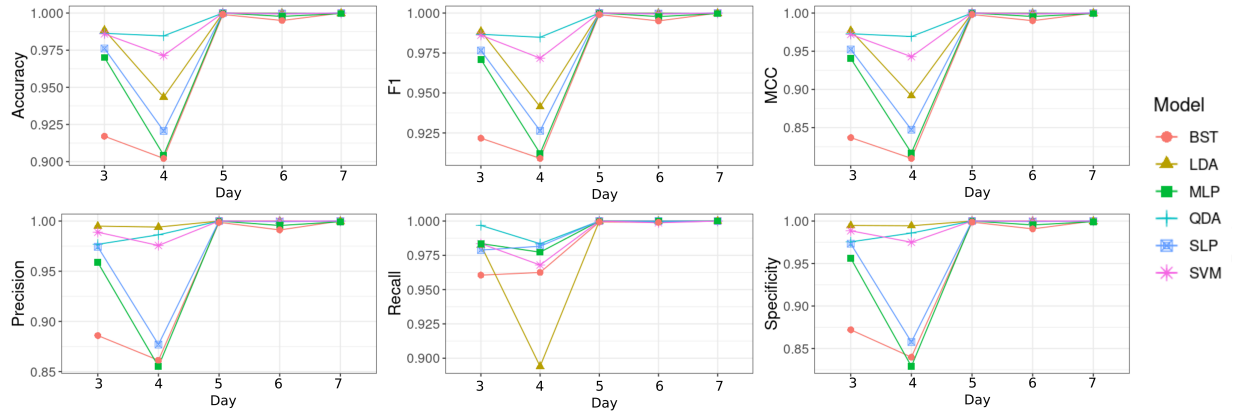


Figure 6-34. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the ‘Tommy Atkins’ e5-to experiment and the Fil+noSG dataset

Comparing the performance of the models on a day-by-day basis, it was observed that the best fitting models were, in order, LDA, QDA, and SVM for the e5-to assay, while the SLP, QDA, and MLP models for the e6-to assay. From the first day of study (day 3), these models achieved scores above 0.94 on all metrics, excluding the recall metric on day 4 for the LDA model. In addition, it is also noteworthy that they did not present considerable differences in performance from day 5 onwards, indicating that any of the listed models could perform the foliar detection of anthracnose-diseased samples without difficulty.

On the one hand, in the e6-to assay, the SLP model with $neurons = 200$, and $epochs = 150$ was the best classifier from the first day of the study, indicating an enormous potential

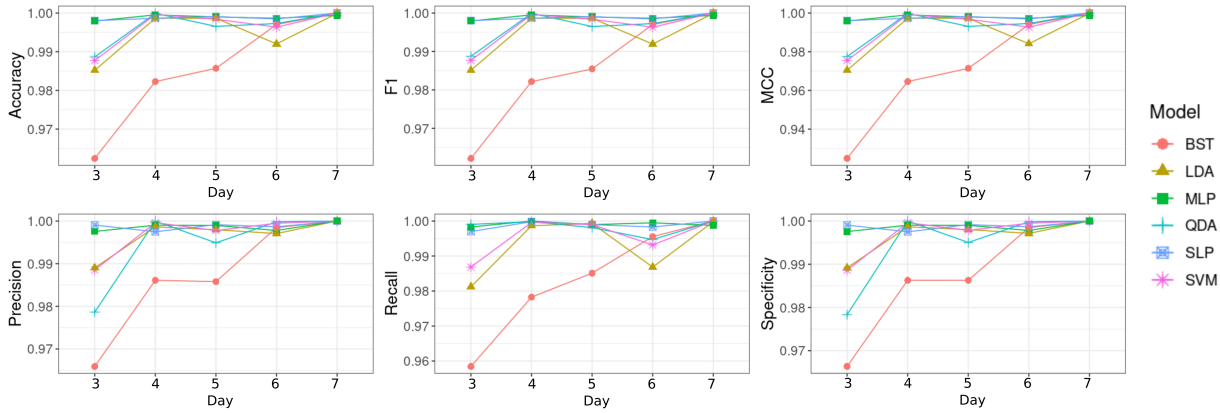


Figure 6-35. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the ‘Tommy Atkins’ e6-to experiment and the Fil+noSG dataset

to detect diseased samples from early stages of anthracnose development. The QDA model was also highlighted as it obtained metrics similar to those of SLP and MLP, demonstrating the usefulness of this model for the discrimination of sound and diseased spectral samples.

Therefore, although in the e6-to experiment, the best model was SLP (*neurons* = 200, *epochs* = 150), the classification model with the best fit and potential for the *cv.* ‘Tommy Atkins’ in mango leaves was QDA, reporting scores above 0.969 in all performance metrics. Additionally, because this technique was, per se, a mathematical model, it offers the advantage of being calibrated, validated and readjusted in a shorter computational time, as opposed to the training time employed by an SVM model or a neural network (such as SLP).

AI-based models applied to ‘Keitt’ assays

Figures 6-36 and 6-37 present the performance metrics of the QDA, LDA, SLP, MLP, SVM and BST biclass models, per day, for the e7-kt and e8-kt experiments, respectively. For this cultivar, a considerable decrease in the accuracy, specificity and precision metrics of the implemented models was found on day 4 in the e7-kt assay (around 12 %) and day 5 in the e8-kt assay (around 1.5 %). However, from the following days for each assay (day 5 in e7-kt and day 6 in e8-kt), the models managed to fit the data and increase their performance with subsequent days, showing that this drawback has been overcome. In fact, from day 7 onwards, the models achieved outstanding and similar performance metrics to each other, so any model used from this day onwards was able to perform well the task of distinguishing healthy samples from anthracnose-infected ones.

On the other hand, when analysing the behaviour of the metrics day by day, it was found that the best models per assay were: QDA, LDA and SVM for the e7-kt and QDA, SLP and MLP for the e8-kt assay. From Figures 6-36 and 6-37, it is observed that the discriminant models (QDA and LDA) stood out on days 3 and 4, showing their ability in the detection

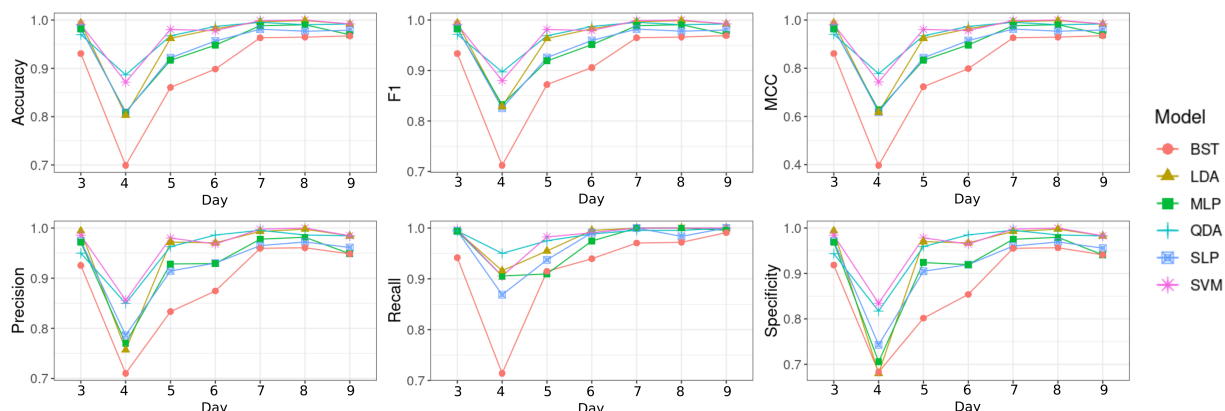


Figure 6-36. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e7-kt ‘Keitt’ assay and the Fil+noSG dataset

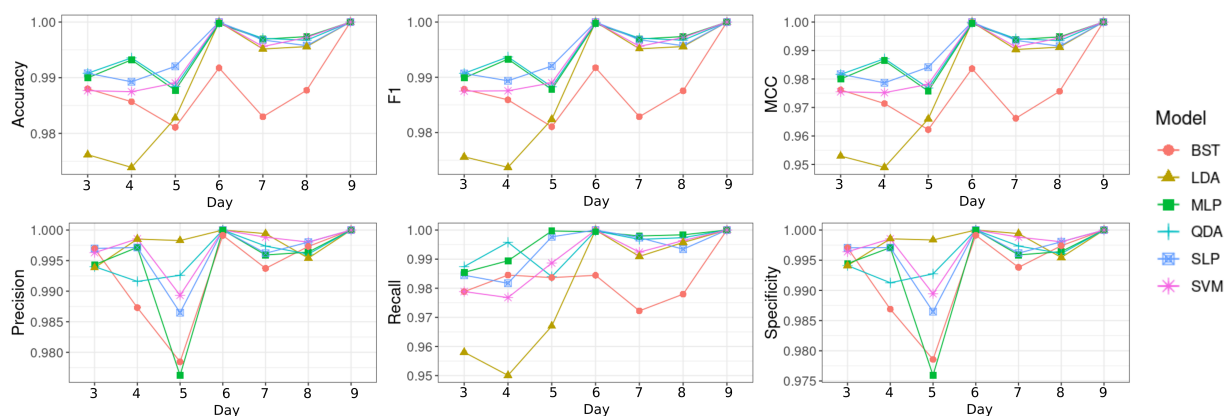


Figure 6-37. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e8-kt ‘Keitt’ assay and the Fil+noSG dataset

(discrimination) of anthracnose diseased samples. In addition, it can be noted that from day 6 onwards, any of the models mentioned above could detect diseased samples without major difficulty, demonstrating the potential and ability of the models trained in this research.

Based on the results outlined above, the best classification model, and therefore the best foliar detection model, for the *cv.* ‘Keitt’ was QDA. This model once showed the enormous potential and the discriminant ability to differentiate healthy samples from anthracnose-diseased samples from the first days of infection. This feature was an important point as one of the objectives of this research was to determine the feasibility of detecting anthracnose in mango, so the possibility of carrying out this task with spectral data with a fast-calibrating mathematical model was a relevant finding of this study.

AI-based models applied to ‘Kent’ essay

Figure 6-38 presents the performance metrics of the QDA, LDA, SLP, MLP, SVM and BST

biclass models, per day, for the e9-kn assay. The results for the *cv.* ‘Kent’ showed considerable changes compared to the other cultivars. Firstly, all models obtained good performance metrics from the first day of the study (day 2) with scores above 0.9. Secondly, despite the slight decrease in the performance of the models on day 4, this differs from the decreases reported for the ‘Tommy Atkins’ and ‘Keitt’ cultivars, as the amplitude of the decrease was less than 0.07 in the metrics, while in the other cultivars it was higher.

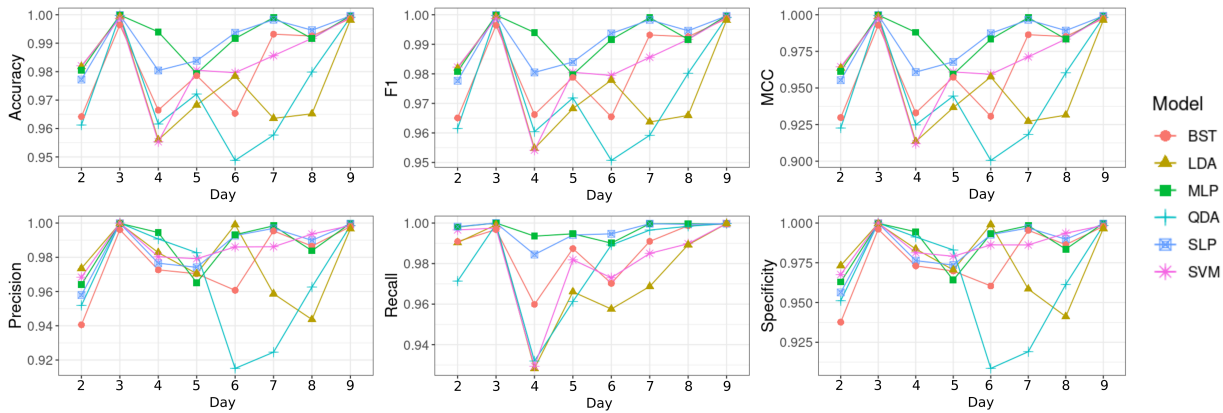


Figure 6-38. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented for the e9-kn experiment of the *cv.* ‘Kent’ and the Fil+noSG dataset

Regarding the best classification models for this cultivar, it was possible to determine that the models with the best performance throughout the experiment were SLP, MLP and SVM, with the SLP model standing out mainly on days 3, 5 and 6. The SLP model was trained with the hyperparameters *neurons* = 200, *epochs* = 100, while the MLP model with *neurons* = 200, *epochs* = 100 and *layers* = 3. As a particular finding for this cultivar, the e9-kn assay was one of all the mango leaf assays that evidenced the supremacy of neural networks in classifying sound and anthracnose-diseased samples. As the second-best model was ranked the MLP model, although it reported a slight drop in its metrics on day 5. However, from day 7 onwards the differences between the best models were minimal (less than 1 %) so that the use of any of the neural models could perform the foliar detection task robustly and reliably at late stages of anthracnose infection.

On the other hand, it should be noted that the discriminant models (QDA and LDA) did stand out in the first days of the experiment (days 2, 3 and 4), which is why these models (especially QDA) could be taken as a reference in case a robust and fast calibrated (training) model for the early foliar anthracnose detection in the *cv.* ‘Kent’ is required: the metrics of these models did not fall below 0.9, so these models qualify as outstanding classifiers.

However, as this study was based on the design of a classification model whose performance was the best fit to the data regardless of the day of study, the SLP model (*neurons* = 200,

$epochs = 100$) was the best-performing technique for this particular cultivar. Finally, it was reported that the hyperparameters per day of the study did not vary considerably. Therefore, the configuration mentioned above was chosen for the selected model.

Best biclass model to detect anthracnose symptoms in mango leaf assays

In the previous sections, the analysis per assay and cultivar of the best classification models was presented, with the QDA, SLP, SVM and MLP models being the most outstanding. In addition, the QDA model was established as the best classifier for the ‘Tommy Atkins’ and ‘Keitt’ cultivars, while SLP($neurons = 200$, $epochs = 100$) was chosen for the *cv.* ‘Kent’. Nonetheless, in this section, the AUC metric, calculated from the ROC curve, of the QDA and SLP models is analysed to establish which model has the best discriminatory power, i.e., which was closer to the ideal classifier ($AUC = 1$).

Table 6-10 presents the AUC metric for the QDA and SLP models in the mango leaf experiments. From these results, the potential of the selected models in these experiments was evidenced since the AUC metric obtained scores higher than 0.970, excluding the SLP model in the e7-kt assay that on day 4 obtained 0.880 (score equally representative). Furthermore, the AUC metric and the ROC curve indicate that the QDA model had the potential to discriminate the defined classes in the e9-kn assay, even though it was not among the three best models according to the results presented in section 6.2.7. Thus, the potential of the calibrated QDA model and the possibility of improving its performance by adjusting the decision threshold based on the ROC curve cannot be ignored.

Table 6-10. Area under the ROC curve (AUC) of the QDA and SLP models for the e5-to, e6-to, e7-kt, e8-kt and e9-kn leaf assays, and by day of study

Diseased class	Assay e5-to		Assay e6-to		Assay e7-kt		Assay e8-kt		Assay e9-kn	
	QDA	SLP	QDA	SLP	QDA	SLP	QDA	SLP	QDA	SLP
Day 2									0.984	0.995
Day 3	0.999	0.998	1.000	1.000	0.995	0.999	0.995	0.998	1.000	1.000
Day 4	0.999	0.986	1.000	1.000	0.972	0.880	1.000	0.999	0.994	1.000
Day 5	1.000	1.000	0.999	1.000	0.992	0.972	0.994	1.000	0.996	0.999
Day 6	1.000	1.000	1.000	1.000	0.998	0.992	1.000	1.000	0.995	0.998
Day 7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000
Day 8					1.000	1.000	1.000	1.000	1.000	0.996
Day 9					0.998	0.989	1.000	1.000	1.000	1.000

Similarly, the metrics obtained by the SLP model showed that it has done a remarkable task in all the experiments, despite not being the best model in the ‘Tommy Atkins’ and ‘Keitt’ cultivars. However, a weakness of this neural model was that the training and validation

time compared to the QDA model was considerably longer: the learning process of neural networks is iterative, layer by layer and proportional (in computational time) to the volume of data.

Consequently, as a result of the analysis of the performance of the filters and models implemented for the foliar detection of anthracnose in the ‘Tommy Atkins’, ‘Keitt’ and ‘Kent’ mango cultivars, it was determined that the QDA model with the outlier filter - Fil+noSG- was the best option for the development of a robust solution to support the diagnosis of this disease. The main advantage of this model is that its calibration can be done in real-time. Likewise, as an alternative option, it was determined that the SLP model (*neurons* = 200, *epochs* = 100) presented an outstanding performance that also supported the diagnostic task of anthracnose in mango leaves; however, its limitation lies in the training and calibration times, as it requires more time than the discriminant model.

6.2.8. Multiclass models for detection of anthracnose symptoms in mango leaves

12 classification models, presented in section 4.1, were trained with the data presented in Table 6-4 for the Fil+noSG dataset and the mango leaf assays e5-to, e6-to, e7-kt, e8-kt and e9-kn. Following the flowchart defined in Figure 4-2, the model training was performed and the results were consolidated to determine the potential of multiclass AI models for early detection of foliar anthracnose symptoms in 3 mango cultivars.

In addition, the models reported in this section corresponded to the neural models (SLP and MLP), discriminant models (LDA and QDA) and SVM since they achieved the best performance metrics in most leaf experiments, metrics above 0.9 (except for the e7-kt assay). The other models can be consulted in Appendix B.2.

AI-based multiclass model applied to ‘Tommy Atkins’ assays

Tables 6-11 and 6-12 present the results of the classification models implemented for the ‘Tommy Atkins’ assays on mango leaves (e5-to and e6-to). From these tables, it can be noted that the models obtained metrics above 0.9 for all metrics in both assays. This indicated their overall performance in discriminating the defined classes was outstanding, especially in the recall metric.

Likewise, the accuracy, F1, and MCC metrics showed that the models managed to fit the data even when there were imbalances between their classes, which is very recurrent in multi-class models since they do not usually have the same number of samples per class. In addition, the specificity metric showed that the QDA, SVM, SLP, MLP and LDA models obtained higher scores than the rest of the metrics, with 0.98 being the lower limit in this

metric for both experiments. This result indicated that for the classification of the negative class (which mostly corresponds to the diseased class), the model could detect the presence of a diseased sample regardless of its stage of development.

The most prominent models for this cultivar were QDA and SVM, which performed best in the e5-to and e6-to assays, respectively, and preserved their performance metrics regardless of the assay. However, the SLP model with hyperparameters *neurons* = 50 and *epochs* = 450 also stood out as an outstanding classification method in this cultivar.

Table 6-11. Performance metrics of the classification models for the ‘Tommy Atkins’ e5-to experiment

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
QDA	0.985	0.985	0.997	0.985	0.982	0.986
SVM	0.985	0.984	0.997	0.984	0.981	0.985
SLP	0.962	0.961	0.992	0.961	0.954	0.962
LDA	0.957	0.957	0.991	0.956	0.949	0.959
MLP	0.952	0.950	0.990	0.950	0.941	0.953

Table 6-12. Performance metrics of the classification models for the ‘Tommy Atkins’ e6-to experiment

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
SVM	0.948	0.947	0.990	0.947	0.938	0.950
SLP	0.947	0.947	0.990	0.945	0.937	0.949
QDA	0.940	0.939	0.988	0.938	0.927	0.940
MLP	0.932	0.931	0.987	0.928	0.919	0.938
LDA	0.927	0.927	0.986	0.926	0.912	0.928

On the other hand, when analysing the confusion matrices of the QDA and SVM models (Figure 6-39), it was observed that the QDA model presented lower classification error than the SVM model in both experiments. For this study, the presence of data classified in the first row and first column of the confusion matrices between days 3 and 7 meant that the model misclassified a datum, being the data in the first row (error type II) the most critical as the model classified diseased samples as sound (error type I is located in the first column). Therefore, when comparing the total number of misclassified samples in the e5-to and e6-to assays, the QDA model obtained 165 and 100, respectively, while the SVM model obtained 200 and 161.

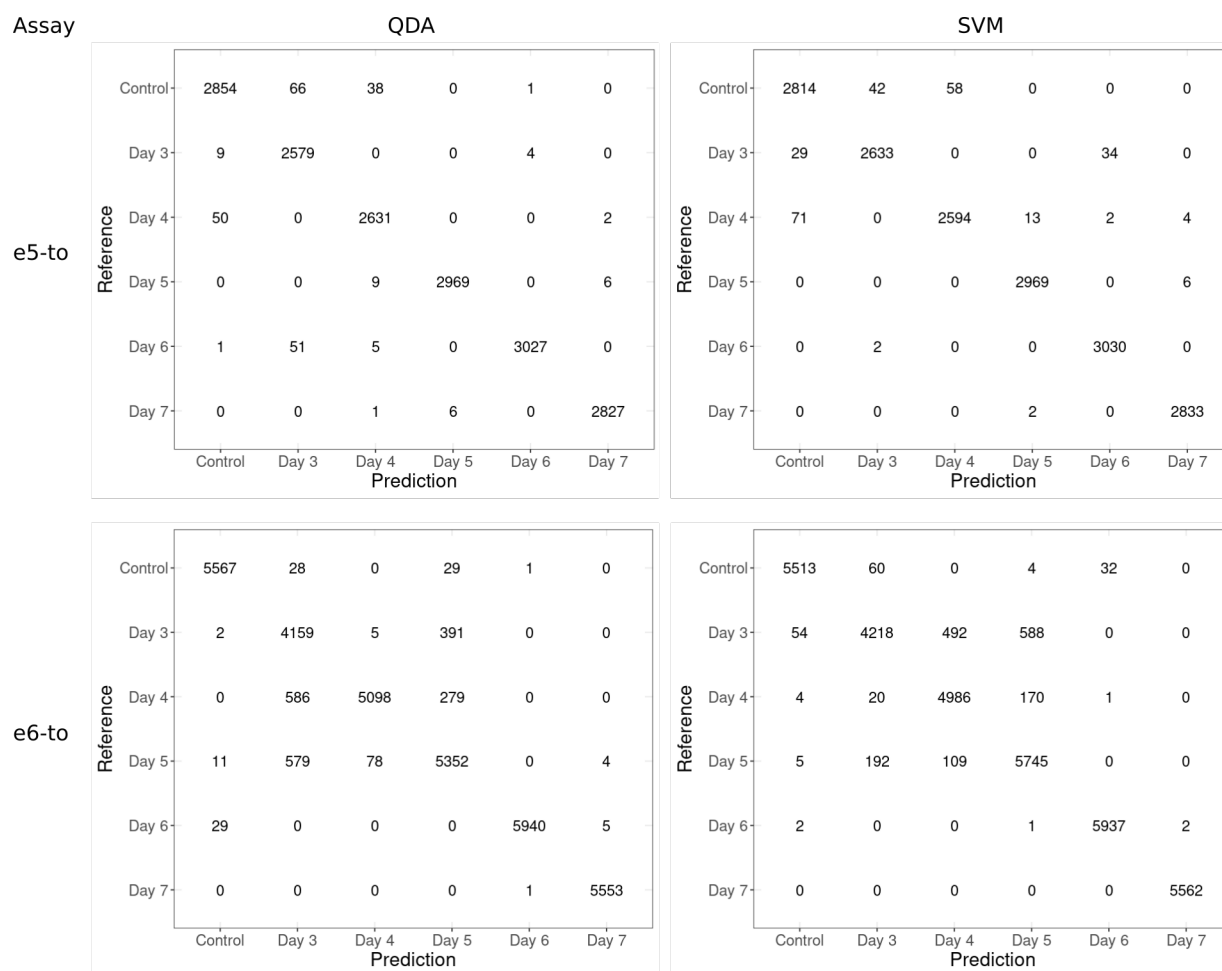


Figure 6-39. Confusion matrices of the QDA and SVM classification models for the e5-to and e6-to assays

Consequently, and also considering the results obtained for this cultivar in the biclass model (section 6.2.6), the QDA model was chosen as the foliar anthracnose detection model of the *cv.* ‘Tommy Atkins’. This model showed the best performance in the biclass and multiclass classification and has the advantage of being readjusted and calibrated in less time than SVM or neural models.

Multiclass AI-based models applied to ‘Keitt’ assays

Tables 6-13 and 6-14 present the results of the classification models implemented for the e7-kt and e8-kt experiments, respectively. From these tables, notable differences were observed in the performance metrics of the best models between the e7-kt and e8-kt experiments: while for the e7-kt assay models, all models were below 0.9, for the e8-kt they were positioned above this threshold. This situation can be explained by the limited data availability in the e7-kt assay, which suggested that multiclass models require a large amount of data to achieve good results, especially the ANNs. This drawback can be contrasted with the metrics in the e8-kt assay, which resulted in better model scores since more data were available.

Furthermore, as reported in the analysis of the biclass models of this cultivar (section 6.2.7), for the e7-kt assay there was a drop in the performance of the models on day 4, so it was foreseeable that in the calibration of the multiclass models (which integrate all the defined classes), there was a performance degradation of the models as the classification error of the class on day 4 was transferred to the rest of the classes.

A particular result of these experiments was that the neural models (SLP and MLP) were sensitive to limited data availability (e7-kt assay), which revealed a weakness compared to the discriminant models, e.g. for use in detection systems where data recording is limited. Additionally, for the e8-kt assay, it was clear that the discriminant models outperformed the neural models, as the QDA model outperformed the neural models, the LDA model obtained similar performance metrics to the SLP model and outperformed the MLP model.

Table 6-13. Performance metrics of the classification models for the *cv.* ‘Keitt’ e7-kt assay

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
SVM	0.861	0.863	0.980	0.862	0.845	0.869
QDA	0.831	0.831	0.976	0.835	0.815	0.851
LDA	0.806	0.810	0.972	0.805	0.791	0.841
SLP	0.751	0.751	0.964	0.744	0.735	0.819
MLP	0.726	0.728	0.961	0.718	0.696	0.760

Table 6-14. Performance metrics of the *cv.* ‘Keitt’ e8-kt assay classification models

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
SVM	0.949	0.941	0.993	0.943	0.937	0.946
QDA	0.944	0.934	0.992	0.937	0.930	0.943
SLP	0.935	0.933	0.991	0.931	0.923	0.933
LDA	0.920	0.914	0.989	0.914	0.904	0.919
MLP	0.900	0.895	0.986	0.893	0.885	0.910

When analysing the performance metrics of the two best models (SVM and QDA) it was observed that their performance in the e7-kt assay was good (metrics below 0.9) and outstanding in the e8-kt experiment (metrics above 0.9). These models presented good accuracy, F1, MCC, precision and recall metrics, indicating their potential to adapt to data sets with class imbalances and a reduced number of data (as reported in the e7-kt assay). The specificity metric showed that these models obtained scores above 0.975, indicating that for the

negative class classification (which mostly corresponds to the diseased class), models could detect when a sample was infected regardless of the stage of anthracnose development.

However, when analysing the confusion matrices of the QDA and SVM models, see Figure 6-40, a similar behaviour to that reported for the *cv.* ‘Tommy Atkins’ (section 6.2.8) was observed: the QDA model achieved a lower type I and type II classification error in the e8-kt assay compared to the SVM model (124 vs 138 misclassified data, respectively), while for the e7-kt assay both models obtained similar results (127 vs 126 misclassified data, respectively). This result highlighted that, despite SVM having higher performance metrics, the QDA model performed better in correctly detecting sound and diseased samples with foliar anthracnose.

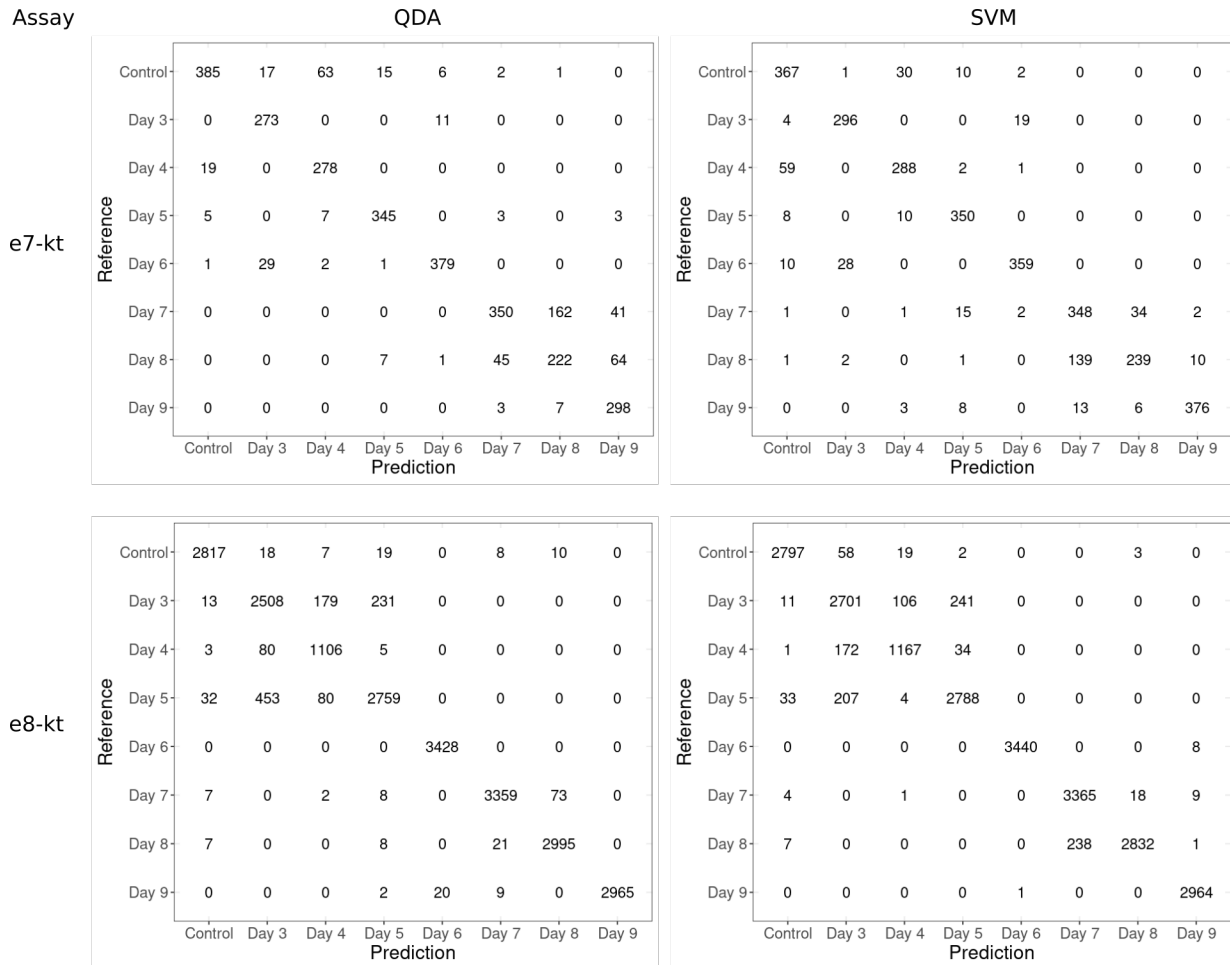


Figure 6-40. Confusion matrices of the QDA and SVM classification models for the e7-kt and e8-kt experiments

Therefore, the QDA model with the outlier filter (Fil+noSG) was selected as the most suitable model for the multiclass detection of anthracnose foliar symptoms in the ‘Keitt’ mango

cultivar. This result confirms the analysis and selection of the biclass model in section 6.2.7, where the QDA model was chosen as the best-fitting model. Also, this result revealed the potential of discriminant models for developing solutions to support the diagnosis of anthracnose in mango.

Multiclass AI-based models applied to ‘Kent’ assay

Table 6-15 reports the results associated with the classifiers for the e9-kn assay. Unlike the rest of the cultivars, where the discriminant models dominated, it was evident that the neural networks and SVM were the three best classification models in this cultivar. Specifically, the best model for this cultivar was SLP with accuracy = 0.981, recall = 0.981, specificity = 0.998, F1 = 0.981, MCC = 0.979, precision = 0.981 and hyperparameters *epochs* = 455, *neurons* = 50. The scores obtained with this model indicate that the simple perceptron perfectly fits the classes defined for this cultivar. This result suggests that this model could robustly detect anthracnose-diseased samples on ‘Kent’ mango leaves from the early stages of development.

Table 6-15. Performance metrics of the *cv.* ‘Kent’ e9-kn assay classification models

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
SLP	0.981	0.981	0.998	0.981	0.979	0.981
SVM	0.963	0.962	0.995	0.962	0.958	0.962
MLP	0.933	0.931	0.992	0.933	0.928	0.946
QDA	0.920	0.921	0.990	0.920	0.912	0.923
LDA	0.903	0.903	0.988	0.900	0.889	0.902

Furthermore, when comparing the confusion matrices of the SLP and SVM models, see Figure 6-41, it was found that the SLP model obtained considerably lower classification errors than the SVM model: for the type I error (first column) SLP obtained 182 errors vs. 268 for the SVM model; while for the type II error (first row), the most severe error for the model, SLP obtained 49 vs. 330 for the SVM model.

On the other hand, it is worth mentioning that for this cultivar the QDA model, which had been selected for the ‘Tommy Atkins’ and ‘Keitt’ cultivars, failed to overcome the goodness of fit of the neural networks and was positioned in 4th place with accuracy = 0.920, recall = 0.921, specificity = 0.990, F1 = 0.920, MCC = 0.912 and precision = 0.923. This result, although far from the metrics obtained by the SLP model in this cultivar, also evidenced the potential of the QDA model in the detection of anthracnose symptoms independently of the cultivar, as this model obtained remarkable results for all 3 mango leaf cultivars.

Assay		SVM										SLP									
e9-kn	Reference	Control	3154	28	1	184	34	54	15	14	0	Control	3435	16	0	4	0	20	4	5	0
		Day 2	39	3321	0	0	89	0	2	0	0	Day 2	90	3344	0	0	0	0	17	0	0
		Day 3	3	0	3295	92	14	0	0	0	46	Day 3	0	0	3278	117	12	11	0	32	0
		Day 4	63	0	74	3238	6	0	0	0	22	Day 4	8	0	29	3353	1	1	0	11	0
		Day 5	45	1	18	36	3036	2	12	1	0	Day 5	33	0	15	20	3070	8	4	0	1
		Day 6	49	0	1	0	6	3341	0	1	0	Day 6	47	0	1	1	3	3345	0	1	0
		Day 7	42	1	0	8	46	0	3667	19	0	Day 7	2	4	0	1	6	0	3767	1	2
		Day 8	27	0	51	27	1	2	1	3686	0	Day 8	2	0	39	25	0	1	1	3726	1
		Day 9	0	0	0	0	0	0	0	0	3818	Day 9	0	0	0	0	0	0	1	1	3816
		Prediction										Prediction									

Figure 6-41. Confusion matrices of the SVM and SLP classification models for the e9-kn assay

Nevertheless, considering the main objective of designing the best classification model for each cultivar, with the *cv.* ‘Kent’ it was found that the best classification model was SLP with $epochs = 100$, $neurons = 200$, $nfolds = 3$ and the outlier filter (Fil+noSG). This result agrees with the analysis done in the biclass variant, thus confirming the potential of the simple perceptron trained and calibrated in this assay for this particular mango cultivar.

6.2.9. Validation of early detection models of anthracnose in mango samples at surface level

In sections 6.2.5 and 6.2.8, it was established the feasibility of designing a classification model based on ML and hyperspectral analysis, able to discriminate different stages of anthracnose development in mango fruits and leaves, respectively. As a main result, it was possible to establish the early detection from day 2 after inoculation on fruits, while from day 3 after inoculation on leaves with a confidence level over 90 % in all performance metrics. Considering that the point of visible detection of lesions compatible with anthracnose was between day 4 and 5 post-inoculation (when the lesions exceeded 1 *mm* radius), a pre-visual detection window of about 2 days (48 hours) and 1 day (24 hours) for mango fruit and leaf samples, respectively, was achieved.

However, these models were trained with spectral signatures, which implied that the detection was performed at a punctual level, limiting their use in the spatial component when hyperspectral imaging is used. Therefore, in addition to the development of the spectral model for early anthracnose detection in mango, the models obtained for each cultivar were tested at the surface level of mango samples to verify their performance and to determine whether this pre-visual detection window of symptoms was extensible to a complete mango fruit or leaf. In this way, this experiment aimed to verify the shape and size of the detected

lesions (as a human would do when inspecting the whole fruit) and whether they corresponded to the day of inoculation as defined in the labelling data.

Considering the early detection window found in the assays (days 2 and 3 after inoculation), the trained models were applied on all the hyperspectral images of the test group samples of each assay in the first 3 acquisition days (days 2, 3 and 4 for fruits and days 3, 4 and 5 for leaves). The preprocessing and filtering procedure established for the spectral data in each assay and mango cultivar was applied to all the hyperspectral images, given that a hyperspectral image is a composition, or arrangement, of spectral signatures. Additionally, in the final prediction stage of the models, a mean filter with a 3×3 window (or in some cases 5×5) was used to reduce noise and improve the accuracy and quality of the models when predicting the size and shape of the lesions on the samples. Larger windows than those mentioned above, while reducing the incidence of noise on the surface of the samples, proved to be very restrictive with the final model prediction in the first days of the assays, thus reducing the overall performance in detecting early anthracnose lesions.

Figures **6-42** and **6-43** show the results of a control sample from the test group for the fruit and leaf assays, respectively, while Figures **6-44** to **6-47** show an inoculated sample from the test group for the fruit and leaf assays. Each figure includes a pseudo-RGB image and a prediction image resulting from the selected classification model for each cultivar. In addition, each image provides the scale in cm and the colour scale used to represent the prediction results of the detection model, being the black colour used for the background and deleted pixels, the green colour to represent the sound condition and the other colours to represent the different developmental stages of anthracnose.

From the results obtained with the control samples, Figures **6-42** and **6-43**, the remarkable performance of the detection models in correctly classifying anthracnose-free samples can be observed: the green colour predominates over the induced lesions in the fruits, while in the leaves the induced lesions passed unnoticed by the detection model (their size and spectral variation were so small that the model did not perceive them). Likewise, these results verified the assumption that the spectral signature of control samples is invariant to the day of acquisition, as the samples of this group presented similar prediction results throughout the study days.

However, false anthracnose symptoms were also detected on the borders of fruits and leaves. This behaviour can be explained by the fact that the diffuse reflectance in these regions is different from that recorded in the central region of the fruits and the equatorial region of the leaves because of the hemispherical surface of the fruits and the folds of the leaves when they were placed in the hyperspectral acquisition dome. This meant that the surface of the samples was neither flat nor equidistant from the hyperspectral camera lens. Hence,

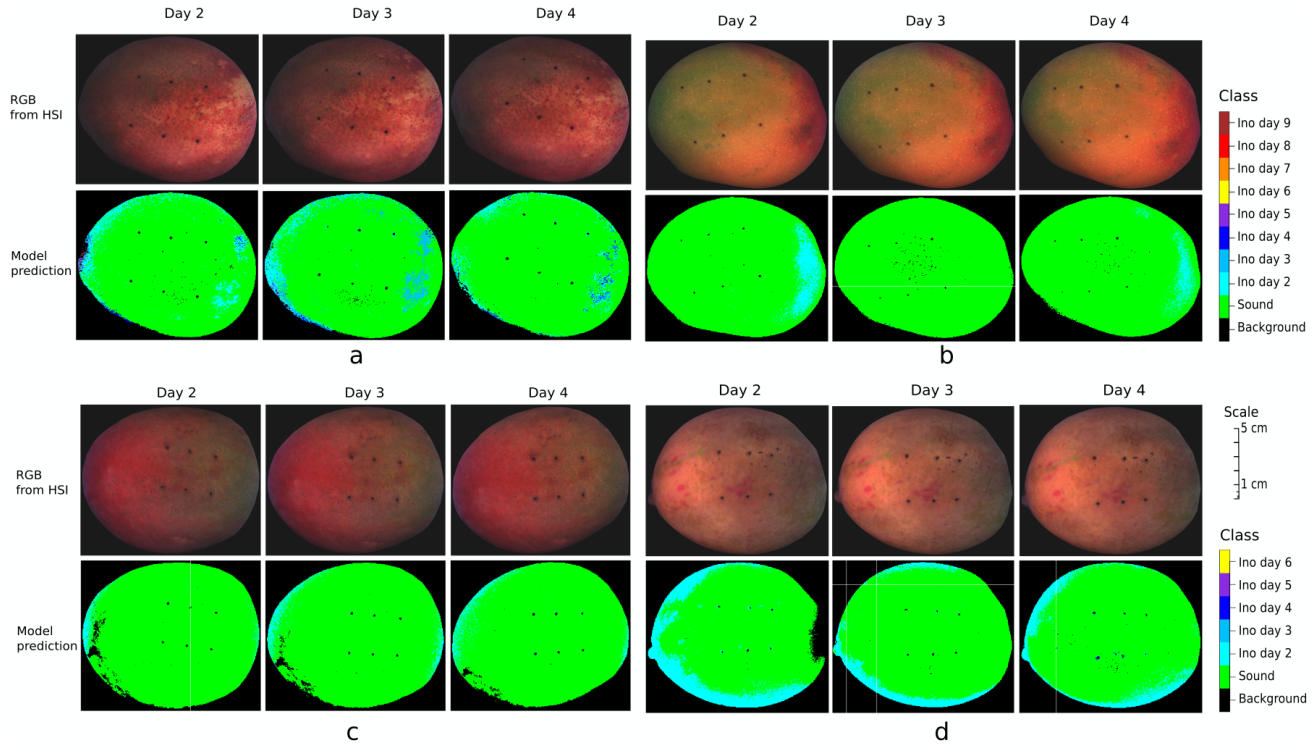


Figure 6-42. Model prediction result of the control samples for the a) ‘Keitt’, b) ‘Kent’, c) ‘Osteen’ and d) ‘Keitt’ fruit assays

the angles of the light reflection on the surface varied, causing distortions in the regions described above. Furthermore, due to the structure of the hyperspectral acquisition dome and the size and shape of the calibration white panel, the region of interest, corresponding to the localised inoculation points of the anthracnose on the samples, was established to allow the spectral analysis of the symptoms associated with the disease.

As a potential solution to correct this deficiency, the integration of a 3D model of the surface to correct the reflectance or the three-dimensional spectral scanning of the samples could be a good approach to explore to improve the results of the detection models in the described regions. However, for the objectives of this research (early detection and spectral characterisation of anthracnose symptoms in mango), this drawback did not affect the results and findings obtained since no analysis was performed on these regions.

On the other hand, the prediction in the inoculated samples, Figures 6-44 and 6-45, showed a remarkable result in all the detection models of the fruit assays. In these figures, for more clarity, an amplified comparison of 3 ROIs (anthracnose lesions) per sample on each of the days of acquisition is presented; each of these ROIs represents an area of 1cm^2 .

Firstly, the anthracnose lesions detected on day 4, the day they are completely distinguish-

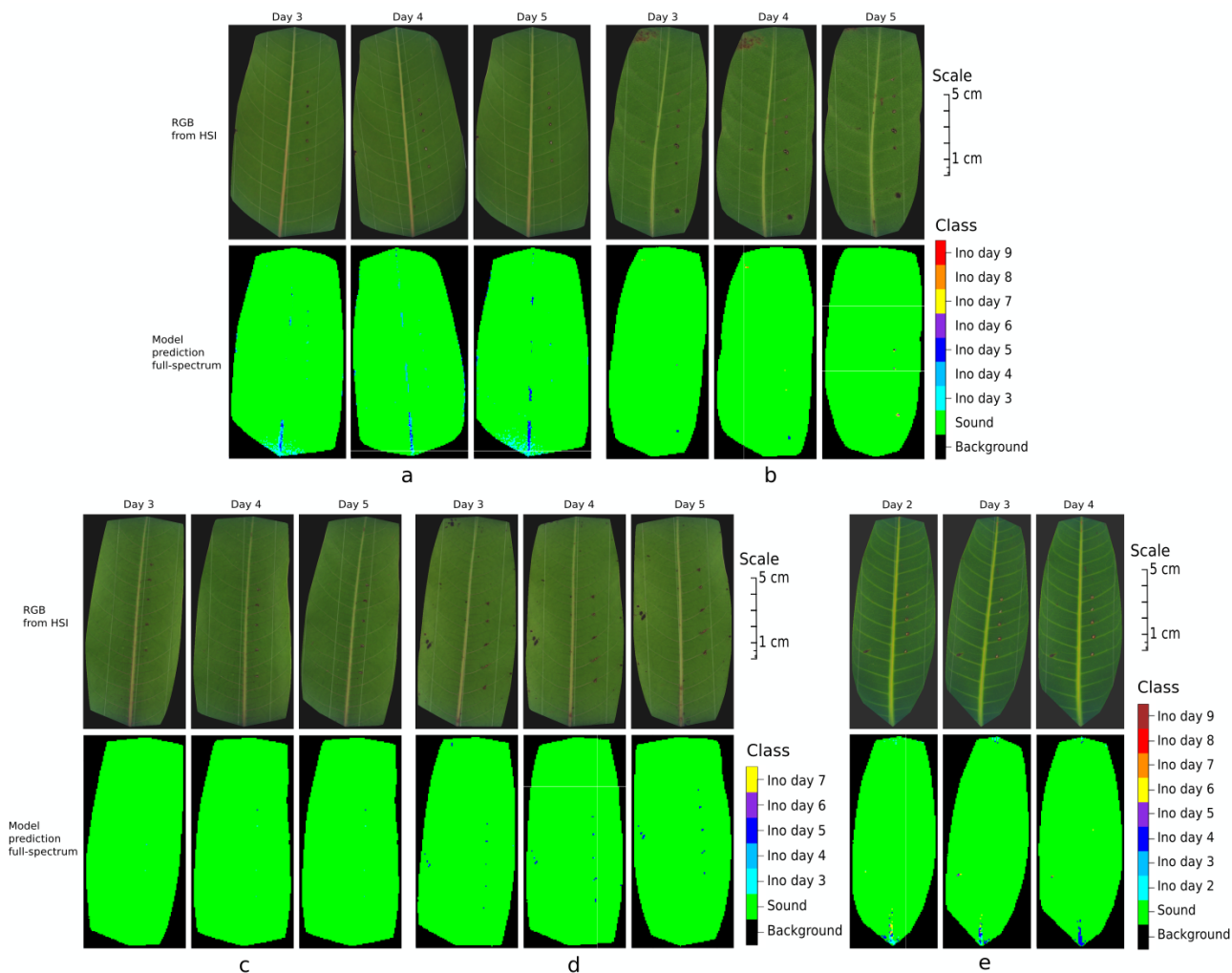


Figure 6-43. Model prediction result of the control samples for the ‘Keitt’ (a and b), ‘Tommy Atkins’ (c and d) and ‘Kent’ (d) leaf assays

able by the human eye, are consistent in size and classification (according to the colour scale) with the pseudo-RGB image. Secondly, in fruit assays, it is notable that from day 2 onward, the models completely detect anthracnose lesions, while in the RGB image, no visible symptoms were reported except for surgical punch lesions. This demonstrated the potential of these models to assist in the automatic detection of anthracnose symptoms in mango and the feasibility of early (previsual) detection of this disease. As a third point, it was found that on all days the lesions detected by the models cover a larger area than the necrotic lesion visible in the RGB image, therefore, it can be deduced that this model can predict in advance the growth (shape, size and location) of anthracnose lesions on fruits. In the 3 zoomed ROIs, it can be seen that the size of the anthracnose lesions covered a larger area than the lesion visible in the RGB image.

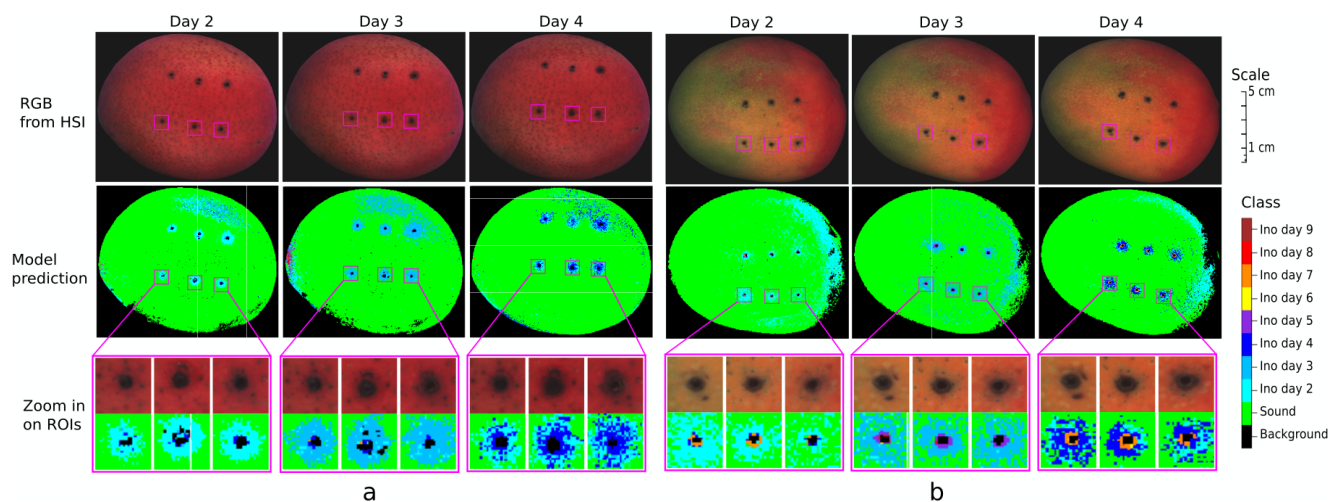


Figure 6-44. Model prediction result of the inoculated samples for the a) ‘Keitt’ and b) ‘Kent’ fruit assays

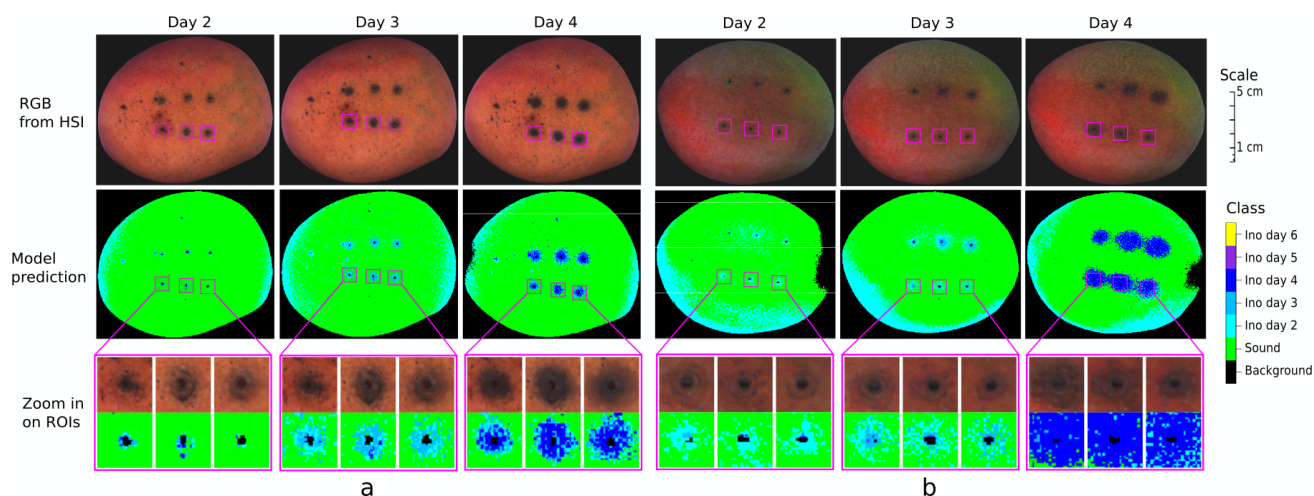


Figure 6-45. Model prediction result of the inoculated samples for the a) ‘Osteen’ and b) ‘Keitt’ fruit assays

Despite the limited predicted lesion size in the *cv.* ‘Osteen’ for day 2 was similar to those defined by the surgical punch. At the spectral level, the detection model could differentiate the different stages of anthracnose development (see section 6.2.5). From day 3 onwards, the predictive power of the selected model can be evidenced, as it was reported for the other two cultivars. Therefore, in the *cv.* ‘Osteen’, the window of time to detect early anthracnose symptoms was 36-24 hours before the symptoms were completely distinguishable.

For the leaf assays, Figures 6-46 and 6-47, the patterns and results reported for the fruit assays were also observed, although the anthracnose lesions in the pseudo-RGB image were visible from the first day of study. However, the small size of these visible lesions makes their adequate detection or discrimination difficult for the human eye. Hence, to confirm the diag-

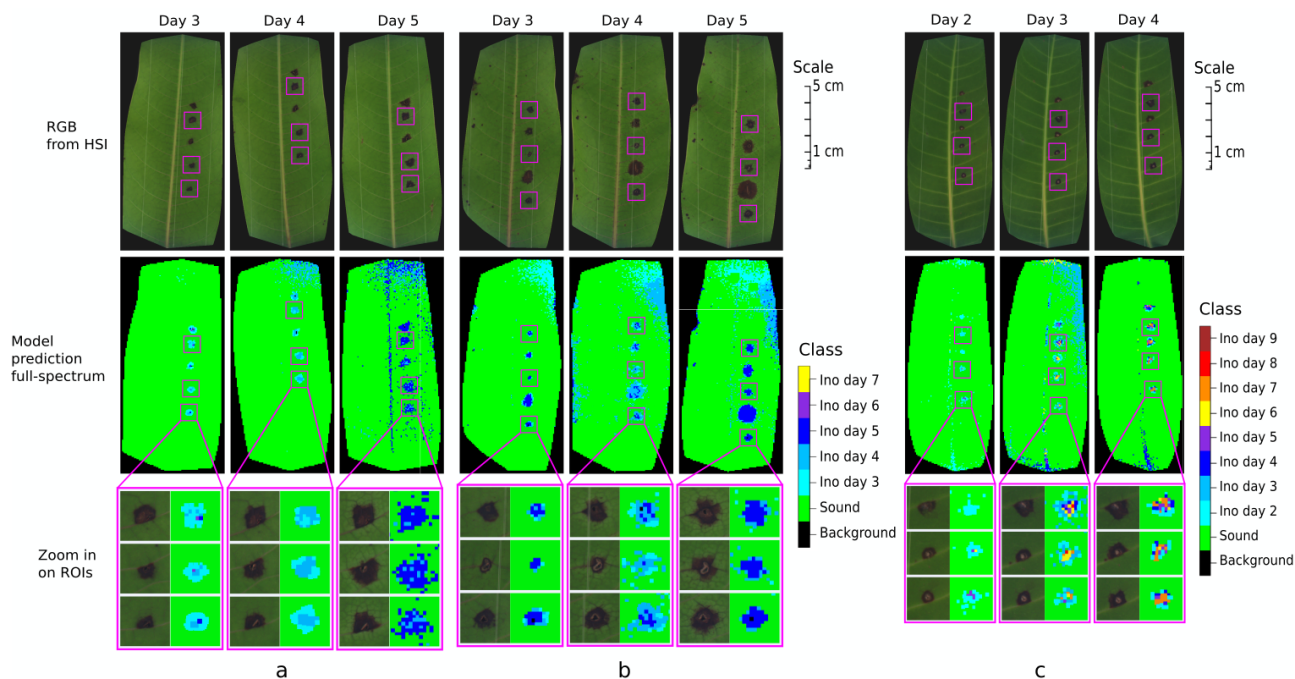


Figure 6-46. Model prediction result of the inoculated samples for the ‘Tommy Atkins’ and ‘Kent’ leaf assays: a) a5-to, b) a6-to and c) a9-kn

nosis of anthracnose symptoms, taking into account this small stage, it is necessary to rely on its evolution (temporal growth), which implies time and disease spread. This drawback could be overcome by the detection model designed for each cultivar, as they showed temporal and spectral consistency in correctly classifying these small lesions from the first days of infection.

Furthermore, as can be seen in the RGB images, the nylon strings used to flatten and manipulate the leaves did not considerably impact the model prediction, as no alterations in the model prediction were observed in these regions (both in control and inoculated samples). Therefore, using this magnetised holder to acquire hyperspectral images was a valid and useful tool. In addition, as a noteworthy result, the anthracnose lesions detected by the models were slightly larger than those visible in the RGB image. Hence, the detection from an early stage (on leaves and from the 2nd or 3rd day of inoculation) stood out as a valuable feature of the models designed and trained in this research.

As a shortcoming reported in the leaf experiments, the models performed detections on the main vein of the leaves, which turned out to be an inaccurate result because at the stage of disease development, no visible symptoms of the disease were visible on this area. In a later stage of anthracnose growth (stage beyond the hyperspectral imaging window) it was reported that the pathogen colonised this portion of the leaf (as it has the most abundant nutrients), however, for the days detailed in the figures this pattern was still not visible, so this was the reason to consider it as an inaccurate result. Besides, it is worth mentioning

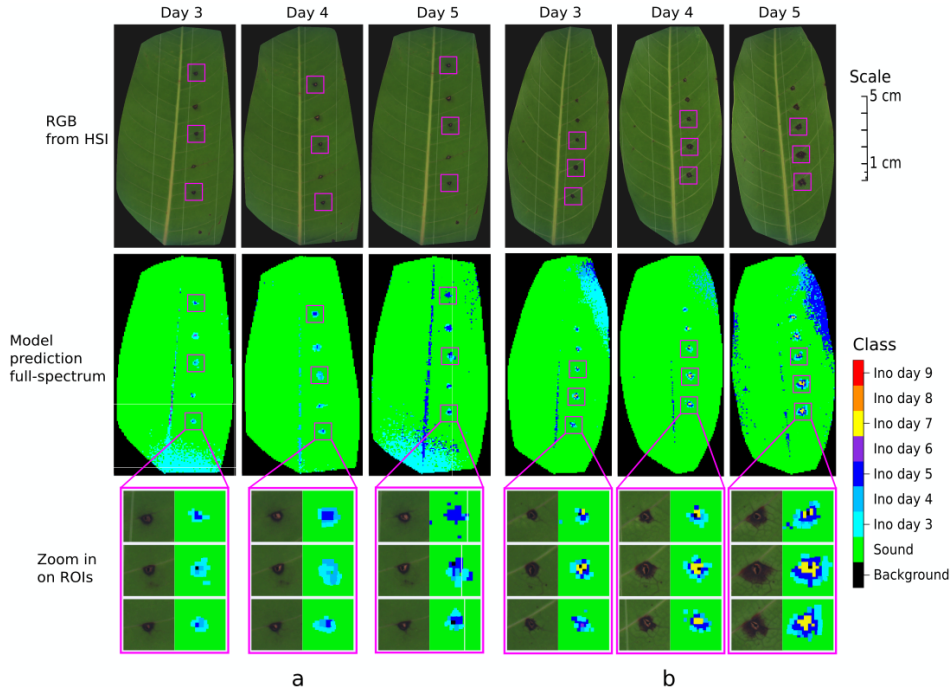


Figure 6-47. Model prediction result of the inoculated samples for the ‘Keitt’ leaf assays: a) a7-kt and b) a8-kt

that for the control samples (see Figure 6-43) this shortcoming was not observed, thus, this result could be considered consistent and relevant for the foliar detection of anthracnose in mango leaves, but further experiments should be carried out to check the validity of this hypothesis or to explain with certainty the origin of this result.

6.3. Identification of relevant wavelengths

As it was analysed in sections 6.2.5 and 6.2.8 with the full spectrum and the classification models, some models did not show acceptable performance (metrics below 0.7) with the datasets and full-spectrum: PLSDA, CART, Naive Bayes, BST, GBM and XGB. Then, their performance with fewer input variables (selected wavelengths) was not considered in this study, as less data for the models directly impacts their results, and only the remaining 6 models were analysed to design and select the best-fitted detection model with a reduced group of wavelengths. Likewise, as filters influence classification model performance, they may also affect the feature selection techniques. Therefore, the assessment of the 4 datasets was also included in this study to determine which combination (filter + feature selector + classification model) provides the best results and performance.

Also, the feature selection methods were evaluated using the training and test groups defined in section 6.2.1 as they were used to train and compare the full-spectrum classification

models. Additionally, the approach implemented in this feature selection study was the multiclass variant as it found it was possible to detect anthracnose symptoms from early to late stages spectrally. Hence, the study was focused on verifying the existence of relevant wavelengths closely related to anthracnose symptoms in mango and had no relation with the feature selector or the classification models.

Finally, as it was mentioned in section 5.2, feature selectors were implemented in R, using the IDE RStudio v. 2022.02.3+492 on Ubuntu Bionic 20.04.5 LTS x64 with AMD® Ryzen 7 5800h processor, 64 Gb RAM + 64 Gb swap RAM and NVIDIA GeForce RTX 3050i 4Gb GPU, and following the official distributions and guidelines of each author.

6.3.1. Feature selectors applied to the spectral characterisation of anthracnose on mango fruit

The feature selectors in section 5.1 were trained with the data in Table 6-4 for the mango fruit assays. Following the flowchart defined in Figure 5-3, the selectors were calibrated, and then the wavelength groups extracted by each technique were consolidated and analysed.

Feature selectors applied to ‘Keitt’ fruit assays

Figures 6-48 and 6-49 present the selected wavelengths in the e1-kt and e4-kt experiments from the 6 feature selectors (FS) and the 4 defined datasets. As a first insight, most selectors’ tendency to extract wavelengths in the NIR region, 750 - 970 nm: while in the e1-kt experiment, the wavelength distribution was in this whole range, in the e4-kt assay, the regions 750 - 810 nm (*NIR1*) and 910 - 970 nm (*NIR2*) were highlighted. The *NIR1* region has been reported in the literature as a region with a high correlation with carbohydrate, TSS and cellulose content [43,117,152], while the *NIR2* region correlates with the presence of sugars (910 - 940 nm) [136,142,152] and a water absorption peak (950 - 980 nm) [43,59,162]. Therefore, the presence of these regions in the sets extracted by the selectors was related to biochemical changes induced by the infection: the pathogen produces a cell membrane rupture that causes water release and degradation of carbohydrates, cellulose and sugars.

Secondly, the SLP4FS and PCA4FS selectors showed 2 regions of the visible spectrum that correlate with anthracnose development as they produce changes in fruit pigmentation (necrotic lesions) in late stages. The regions 450 - 460 nm (*VIS1*) and 670 - 690 nm (*VIS2*) correlate with the pigments chlorophyll b and chlorophyll a, respectively, showing that the degradation of these pigments represents a spectral change of interest to characterise anthracnose symptoms. However, for this mango cultivar, the degradation of other pigments in the fruit, such as carotenes or anthocyanins, did not mean a considerable change to spectrally characterise the defined classes (and thus the different stages of anthracnose infection).

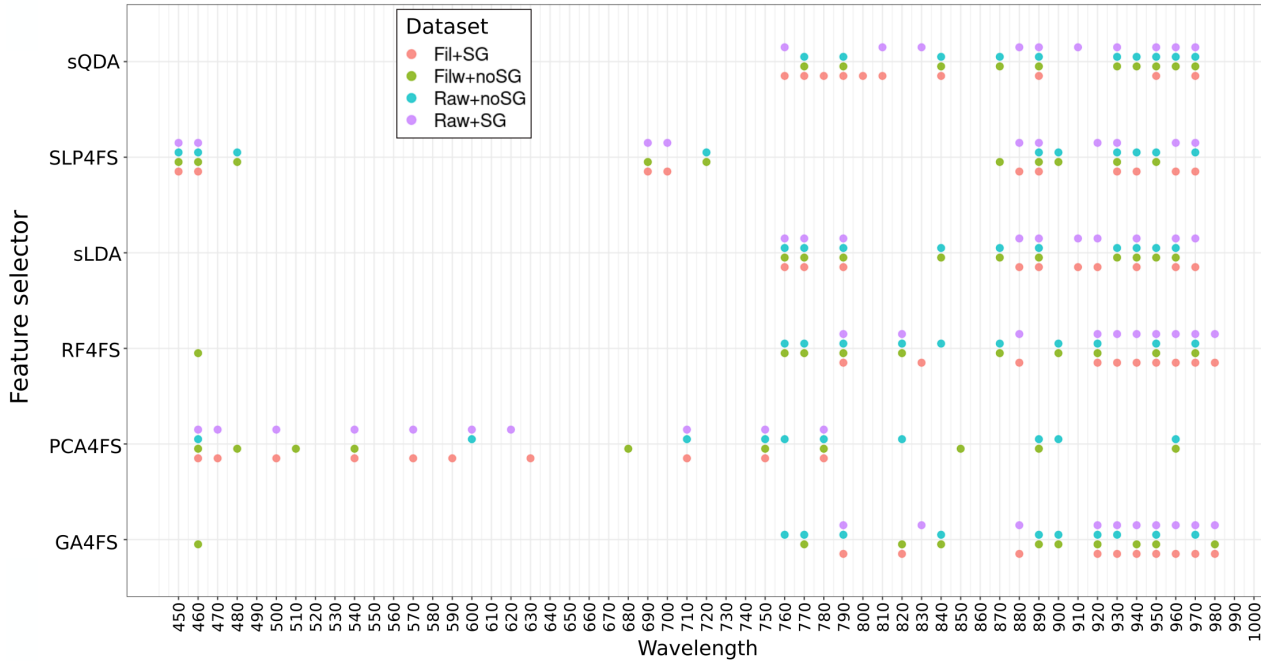


Figure 6-48. Group of wavelengths selected by the feature selectors, per defined dataset, for the e1-kt assay of the ‘Keitt’ mango fruit cultivar

When evaluating the influence of the filters (datasets) on the wavelength selection, using the SG filter induced the greatest variation in the results among the selectors. In the wavelength groups selected with (Raw+SG and Fil+SG) and without (Raw+noSG and Fil+noSG) the SG filter varied considerably in most selectors of the e1-kt assay. In contrast, when the outlier filter was applied in this assay, the variation in these sets was not considerable, being the PCA4FS selector the only exception.

Likewise, when evaluating the influence of the filters in the e4-kt assay, most of the selectors had stable behaviour, showing stability in their estimations despite the application of preprocessing filters. In this experiment, the RF4FS, SLP4FS and GA4FS selectors showed the least variation of the selected sets, obtaining more than 50 % of the same wavelengths in their sets.

Feature selectors applied to ‘Kent’ fruit assay

Figure 6-50 shows the wavelength sets extracted by the 6 FS in the 4 datasets defined for the e2-kn assay of the ‘Kent’ mango fruit cultivar. From the results, the effect of the SG filter in the wavelength selection was again denoted: their use produced a marked differentiation in the resulting spectral band sets. In contrast, the outlier filter produced the opposite effect (stability in the performance of the selectors). Similar to the *cv.* ‘Keitt’, the RF4FS and GA4FS selectors showed the least variation in wavelength selection, while the SQDA selector showed the greatest variation in wavelength selection. Additionally, the



Figure 6-49. Group of wavelengths selected by the feature selectors, per defined dataset, for the e4-kt assay of the ‘Keitt’ mango fruit cultivar

PCA4FS selector had the largest wavelength distribution across the Vis-NIR range under study, while the rest of the selectors were concentrated in the red-edge (RE) and NIR regions.

As a common point with the *cv.* ‘Keitt’, the *NIR1* and *NIR2* regions were relevant for the characterisation of the classes defined for this assay, indicating the strong spectral correlation with the degradation of carbohydrates (starches and sugars), cell membrane rupture and water release under the epicarp of the fruit. As a difference of this cultivar, the degradation of chlorophyll *a* (*VIS2* region) was a relevant region for most of the selectors, while the *VIS1* region (related to chlorophyll *b* degradation) presented a punctual relevance for the selectors PCA4FS and SQDA.

Additionally, for the SLP4FS, SLDA and PCA4FS selectors, the region related to carotenes, specifically β -carotene, was listed between 470-480 nm (*VIS3*), which revealed the importance of the degradation of this pigment in mango fruits for these selectors. As a particularity of the PCA4FS selector, the anthocyanin pigment was chosen within the 4 implemented datasets, region between 520 - 530 nm (*VIS4*), indicating that for this particular selector, the degradation of the 4 pigments (chlorophyll *a*, chlorophyll *b*, β -carotene and anthocyanin) were relevant.

As it could be observed, for the *cv.* ‘Kent’, there was a tendency to characterise anthracnose symptoms through the degradation of pigments in the fruit, however, the internal biochemi-



Figure 6-50. Groups of wavelengths selected by the feature selectors, per defined dataset, for the e2-kn assay of the ‘Kent’ mango fruit cultivar

cal changes induced by the pathogen (and correlated with the NIR region) were also relevant.

Feature selectors applied to ‘Osteen’ fruit assay

Figure 6-51 shows the wavelength groups selected by the 6 FS in the 4 datasets defined for the e3-os assay of the *cv.* ‘Osteen’ on mango fruits. The FS extracted relevant wavelengths from the NIR spectral range, specifically in the *NIR1* (740 - 810 nm) and *NIR2* (910 - 970 nm) regions. Similar to the *cv.* ‘Keitt’, it was observed that the selectors found major interest in the NIR range over the visible range to characterise the classes defined in this assay, thus showing a clear tendency to spectrally distinguish the internal biochemical changes produced by the infection over the external changes (fruit pigment degradation). The biochemical changes induced by the infection in the fruit are the carbohydrate degradation (mainly starches and sugars), the cellulose degradation and the cell membrane rupture leading to water release under the epicarp of the fruit.

However, in the SLP4FS, GA4FS and RF4FS selectors, the *VIS1* region (450 - 460 nm) was shown to be important as it was extracted by these selectors in all 4 datasets, evidencing the relevance of chlorophyll *b* degradation in the anthracnose symptom characterisation in fruits of this mango cultivar.

A peculiarity of the feature selectors in the *cv.* ‘Osteen’ was the stability of the results with the use of preprocessing filters: the SLP4FS, GA4FS and RF4FS selectors maintained more

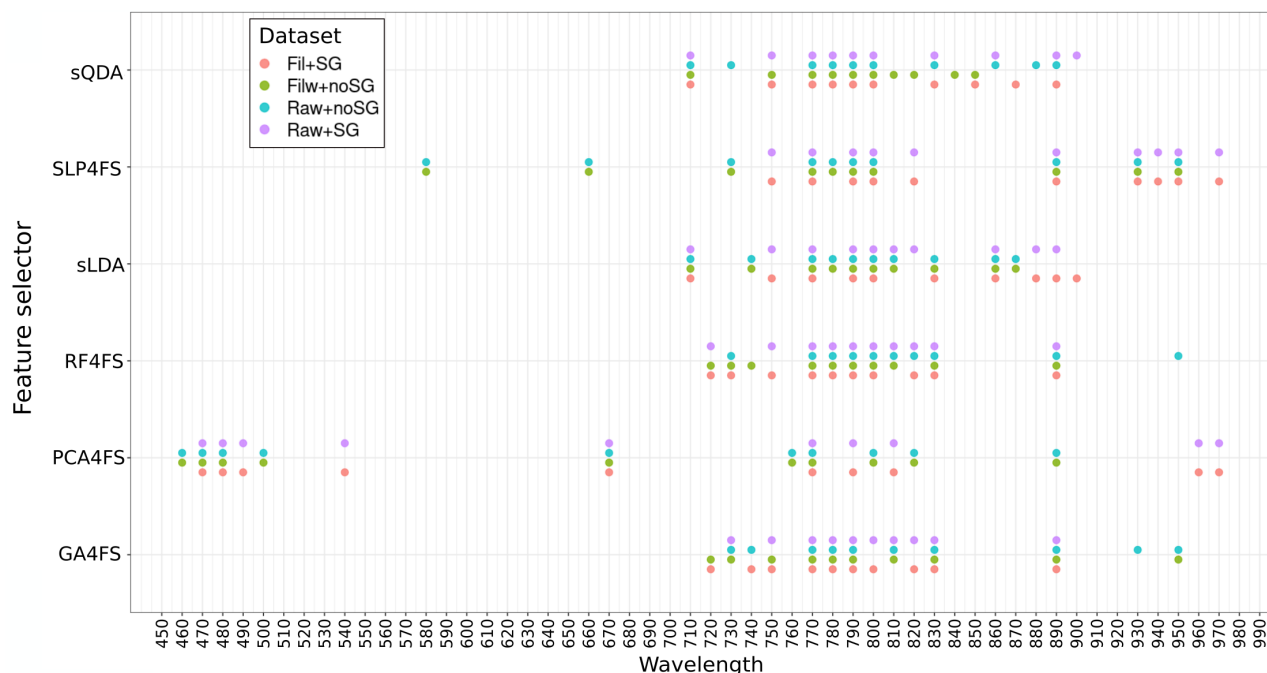


Figure 6-51. Group of wavelengths selected by the feature selectors, per defined dataset, for the e3-os assay of the ‘Osteen’ mango fruit cultivar

than 50 % of the selected wavelengths, while the SQDA and SLDA selectors preserved about 40 % of these wavelengths. The result obtained in this analysis, and validated on the other mango cultivars, shows that the use or not of the preprocessing filters did not have a direct and relevant influence on the feature selection. Therefore, to validate the use or not of a filter in the feature selection step, the performance of the best-fit classification model with the relevant wavelength group extracted by each dataset (filter) was computed.

6.3.2. Multiclass models for anthracnose detection on mango fruits

The results associated with the performance metrics of the 10 best classification models with the feature selectors (FS) and filters (4 datasets) are presented in Tables 6-16 - 6-19 for the e1-kt, e2-kn, e3-os and e4-kt assays, respectively. Hereinafter, the model trained with the selected wavelengths will be referred to as a “selected wavelength model” when comparisons are introduced with full-spectrum models. The rest of the models trained and calibrated with the datasets and feature selectors implemented in this research can be consulted in Appendix A.3.

Tables 6-16 - 6-19 show, firstly, the dominance of the neural models among the best-performing techniques for the discrimination of the classes defined in the assay e1-kt for the ‘Keitt’ mango fruit cultivar. These models obtained outstanding performance metrics (scores above 0.8) for all experiments. Within these metrics, specificity stood out with scores above

0.96, which evidenced that all these models had an outstanding ability to discriminate the negative class, which for this research corresponded to the classes representing the different developmental stages (symptoms) of anthracnose in mango.

The best classification model for the 4 experiments was SLP, with metrics above 0.85 (except for MCC in the e4-kt study), stating its ability to learn from datasets with a limited number of input variables (wavelengths). The hyperparameters for the SLP models trained with the 10 most relevant wavelengths were: *neurons* = 200, *epochs* = 382 for the e1-kt, *neurons* = 200, *epochs* = 452 for the e2-kn, *neurons* = 50, *epochs* = 500 for the e3-os, *neurons* = 200, *epochs* = 309 for the e4-kt. From these results, it can be deduced that, as there were no significant variations in the parameters that most affected the model, SLP proved to be a robust and stable technique against variations in the input datasets. Therefore, this technique can guarantee anthracnose symptom detection in mango fruits for developing neuro-spectral detection models with a limited number of spectral bands.

Table 6-16. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e1-kt assay of the *cv.* ‘Keitt’ on mango fruit

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.866	0.867	0.983	0.866	0.851	0.869	Raw+noSG	SLP4FS
MLP	0.866	0.866	0.983	0.866	0.850	0.870	Raw+noSG	SLP4FS
MLP	0.865	0.865	0.983	0.865	0.848	0.866	Raw+SG	SLP4FS
MLP	0.847	0.848	0.981	0.848	0.831	0.853	Raw+SG	SLDA
MLP	0.846	0.847	0.981	0.847	0.829	0.850	Raw+SG	SQDA
SLP	0.843	0.844	0.980	0.843	0.825	0.847	Raw+SG	SQDA
MLP	0.832	0.835	0.979	0.828	0.815	0.848	Raw+noSG	SLDA
MLP	0.831	0.833	0.979	0.829	0.814	0.843	Raw+noSG	GA4FS
SLP	0.830	0.831	0.979	0.828	0.811	0.838	Raw+noSG	GA4FS
MLP	0.825	0.826	0.978	0.824	0.805	0.829	Raw+SG	GA4FS

Table 6-17. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e2-kn assay of the *cv.* ‘Kent’ on mango fruit

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.884	0.884	0.985	0.885	0.871	0.887	Raw+SG	SQDA
SLP	0.872	0.873	0.984	0.873	0.859	0.878	Raw+noSG	SQDA
MLP	0.871	0.872	0.984	0.871	0.857	0.877	Raw+noSG	SQDA
MLP	0.858	0.857	0.982	0.856	0.845	0.875	Raw+noSG	SLP4FS
SLP	0.842	0.844	0.980	0.837	0.825	0.854	Raw+noSG	SLP4FS
MLP	0.839	0.840	0.980	0.837	0.822	0.850	Raw+SG	SLP4FS
SVM	0.839	0.840	0.980	0.840	0.820	0.841	Fil+SG	SQDA
SLP	0.832	0.832	0.979	0.829	0.814	0.844	Raw+noSG	GA4FS
SVM	0.831	0.832	0.979	0.832	0.811	0.833	Fil+noSG	SLP4FS
MLP	0.828	0.829	0.979	0.815	0.815	0.872	Fil+noSG	SLP4FS

Furthermore, when comparing these models with the full-spectrum version for each experiment (presented in section 6.2.5), it was observed that the reduction in performance metrics did not exceed 10 %, while the reduction in wavelengths exceeded 80 % (going from 54 to 10 wavelengths). This result also validated the premise of the existence of redundant wavelengths (which did not contribute to the characterisation of classes), so the use of FS is necessary 1) to reduce computational cost in model training and calibration to avoid possible overfitting, 2) to reduce costs in the design of AI-based solutions and 3) to enhance the abilities of AI-based models to adapt to data with a limited number of input variables.

Table 6-18. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e3-os assay of the *cv.* ‘Osteen’ on mango fruit

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.895	0.895	0.979	0.895	0.875	0.897	Raw+SG	SLP4FS
MLP	0.884	0.884	0.977	0.884	0.862	0.888	Raw+SG	SLP4FS
QDA	0.855	0.855	0.971	0.855	0.826	0.855	Fil+SG	SLP4FS
MLP	0.852	0.852	0.970	0.852	0.823	0.854	Raw+SG	SQDA
SLP	0.850	0.850	0.970	0.850	0.820	0.850	Raw+SG	SQDA
SVM	0.846	0.847	0.969	0.847	0.816	0.847	Fil+SG	SLP4FS
QDA	0.845	0.845	0.969	0.845	0.814	0.845	Raw+SG	SLP4FS
SVM	0.844	0.844	0.969	0.844	0.813	0.844	Raw+noSG	GA4FS
QDA	0.842	0.842	0.968	0.842	0.811	0.843	Raw+noSG	GA4FS
SVM	0.837	0.837	0.967	0.837	0.805	0.837	Raw+SG	SLP4FS

Similarly, from the results reported in the mentioned tables, it was observed that the raw datasets allowed the highest score of the classification models, which indicated that for the reduced version (in wavelengths) the non-use of preprocessing filters allowed better spectral characterisation of the classes defined for each experiment. This finding contrasts with the result obtained for the full-spectrum version, where the outlier filter (Fil+noSG dataset) allowed the best model performance.

Table 6-19. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e4-kt assay of the *cv.* ‘Keitt’ on mango fruit

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.865	0.865	0.973	0.864	0.839	0.867	Raw+noSG	SLP4FS
SLP	0.847	0.848	0.969	0.849	0.821	0.858	Raw+noSG	SQDA
SLP	0.839	0.840	0.968	0.840	0.810	0.846	Raw+SG	SLDA
MLP	0.837	0.839	0.967	0.839	0.811	0.851	Raw+noSG	SLDA
MLP	0.835	0.836	0.967	0.835	0.804	0.840	Raw+SG	SLDA
MLP	0.829	0.832	0.966	0.831	0.802	0.844	Raw+noSG	SQDA
MLP	0.824	0.824	0.965	0.823	0.789	0.826	Raw+SG	SLP4FS
SVM	0.816	0.817	0.963	0.817	0.780	0.817	Fil+noSG	SQDA
RF	0.813	0.813	0.963	0.813	0.776	0.814	Fil+noSG	SQDA
SVM	0.812	0.812	0.962	0.812	0.775	0.813	Fil+SG	SQDA

On the other hand, the feature selectors with the best ranking models were SLP4FS and SQDA, with SLP4FS being ranked in the top positions for the e1-kn, e3-os and e4-kn assays, while SQDA dominated in the e2-kn study. The GA4FS technique was found in all experiments as a third feature selector, while the SLDA selector was found as an alternative in the *cv.* ‘Keitt’. This dominance of the neural selector is partly explained by the ability of neural networks to adapt to the data and their combination with neural classification models. Therefore, the combination of the neural feature selector + neural classification model was expected to give the best results in this research.

However, this combination was not the best performing for all cultivars, as in the e2-kn assay, it was evident that the selector that enhanced the neural model performance was SQDA. This result also revealed the usefulness of discriminant models for data prediction/classification and class characterisation with a reduced set of variables. Even for the *cv.* ‘Keitt’, e1-kt and e4-kt assays, the SQDA and SLDA selectors were also found to be alternatives to the neural selector.

Finally, since this study addressed the relationship of spectral information with anthracnose symptomatology in mango fruit, selecting the best set of wavelengths and classification model per assay is presented in detail in section 6.3.3. This section discusses the physiological and biological aspects of the pathogen-induced infectious process in mango fruit that enrich and validate the choice of one set and classification model over another.

6.3.3. Relevant wavelength selection for spectral characterisation of anthracnose on mango fruit

The wavelength sets of the most prominent feature selectors per test are shown in Table 6-20. This table shows the wavelengths in order of importance, with #1 being the most relevant and 10 being the least important within each set. Additionally, the datasets selected for each selector were made according to their position and prevalence in Tables 6-16 - 6-19.

Table 6-20. Sets of the 10 most representative wavelengths by best-performing feature selector in each assay conducted for the three mango fruit cultivars

Rank	e1-kt			e2-kn			e3-os			e4-kt		
	SLP4FS	SQDA	SLDA	SLP4FS	SQDA	GA4FS	SLP4FS	SQDA	GA4FS	SLP4FS	SQDA	SLDA
1	950	810	760	950	800	950	770	710	770	450	760	770
2	460	880	790	720	950	920	790	750	890	760	920	920
3	900	890	770	710	970	900	800	770	810	950	860	940
4	450	910	910	900	920	770	890	780	790	930	950	970
5	930	970	920	470	930	790	750	790	780	700	960	930
6	940	960	890	700	910	720	820	800	730	710	930	910
7	480	930	880	690	760	970	930	830	830	920	790	780
8	890	950	940	760	680	820	950	860	950	940	940	790
9	970	760	970	770	740	980	970	890	740	810	740	800
10	720	830	960	670	450	730	940	900	930	460	970	980

As a first finding, the most prevalent range within the sets was between 730 - 910 nm, which, as reported in studies [117,152], correlates with the TSS that predicts the presence of carbohydrates (mainly starches and sugars). In [136], a ripening index in mango with 15 wavelengths was established, which was mostly composed of this spectral range. Particularly, for the *cv.* 'Osteen', assay e3-os, this range occupied the first 6 positions of the 3 selectors and obtained more than 60 % of the relevant variables per selector. Similarly, this range concentrated most of the relevant wavelengths for the selectors based on discriminant analysis (SLDA and SQDA). For instance, the SLDA and SQDA selectors of the e1-kt test shared 6 of the 10 most important wavelengths.

Similarly, a spectral range in most selectors, related to the quantification of sugars in fruits, was 920 - 940 nm. Although sugars fall into the carbohydrate group, they can be detected more easily by the water accumulation (or release) [136,152]. Therefore, water accumulation in the fruit will highlight the spectral range 920 - 940 nm to indicate the presence (or degradation) of sugars. As seen, in all the selectors of the *cv.* 'Keitt' (e1-kt and e4-kt assays), this spectral range was found, while in the 'Kent' and 'Osteen' cultivars, it was only present in 2 of the 3 selectors with 1 or 2 wavelengths.

On the other hand, symptoms closely related to anthracnose in mango are water release under the fruit peel and pigment degradation such as chlorophyll (*a* and *b*), carotenes (β -carotene) and anthocyanins. The spectral range characterising the water absorption peak is 950 - 980 nm. This range was located in the first positions of the 3 selectors of the *cv.* 'Kent' (e2-kn test), while in all the selectors of the *cv.* 'Keitt', it was present with at least 1 representative wavelength. For the *cv.* 'Osteen', this range was only listed in 2 of the 3 selected selectors, SLP4FS and GA4FS.

Concerning pigment degradation, it was observed that the spectral ranges associated with chlorophyll and carotenes were only found in the SLP4FS selectors of the 'Keitt' and 'Kent' cultivars, and the SQDA selector of the *cv.* 'Kent'. For the *cv.* 'Keitt', e1-kt and e4-kt assays, the SLP4FS selector included wavelengths related to pigments chlorophyll *b*, 460 nm, and carotenes, 450 and 480 nm. For the *cv.* 'Kent', the SLP4FS and SQDA selectors indicated the relevance of spectral changes (degradation) of pigments chlorophyll *a* (670 - 690 nm) and *b* (470 nm), and chlorophyll *a* (680 nm) and carotenes (450 nm), respectively. However, for the *cv.* 'Osteen', no wavelengths related to these pigments were reported. These results revealed the dependence on fruit cultivar for the anthracnose symptoms characterisation because not all fruits ripen in the same way, and not all mango cultivars have the same pigment composition in their peel.

Therefore, and considering the previous exposition of spectral ranges related to anthracnose symptoms in mango fruits, the selection of the best set of wavelengths per cultivar was

carried out considering the highest presence of symptoms within the sets, i.e. the highest presence of the aforementioned spectral ranges within the 10 wavelengths of each selector:

- For the *cv.* ‘Kent’, the selected wavelength group corresponded to the one extracted by the SQDA selector. All 10 wavelengths were related to symptoms produced by anthracnose, with water release and carbohydrate degradation (starches and sugars) being the most relevant symptoms. Additionally, this selector included within its set the changes produced by the degradation of chlorophyll *a* and carotenes, which evidenced the importance of these pigments in the spectral characterisation of anthracnose in this cultivar. Therefore, the best detection model with this selector was SLP with *neurons* = 200, *epochs* = 452, *nfolds* = 3 and Raw+SG dataset. As mentioned in section 6.3.2, this detection model obtained the best performance among all the implemented classification models, reaffirming the importance and robustness of the wavelength set extracted by the SQDA selector.
- For the *cv.* ‘Keitt’, it was found that the SLP4FS selector adequately represented the most representative anthracnose symptoms for this cultivar: 5 spectral ranges related to peak water absorption, carbohydrate degradation (starches and sugars) and pigment degradation (chlorophyll *b* and carotenes) were represented within its sets. Even though the sets of both assays differed, it is evident that the feature selector extracted the wavelengths 460, 460, 930, 940 and 950 nm from both experiments, indicating that 50% of relevant wavelengths for the *cv.* ‘Keitt’ assays were common, while the rest were distributed in other regions also related to spectral ranges associated with anthracnose symptoms.

Therefore, the best-performing classification model with this selector was SLP, combined with the Raw+noSG dataset and the hyperparameters *neurons* = 200, *epochs* = 382 or 309, *nfolds* = 3. Similar to the *cv.* ‘Kent’, these models corresponded to the best-performing models according to Tables **6-16** and **6-19** and demonstrated the feasibility of characterizing the anthracnose symptomatology in mango fruits no matter the experimental conditions: both assays differed in the number of days under study (9 vs. 7), the number of samples (55 vs. 100) and harvesting season. Therefore, the findings reported in this research were relevant since they showed the non-dependence on the seasonality or fruit quantity to perform the anthracnose diagnosis with an AI-based model and a group of relevant spectral bands.

- For the *cv.* ‘Osteen’, experiment e3-os, given the focus of the feature selectors in the NIR region, the selector that best represented the anthracnose symptoms for this cultivar was SLP4FS. All wavelengths of this selector were related to spectral ranges of interest for anthracnose, with the symptoms of carbohydrate degradation (mainly starches and sugars) and water release being the most present in this set. The latter

also had the particularity that all these symptoms were not visible because they were in a range of the spectrum not visible to the human eye. The other selectors (SQDA and GA4FS), although they stood out with the classification models implemented, focused their attention on the range 730 - 910 nm, minimising the anthracnose symptoms, such as sugar degradation and water release.

Like the other cultivars, the best-performing classification model with this selector was SLP, combined with the Raw+SG dataset and the hyperparameters *neurons* = 50, *epochs* = 500, *nfolds* = 3. This result confirmed the connection between the anthracnose symptoms represented in the relevant wavelength set and the detection models' performance, as the selectors that performed a better spectral characterisation of the symptoms enhanced the predictive ability of the classifiers with remarkable performance metrics.

6.3.4. Feature selectors applied to the spectral characterisation of foliar anthracnose on mango leaves

The feature selectors in section 5.1 were trained with the data in Table 6-4 for the mango leaf experiments. Following the flowchart defined in Figure 5-3, the selectors were calibrated, and then the wavelength groups extracted by each technique were consolidated and analysed.

Feature selectors applied to 'Tommy Atkins' leaf assays

Figures 6-52 and 6-53 show the wavelength groups selected in the e5-to and e6-to tests, respectively. These figures compile the results obtained by the 6 FS in the 4 defined datasets. From these figures, there were 2 spectral ranges of interest for the selectors: a region located in the visible spectrum, 450 - 470 nm, and a region located in the NIR spectrum, 940 - 980 nm. The first region, 450 - 480 nm, is related to the presence of pigments chlorophyll *b* (with a peak reflectance at 450 - 460 nm) and carotenes (with a peak reflectance at 470 - 480 nm). These pigments are responsible for giving the greenish and yellowish tones to the leaves and fruits, although in mango cultivars, carotenes are more present in the fruits than in the leaves. The 940 - 980 nm region is mainly related to the presence of water, which corresponds to the water release under the leaf cuticle due to the cell membrane rupture that causes the infectious process induced by *Colletotrichum sp.* These regions were also reported in the 'Keitt' and 'Kent' cultivars, so it can be stated that foliar anthracnose at the spectral level is closely related to pigment degradation and surface water release (accumulation).

In addition, the 520 - 540 nm, 720 nm and 900 - 920 nm spectral ranges were also relevant for specific selectors in the two assays for this cultivar. These regions are related to the presence of anthocyanins, carbohydrates and sugars, respectively. However, given the limited availability of nutrients in mango leaves, their relevance at the spectral level is limited, therefore,

the importance in the spectral characterisation of foliar anthracnose in this cultivar will depend on their contribution to the calibration and training of the implemented classification models: if these sets enhance the class discrimination (detection of the different stages of anthracnose), the relevance of these regions will be the differential over other selectors that did not include them.

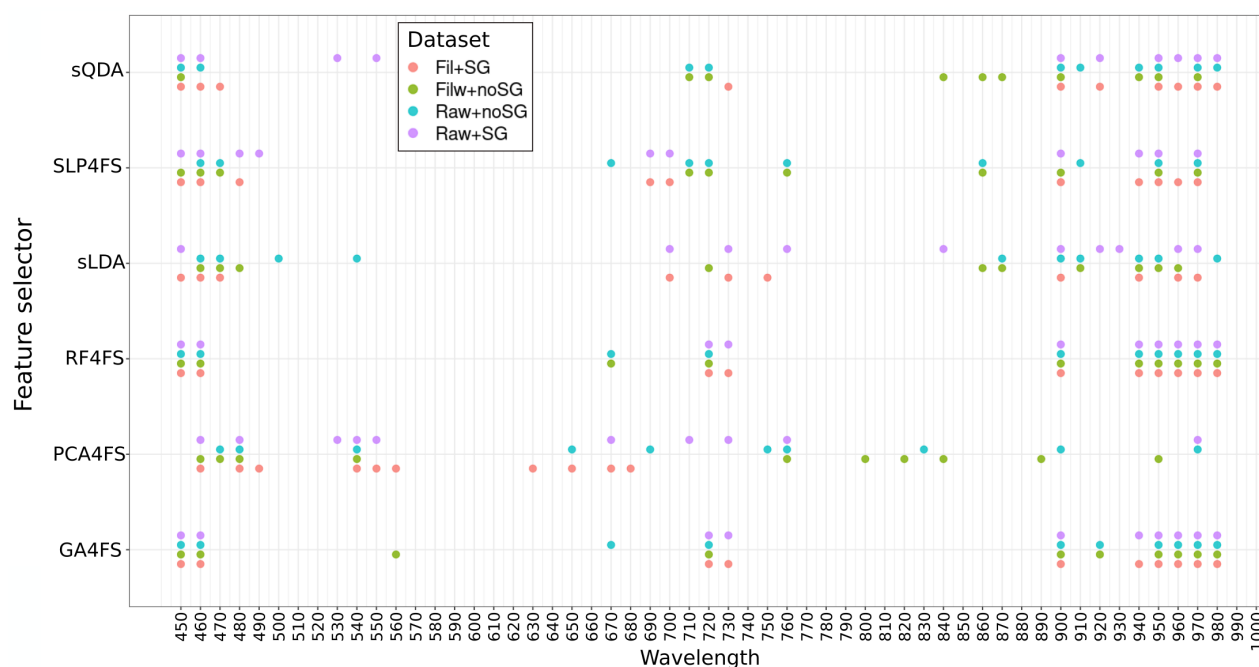


Figure 6-52. Group of wavelengths selected by the feature selectors, per defined dataset, for the ‘Tommy Atkins’ e5-to assay on mango leaves

Among the selectors that reported less variation in their results when using the preprocessing filters were RF4FS and GA4FS, as these techniques preserved more than 50 % of their variables regardless of the 4 datasets tested. Likewise, SLP4FS and SQDA presented an intermediate variation, i.e., these techniques managed to keep about 40 % of the variables unaltered.

Feature selectors applied to ‘Keitt’ leaf assays

Figures 6-54 and 6-55 present the selected wavelength groups in the e7-kt and e8-kt tests, respectively. These figures compile the results obtained by the 6 FS in the 4 defined datasets. For both ‘Keitt’ assays, the NIR spectral region, 900 - 980 nm, was highlighted as the most relevant range for the class characterisation in this cultivar. Indeed, more than 50 % of the selected wavelengths by the selectors were located in this portion of the spectrum. Moreover, the 450 - 460 nm range was denoted by its intermediate relevance, as half of the selectors included it in their sets. These two regions, as reported for the other cultivars, are closely



Figure 6-53. Group of wavelengths selected by the feature selectors, per defined dataset, for the ‘Tommy Atkins’ e6-to assay on mango leaves

related to the sugar degradation (limited in leaves), the water release due to the cell membrane rupture in the leaf and the degradation of chlorophyll *b* (pigment also responsible for the photosynthetic activity in leaves).

Similarly, there were regions related to other pigments and compounds in leaves, such as the 470 - 480 nm, 670 - 680 nm and 760 - 780 nm ranges; however, their incidence was lower, as no strong selection pattern was reported. Therefore, when the implemented classification models in this study are evaluated, the usefulness and relevance of these regions for the spectral characterisation of foliar anthracnose on ‘Keitt’ mango will be confirmed.

On the other hand, when evaluating the individual performance of the selectors, the RF4FS and GA4FS techniques were the most stable in the wavelength selection regardless of the filter (dataset) used: more than 50 % of their variables were preserved in the 4 datasets evaluated. SLP4FS showed this behaviour in the e7-kt assay, while for the e8-kt one, this percentage was reduced to about 40 %.

Feature selectors applied to ‘Kent’ leaf assay

Figure 6-56 shows the wavelength groups selected by the 6 FS in the 4 datasets defined for the e9-kn test of the *cv.* ‘Kent’ on mango leaves. From this figure, it was observed that the RF4FS, GA4FS and SLP4FS selectors were the most stable in the wavelength selection regardless of the filter (dataset) used: more than 50 % of their variables were maintained in



Figure 6-54. Group of wavelengths selected by the feature selectors, per defined dataset, for the e7-kt ‘Keitt’ mango leaf assay

the 4 datasets evaluated. However, for SQDA, SLDA and PCA4FS, this same behaviour was not reported, which did not allow a conclusive result on the influence of the preprocessing filters. In fact, contrary to the fruit and other cultivar assays, where the SG filter induced changes in selection, for the *cv.* ‘Kent’ in the DA-based selectors, there was an association of the outlier filter (Fil+noSG and Fil+SG) with wavelength selection: wavelengths with this filter varied compared to the datasets that did not use it (Raw+noSG and Raw+SG) while remaining stable (equal) in the datasets that used it.

On the other hand, when analysing the extracted wavelength sets, it was found that the region with the highest relevance for the selectors was located between 920 - 970 nm (called *NIR2* in section 6.3.1). This region, as mentioned above, is directly related to the presence of sugars and water accumulation. The next most relevant region is the 450 - 460 nm region (called *VIS1* in section 6.3.1), which was present in the RF4FS, GA4FS PCA4FS and SLP4FS selectors. This region is closely related to the degradation of chlorophyll b, the pigment responsible for giving the greenish hue to leaves and for most of the photosynthetic activity in the tree, so it is understood that the infectious process limits the capacity of the infected leaf for the production of energy and nutrients needed by the tree.

Similarly, within the sets reported for this cultivar, there was a wavelength, 720 nm, and a spectral range, 760 - 770 nm, of medium importance for the selectors. As reported in section 6.3.1, this region is related to cellulose and carbohydrate degradation in plant material.



Figure 6-55. Group of wavelengths selected by the feature selectors, per defined dataset, for the e8-kt ‘Keitt’ mango leaf assay

Therefore, its relevance will be determined purely by its influence on the performance of the classification models to discriminate the classes defined for this cultivar: as these regions were only found in particular sets, the combination of these with the classification model may result in a unique and reliable foliar anthracnose detection model for the ‘Kent’ mango cultivar.

6.3.5. Multiclass models for foliar anthracnose detection on mango

The models presented in section 4.1 were trained with the data presented in Table 6-4 for the mango leaf assays. Following the flowchart defined in Figure 5-3, the model training was carried out, and the results were consolidated and analysed to determine the potential of AI models for early detection of anthracnose symptoms in leaves of the 3 mango cultivars. Furthermore, in this section only the 10 best classification models are presented (see Tables 6-21 - 6-25), while the rest of the models can be consulted in Appendix B.3

Selected wavelength multiclass models for anthracnose symptom detection on ‘Tommy Atkins’ mango leaves

Tables 6-21 and 6-22 present the results associated with the performance metrics of the 10 best classification models for the e5-to and e6-to tests of the *cv.* ‘Tommy Atkins’ on mango leaves. From these tables, models with outstanding performance were obtained in the e5-to assay, with metrics above 0.9, and notable performance in the e6-to assay, with



Figure 6-56. Group of wavelengths selected by the feature selectors, per defined dataset, for the e9-kn test of the ‘Kent’ mango leaf cultivar

metrics between 0.83 and 0.9. Specifically, the best models for this cultivar were the neural models, with MLP (*neurons* = 200, *layers* = 3) being the best classifier for the e5-to assay and SLP (*neurons* = 100, *epochs* = 500) for the e6-to one.

Table 6-21. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e5-to *cv.* ‘Tommy Atkins’ experiment on mango leaves

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
MLP	0.945	0.944	0.989	0.945	0.934	0.947	Fil+noSG	SLP4FS
QDA	0.942	0.940	0.988	0.941	0.929	0.942	Fil+noSG	SLP4FS
SVM	0.941	0.940	0.988	0.940	0.928	0.941	Fil+SG	SLDA
QDA	0.933	0.931	0.987	0.932	0.919	0.934	Fil+noSG	SLDA
MLP	0.933	0.933	0.987	0.932	0.919	0.933	Fil+SG	SLDA
SVM	0.932	0.931	0.987	0.931	0.919	0.934	Fil+noSG	SLP4FS
QDA	0.930	0.929	0.986	0.930	0.917	0.933	Fil+noSG	SQDA
QDA	0.927	0.926	0.985	0.927	0.912	0.928	Fil+SG	SQDA
SLP	0.925	0.923	0.985	0.924	0.911	0.931	Fil+noSG	SQDA
QDA	0.923	0.922	0.985	0.923	0.908	0.924	Fil+SG	SLDA

Following the neural models was the QDA model, with scores of accuracy = 0.942, recall = 0.940 and specificity = 0.988 in the e5-to assay and accuracy = 0.875, recall = 0.875 and specificity = 0.975 in the e6-to one. The result of this discriminant model stood out since the metrics obtained were similar (difference less than 1 %) to the neural model metrics and

validated the great performance obtained by this model in the full-spectrum version. On the other hand, the SVM model, which presented outstanding behaviour with the full-spectrum, did not appear among the top positions for this study, revealing a possible weakness of this technique when classifying data with reduced inputs (wavelengths).

When analysing the feature selectors, it was evident that the best techniques were SLP4FS and SQDA for the e5-to and e6-to assays, respectively, showing the disparity in the results obtained. However, the same cultivar of mango leaves was being evaluated under the same experimental conditions. However, as alternative selectors to those mentioned above, the SLDA technique was able to position classification models within the first positions for both assays, revealing the potential of discriminant models (SQDA and SLDA) for the wavelength selection.

Table 6-22. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e6-to *cv.* ‘Tommy Atkins’ experiment on mango leaves

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.887	0.887	0.977	0.888	0.868	0.895	Raw+noSG	SQDA
QDA	0.875	0.875	0.975	0.874	0.851	0.878	Raw+noSG	SQDA
SLP	0.875	0.875	0.975	0.875	0.854	0.885	Raw+SG	SLDA
SLP	0.873	0.873	0.975	0.874	0.851	0.882	Raw+SG	SQDA
SVM	0.871	0.872	0.974	0.870	0.846	0.873	Fil+noSG	SLP4FS
MLP	0.865	0.865	0.973	0.867	0.841	0.873	Raw+noSG	SLDA
SLP	0.865	0.865	0.973	0.863	0.839	0.869	Raw+noSG	SLDA
SLP	0.865	0.868	0.973	0.864	0.846	0.885	Fil+noSG	SQDA
QDA	0.864	0.864	0.973	0.862	0.837	0.865	Fil+noSG	SLDA
SVM	0.863	0.861	0.973	0.858	0.835	0.867	Fil+noSG	SLDA

Given the above explanation, and taking into account the diversity of selectors (SLP4FS, SQDA and SLDA) and datasets (Raw+noSG and Fil+noSG) with the potential to discriminate the classes defined for the *cv.* ‘Tommy Atkins’, the choice of the best selector, and hence the best set of wavelengths, is presented and discussed in section 6.3.6.

Selected wavelength multiclass models for anthracnose symptom detection on ‘Keitt’ mango leaves

The results associated with the performance metrics of the 10 best classification models are presented in Tables 6-23 and 6-24 for the e7-kt and e8-kt experiments, respectively, of the ‘Keitt’ mango cultivar. Contrary to the *cv.* ‘Tommy Atkins’, for the *cv.* ‘Keitt’ assays, models obtained a notable performance in the e8-kt study (metrics ≥ 0.83) and an acceptable performance in the e7-kt experiment, with only two models exceeding the 0.8 threshold. This result revealed a clear difference in the foliar anthracnose spectral characterisation with a set of relevant wavelengths, as classification models did not meet the requirements for a good detection model in the e7-kt assay. This issue may be due to the limited data availability for

this assay, since this drawback was also reported in its full-spectrum analysis, section 6.2.8.

Table 6-23. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e7-kt assay of the *cv.* ‘Keitt’ on mango leaves

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SVM	0.807	0.805	0.972	0.803	0.782	0.820	Raw+noSG	SLDA
QDA	0.805	0.804	0.972	0.805	0.780	0.816	Raw+noSG	SLDA
QDA	0.802	0.802	0.972	0.799	0.775	0.808	Fil+SG	SQDA
SVM	0.797	0.798	0.971	0.791	0.771	0.811	Fil+SG	SQDA
QDA	0.795	0.793	0.971	0.793	0.767	0.805	Raw+SG	SQDA
SVM	0.791	0.789	0.970	0.785	0.761	0.798	Raw+SG	SQDA
SVM	0.780	0.778	0.969	0.773	0.748	0.788	Raw+SG	SLDA
SVM	0.770	0.769	0.967	0.765	0.740	0.786	Raw+noSG	SQDA
QDA	0.769	0.767	0.967	0.766	0.737	0.775	Raw+SG	SLDA
SVM	0.767	0.767	0.967	0.764	0.742	0.795	Fil+noSG	SLDA

When comparing the full-spectrum versus the selected wavelength models, the reduction in performance metrics in the latter did not exceed 6 %: for the e7-kt assay, the full-spectrum QDA model obtained an accuracy = 0.831, recall = 0.831 and specificity = 0.976, versus the selected wavelength SVM model with an accuracy = 0.807, recall = 0.805 and specificity = 0.972; for the e8-kt assay, the full-spectrum QDA model obtained an accuracy = 0.944, recall = 0.934 and specificity = 0.992, versus the selected wavelength SLP model with an accuracy = 0.891, recall = 0.871 and specificity = 0.984. Therefore, the variations in performance metrics resulting from the wavelength reduction were considered to be within the expected result for this study (reductions in performance metrics below 6 %).

Table 6-24. Performance metrics of the 10 best classification models with the most fitting feature selector and dataset for the e8-kt assay of the *cv.* ‘Keitt’ on mango leaves

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.891	0.871	0.984	0.874	0.861	0.884	Raw+noSG	GA4FS
SVM	0.887	0.877	0.984	0.873	0.858	0.871	Fil+noSG	SQDA
MLP	0.875	0.863	0.982	0.861	0.845	0.863	Raw+noSG	SQDA
MLP	0.874	0.858	0.982	0.859	0.841	0.861	Raw+noSG	GA4FS
SVM	0.873	0.851	0.982	0.854	0.836	0.859	Fil+noSG	GA4FS
SLP	0.873	0.866	0.982	0.861	0.846	0.868	Raw+noSG	SQDA
SLP	0.872	0.865	0.982	0.861	0.850	0.878	Raw+SG	SQDA
QDA	0.870	0.863	0.982	0.859	0.841	0.857	Fil+noSG	SQDA
SVM	0.863	0.842	0.981	0.843	0.823	0.844	Fil+SG	SQDA
SVM	0.863	0.843	0.980	0.844	0.825	0.847	Fil+noSG	RF4FS

On the other hand, within the feature selectors, it was evident that the best techniques were SLDA and GA4FS for the e7-kt and e8-kt assays, respectively. However, the discriminant models (SLDA and SQDA) were the dominant ones for both studies: while in the e7-kt all

the selectors listed corresponded to discriminant models, for the e8-kt assay, 6 of the 10 selectors were discriminant models. The result obtained in the e7-kt assay demonstrated the usefulness of discriminant models for characterising data with a small number of data and variables, which makes them useful when data availability is a constraint, but descriptions and analyses are required. However, the GA4FS selector reported a good performance in the e8-kt experiment, being the first study in which this class of algorithms stood out for its potential in variable selection.

Consequently, the GA4FS selector would be evaluated with the discriminant selectors to establish the best set of wavelengths to spectrally characterise the anthracnose symptoms represented in the classes defined in both ‘Keitt’ mango leaf experiments. This analysis is presented in section 6.3.6, where the relation of the wavelengths of these selectors with the anthracnose symptomatology of ‘Keitt’ mango leaves is discussed.

Selected wavelength multiclass models for anthracnose symptom detection on ‘Kent’ mango leaves

The results associated with the performance metrics of the 10 best classification models for the *cv.* ‘Kent’, test e9-kn, are presented in Table 6-25. It was observed that the neural models were in the top positions with performance metrics above 0.84 for the first 3 listed models. In addition, it was evident that the best model was SLP with the Fil+noSG dataset, being the same result for this study in the full-spectrum version.

Table 6-25. Performance metrics of the top 10 classification models with the best fitting feature selector and dataset for the e9-kn test of the ‘Kent’ mango leaf cultivar

Model	Accuracy	Recall	Specificity	F1	MCC	Precision	Dataset	Feature selector
SLP	0.865	0.860	0.983	0.860	0.851	0.887	Fil+noSG	SLP4FS
MLP	0.857	0.856	0.982	0.858	0.842	0.864	Raw+noSG	GA4FS
SLP	0.853	0.854	0.982	0.857	0.845	0.879	Raw+noSG	SLDA
SVM	0.851	0.849	0.981	0.850	0.833	0.857	Fil+noSG	GA4FS
SLP	0.851	0.850	0.981	0.847	0.836	0.868	Raw+noSG	SLP4FS
SLP	0.850	0.847	0.981	0.846	0.836	0.871	Fil+noSG	GA4FS
SVM	0.850	0.848	0.981	0.848	0.830	0.851	Fil+noSG	SLDA
SVM	0.847	0.846	0.981	0.846	0.829	0.853	Fil+noSG	SLP4FS
SVM	0.845	0.845	0.981	0.844	0.826	0.847	Fil+noSG	SQDA
SVM	0.844	0.842	0.981	0.843	0.826	0.851	Fil+noSG	RF4FS

As a difference between the full and selected wavelength models, the disparity in the metrics did not exceed 12 %: the full-spectrum SLP model obtained an accuracy = 0.981, recall = 0.981, specificity = 0.998, while the selected wavelength SLP model achieved an accuracy = 0.865, recall = 0.860, specificity = 0.983. Therefore, it can be stated that the reduction in input variables led to a decrease in performance metrics. However, the proportionality of these was not equivalent: more than 80 % in variables versus 12 % in metrics.

Consequently, and given that the SLP4FS and GA4FS selectors were placed in the first 2 positions with neural models with less than 1 % difference in the metrics, the selection of the best wavelength set (extracted by the selectors) will be done considering their association with the anthracnose symptomatology. As mentioned in previous sections, this analysis is presented in section 6.3.6.

6.3.6. Relevant wavelength selection for spectral characterisation of anthracnose on mango leaves

The list of relevant wavelengths for the feature selectors per test is shown in Table 6-26. In this table, the wavelengths were listed in order of importance from 1 to 10, with position 1 being the wavelength of highest relevance and 10 being the wavelength of lowest importance within each set. Additionally, the datasets selected for each selector were made according to their position and prevalence in Tables 6-21-6-25. For the *cv.* ‘Keitt’ assays only 2 selectors were listed as these were the most dominant, see section 6.3.5.

Table 6-26. Sets of the 10 most representative wavelengths by best-performing feature selector in each assay conducted for the 3 mango leaf cultivars

Rank	e5-to			e6-to			e7-kt		e8-kt		e9-kn		
	SLP4FS	SLDA	SQDA	SLP4FS	SLDA	SQDA	SLDA	SQDA	SQDA	GA4FS	SLP4FS	SLDA	GA4FS
1	950	960	950	950	970	530	980	980	770	950	950	550	950
2	460	970	940	450	960	450	940	950	950	450	720	950	720
3	470	900	970	710	920	470	950	900	920	920	460	970	930
4	710	730	870	770	790	950	910	970	970	910	760	920	970
5	900	750	860	940	780	940	900	920	910	900	930	930	920
6	720	700	900	460	940	900	970	790	900	970	670	720	460
7	450	940	710	470	770	920	920	930	860	720	450	710	450
8	860	450	720	760	450	670	630	840	790	980	690	760	770
9	760	460	840	670	470	650	800	770	980	890	580	690	760
10	970	470	450	800	680	970	840	860	800	940	920	670	980

The most recurrent spectral ranges in all experiments were 730 - 910 nm and 950 - 980 nm. These ranges are related to carbohydrate and cellulose degradation and water accumulation/release, respectively. These spectral ranges have already been reported in the literature as regions of interest to quantify changes in TSS content and water absorption [117,152,161], so their relation with anthracnose is that the pathogen induces the cell membrane rupture in leaves to initiate its infectious cycle (see section 1.1.2), resulting in the degradation of cellulose/available nutrients (carbohydrates), the release of surface water and the appearance of the characteristic necrotic lesions of anthracnose.

As a particularity, from the sets listed in Table 6-26, it was observed that the spectral range 950 - 980 nm, related to the water absorption peak, was located within the first positions in the sets (in 10 of the 13 selectors), indicating that this region was the most important for

the foliar anthracnose characterisation.

A symptom closely related to anthracnose in mango is the degradation of chlorophyll pigments (*a* and *b*), carotenes (β -carotene) and, rarer on leaves, anthocyanins. The regions of the spectrum that characterise these ranges were located at: 460 - 470 nm and 630 - 640 nm for chlorophyll *b*; 430 nm and 670 - 690 nm for chlorophyll *a*; 450 nm and 480 nm for β -carotene; and 510 - 530 nm for anthocyanins. Chlorophyll was the most prevalent in leaves among these pigments, since it was listed in 9 of the 13 selectors and was the most important in the ‘Kent’ and ‘Tommy Atkins’ cultivars. Similarly, the reflectance peak associated with carotenes was also present in 9 of the 13 selectors, being recurrent in the *cv.* ‘Tommy Atkins’, followed by the ‘Kent’ and ‘Keitt’ cultivars. In addition, as a particular finding, carotenes reported changes at the spectral level along with chlorophyll *b*: their reflectance peaks are closely linked and, due to the pigment degradation produced by the pathogen, their changes were reported simultaneously on leaves.

On the other hand, in the *cv.* ‘Tommy Atkins’, the SQDA technique selected the 530 nm wavelength as the most important, indicating that the change in anthocyanin content was a symptom of relevance for this selector in the e6-to assay; however, the individual presence of this pigment in this single selector also indicated that the reported symptom may not be relevant for the overall symptom characterisation of this cultivar.

In addition, the spectral region 700 - 720 nm, which is usually associated with chlorophyll *a* content and cellulose degradation [59, 74], was also of interest for the selectors and was present in 6 of the 13 selectors. However, the spectral profile of this pigment presents a peak reflectance in the range 670 - 690 nm and this region is located in the red-edge range; therefore, in this research, it was not considered as part of any degradation symptom or pigment as the characterisation of this region, due to the pronounced change in reflectance, makes its association or correlation with physiological processes in the plant material difficult.

In consequence, and considering the previous exposition of spectral ranges related to anthracnose symptoms in mango leaves, the selection of the best set of wavelengths per cultivar was made according to the highest presence of symptoms within the sets, i.e. the highest presence of spectral ranges within the 10 wavelengths of each selector:

- For the *cv.* ‘Tommy Atkins’, the selector that best characterised the classes defined in both studies of this cultivar was SLP4FS. In combination with the Fil+noSG dataset, the sets of this selector shared 60 % of the selected variables, while all the selected wavelengths belonged to spectral regions of interest. Moreover, as a particularity of this selector, the variable extraction focused on two of the most relevant foliar anthracnose symptoms: water accumulation caused by cell membrane rupture and pigment

degradation. While for the e5-to assay, the regions related to chlorophyll *b* and carotene pigments were reported; in the e6-to assay, chlorophyll *b*, chlorophyll *a* and carotenes were reported.

Similarly, the result achieved by this selector in combination with the classification models ranked first for each assay (see Tables **6-21** and **6-22**) and agreed with the dataset selected for the full-spectrum version, Fil+noSG: MLP ($neurons = 200$, $layers = 3$) for the first assay and SVM ($cost = 1$) for the second assay.

- For the *cv.* ‘Keitt’, the selectors that reported the best spectral description of foliar anthracnose symptoms were SLDA and GA4FS, reporting 4 regions of interest for each selector and sharing 50 % of the extracted wavelengths: 900, 920, 950, 970 and 980 nm. As in this cultivar, the spectral range of greatest presence was the NIR region, the anthracnose symptoms related to the carbohydrate and cellulose degradation and water accumulation stood out; however, these selectors also managed to include within their sets regions of the spectrum related to pigment degradation (chlorophyll *b* and carotenes).

As with the *cv.* ‘Tommy Atkins’, the classification models trained with these selectors obtained the best model performance in combination with the dataset Raw+noSG, thus confirming the importance of including changes related to leaf pigments in the group of relevant wavelengths. The best classification models were SVM with $cost = 10$, $gamma = 0.1$ and SLP with $neurons = 100$, $epochs = 468$ for the e7-kt and e8-kt tests, respectively.

- For the *cv.* ‘Kent’, the wavelength set corresponded to that extracted by the SLP4FS selector. All 10 wavelengths were related to foliar anthracnose, with water accumulation being the most relevant. Moreover, this selector was the only one that included in a single set all pigments related to chlorophyll (*a* and *b*) and carotenes, indicating their importance in the spectral characterisation of foliar anthracnose on ‘Kent’ mango (pigment degradation).

Therefore, the detection model that performed best with this selector was SLP with $neurons = 100$, $epochs = 370$ and dataset Fil+noSG, reaffirming the performance of this model and dataset that also figured in the design of the full-spectrum detection model (see section 6.2.8). As mentioned in section 6.3.5, this detection model obtained the best performance among all the classification models implemented, confirming the importance and robustness of the wavelength set extracted by the SLP4FS selector.

7. Discussion

The main objective of this research was to answer the question about the existence of a relationship (or connection) between spectral information and anthracnose-associated symptoms in mango. Also, to evaluate the feasibility of detecting these anthracnose symptoms early and establish a set of wavelengths that could characterise anthracnose symptoms in mango through feature selection techniques and machine learning models. The primary motivation was based on **1) generating new knowledge to enable the development of new detection systems using state-of-the-art technological tools at the service of the agricultural sector and 2) assisting in the reliable and effective diagnosis of anthracnose early to support on-site decision-making**. Early diagnosis, understood as detection at the previsual stage (not visible to the human eye) on fruits or reliable detection on the tree (mainly leaves), can help to stop or contain the spread of anthracnose on harvested or unharvested fruits, allowing the control of the disease in mango crops.

In the literature, studies have reported advances in the characterisation, control and detection of this disease in mango and other fruit crops. In mango, these works have addressed the assessment of virulence and prevalence of the disease [24], the characterisation of anthracnose symptoms in mango [66], the development of CV systems using UV-RGB images for its detection [123, 132], the use of ML and DL techniques [126, 165] and, recently, the assessment of symptoms through spectral data (spectroscopy) [30]. All these studies have allowed a better understanding of disease development in this climacteric fruit; however, their focus has been mainly on the development of late-stage inspection systems, emulating the expert human eye in the detection of visible symptoms of anthracnose, and not on assessing the feasibility of technological tools capable of detecting the presence of symptoms in early stages. Only the works reported by [30, 149, 150] evaluated the potential of spectral information to determine the feasibility of early detection in mango fruits. However, being subject to the use of spot spectral signatures, as in the case of Cabrera et al. [30], the results do not allow for establishing with precision the stage of development of spectrally detectable anthracnose symptoms, nor select a group of wavelengths to characterise the early anthracnose symptoms in mango.

Anthracnose remains latent until the environmental and nutritional conditions of the host are suitable for its development. This behaviour implies that the changes caused at the physiological level when developing are not perceptible to the naked eye. Therefore, by fo-

cusing on the study of anthracnose from inoculation to visual development, it was possible to establish the feasibility of detecting anthracnose symptoms in the pre-visual stage in the studied mango cultivars. With the classes defined in this study (one class per day after inoculation) and the ROI selection strategy, it was possible to detect the infection from 48 and 72 hours (the second and third day) after inoculation on fruits and leaves, respectively, see Figures 7-1-7-2 and 7-3-7-4.

Figures 7-1 and 7-2 depict fruits of the ‘Osteen’ and ‘Keitt’ cultivars, respectively, with an RGB image and the 3 main wavelengths extracted for each cultivar on days 2, 3 and 4, while Figures 7-4 and 7-3 show leaves of the ‘Kent’ and ‘Tommy Atkins’ cultivars with the same image layout. In these figures, the 6 inoculation points are shown in black for the fruits, while in the leaves, the 5 inoculation points are demarcated on one side of the leaves.

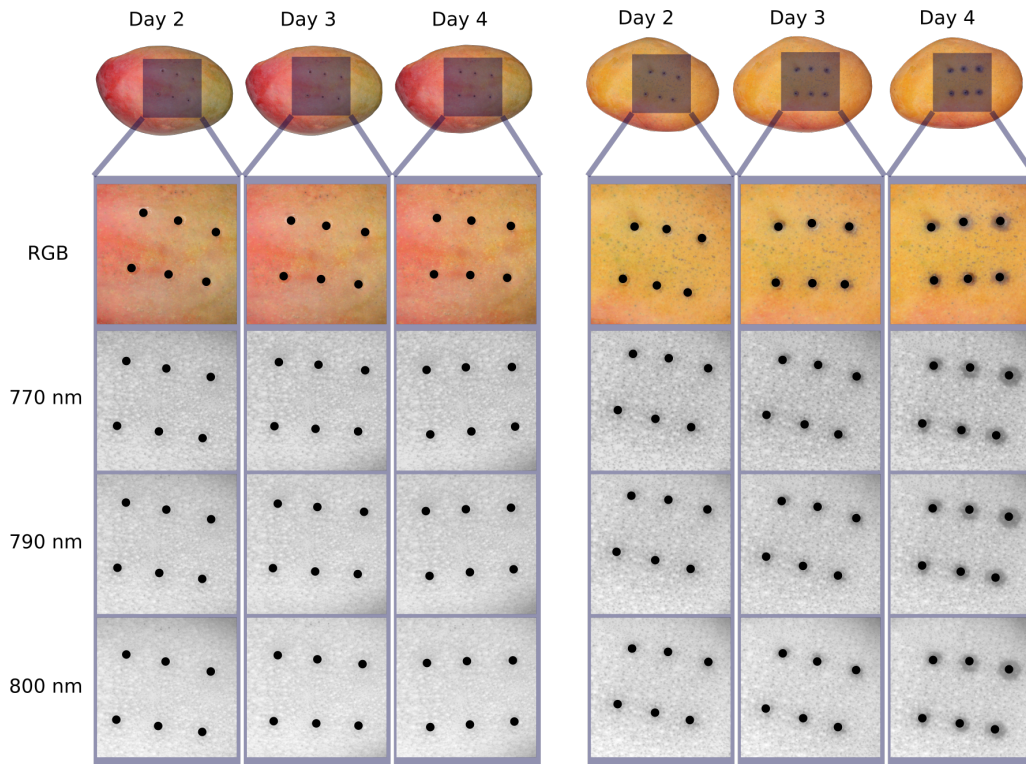


Figure 7-1. Temporal evolution (from day 2 to day 4) of ‘Osteen’ mango fruits in colour image (RGB) and the 3 most relevant wavelengths (770, 790 and 800 nm)

From Figures 7-1 and 7-2, it can be observed that in the control fruits of the mentioned cultivars, there are no differences in the colour image as well as in the monochrome images of the relevant wavelengths, despite the pigment variations in the RGB image. In contrast, the anthracnose lesions progress on the diseased fruits and become fully visible on day 4. In the ‘Keitt’ cultivar, there are dark halos around the inoculation points, indicating the presence of the disease. In contrast, in the ‘Osteen’ cultivar, there are changes in texture in

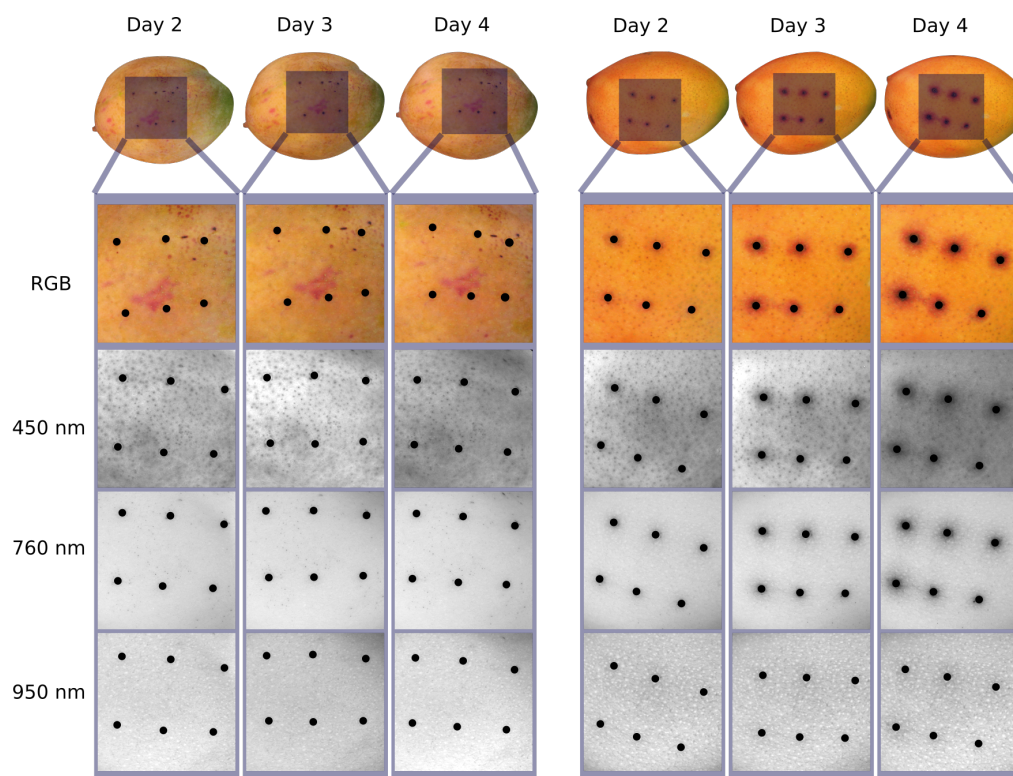


Figure 7-2. Temporal evolution (from day 2 to day 4) of ‘Keitt’ mango fruits in colour image (RGB) and the 3 most relevant wavelengths (450, 760 and 950 nm)

the exposed area and more noticeable dark halos from day 3 onwards. These halos, initial symptoms of anthracnose, are not visible in the RGB image of the ‘Osteen’ cultivar, while in the ‘Keitt’ cultivar, they are denoted with a reddish hue. However, for this cultivar, a reddish pigmentation present in the control fruit is not visible in the monochromatic images at wavelengths 450, 760 and 950 nm, indicating the difference at the spectral level that the model chosen for this cultivar can detect. On the other hand, for day 4, where symptoms tend to start to be distinguishable, it was observed that not all fruits have distinguishable anthracnose lesions; therefore, day 5 (after inoculation) was defined as the day when the lesion becomes distinguishable by the human eye (by size and colour).

On the other hand, from Figures 7-4 and 7-3, it is observed that in the control leaves of both cultivars, there is no change in colouring or changes in the spectra of the 3 selected wavelengths. Likewise, it is observed that in the ‘Kent’ cultivar, the anthracnose symptoms appear as necrotic lesions from day 2 of inoculation; however, due to the size reported, these lesions are not recognisable to the naked eye. From day 4 onwards, symptoms can be verified when these lesions reach a larger size on the leaf. In the same way, in the ‘Tommy Atkins’ cultivar, this temporal development of symptoms is present with the difference that in this cultivar, in the selected wavelengths, a larger damage area is detected (in the 950 nm

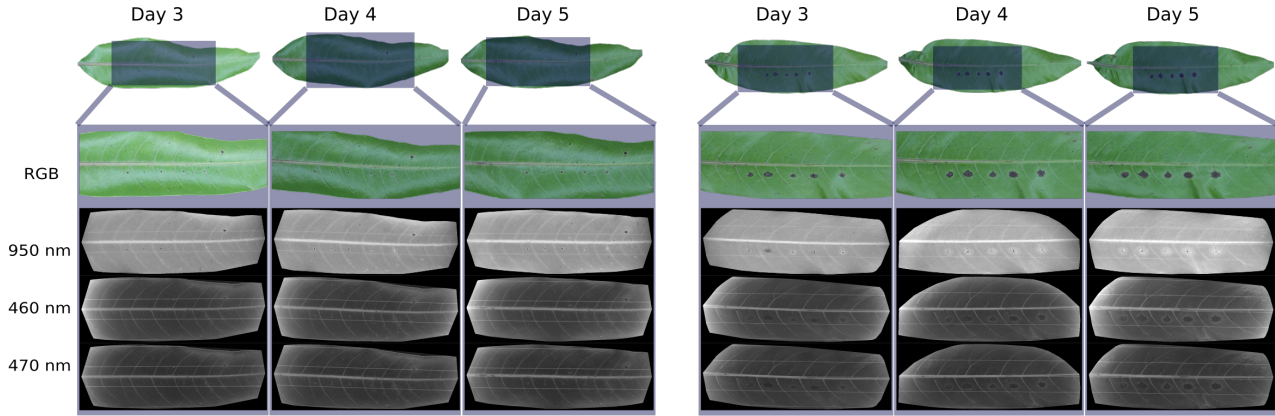


Figure 7-3. Temporal evolution (from day 2 to day 4) of ‘Tommy Atkins’ mango leaves in colour image (RGB) and the 3 most relevant wavelengths (950, 460 and 470 nm)

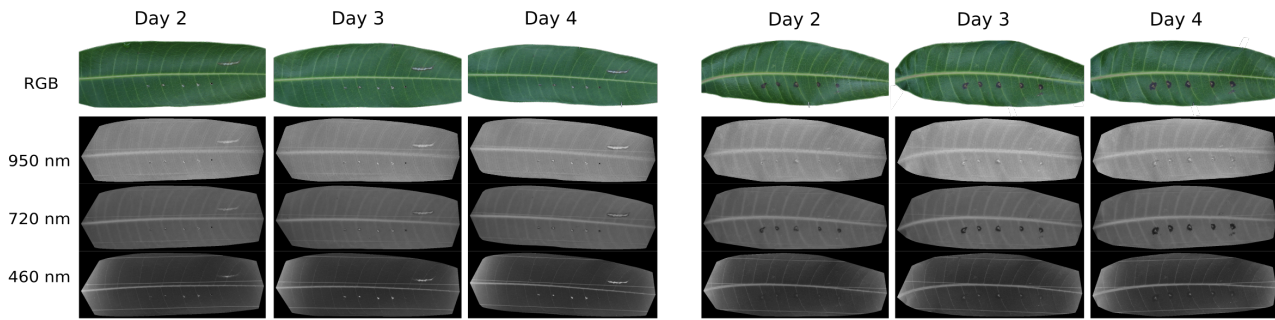


Figure 7-4. Temporal evolution (from day 2 to day 4) of ‘Kent’ mango leaves in colour image (RGB) and the 3 most relevant wavelengths (950, 720 and 460 nm)

wavelength) and a halo that contrasts with the anthracnose lesion in the 460 and 450 nm wavelengths, which indicates in a pre-visual way where the anthracnose lesions will continue to grow. In leaves, anthracnose lesions become distinguishable from day 5 onwards.

Since the fruit quality is evaluated after harvesting, potential applications to implement AI-based spectral solutions are feasible. Considering that the conditions of fruit inspection can be controlled, the correct acquisition of the spectral information to diagnose this disease early can be assured. However, for a practical application of the results obtained in this research on leaves, developing a system (or protocol) that adapts the spectral sensor to field conditions is required to guarantee the quality of the spectral data collected. In [101], this challenge of transferring spectral analysis from the laboratory to field conditions is presented. Authors consider the physical limitations of the environment (type of crop, type of terrain) and the data acquisition methods (drone, guided vehicle, expert personnel), among other aspects. Additionally, a description of adjustments, potential differences in the collected spectral data, and recommendations for making the transition in data collection optimally to outdoor conditions are also discussed.

Based on the remarkable results obtained with the full-spectrum detection models (450 - 980 nm), it was also possible to analyse and determine a set of 10 relevant wavelengths for each mango cultivar and design a detection model that allowed the discrimination of the classes (representing symptoms at different stages of development) with outstanding performance: in all cultivars a detection model with metrics above 0.85, and even above 0.98 for the specificity metric, was obtained. The difference in performance of the models between full-spectrum (54 wavelengths) and selected bands (10 wavelengths) detection models was less than 12% in all tests. However, the reduction in wavelengths was around 81 %, so it can be stated that with these relevant wavelength sets, the detection models adequately fit the collected spectral data. Furthermore, the results reported with these models also indicated the existence of spectral regions irrelevant to the model performance since they were not related to the spectral changes induced by the pathogen (for instance, the region 570 - 640 nm). These results confirmed the need for the inclusion of two or more wavelengths from the same region of the spectrum to maintain good model performance, since anthracnose symptoms, as discussed in sections 6.3.3 and 6.3.6, induce changes in considerably large portions of the spectrum (especially the NIR region related to carbohydrate degradation).

The wavelengths selected for each cultivar, although related to physiological and biochemical changes referenced in the literature [59,74,117,152,162], presented variations that were specific to each cultivar since the order of importance of the wavelengths changed. Figures 7-5 and 7-6 depict the mean spectra of the classes defined for the cultivars 'Osteen' and 'Keitt' in fruits and 'Kent' and 'Tommy Atkins' in leaves of mango, with the sets of wavelengths selected for these cultivars. In these figures it is possible to detail the different ranges extracted for each cultivar and the remarkable differences in the characterisation of anthracnose both in fruits and leaves: while in fruits there is a higher concentration in the NIR region, for leaves these sets are distributed along the range 450 - 980 nm as pigments (chlorophyll *a* and *b* and anthocyanins), cellulose degradation and water release have different reflectance peaks. Among the most important symptoms for the spectral characterisation of anthracnose, the following were reported:

- **Pigment degradation:** Chlorophyll *a* and *b*, carotenoids (mainly β -carotene), and anthocyanins. The spectral changes reported in the mango fruit assays due to the degradation of these pigments were notable for all cultivars. However, it was evident that this symptom depends on both the cultivar and the physiological maturity of the fruit.

The leaf assays found that the pigments with the most considerable spectral variations were chlorophyll *a* and chlorophyll *b*. These pigments, which are responsible for photosynthetic activity in the tree and for giving the leaves a greenish hue, were the most

affected from an early stage of anthracnose. The degradation of this vital pigment slows down the photosynthetic activity in the leaf, thus affecting the production of energy and nutrients for the tree. Consequently, the pathogen causing anthracnose can grow and spread quickly through the infected plant material, as the defence constituent substances for the tree are limited.

- **Water release/accumulation:** Due to the cell membrane rupture in both leaves and fruits, water is released under the surface, which can be detected on a spectral level from an early stage of anthracnose development. With the rupture of the cell membrane, there is also the release of phenolic acids, which spread quickly due to water accumulation. This event produces the characteristic symptoms of anthracnose: circular necrotic lesions.

The experiments found that this symptom was the most important for the leaf tests, indicating that the spectral changes related to the water absorption peak (due to its accumulation) can be used to determine if the leaf is subjected to an early infectious process. The fruit tests found that this symptom is also relevant for characterising anthracnose symptoms, as it also allows detecting the sugar degradation due to the infection induced by *Colletotrichum sp.*.

- **Carbohydrate degradation:** The degradation of nutrients (starches and cellulose) is an important indicator of anthracnose development in mango. In all assays, both fruit and leaf, it was reported that the spectral range 730 - 910 nm was of importance to characterise all classes defined in all assays, which indicates that the spectral changes induced by the disease can be detected early (previsual) since this range is located in a portion of the spectrum not visible to the human eye and allows the designed detection models to make reliable estimates from the first day of study of the disease after its inoculation.

The main difference found between the leaf and fruit assays was on the portion of the spectrum located between 720 - 800 nm, which showed a strong degradation of cellulose and carbohydrates (mainly starches). In contrast, the wavelength distribution is present in the fruit throughout this range.

Sugars, one type of carbohydrate, are also related in the bands selected for each cultivar, being more relevant for the fruits than for the leaves, given that the transformation of starches into sugars takes place with greater emphasis when the fruit begins its organoleptic maturation phase (after harvesting). As a peculiarity, the region that characterises these nutrients is located between 920 - 940 nm, close to the water absorption peak (950 - 980 nm). As reported in the literature, this enhances the detection

and non-destructive estimation of sugar content in mango fruit since water helps the dissolution (and, therefore, quantification) of these nutrients.

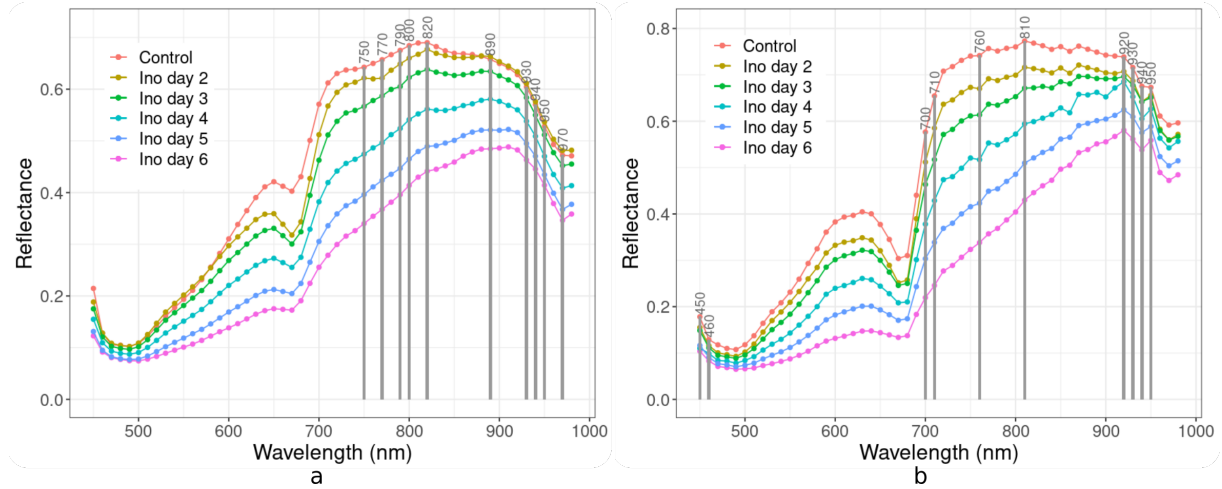


Figure 7-5. Mean spectra of a) ‘Osteen’ and b) ‘Keitt’ mango classes and the 10 most relevant wavelengths for the anthracnose spectral characterisation

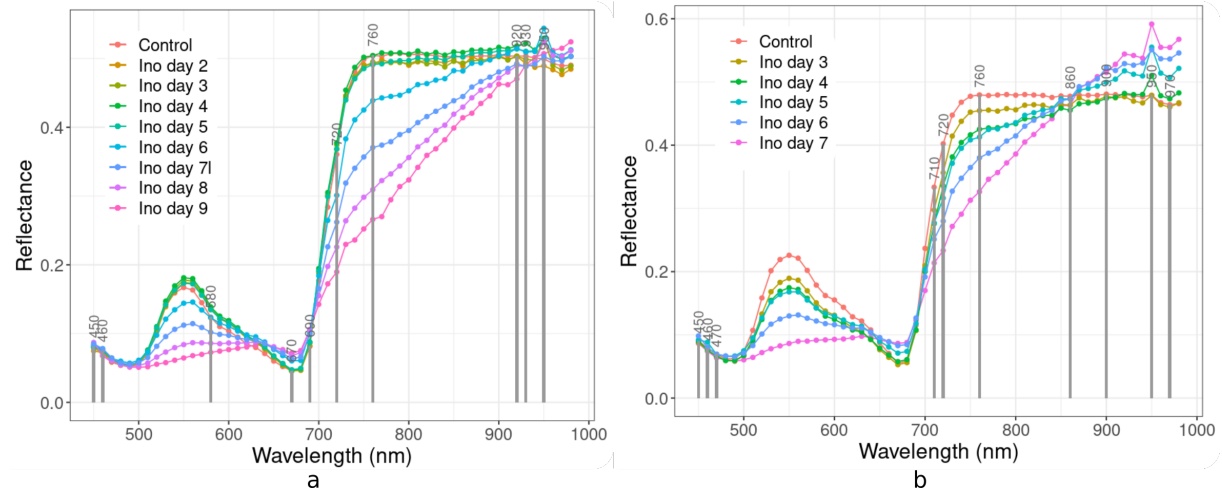


Figure 7-6. Mean spectra of a) ‘Kent’ and b) ‘Tommy Atkins’ mango classes and the 10 most relevant wavelengths for the foliar anthracnose spectral characterisation

On the other hand, since the most relevant symptoms for anthracnose were characterisable in the NIR region, according to the set of wavelengths selected in this research, there was the possibility of confusing spectral variations in this region with other problems of interest in the fruit, such as mechanical damage. Mechanical damage to the fruit leads to fibre breakage and thus to water release, although it differs from anthracnose as it does not lead to the degradation of sugars and starches. Vélez-Rivera et al. [162] studied this issue and found that mechanical damage can be distinguished in the NIR region (700 - 780 nm, 890 - 900

nm and 1070 - 1080 nm). However, these regions showed a slight decrease in the spectrum when mechanical damage occurs, so the shape and amplitude of this variation differ from that caused by anthracnose (7-7).

From 7-7, it can be observed that in the spectral region between 730 and 820 nm, the spectral variations reported for mechanical damage (7-7.a) are similar in amplitude to those obtained with the classes day 2 and 3 of the 'Kent' cultivar (7-7.c). In contrast, for the 'Keitt' cultivar (7-7.b), the reported spectral variation was notably higher between the control and diseased classes, indicating that the expression of early anthracnose symptoms was distinguishable from mechanically induced fruit damage. Moreover, as evidenced in this research, anthracnose caused a pigment degradation in mango fruits and leaves (related to the spectral range located between 450 - 700 nm), so this particular symptom also contributed to the differentiation of anthracnose-induced damage (lesion) from mechanical damage in fruits: from Figure 7-7.a, it can be observed that in the 650 - 700 nm range, there was no spectral variation induced by mechanical damage.

Another important point was that the wavelength sets found in this research's fruit tests differed from those reported in [136] for identifying the ripening stage of mango fruits. Although a relation between anthracnose development and the organoleptic ripening of mango has been established, the spectral variations found to differentiate an infected fruit from a naturally ripening fruit were notable. The authors in [136] established a ripening index with a set of 15 wavelengths in the NIR region (730, 756, 778, 794, 800, 844, 884, 894, 918, 940, 976, 984, 992, 1040 and 1046 nm). These wavelengths are closely related to the functional groups -CH (sugars) and -OH (water) [43,117,136,152], components highly correlated with anthracnose symptoms. However, the spectral variation produced by the ripening process is minimal as the moisture content, reflected in the water absorption peak, remains constant in the absence of an abrupt water loss such as that caused by anthracnose-induced cell membrane rupture.

Additionally, depending on the fruit cultivar, there is a transformation of starches into sugars that may be reflected in a slight decrease in reflectance in the range 730 - 850 nm, while in the 920 - 940 nm region, there is no considerable change in reflectance as there is no sugar degradation or increase in water content under the peel that would make the presence of these nutrients distinguishable. Figure 7-8 shows the mean signatures of the control and inoculated classes per day of acquisition of the 'Keitt', 'Kent' and 'Osteen' cultivars. In this figure, it can be observed that the control classes (per acquisition day) showed minimal variations in the range of 730 - 850 nm, which corresponded to what is reported in the literature concerning the ripening process of mango fruits. However, it was evident that the inoculated classes showed a higher spectral variation from day 2 onwards (except the 'Osteen' cultivar, Figure 7-8.c), indicating that the infectious process of anthracnose develops and induces a

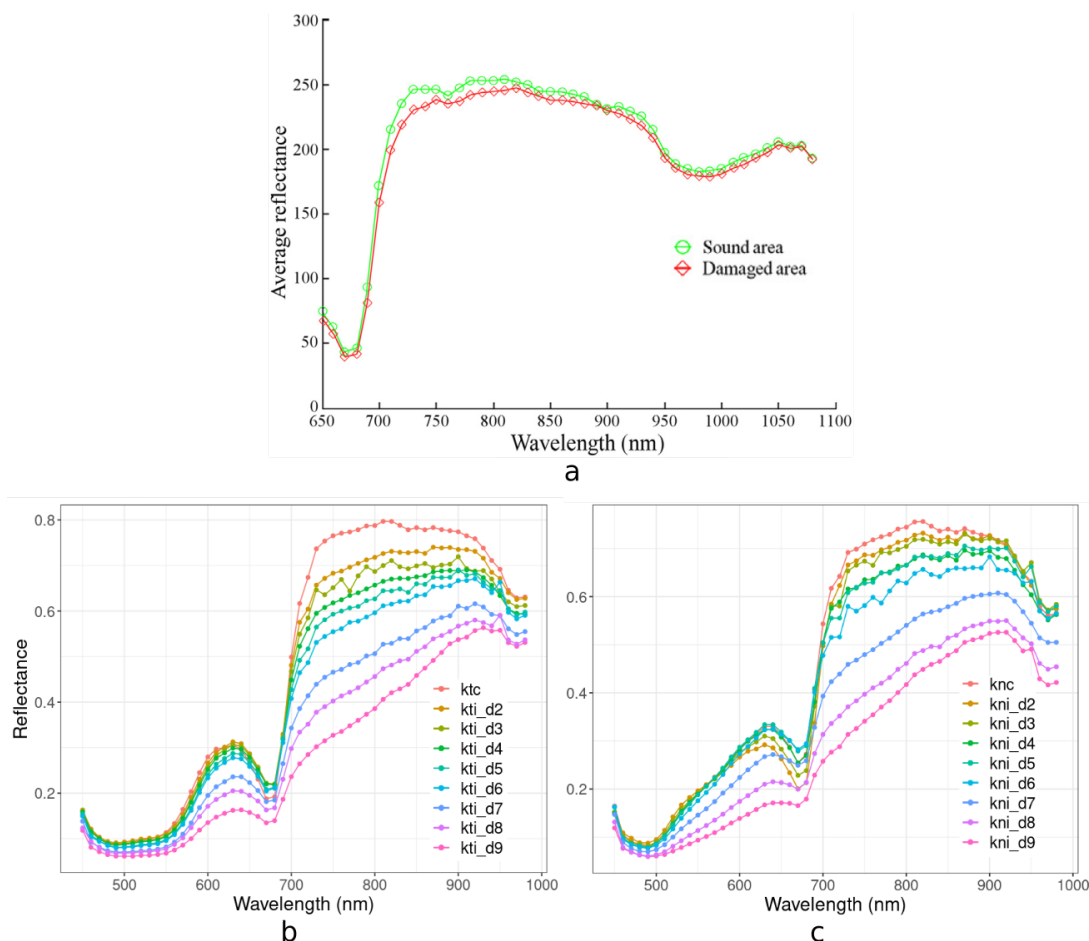


Figure 7-7. Mean spectral signature of the classes defined for mango fruit assays with a) mechanical damage (taken from [162]); b) *Colletotrichum gloeosporioides* inoculation on the 'Keitt' cultivar; and c) *Colletotrichum gloeosporioides* inoculation on the 'Kent' cultivar

strong decay in the reflectance of the NIR region. Likewise, it is noted that in the visible region related to chlorophyll and carotenoid pigments, ranges 600 - 690 nm and 450 - 460 nm, the spectral behaviour explained above was also present: the spectral variation among control classes was minimal, but between control and diseased classes, it was higher and more distinguishable.

From the foliar assays on the 'Tommy Atkins', 'Keitt', and 'Kent' mango cultivars, it is important to note that the findings validated the study presented in [69], which focused on the spectral detection of anthracnose symptoms on strawberries. The development of leaf symptoms and the spectral characterisation of these symptoms were similar to those obtained in the mango assays of this research. Likewise, the hypothesis proposed in [74] that established a possible correlation of the visible spectrum with the degradation of photosynthetic pigments was also proved: the degradation of chlorophyll *a* and *b*, carotenoids and, more limited, anthocyanins are anthracnose symptoms that can be detected at an early stage

and characterised at spectral level since the reflectance decreases as the induced infection evolves in the leaf, see Figures 7-4 and 7-3.

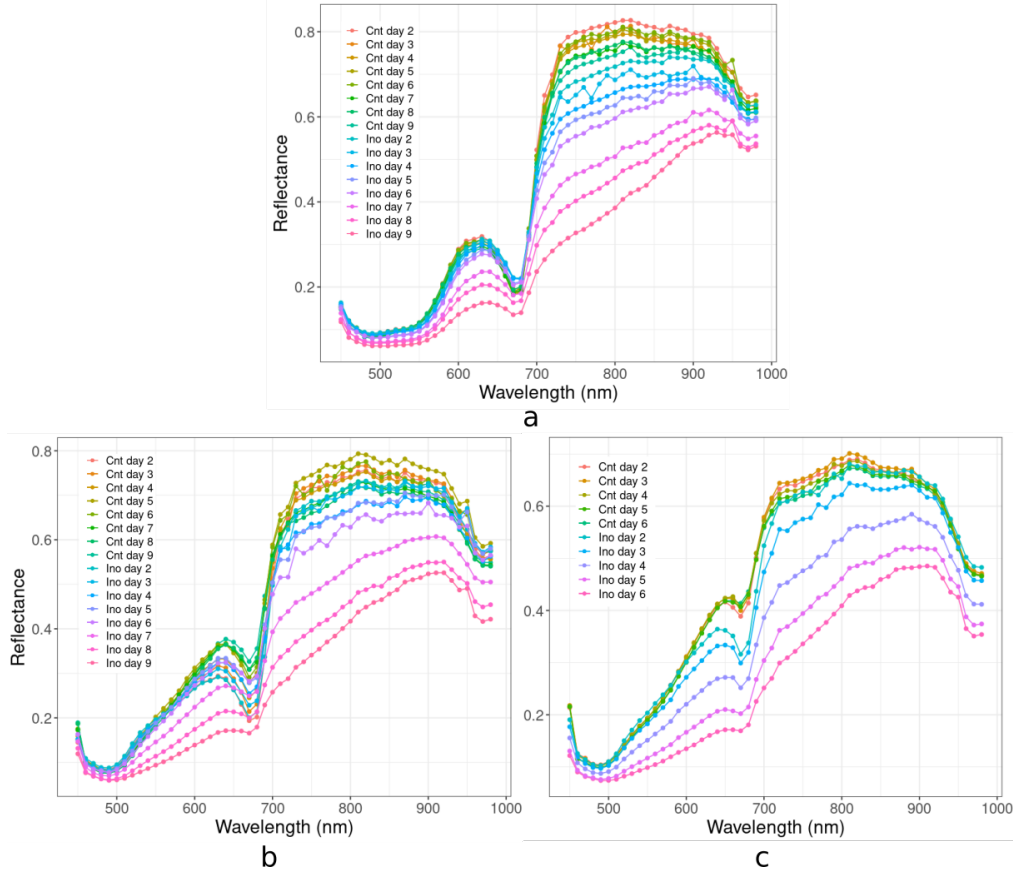


Figure 7-8. Mean spectral signature of the control and inoculated classes per day of acquisition for the ‘Keitt’ (a), ‘Kent’ (b) and ‘Osteen’ (c) assays with mango fruits

The leaf assays also tested the susceptibility (or resistance) to anthracnose development described in [23, 45, 127]. The ‘Kent’ cultivar was shown to be the most susceptible by the incidence reached and the size of the lesions at the end of each experiment, while the ‘Tommy Atkins’ and ‘Keitt’ cultivars reported a significantly smaller lesion size, indicating that these cultivars showed moderate resistance by delaying the infectious process (the ‘Keitt’ cultivar reported a slightly higher resistance than ‘Tommy Atkins’). However, at the spectral level, this variation in the susceptibility of the cultivars was not evident, as common spectral regions were found in all three mango leaf cultivars, allowing the characterisation of foliar anthracnose in mango. These foliar symptoms, both with the full spectrum and the 10 most relevant wavelengths, could be detected from early stages with a good level of accuracy, sensitivity and specificity, according to the performance metrics achieved by the detection models (see sections 6.2 and 6.3). Hence, their use in real monitoring applications would help in the intervention and decision-making in situ quickly and reliably.

Despite the results and advances reported in the literature, the problem of spectral and temporal detection of anthracnose has not been previously addressed, as it was in this research. Works related to the spectral assessment of leaf anthracnose on mango have not been reported. Therefore, the results presented in this research make an essential contribution to the understanding, characterisation and early detection of anthracnose in mango by using spectral technology and machine learning modelling as:

- It demonstrated the feasibility of the identification and spectral characterisation of symptoms associated with anthracnose in mango, making possible the non-destructive and non-invasive assessment of this disease: the developed detection models were able to discriminate the defi classes with performance metrics above 0.95 (95 %) in all cultivars.
- It demonstrated the potential of AI-based models, particularly neural networks and discriminant models, for the discrimination and early detection of anthracnose symptoms in both fruits and leaves of 3 commercially desirable mango cultivars: it was possible to detect anthracnose symptoms between 48 and 72 hours (depending on the cultivar) before they are visible to the human eye.
- It identified wavelengths closely related to anthracnose symptoms in mango. Although the spectral regions related to anthracnose varied between cultivars (both in fruits and leaves), it was found that these regions always appeared among the most important for the spectral characterisation of anthracnose: the regions 450 - 480 nm, 650 - 680 nm, 730 - 810, 890 - 940 and 950 - 970 nm were chosen for all cultivars and assays conducted in this research. All these regions are associated with anthracnose symptoms, including pigment degradation, degradation of starches and sugars (the major carbohydrates), and water release or accumulation.
- It assessed different mango cultivars with various physiological and visual features (fruits and leaves), showing the existence of common and particular anthracnose symptom patterns for each cultivar. For mango leaf cultivars, the 'Keitt' and 'Tommy Atkins' cultivars showed moderate resistance to developing foliar symptoms, whereas the 'Kent' cultivar exhibited higher susceptibility. No difference in the anthracnose symptoms was reported for the fruit cultivars ('Kent', 'Keitt' and 'Osteen'), showing that the defence against the disease lies in the mango cultivar.

Finally, the combination of the spectral regions and wavelengths found in this study allows us to detect anthracnose symptoms in mango fruits and to differentiate them from mechanical damage, the fruit ripening stage and other common problems in mango fruits. As evidenced throughout this dissertation, physiological and biochemical changes in both fruits and leaves occur as anthracnose infection develops; however, these changes (symptoms) can be detected

spectrally in an early or predictive manner, allowing for early intervention in the crop to control this disease. Even though there are differences in the development of anthracnose depending on the mango cultivar, with the detection models developed and the wavelength sets developed, a diagnosis of the disease can be made with a high level of accuracy and reliability, demonstrating the potential of the results derived from this research.

8. Conclusion

8.1. Conclusions

The results generated by this doctoral thesis demonstrated the feasibility of the early detection of anthracnose-associated symptoms in three cultivars of mango fruits and leaves. The classification models designed, both with the full Vis-NIR spectrum and with the relevant wavelength sets, obtained outstanding performance metrics, allowing us to state that anthracnose symptom detection in mangoes from the early stages of development is possible.

With the implementation of AI-based classification models, it was possible to establish that neural networks (mainly the single-layer perceptron -SLP-) and discriminant models (QDA) were the best-fitting models for the spectral data from the mango samples collected and studied in this research. This result revealed the potential of this class of models, currently considered standard techniques, for discriminating spectral data to detect crop diseases. Hence, the approach adopted in this work demonstrated the validity and potential application of these techniques for addressing latent problems and needs in the agricultural sector.

All the designed full-spectrum models for all the mango cultivars obtained an average performance metric above 0.95, indicating an outstanding performance level and allowing the detection of anthracnose symptoms from the first 48-72 hours after the inoculation of the pathogen in the samples. Furthermore, with the tests carried out in this research, it could be determined that under established experimental conditions, the first visible anthracnose lesions appear from the fifth day after the pathogen inoculation, so with these models, there is a disease prediction time between 48-72 hours before the human eye can see the symptoms. These findings allow us to state that it is possible to make an early diagnosis of anthracnose disease with the Vis-NIR spectral information and the AI models implemented in this research and, therefore, to intervene or make early decisions to contain and control the spread of the disease to batches of anthracnose-free mango fruits.

Two resulting approaches from the designed detection models equally allow for disease detection: the biclass and multi-class models. For the biclass models, which compare the control class with each diseased class, QDA models performed best regardless of the mango cultivar. In contrast, for the multi-class models, neural networks were the best classifiers. However, it should be mentioned that in this variant (multi-class models), the discriminant

models always obtained scores close to those obtained by the neural networks. The value of finding the enormous potential of discriminant models for detecting anthracnose symptoms in mango trees lies in the fact that these mathematical models allow a fast calibration and deployment to model the hyperspectral data collected in real-time applications robustly.

In contrast, neural networks require extensive training and validation times, limiting their adoption in applications where constant updating of the training database or recalibration of the models is required. For end solutions, where the data collection and training stages have been completed, neural networks emerge as potential solutions to design detection models. Therefore, the neural models obtained in this research allow for the detection of anthracnose in mango, as the spectral database used to train these networks was composed of more than 230,000 spectral signatures on average.

On the other hand, for the models based on the selected bands, the average performance metrics remained above 0.85, indicating that the suitability of the selected sets (10 wavelengths per cultivar) for the spectral characterisation of the defined classes and, therefore, the symptoms associated with anthracnose from the early stages of disease development. As demonstrated in this research, the selected wavelengths are closely related to disease symptoms that are not distinguishable by the human eye: the biochemical changes produced (degradation of carbohydrates, degradation of photosynthetic pigments and release/accumulation of water) in the inner layers of fruits and leaves are anthracnose symptoms that become visible only when the degradation of plant material produces the well-known circular necrotic lesions (symptoms by which the disease is recognisable by an expert). Therefore, it can be affirmed that there are spectral regions closely related to anthracnose symptoms in mango, being the wavelength sets established in this research representative and relevant for the selected mango cultivars ('Keitt', 'Kent' and 'Osteen' in fruits and 'Keitt', 'Kent' and 'Tommy Atkins' in leaves). Also, it can be stated that spectral sensors are required for the correct and early diagnosis of anthracnose in mango since they provide suitable information to detect the early anthracnose symptoms.

The results obtained at the spectral level and the early detection of anthracnose in mango (mentioned in previous paragraphs) were possible thanks to the methodological approach applied during the laboratory experimentation since the selection of mango cultivars and samples, the inoculation strategy and the acquisition of hyperspectral images allowed the success of the objectives proposed in this research. Likewise, the spectral analysis carried out, the preprocessing of the collected data and the ROI extraction strategy allowed for the obtaining of quality spectral data that adequately represented the symptoms of anthracnose in its different stages of development (early and late stages).

Similarly, the AI design and modelling established in this research allowed the design, cali-

bration, and objective comparison of the different classification models to select the best-fit models for the data collected for each mango cultivar. Furthermore, the implementation and evaluation of feature selectors were presented as a relevant tool for selecting wavelength sets closely related to anthracnose symptoms in mangoes. Therefore, their use in spectral studies is necessary to establish potential correlations between the spectral information and the physiological (or biochemical) phenomenon under study.

8.2. Planned work

Given that this research determined the possibility of early anthracnose detection in mango and the existence of relevant wavelengths for the characterisation of the symptoms of this disease, as a future work, the extension and validation of these results to other mango cultivars is proposed, potentially to the 10 most commercialised mango cultivars to establish the feasibility of a general anthracnose detection in mango (independent of the cultivar used). Similarly, the inclusion of 3D models for fruit reconstruction and improved reflectance correction over the entire fruit surface is a further work proposed to assess the whole captured fruit surface in a single hyperspectral image. The 3D reconstruction integrated with the anthracnose detection model would set a benchmark for the non-destructive spectral assessment of fruit with irregular surfaces. With this, deploying compact and reliable technological solutions will improve current fruit inspection and quality control systems.

Furthermore, the experimentation in field conditions with the mango cultivars studied in this research should be carried out, as some technological limitations and challenges should be explored to ensure that the quality of the hyperspectral images will be similar to those obtained in the laboratory, thus allowing the extension and validation of the results of the classification models obtained in this research to field conditions. Research lines related to the spectral image collection strategy (hyperspectral or multispectral) on mango trees, the radiometric correction of field spectral images (correction of reflectance by angle of light incidence of light and of the leaf concerning the sensor) and the assessment of anthracnose symptoms on trees (leaves not detached) are topics of interest that are raised as relevant future work to overcome the challenges related to real field conditions.

On the other hand, as future work, classification models based on deep learning and morphological analysis are proposed to determine the best approach for the spectral detection of anthracnose in mango. Also, exploring the electromagnetic spectrum, beyond 1000 nm, is a work to be done as the NIR and mid-infrared ranges have also proven to be of interest for the characterisation of plant material and the quantification of e.g. water, minerals. However, to date, the detection models obtained in this study set a benchmark in the state of the art, as the performance metrics reported for both full-spectrum and selected wavelength models were outstanding for all assessed cultivars.

A. Appendix: Results of mango fruit assays

A.1. Biclass results for the mango fruit assays

A.1.1. Performance metrics results per dataset

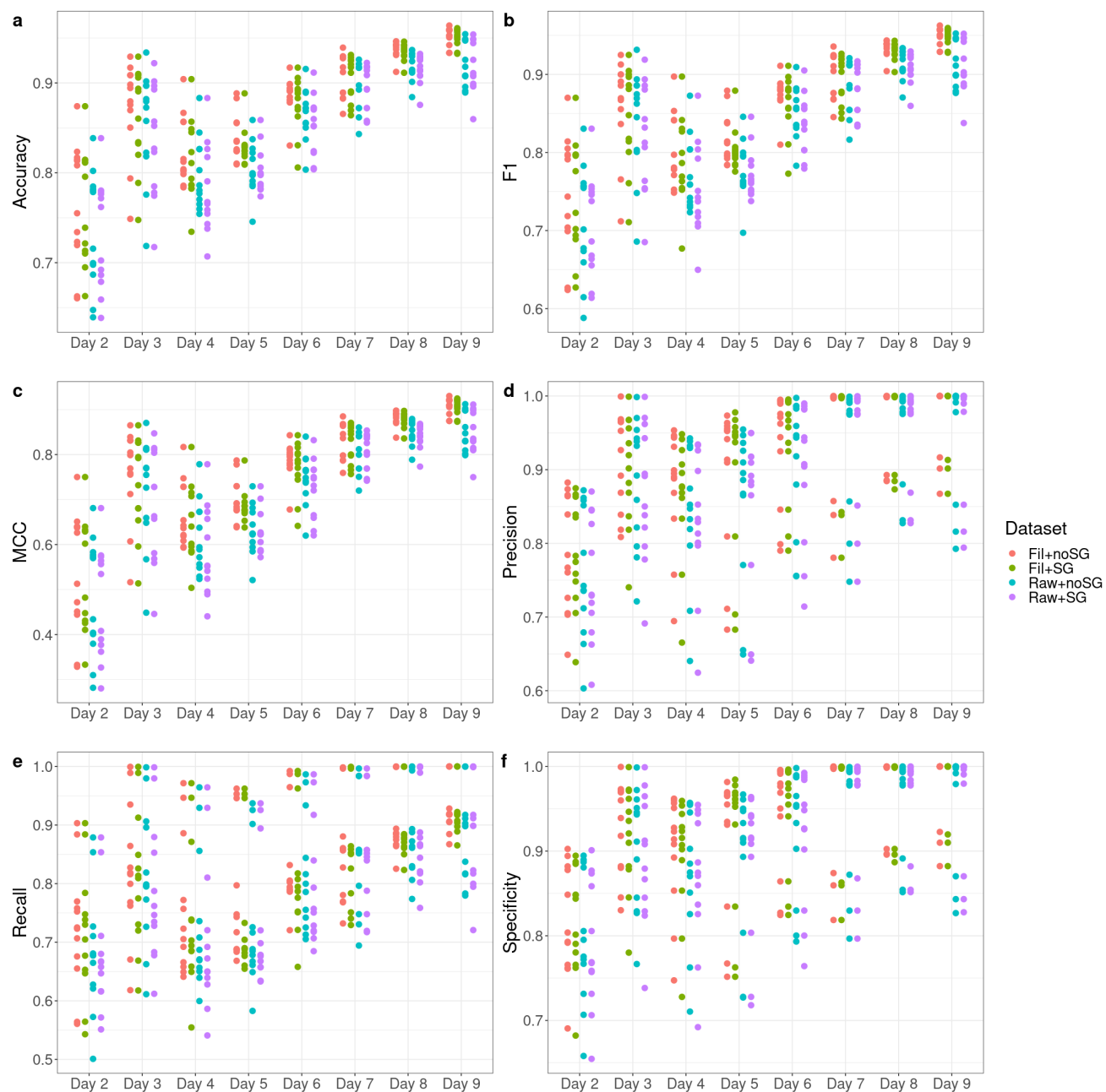


Figure A-1. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e1-kt assay)

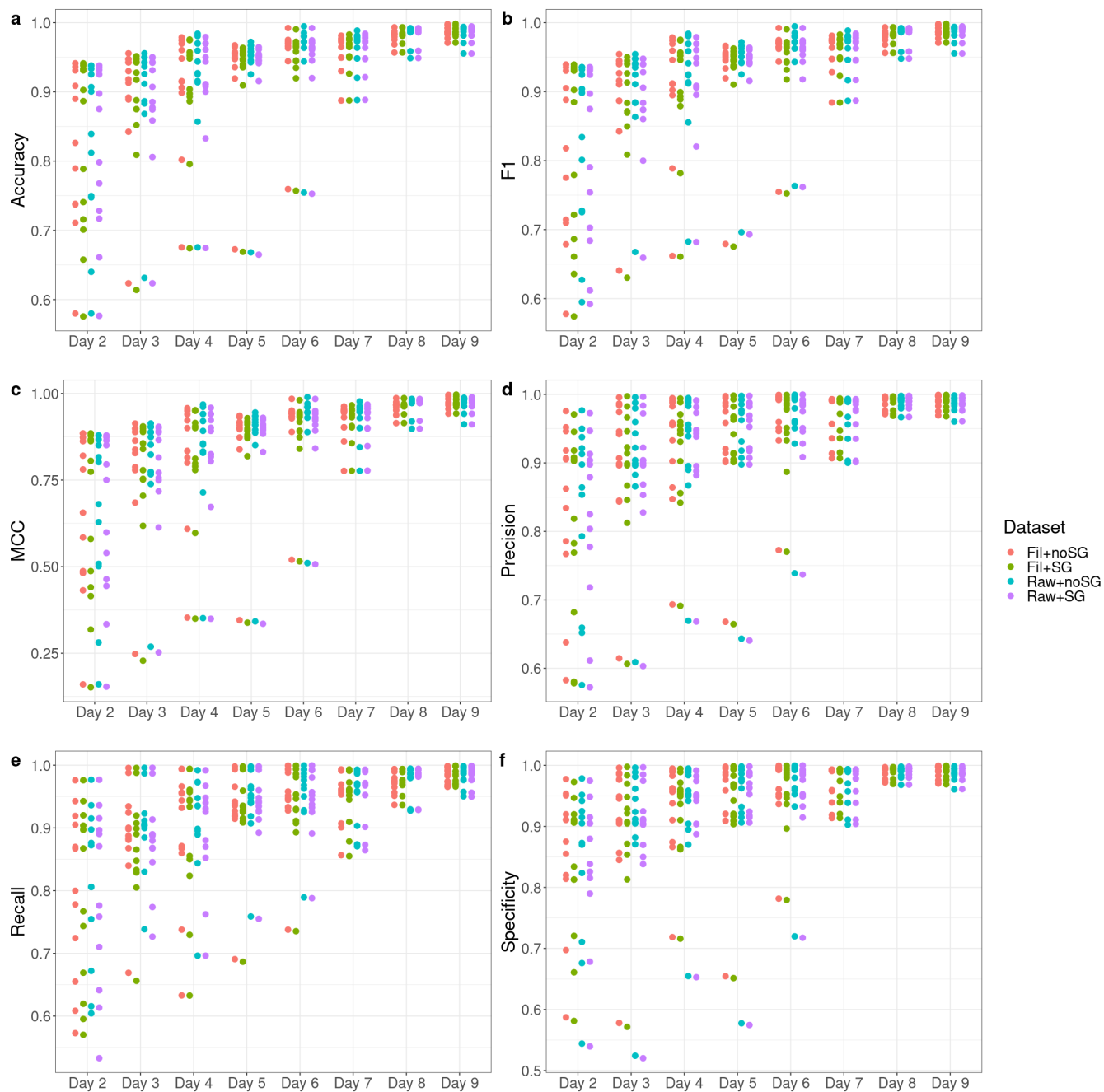


Figure A-2. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Kent variety (e2-kn assay)

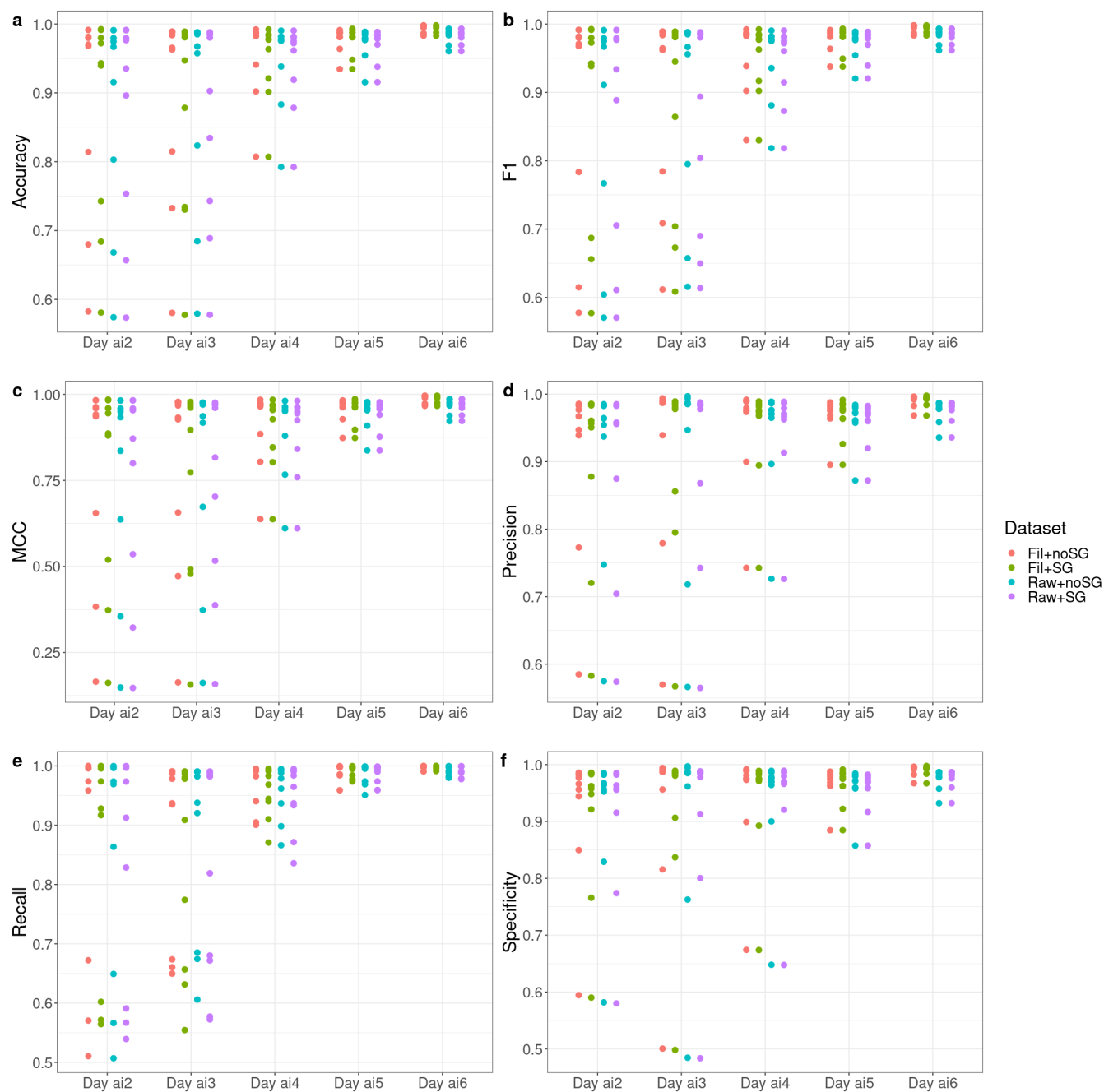


Figure A-3. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Osteen variety (e3-os assay)

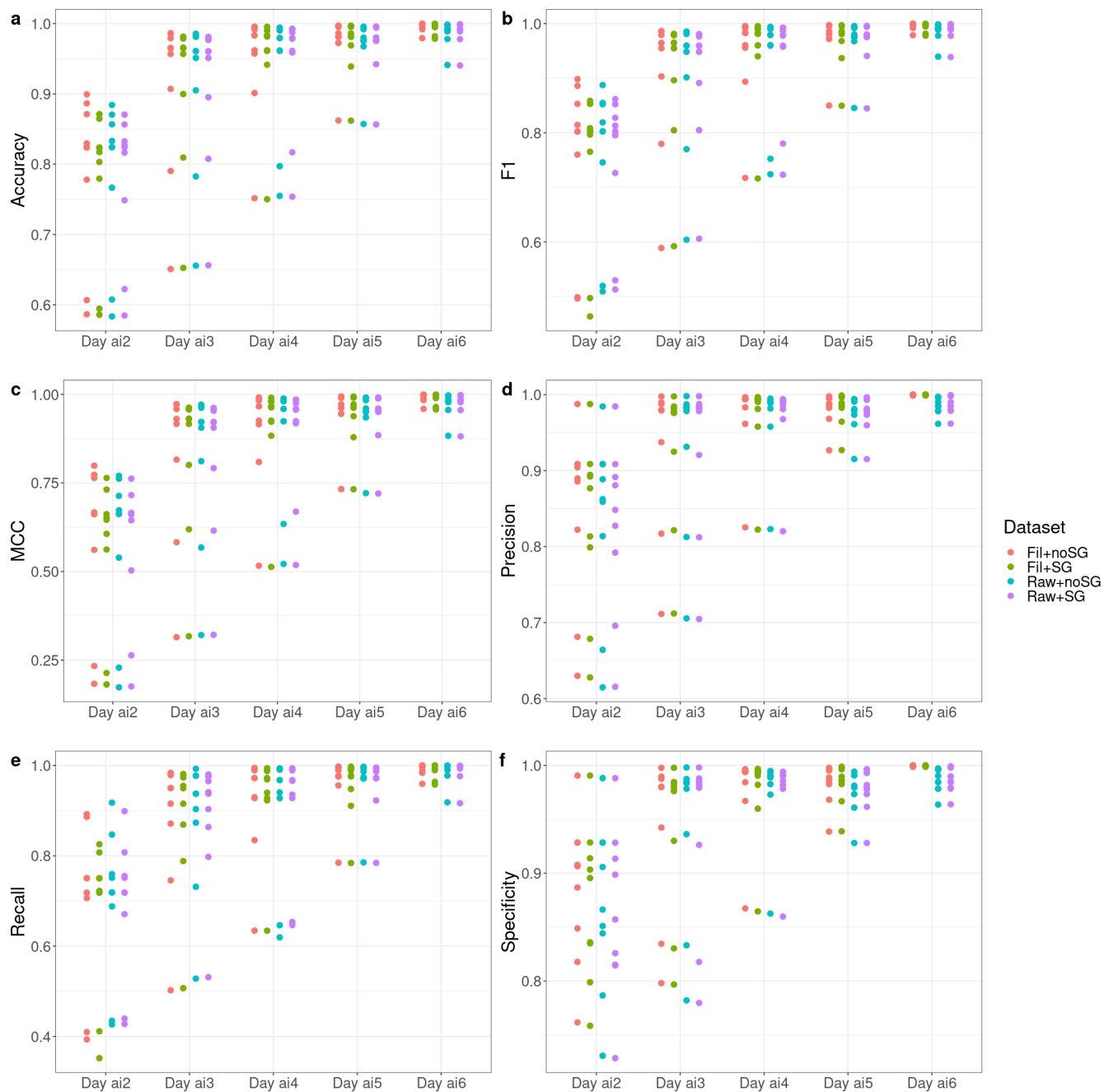


Figure A-4. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e4-kt assay)

A.1.2. Performance metrics results by classification model

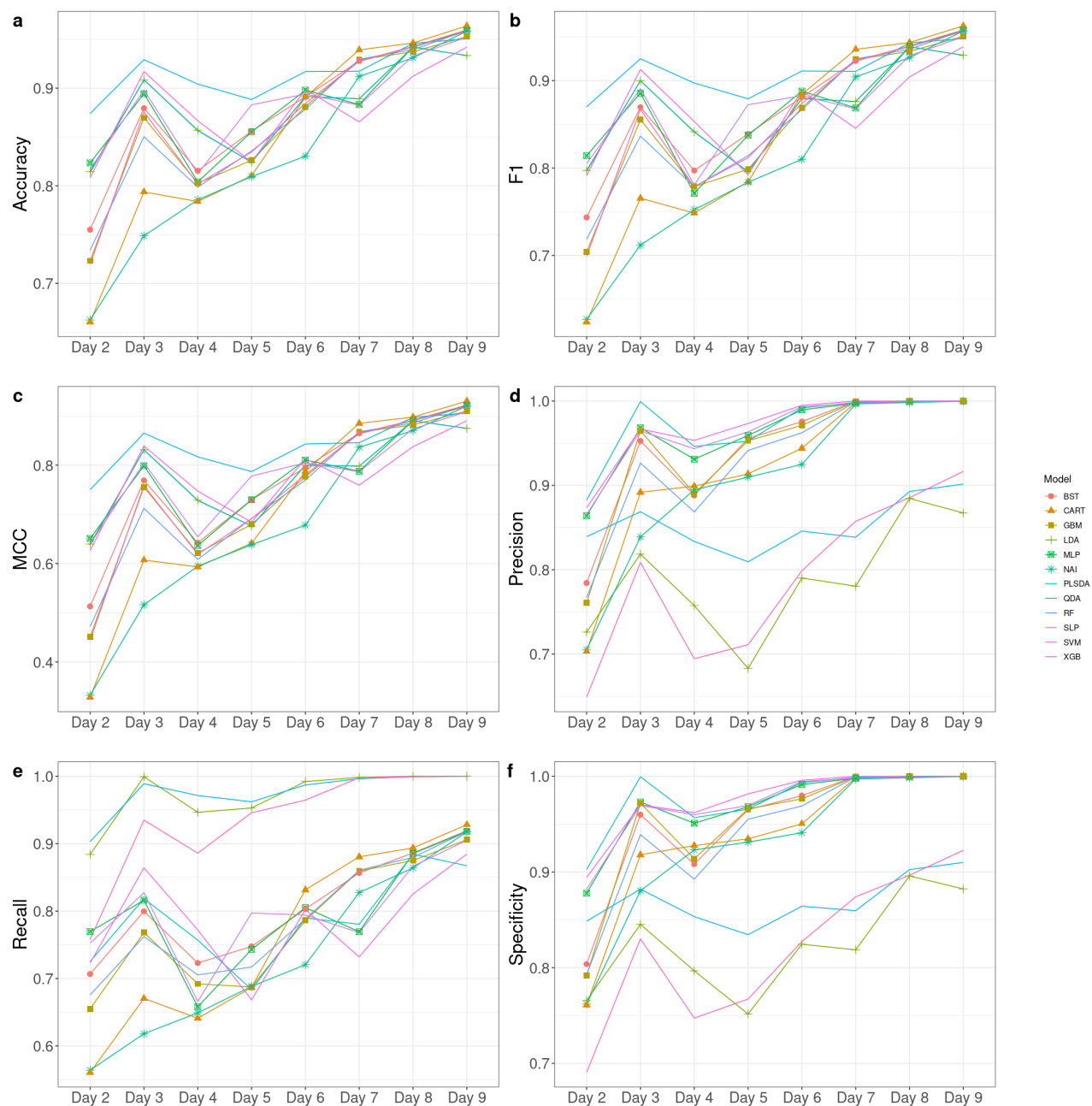


Figure A-5. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e1-kt assay)

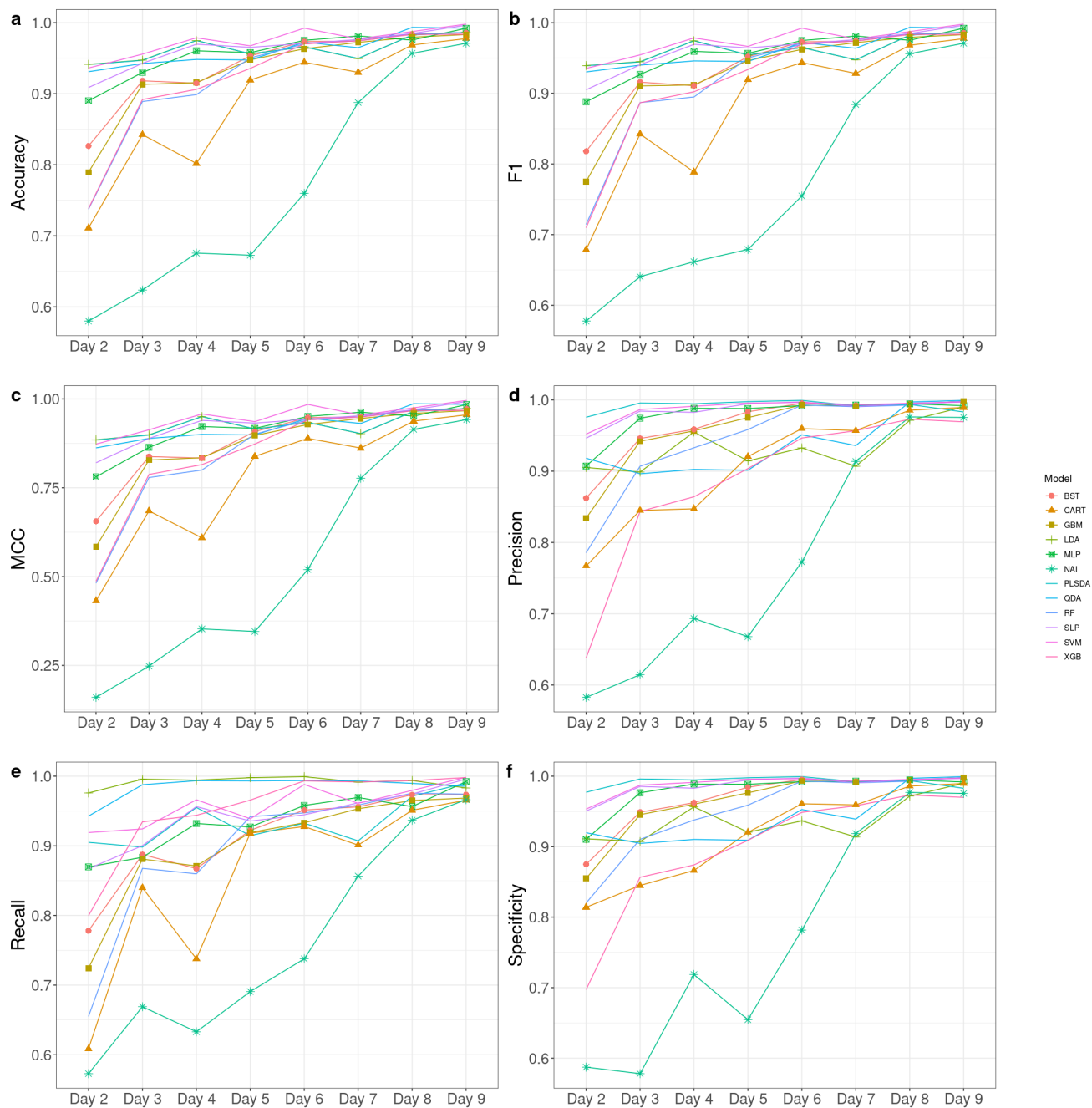


Figure A-6. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Kent variety (e2-kn assay)

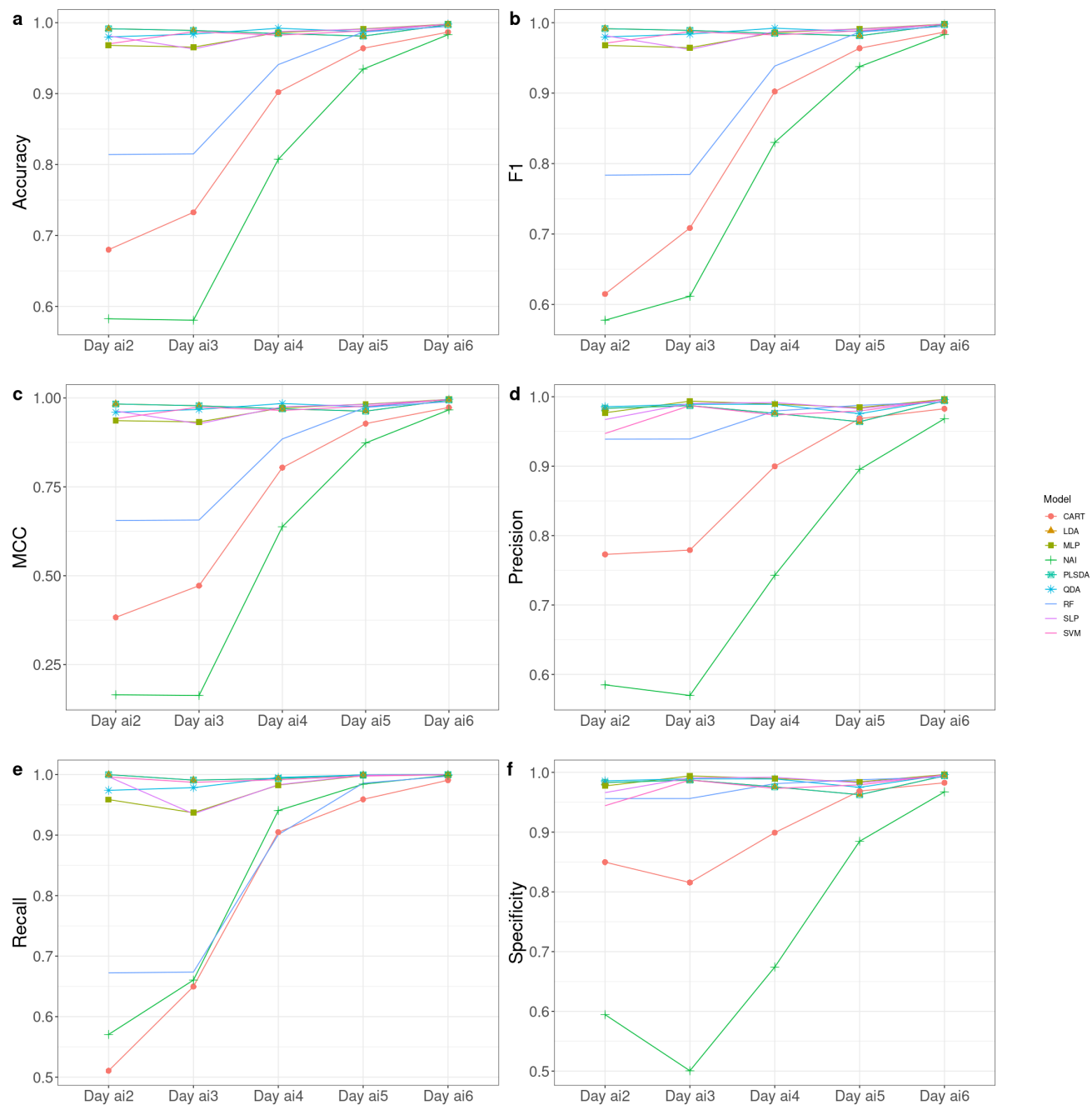


Figure A-7. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Osteen variety (e3-os assay)

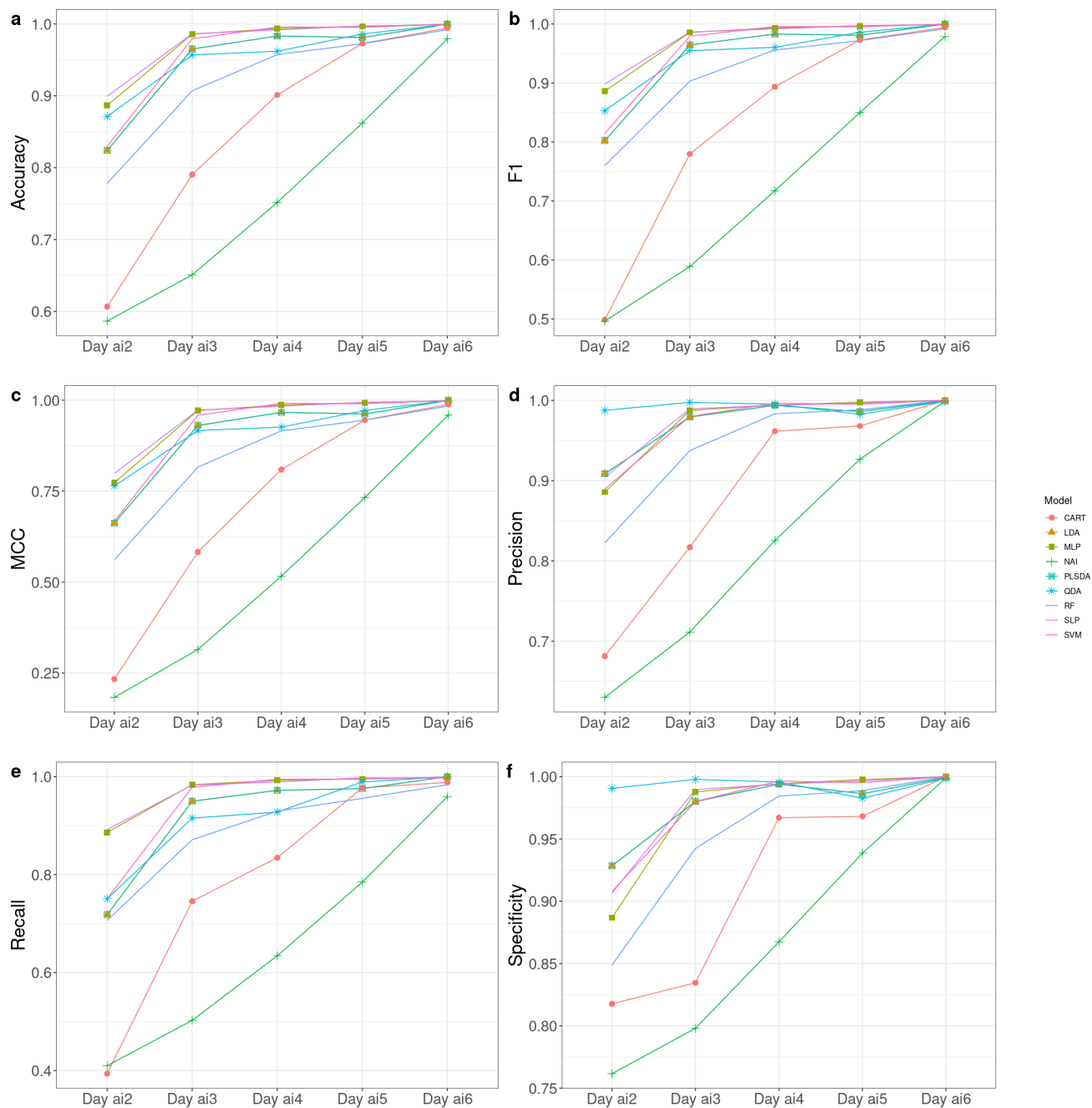


Figure A-8. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e4-kt assay)

A.2. Multiclass results for the mango fruit assays

Table A-1. Performance metrics of the dataset classification models for the Keitt variety e1-kt assay.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
LDA	0.871	0.872	0.984	0.862	0.850	0.867
PLSDA	0.868	0.868	0.983	0.860	0.847	0.864
RF	0.838	0.838	0.980	0.838	0.818	0.840
GBM	0.748	0.748	0.969	0.749	0.719	0.755
XGB	0.743	0.757	0.968	0.744	0.716	0.744
BST	0.409	0.409	0.926	0.389	0.000	0.000
CART	0.390	0.390	0.924	0.367	0.000	0.000
NAI	0.277	0.277	0.910	0.226	0.159	0.235

Table A-2. Performance metrics of the dataset classification models for the Keitt variety e4-kt assay.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
LDA	0.856	0.858	0.971	0.856	0.830	0.861
RF	0.842	0.840	0.968	0.840	0.809	0.841
PLSDA	0.825	0.827	0.965	0.823	0.791	0.827
GBM	0.782	0.781	0.957	0.781	0.738	0.782
XGB	0.766	0.765	0.954	0.762	0.717	0.764
BST	0.531	0.529	0.906	0.514	0.431	0.524
CART	0.527	0.525	0.906	0.510	0.426	0.520
NAI	0.443	0.439	0.889	0.412	0.313	0.412

Table A-3. Performance metrics of the detection models for the Kent variety e2-kn assay.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
RF	0.884	0.884	0.985	0.884	0.869	0.884
LDA	0.873	0.875	0.984	0.866	0.856	0.877
PLSDA	0.864	0.866	0.983	0.860	0.848	0.868
XGB	0.807	0.809	0.976	0.808	0.784	0.808
GBM	0.798	0.798	0.975	0.798	0.774	0.800
BST	0.395	0.393	0.924	0.338	0.000	0.000
CART	0.392	0.390	0.924	0.332	0.000	0.000
NAI	0.339	0.338	0.917	0.311	0.247	0.338

Table A-4. Performance metrics of the classification models for the Osteen e3-os assay.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
RF	0.884	0.884	0.985	0.884	0.869	0.884
LDA	0.873	0.875	0.984	0.866	0.856	0.877
PLSDA	0.864	0.866	0.983	0.860	0.848	0.868
XGB	0.807	0.809	0.976	0.808	0.784	0.808
GBM	0.798	0.798	0.975	0.798	0.774	0.800
BST	0.395	0.393	0.924	0.338	0.000	0.000
CART	0.392	0.390	0.924	0.332	0.000	0.000
NAI	0.339	0.338	0.917	0.311	0.247	0.338

A.3. Multiclass results for the mango fruit assays with selected bands

Due to the number of classification models and results obtained per dataset and feature selector, the complete list of classification models obtained per assay is provided as complementary material. It can be requested by emailing cvelasquez@unal.edu.co or consulted at [Complementary material](#).

B. Appendix: Results of mango leaf assays

B.1. Biclass results for the mango leaf assays

B.1.1. Performance metrics results per dataset

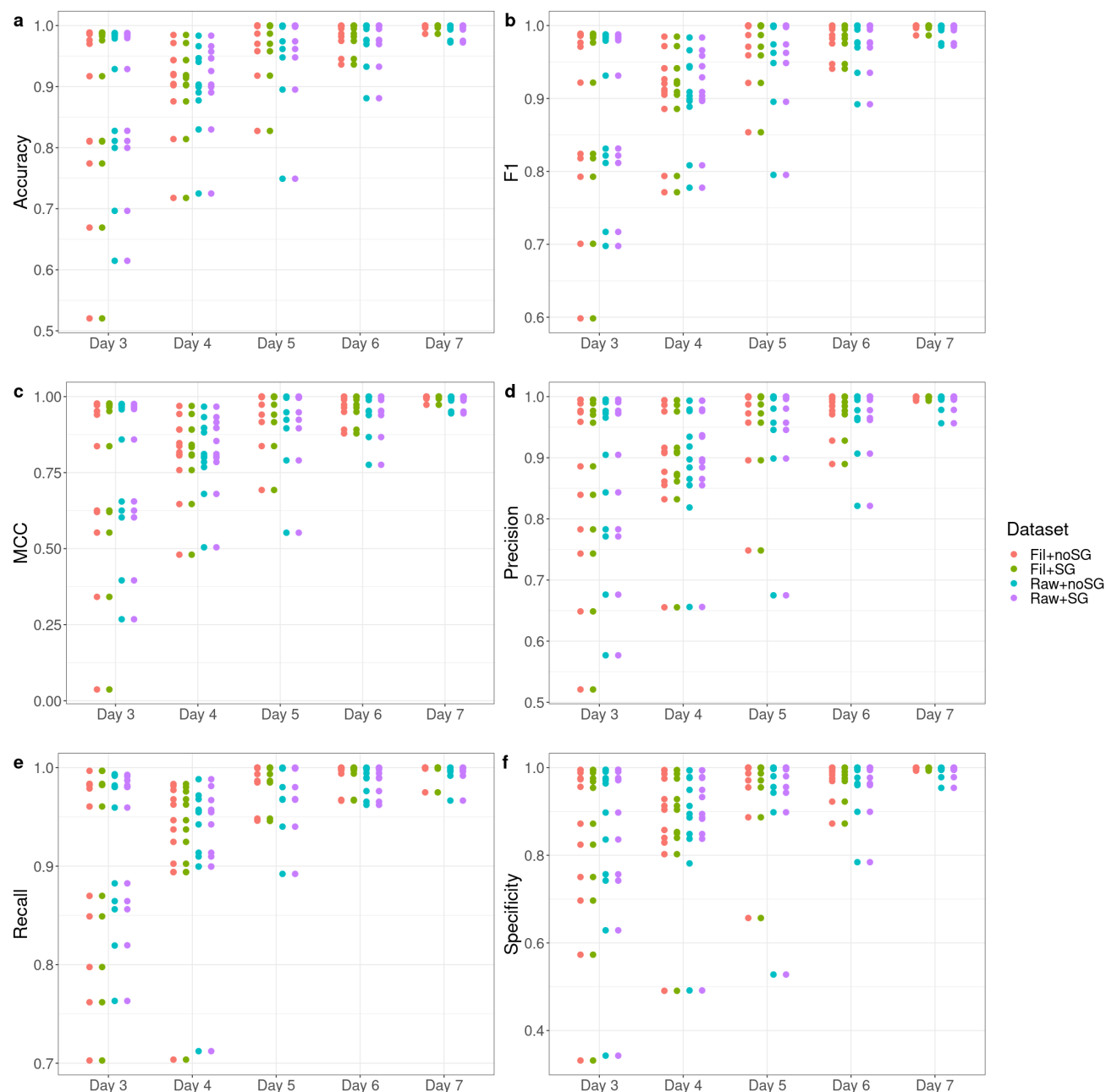


Figure B-1. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Tommy Atkins variety (e5-to assay)

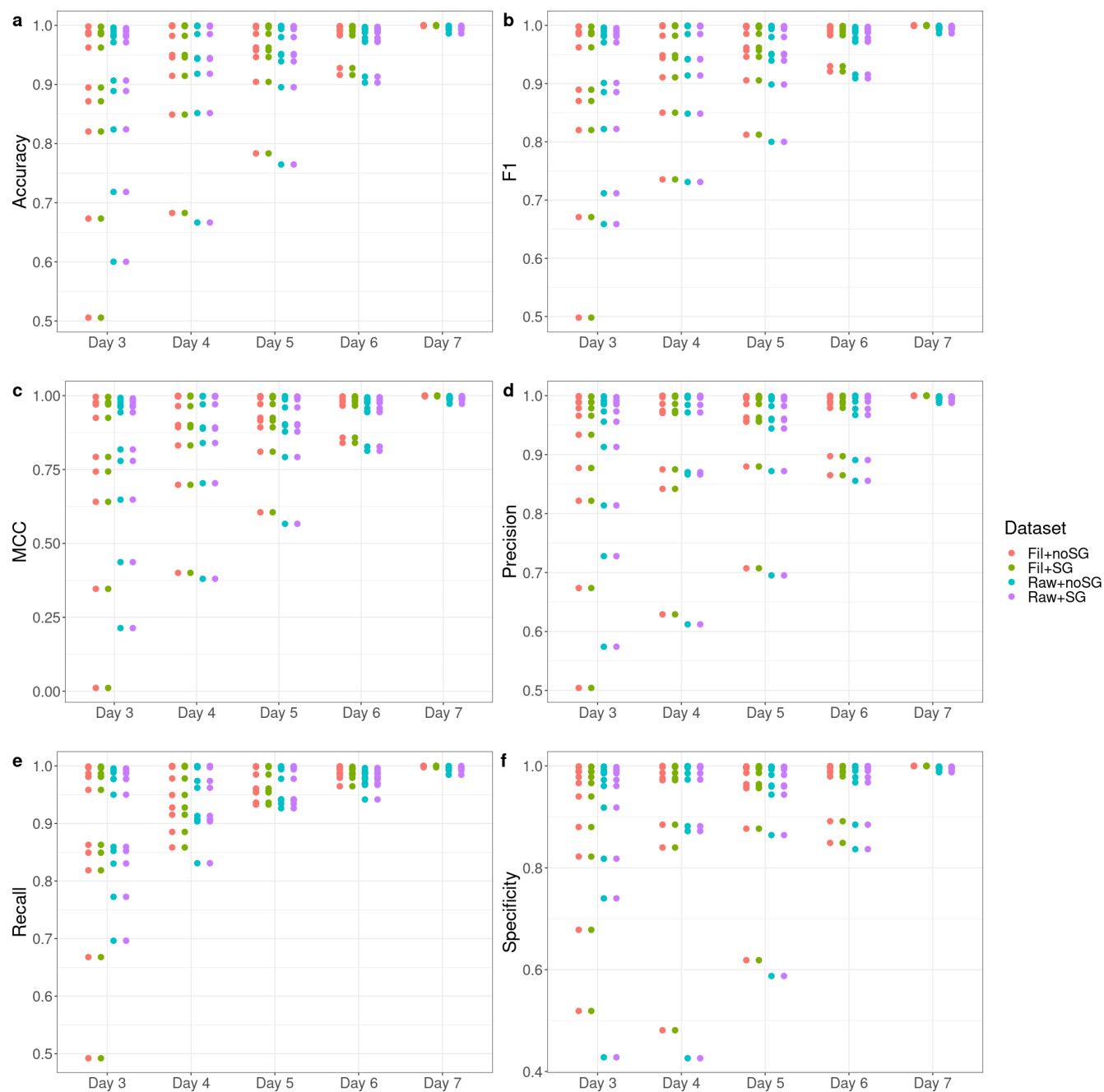


Figure B-2. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Tommy Atkins variety (e6-to assay)

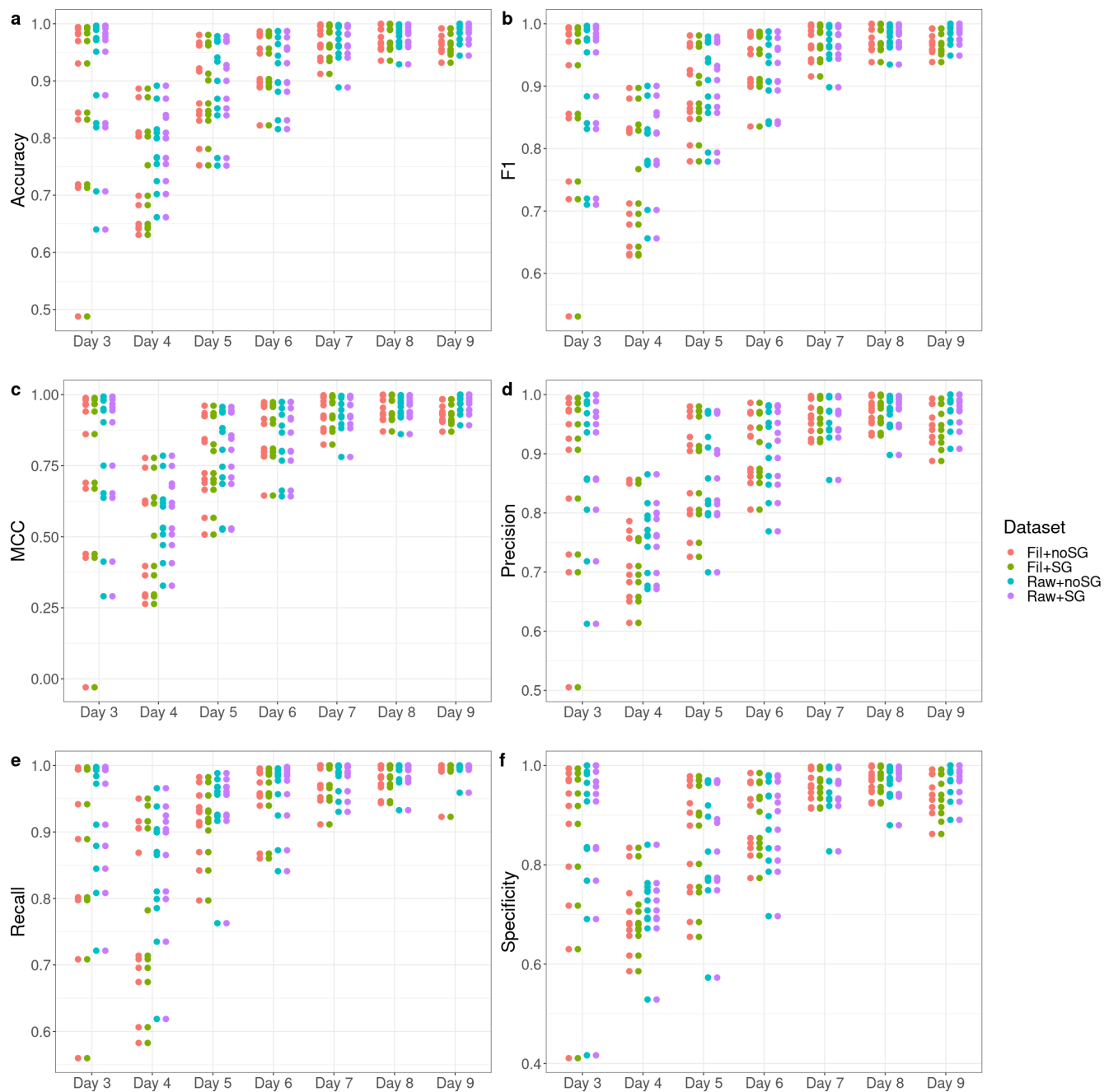


Figure B-3. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e7-kt assay)

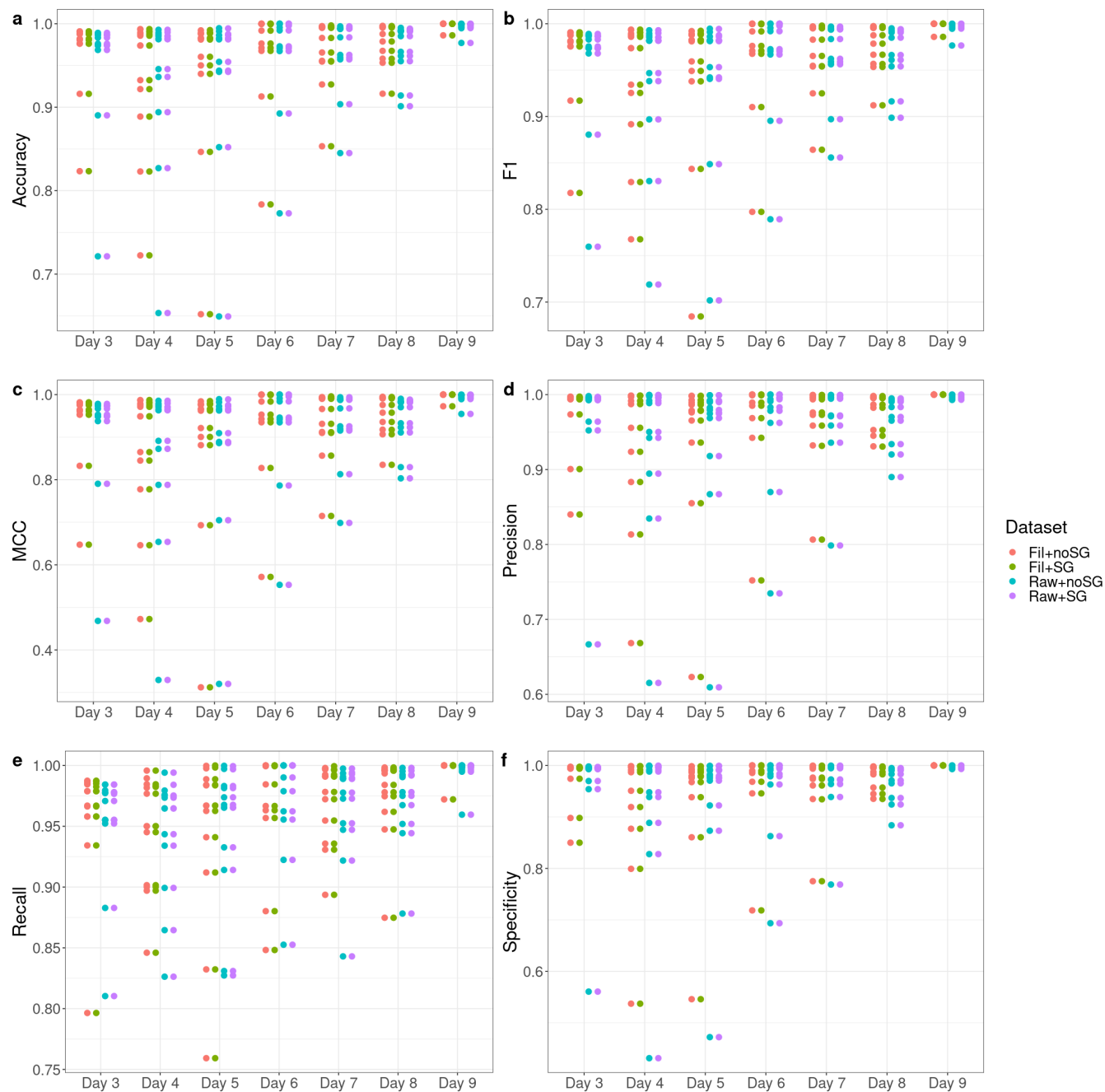


Figure B-4. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Keitt variety (e8-kt assay)

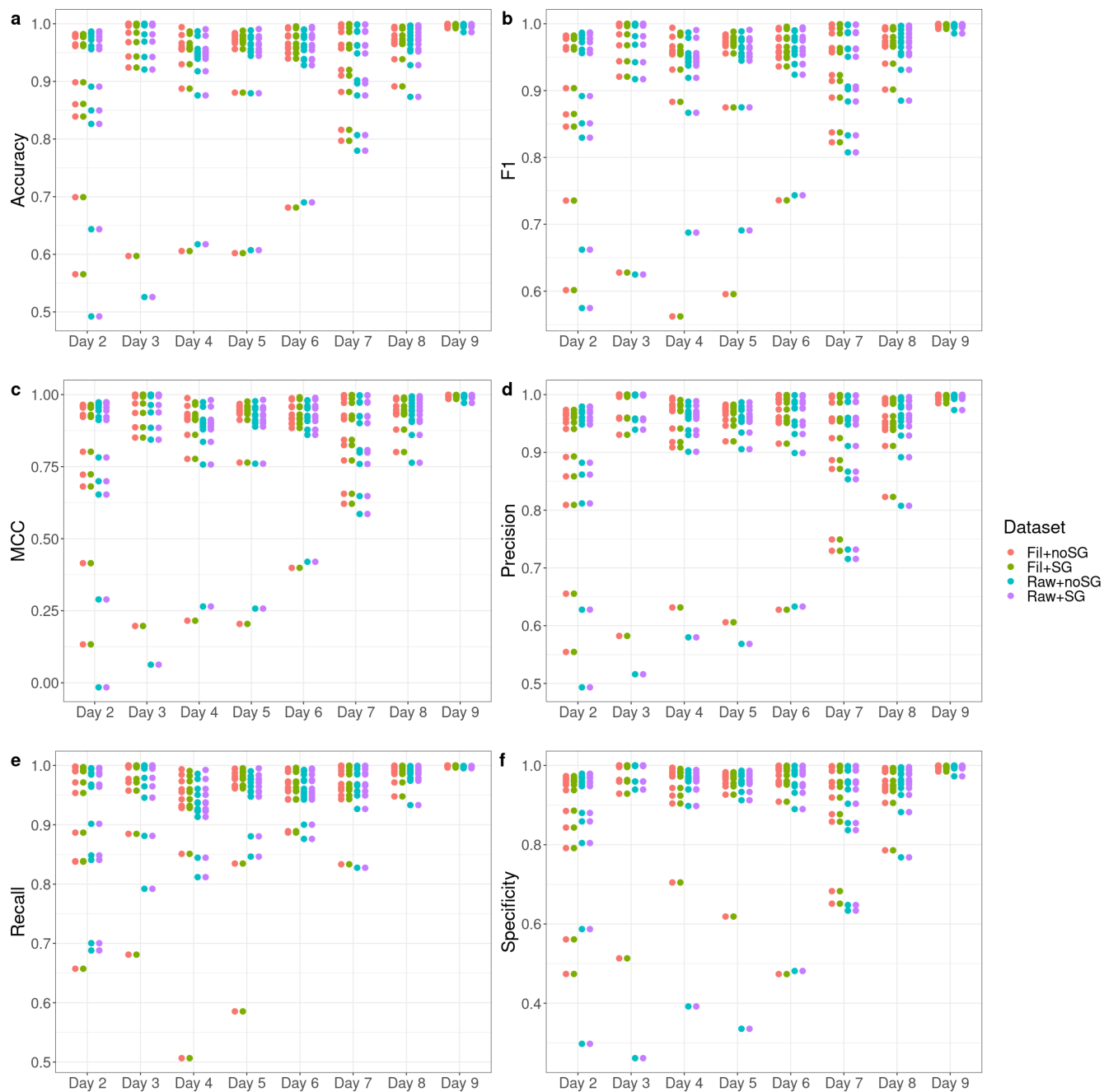


Figure B-5. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented by dataset for the Kent variety (e9-kn assay)

B.1.2. Performance metrics results by classification model

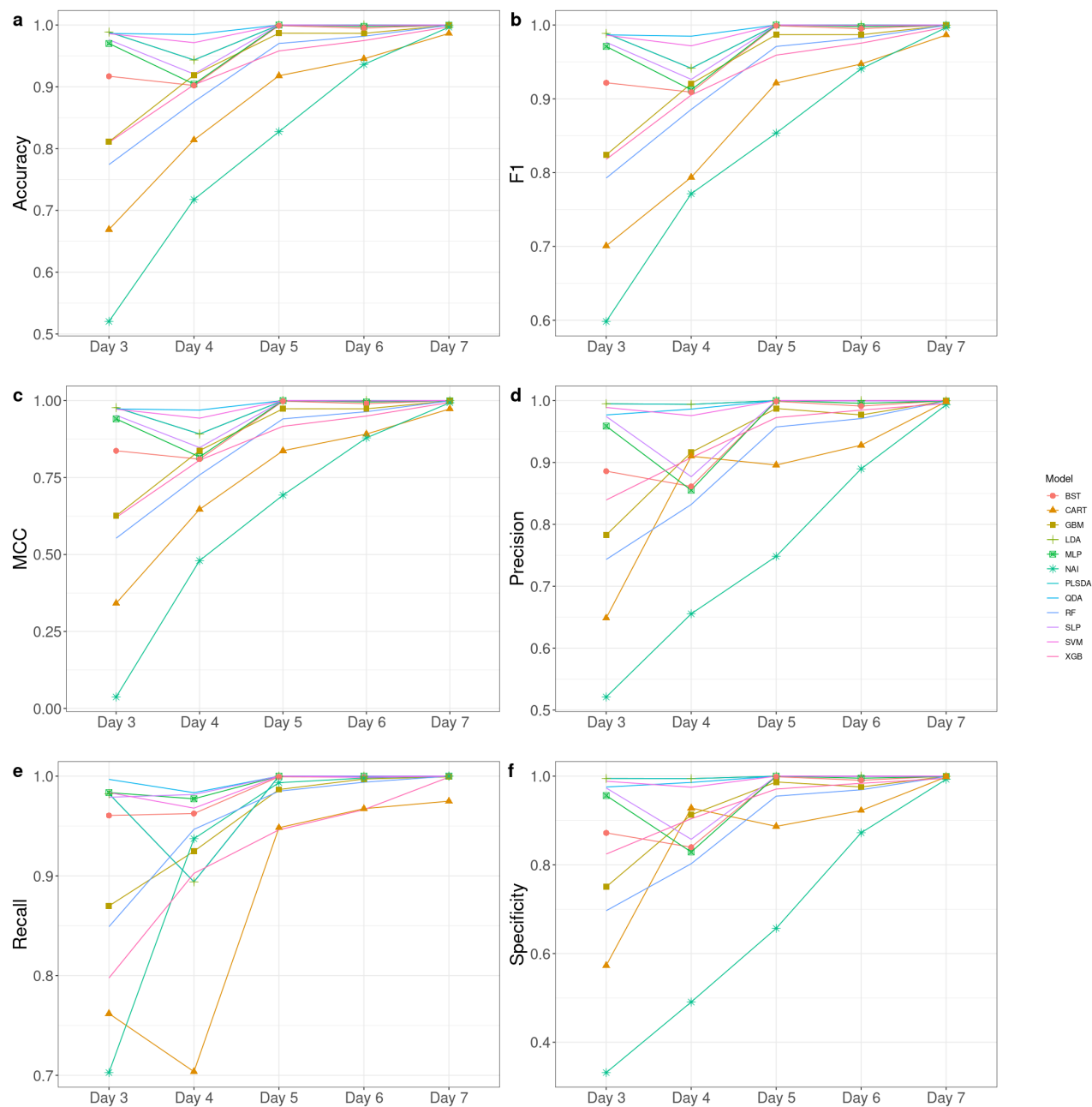


Figure B-6. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Tommy Atkins variety (e5-to assay)

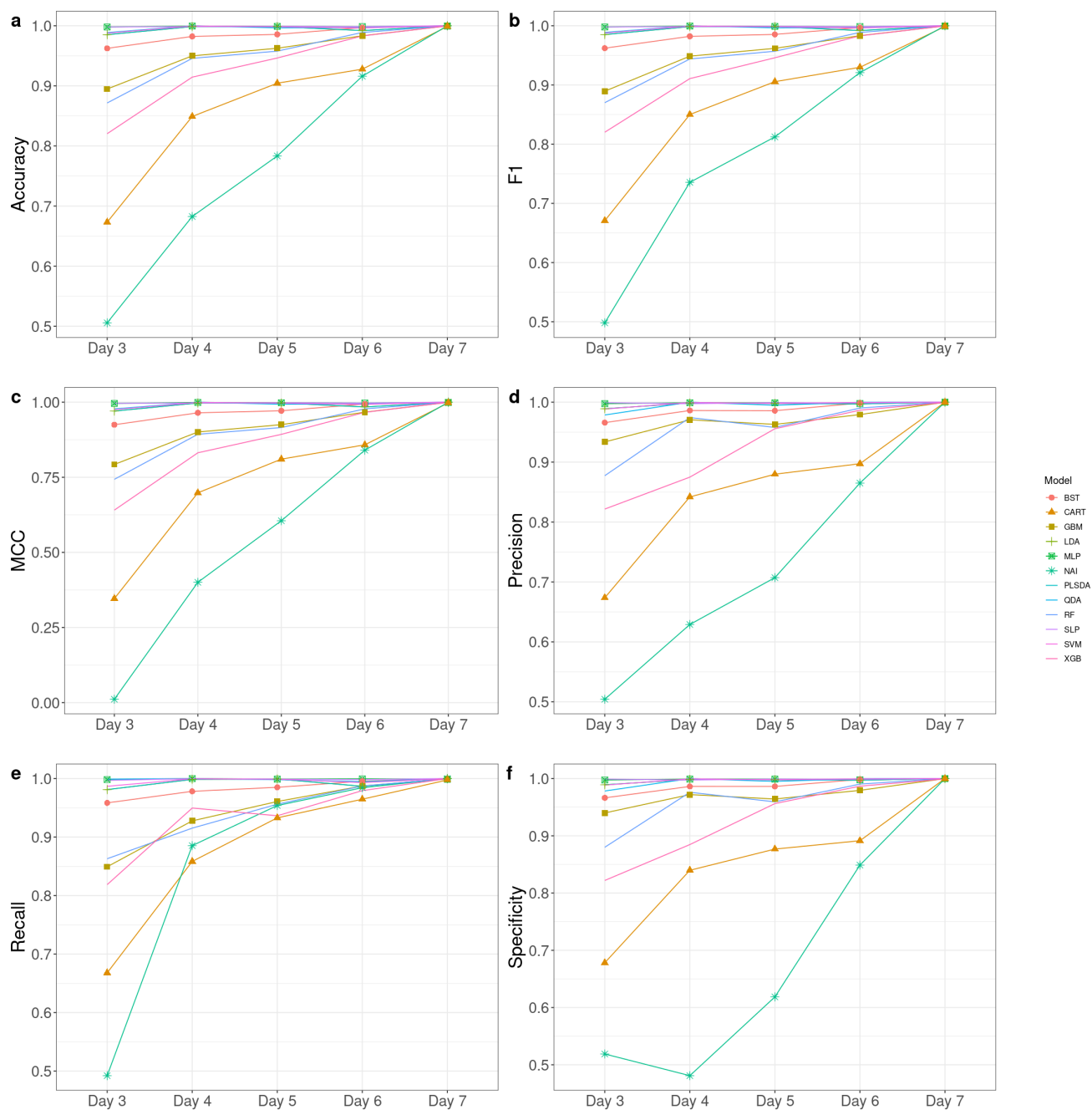


Figure B-7. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Tommy Atkins variety (e6-to assay)

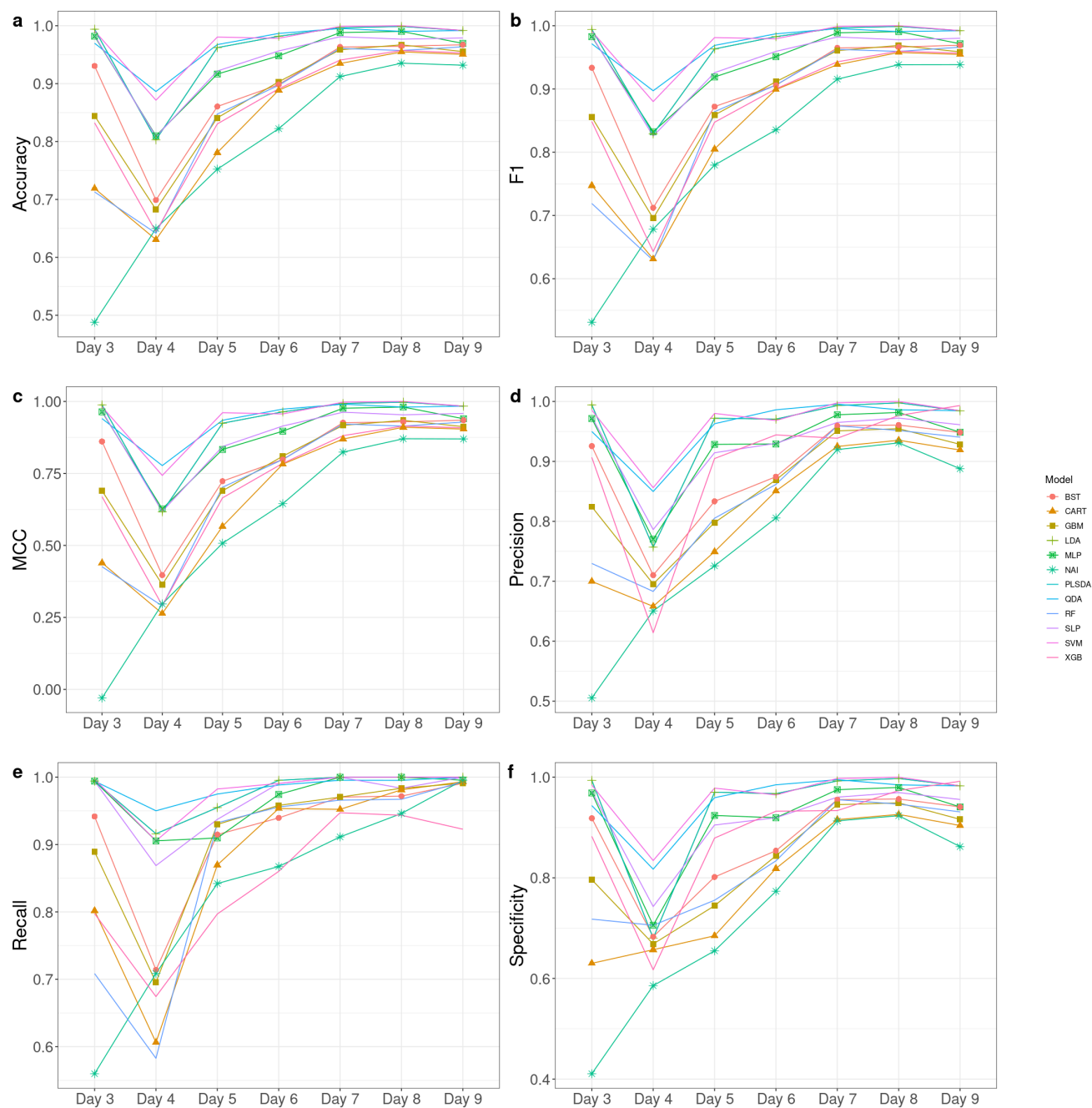


Figure B-8. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e7-kt assay)

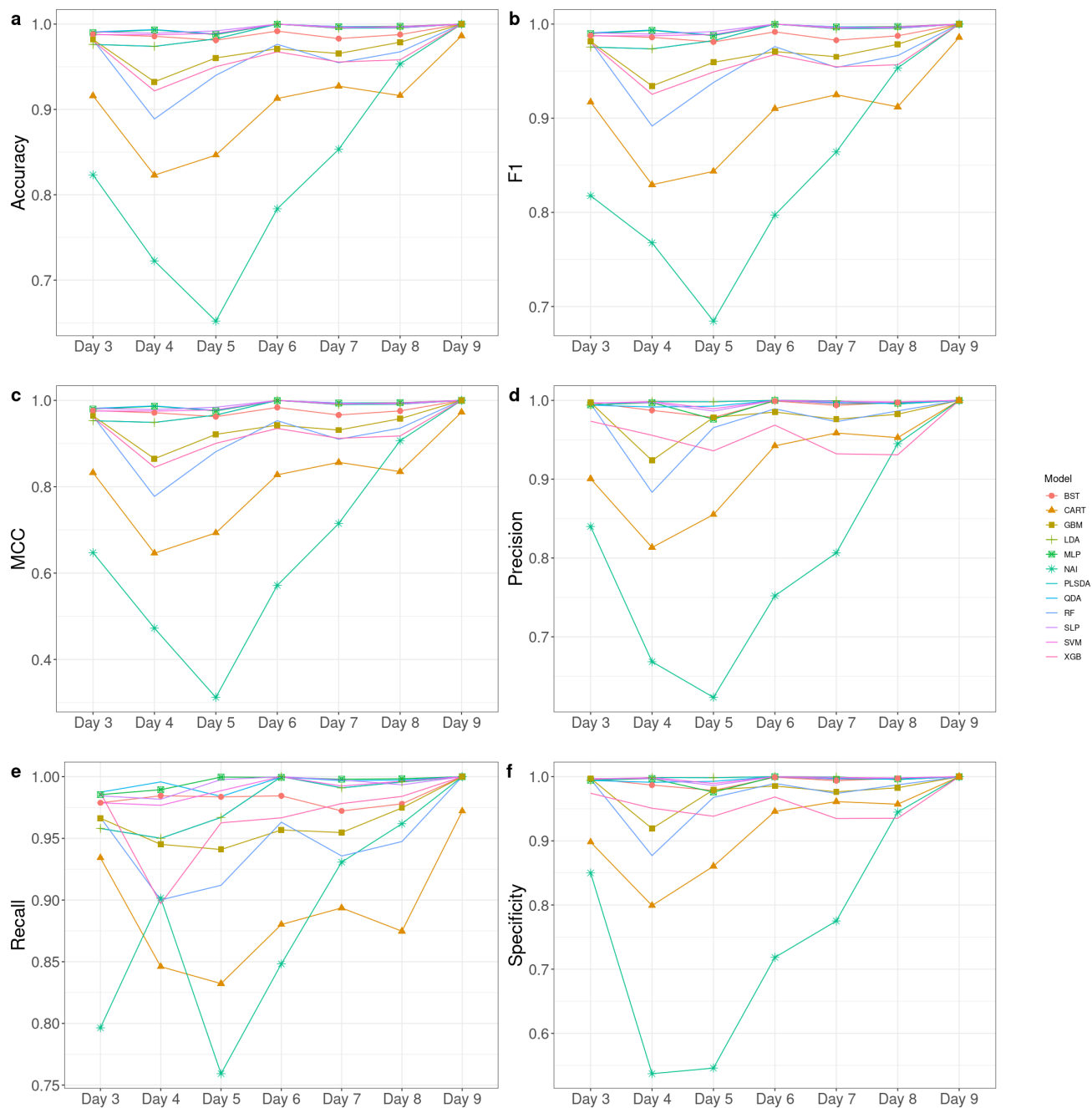


Figure B-9. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Keitt variety (e8-kt assay)

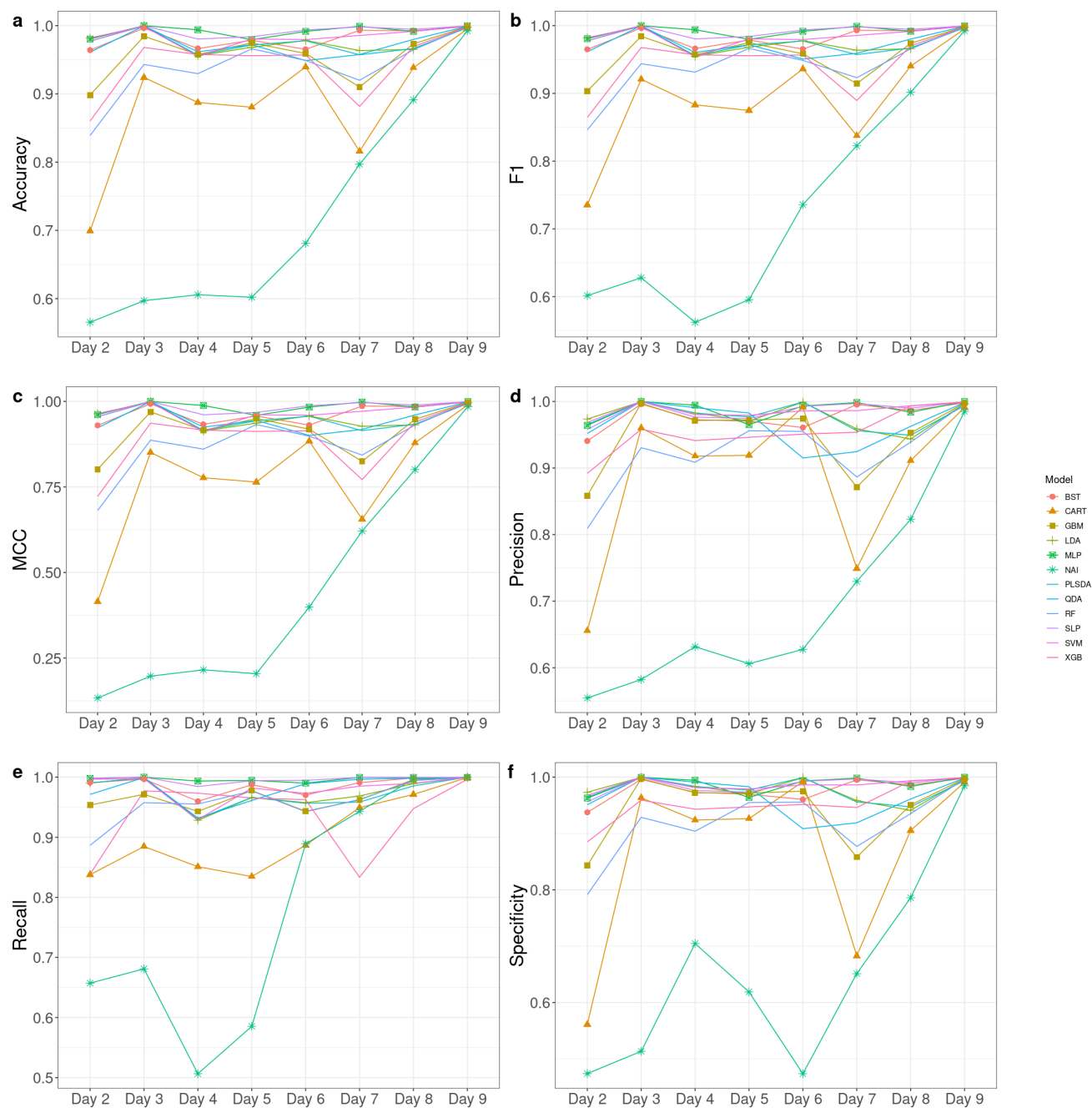


Figure B-10. Performance metrics (a. Accuracy, b. F1, c. MCC, d. Precision, e. Recall and f. Specificity) of the classification models implemented with the Fil+noSG dataset for the Kent variety (e9-kn assay)

B.2. Multiclass results for the mango leaf assays

Table B-1. Performance metrics of the classification models for the Tommy Atkins e5-to experiment.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
PLSDA	0.924	0.922	0.985	0.922	0.908	0.926
GBM	0.883	0.881	0.977	0.882	0.859	0.884
XGB	0.866	0.869	0.973	0.865	0.839	0.863
RF	0.843	0.838	0.968	0.841	0.813	0.854
BST	0.643	0.639	0.929	0.643	0.574	0.654
CART	0.517	0.515	0.903	0.510	0.434	0.563
NAI	0.402	0.400	0.880	0.363	0.276	0.405

Table B-2. Performance metrics of the classification models for the Tommy Atkins e6-to experiment.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
PLSDA	0.898	0.897	0.980	0.895	0.877	0.899
GBM	0.802	0.801	0.961	0.800	0.762	0.803
RF	0.800	0.798	0.960	0.798	0.760	0.803
XGB	0.785	0.785	0.957	0.782	0.740	0.783
BST	0.629	0.630	0.926	0.610	0.551	0.629
CART	0.532	0.538	0.907	0.506	0.439	0.540
NAI	0.390	0.397	0.878	0.341	0.262	0.393

Table B-3. Performance metrics of the classification models for the Keitt variety e7-kt assay.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
PLSDA	0.781	0.788	0.969	0.775	0.753	0.789
GBM	0.545	0.551	0.935	0.539	0.490	0.577
XGB	0.478	0.515	0.926	0.469	0.415	0.486
RF	0.411	0.417	0.916	0.399	0.333	0.438
BST	0.300	0.298	0.900	0.260	0.210	0.371
CART	0.287	0.286	0.898	0.241	0.187	0.329
NAI	0.235	0.231	0.891	0.214	0.130	0.264

Table B-4. Performance metrics of the Keitt variety e8-kt assay classification models.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
PLSDA	0.908	0.897	0.987	0.900	0.888	0.907
RF	0.794	0.776	0.970	0.780	0.756	0.802
GBM	0.784	0.764	0.969	0.765	0.742	0.791
XGB	0.783	0.789	0.969	0.768	0.743	0.765
BST	0.534	0.515	0.933	0.494	0.463	0.573
CART	0.500	0.476	0.928	0.449	0.414	0.523
NAI	0.399	0.382	0.914	0.324	0.280	0.388

Table B-5. Performance metrics of the Kent variety e9-kn assay classification models.

Model	Accuracy	Recall	Specificity	F1	MCC	Precision
PLSDA	0.892	0.891	0.987	0.887	0.875	0.888
GBM	0.821	0.823	0.978	0.820	0.800	0.826
XGB	0.815	0.823	0.977	0.815	0.794	0.815
RF	0.814	0.814	0.977	0.813	0.793	0.820
BST	0.400	0.395	0.925	0.355	0.000	0.000
CART	0.379	0.375	0.922	0.334	0.000	0.000
NAI	0.278	0.275	0.910	0.238	0.176	0.278

B.3. Multiclass results for the mango fruit assays with selected bands

Due to the number of classification models and results obtained per dataset and feature selector, the complete list of classification models obtained per assay is provided as complementary material. It can be requested by emailing cvelasquez@unal.edu.co or consulted at [Complementary material](#).

Bibliography

- [1] ABDULRIDHA, J.; EHSANI, R. ; DE CASTRO, A.: Detection and differentiation between laurel wilt disease, phytophthora disease, and salinity damage using a hyperspectral sensing technique. In: *Agriculture*, 6(4) (2016)
- [2] ACOSTA, Maylin; QUIÑONES, Ana; MUNERA, Sandra; DE PAZ, José M. ; BLASCO, José: Rapid Prediction of Nutrient Concentration in Citrus Leaves Using Vis-NIR Spectroscopy. In: *Sensors* 23 (2023), Nr. 14, S. 6530
- [3] ACOSTA, Maylin; RODRÍGUEZ-CARRETERO, Isabel; BLASCO, José; DE PAZ, José M. ; QUIÑONES, Ana: Non-Destructive Appraisal of Macro- and Micronutrients in Persimmon Leaves Using Vis/NIR Hyperspectral Imaging. In: *Agriculture* 13 (2023), Nr. 4, S. 916
- [4] ACOSTA, Maylin; VISCONTI, Fernando; QUIÑONES, Ana; BLASCO, José ; DE PAZ, José M.: Estimation of Macro and Micronutrients in Persimmon (*Diospyros kaki* L.) cv. ‘Rojo Brillante’ Leaves through Vis-NIR Reflectance Spectroscopy. In: *Agronomy* 13 (2023), Nr. 4, S. 1105
- [5] ADEBAYO, Segun E.; HASHIM, Norhashila; ABDAN, Khalina; HANAFI, Marsyita ; MOLLAZADE, Kaveh: Prediction of quality attributes and ripeness classification of bananas using optical properties. In: *Scientia Horticulturae* 212 (2016), S. 171–182
- [6] ADÃO, Telmo; HRUŠKA, Jonáš; PÁDUA, Luís; BESSA, José; PERES, Emanuel; MORAIS, Raul ; SOUSA, Joaquim J.: Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. In: *Remote Sensing* 9, Nr. 11, S. 2017
- [7] AGRONET: *Reporte: Area, Producción y Rendimiento Nacional por Cultivo [online]*. 2024. – Accessed on August 3rd, 2024
- [8] AHMAD, Muhammad: *Literature Review – Traditional to Deep Models*. Università Degli Studi Di Messina, 2021. – 16–51 S
- [9] AIRES, Priscila S.; GAMBARRA-NETO, Francisco F.; COUTINHO, Wirtton M.; DE ARAÚJO, Alderi E.; DA SILVA, G. F.; GOUVEIA, J. P. ; MEDEIROS, Everaldo P.: Near infrared hyperspectral images and pattern recognition techniques used to identify

- etiological agents of cotton anthracnose and ramulosis. In: *Journal of Spectral Imaging* 7 (2018), S. a8
- [10] AKBARI, Hamed; UTO, Kuniaki; KOSUGI, Yukio; KOJIMA, Kazuyuki ; TANAKA, Naofumi: Cancer detection using infrared hyperspectral imaging. In: *Cancer science* 102 (2011), Nr. 4, S. 852–857
- [11] AL-SANABANI, Dheya Galal A.; SOLIHIN, Mahmud I.; PUI, Liew P.; ASTUTI, Winda; ANG, Chun K. ; HONG, Lim W.: Development of non-destructive mango assessment using Handheld Spectroscopy and Machine Learning Regression. In: *Journal of Physics: Conference Series* 1367 (2019), Nr. 1, S. 012030
- [12] ALIBABAEI, K.; GASPAR, P. D.; LIMA, T. M.; CAMPOS, R. M.; GIRAO, I.; MONTEIRO, J. ; LOPES, C. M.: A Review of the Challenges of Using Deep Learning Algorithms to Support Decision-Making in Agricultural Activities. In: *Remote Sensing* 14 (2022), Nr. 3
- [13] ALSALAM, Bilal Hazim Y.; MORTON, Kye; CAMPBELL, Duncan ; GONZALEZ, Felipe: Autonomous UAV with vision based on-board decision making for remote sensing and precision agriculture. In: *2017 IEEE Aerospace Conference*, 2017, S. 1–12
- [14] AMAT RODRIGO, Joaquín: *Machine Learning con R [online]*. 2017. – Accessed on August 3rd, 2024
- [15] ANAGNOSTIS, A.; TAGARAKIS, A.C.; ASIMINARI, G.; PAPAGEORGIOU, E.; KATERIS, D.; MOSHOU, D. ; BOCHTIS, D.: A deep learning approach for anthracnose infected trees classification in walnut orchards. In: *Computers and Electronics in Agriculture* 182 (2021), S. 105998
- [16] ARAKERI, Megha. P.; KUMAR, B. P. V.; BARSAYIA, Shubham ; SAIRAM, H. V.: Computer vision based robotic weed control system for precision agriculture. In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, S. 1201–1205
- [17] ARENDSE, Ebrahiema; FAWOLE, Olaniyi A.; MAGWAZA, Lembe S. ; OPARA, Umezurike L.: Non-destructive prediction of internal and external quality attributes of fruit with thick rind: A review. In: *Journal of Food Engineering* 217 (2018), S. 11–23
- [18] ARIVAZHAGAN, S.; LIGI, S.V.: Mango leaf diseases identification using convolutional neural network. In: *International Journal of Pure and Applied Mathematics* 120 (2018), S. 11067–11079
- [19] ASHRAFUL, A.; KUMAR, A. S. ; MAHTALAT, A.: Morphological characterization of *Colletotrichum gloeosporioides* identified from anthracnose of *mangifera indica* L. In: *Asian Journal of Plant Pathology* 11 (2017), Nr. 3, S. 102–117

- [20] ATTRI, Ishana; AWASTHI, Lalit K.; SHARMA, Teek P. ; RATHEE, Priyanka: A review of deep learning techniques used in agriculture. In: *Ecological Informatics* 77 (2023), S. 102217
- [21] AVENDANO, J.; RAMOS, P.J. ; PRIETO, F.A.: A system for classifying vegetative structures on coffee branches based on videos recorded in the field by a mobile device. In: *Expert Systems with Applications* 88 (2017), S. 178–192. – ISSN 0957–4174
- [22] BASTIDAS-RODRIGUEZ, María X.; POLANIA, Luisa; GRUSON, Adrien ; PRIETO-ORTIZ, Flavio: Deep Learning for fractographic classification in metallic materials. In: *Engineering Failure Analysis* 113 (2020), S. 104532. – ISSN 1350–6307
- [23] BHAGWAT, R. G.; MEHTA, B. P.; PATIL, V. A. ; SHARMA, H.: Screening Of Cultivars/Varieties Against Mango Anthracnose Caused by *Colletotrichum Gloeosporioides*. In: *International Journal of Environmental & Agriculture Research* 1 (2015), Nr. 1, S. 21–23
- [24] BINCADER, S.; PONGPISUTTA, R. ; RATTANAKREETAKUL, C.: Diversity of *Colletotrichum* Species causing Anthracnose Disease from Mango cv. Nam Dork Mai See Tong based on ISSR-PCR. In: *Indian Journal of Agricultural Research* 56 (2022), Nr. 1, S. 81–90
- [25] BRAZILIAN FARMERS: *Mango: 7 varieties of a tropical global passion [online]*. 2022. – Accessed on August 3rd, 2024
- [26] BURGLEWSKI, N.; SRINIVASAGAN, S.; KETTERINGS, Q. ; VAN AARDT, J.: Spatial and Spectral Dependencies of Maize Yield Estimation Using Remote Sensing. In: *Sensors* 24 (2024), Nr. 12
- [27] BURRUECO, Daniel: *Matriz de pesos [online]*. – Accessed on August 3rd, 2024
- [28] BYRNES, Walker; AHLIN, Konrad; RAINS, Glen ; MCMURRAY, Gary: Methodology for Stress Identification in Crop Fields Using 4D Height Data. In: *IFAC-PapersOnLine* 52 (2019), Nr. 30, S. 336–341
- [29] BÉGUÉ, A.; LEBOURGEOIS, V.; BAPPEL, E.; TODOROFF, P.; PELLEGRINO, A.; BAILLARIN, F. ; SIEGMUND, B.: Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI. In: *International Journal of Remote Sensing* 31 (2010), Nr. 20, S. 5391–5407. – ISSN 0143–1161
- [30] CABRERA ARDILA, C. E.; ALBERTO RAMIREZ, L. ; PRIETO ORTIZ, F. A.: Spectral analysis for the early detection of anthracnose in fruits of Sugar Mango (*Mangifera indica*). In: *Computers and Electronics in Agriculture* 173 (2020), S. 105357

- [31] CASTILLO-GIRONES, S.; VAN BELLEGHEM, R.; WOUTERS, N.; MUNERA, S.; BLASCO, J. ; SAEYS, W.: Detection of subsurface bruises in plums using spectral imaging and deep learning with wavelength selection. In: *Postharvest Biology and Technology* 207 (2024), S. 112615
- [32] CASTILLO-GIRONES, S.; VAN BELLEGHEM, R.; WOUTERS, N.; MUNERA, S.; BLASCO, J. ; SAEYS, W.: Detection of subsurface bruises in plums using spectral imaging and deep learning with wavelength selection. In: *Postharvest Biology and Technology* 207 (2024), S. 112615
- [33] CHILLET, M.; ANDRIANJAFINANDRASANA, S.; MINIER, J.; DANTHU, P.; RAMONTA, I. R. ; LECHAUDEL, M.: Alternative methods to control postharvest disease of mango. In: *Acta Horticulturae* 1183 (2017), S. 319–324
- [34] CHRYS, N. A.: Mango Anthracnose Disease: Present Status and Future Research Priorities. In: *Plant Pathology Journal* 5 (2006), Nr. 3, S. 266–273
- [35] COLILLES, Enrique: *Nace en España la Interprofesional del Aguacate y Mango [online]*. 2023. – Accessed on August 3rd, 2024
- [36] CORKIDI, G.; BALDERAS-RUÍZ, K. A.; TABOADA, B.; SERRANO-CARREÓN, L. ; GALINDO, E.: Assessing mango anthracnose using a new three-dimensional image-analysis technique to quantify lesions on fruit. In: 55 55 (2006), Nr. 2, S. 250–257
- [37] CORTÉS, V.; ORTIZ, C.; ALEIXOS, N.; BLASCO, J.; CUBERO, S. ; TALENS, P.: A new internal quality index for mango and its prediction by external visible and near-infrared reflection spectroscopy. In: *Postharvest Biology and Technology* 118 (2016), S. 148–158
- [38] CUI, Lihan; YAN, Lijie; ZHAO1, Xiaohu; YUAN, Lin; JIN, Jing ; ZHANG, Jingcheng: Detection and Discrimination of Tea Plant Stresses Based on Hyperspectral Imaging Technique at a Canopy Level. In: *Phyton-International Journal of Experimental Botany* 90 (2021), Nr. 2, S. 621–634
- [39] DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA, DANE. *El cultivo del mango, Mangifera indica, y su comportamiento frente a las condiciones ambientales y de manejo. Boletín mensual "insumos y factores de asociados a la producción agropecuaria*. 2015
- [40] DOFUOR, A. K.; QUARTEY, N. K.; OSABUTEY, A. F.; ANTWI-AGYAKWA, A. K.; ASANTE, K.; BOATENG, B. O.; ABLORMETI, F. K.; LUTUF, H.; OSEI-OWUSU, J.; OSEI, J. H. N.; EKLOH, W.; LOH, S. K.; HONGER, J. O.; AIDOO, O. F. ; NINSIN, K. D.: Mango anthracnose disease: the current situation and direction for future research. In: *Frontiers in microbiology* 14 (2023), S. 1168203

- [41] DONG, J.; BURNHAM, J. G.; BOOTS, B.; RAINS, G. ; DELLAERT, F.: 4D crop monitoring: Spatio-temporal reconstruction for agriculture. (2017), S. 3878–3885
- [42] DONG, Jing; BURNHAM, John G.; BOOTS, Byron; RAINS, Glen ; DELLAERT, Frank: 4D crop monitoring: Spatio-temporal reconstruction for agriculture. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, S. 3878 – 3885
- [43] DOS SANTOS NETO, João P.; DE ASSIS, Mateus Wagner D.; CASAGRANDE, Izabella P.; CUNHA JÚNIOR, Luis C. ; DE ALMEIDA TEIXEIRA, Gustavo H.: Determination of ‘Palmer’ mango maturity indices using portable near infrared (VIS-NIR) spectrometer. In: *Postharvest Biology and Technology* 130 (2017), S. 75–80
- [44] DÍAZ, J.F.; CORREA, J.C.: Comparación entre árboles de regresión CART y regresión lineal. In: *Comunicaciones en Estadística* 6 (2013), Nr. 2, S. 175–195
- [45] ELLEN GRICE, Kathy R.; EDWARD BALLY, Ian S.; LOUISE WRIGHT, Carole; MADDOX, Cheryldene; ALI, Asjad ; LOUISE DILLON, Natalie: Mango germplasm screening for the identification of sources of tolerance to anthracnose. In: *Australasian Plant Pathology* 52 (2023), S. 27–41
- [46] ESPINEL, O. F. P.; HERNANDEZ, C. A. V. ; ORTIZ, F. A. P.: Detection of anthracnose in mango leaves by hyperspectral image analysis. In: *2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, 2021, S. 1–6
- [47] EVANS, E. A.; BALLEEN, F. H. ; SIDDIQ, M.: *Mango Production, Global Trade, Consumption Trends, and Postharvest Processing and Nutrition*. John Wiley & Sons, Ltd, 2017. – 1–16 S
- [48] FAZARI, Antonio; PELLICER-VALERO, Oscar J.; GÓMEZ-SANCHIS, Juan; BERNARDI, Bruno; CUBERO, Sergio; BENALIA, Souraya; ZIMBALATTI, Giuseppe ; BLASCO, Jose: Application of deep convolutional neural networks for the detection of anthracnose in olives using VIS/NIR hyperspectral images. In: *Computers and Electronics in Agriculture* 187 (2021), S. 106252
- [49] FOOD AND AGRICULTURE ORGANIZATION, FAO. *Major Tropical Fruits Market Review 2022*. 2023
- [50] FOOD AND AGRICULTURE ORGANIZATION, FAO: *Crops production and trade statistics [online]*. 2024. – Accessed on August 3rd, 2024
- [51] GABRIËLS, Suzan H.; MISHRA, Puneet; MENSINK, Manon G.; SPOELSTRA, Patrick ; WOLTERING, Ernst J.: Non-destructive measurement of internal browning in mangoes using visible and near-infrared spectroscopy supported by artificial neural network analysis. In: *Postharvest Biology and Technology* 166 (2020), S. 111206

- [52] GALÁN-SAÚCO, V.: Trends in world mango production and marketing. In: *Acta Horticulturae* 1183 (2017), S. 351–364
- [53] GARCÍA, J.; ABAUNZA, C. A. ; RIVERA, J. E.: *Modelo productivo para el cultivo de mango en el valle del Alto Magdalena para el departamento del Tolima*. Corpoica, 2017. – 104–119 S
- [54] GAÑÁN, L.; ÁLVAREZ, E. ; CASTAÑO-ZAPATA, J.: Identificación genética de aislamientos de *Colletotrichum* spp. causantes de antracnosis en frutos de aguacate, banana, mango y tomate de árbol. In: *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales* 39 (2015), Nr. 152
- [55] GEBISHU, M.; FIKADU, B.; BEKELE, B.; JULE, L. T.; NAGAPRASAD, N. ; RAMASWAMY, K.: Fluorescence and UV/visible spectroscopic investigation of orange and mango fruit juice quality in case of Adama Town. In: *Scientific Reports* 12 (2022), S. 7345
- [56] GEDEON, T.D.: Data Mining of Inputs: Analysing Magnitude and Functional Measures. In: *International Journal of Neural Systems* 8 (1997), Nr. 2, S. 209–218
- [57] GHAZAL, Sumaira; MUNIR, Arslan ; QURESHI, Waqar S.: Computer vision in smart agriculture and precision farming: Techniques and applications. In: *Artificial Intelligence in Agriculture* 13 (2024), S. 64–83
- [58] GIMENEZ, J.; SANSONI, S.; TOSETTI, S.; CAPRARO, F. ; CARELLI, R.: Trunk detection in tree crops using RGB-D images for structure-based ICM-SLAM. In: *Computers and Electronics in Agriculture* 199 (2022), S. 11. – ISSN 0168–1699
- [59] GÓMEZ-SANCHIS, J.; BLASCO, J.; SORIA-OLIVAS, E.; LORENTE, D.; ESCANDELL-MONTERO, P.; MARTÍNEZ-MARTÍNEZ, J.M.; MARTÍNEZ-SOBER, M. ; ALEIXOS, N.: Hyperspectral LCTF-based system for classification of decay in mandarins caused by *Penicillium digitatum* and *Penicillium italicum* using the most relevant bands and non-linear classifiers. In: *Postharvest Biology and Technology* 82 (2013), S. 76–86
- [60] HAMUDA, Esmael; GLAVIN, Martin ; JONES, Edward: A survey of image processing techniques for plant extraction and segmentation in the field. In: *Computers and Electronics in Agriculture* 125 (2016), S. 184–199
- [61] HERBARIO UNIVERSIDAD DISTRITAL, UDBC. *Guía para la Recolección y Preservación de Muestras Botánicas en Campo*. 2009
- [62] HOLTORF, Tim; KNOLL, Florian J. ; HUSSMANN, Stephan: Multimodal image stitching algorithm for weed control applications in organic farming. In: *2016 SAI Computing Conference (SAI)*, 2016

- [63] HUSSAIN, Abid; PU, Hongbin ; SUN, Da-Wen: Innovative nondestructive imaging techniques for ripening and maturity of fruits – A review of recent applications. In: *Trends in Food Science & Technology* 72 (2018), S. 144–152
- [64] ISHIDA, T.; KURIHARA, J.; VIRAY, F. A.; NAMUCO, S. B.; PARINGIT, E. C.; PEREZ, G. J.; TAKAHASHI, Y. ; MARCIANO, J. J.: A novel approach for vegetation classification using UAV-based hyperspectral imaging. In: *Computers and Electronics in Agriculture* 144 (2018), S. 80–85. – ISSN 0168–1699
- [65] ISHIDA, Tetsuro; KURIHARA, Junichi; VIRAY, Fra A.; NAMUCO, Shiello B.; PARINGIT, Enrico C.; PEREZ, Gay J.; TAKAHASHI, Yukihiro ; MARCIANO, Joel J.: A novel approach for vegetation classification using UAV-based hyperspectral imaging. In: *Computers and Electronics in Agriculture* 144 (2018), S. 80–85
- [66] ISMAIL, A.M.; EL-GANAINY, S.M.: Characterization of Colletotrichum species associating with anthracnose disease of mango in Egypt. In: *Journal of Plant Diseases and Protection* 129 (2022), Nr. 2, S. 449–454
- [67] JAFARI, M.; MINAEI, S. ; SAFAIE, N.: Detection of pre-symptomatic rose powdery-mildew and gray-mold diseases based on thermal vision. In: *Infrared Physics & Technology* 85 (2017), S. 170–183
- [68] JHAUMEER LAULLOO, S.; BHOWON, M. G.; SOYFOO, S. ; CHUA, L. S.: Nutritional and Biological Evaluation of Leaves of Mangifera indica from Mauritius. In: *Journal of Chemistry* 2018 (2018), Nr. 1, S. 6869294
- [69] JIANG, Qiyu; WU, Gangshan; TIAN, Chongfeng; LI, Na; YANG, Huan; BAI, Yuhao ; ZHANG, Baohua: Hyperspectral imaging for early identification of strawberry leaves diseases with machine learning and spectral fingerprint features. In: *Infrared Physics & Technology* 118 (2021), S. 103898
- [70] JIANG, Qiyu; WU, Gangshan; TIAN, Chongfeng; LI, Na; YANG, Huan; BAI, Yuhao ; ZHANG, Baohua: Hyperspectral imaging for early identification of strawberry leaves diseases with machine learning and spectral fingerprint features. In: *Infrared Physics & Technology* 118 (2021), S. 103898
- [71] KABIR, Yearul; SHEKHAR, Hossain U. ; SIDHU, J. S.: *Phytochemical Compounds in Functional Properties of Mangoes*. John Wiley & Sons, Ltd, 2017. – 237–254 S
- [72] KAMILARIS, A.; PRENAFETA-BOLDÚ, F. X.: Deep learning in agriculture: A survey. In: *Computers and electronics in agriculture* 147 (2018), S. 70–90
- [73] KARUNANAYAKE, L.C.; SINNIHAH, G.D.; ADIKARAM, N.K.B. ; ABAYASEKARA, C.L.: How constitutive defences affect the development of postharvest anthracnose and stem-end rot in mango fruit. In: *Acta Horticulturae* 1183 (2017), S. 245–250

- [74] KHAN, M. S.; UANDAI, S. B. ; SRINIVASAN, H.: Anthracnose disease diagnosis by image processing, support vector machine and correlation with pigments. In: *Journal of Plant Pathology* 101 (2019), S. 749–751
- [75] KHAN, Muhammad J.; KHAN, Hamid S.; YOUSAF, Adeel; KHURSHID, Khurram ; ABBAS, Asad: Modern Trends in Hyperspectral Image Analysis: A Review. In: *IEEE Access* 6 (2018), S. 14118–14129
- [76] KHANNA, R.; MÖLLER, M.; PFEIFER, J.; LIEBISCH, F.; WALTER, A. ; SIEGWART, R.: Beyond point clouds - 3D mapping and field parameter measurements using UAVs. In: *2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA)*, 2015, S. 1–4
- [77] KHIRADE, S. D.; PATIL, A. B.: Plant disease detection using image processing. In: *2015 International conference on computing communication control and automation* (2015), S. 768–771
- [78] KOLEY, Chhanda; NIRALA, Anil K.: Activity assessment of anthracnose disease infected region of aonla. In: *AIP Conference Proceedings* Bd. 2352, 2021, S. 030013
- [79] KOTWAL, Jameer; KASHYAP, Dr.Ramgopal ; PATHAN, Dr.Shafi: Agricultural plant diseases identification: From traditional approach to deep learning. In: *Materials Today: Proceedings* 80 (2023), S. 344–356
- [80] KUMAR, M.; SAURABH, V.; TOMAR, M.; HASAN, M.; CHANGAN, S.; SASI, M.; MAHESHWARI, C.; PRAJAPATI, U.; SINGH, S.; PRAJAPAT, R. K.; DHUMAL, S.; PUNIA, S.; AMAROWICZ, R. ; MEKHEMAR, M.: Mango (*Mangifera indica* L.) Leaves: Nutritional Composition, Phytochemical Profile, and Health-Promoting Bioactivities. In: *Antioxidants (Basel, Switzerland)* 10 (2021), Nr. 2, S. 99
- [81] KUMAR, Vinod; SINGH, Ravi S.; RAMBABU, Medara ; DUA, Yaman: Deep learning for hyperspectral image classification: A survey. In: *Computer Science Review* 53 (2024), S. 100658
- [82] KUMARI, Pavitra; SINGH, Rajender: Anthracnose of Mango Incited by *Colletotrichum gloeosporioides*: A Comprehensive Review. In: *International Journal of Pure and Applied Bio-science* 5 (2017), Nr. 1, S. 48–56
- [83] KUMARI, Pavitra; SINGH, Rajender ; PUNIA, Rakesh: Evaluation of Different Inoculation Methods for Mango Anthracnose Disease Development. In: *International Journal of Current Microbiology and Applied Sciences (IJCMAS)* 6 (2017), Nr. 11, S. 3028–3032

- [84] LANGFORD, Z. L.; KUMAR, J. ; HOFFMAN, F. M.: Convolutional neural network approach for mapping arctic vegetation using multi-sensor remote sensing fusion. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (2017), S. 322–331
- [85] LECUN, Y.; BENGIO, Y. ; HINTON, G.: Deep learning. In: *Nature* 521 (2015), S. 436–444
- [86] LEE, David W.; BRAMMEIER, Susan ; SMITH, Alan P.: The Selective Advantages of Anthocyanins in Developing Leaves of Mango and Cacao. In: *Biotropica* 19 (1987), Nr. 1, S. 40–49
- [87] LEE, L.C.; LIONG, C.Y. ; JEMAIN, A.A.: Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. In: *A* 143 (2018), Nr. 15, S. 3526–3539
- [88] LI, Cheng; HE, Mengyu; CAI, Zeyi; QI, Hengnian; ZHANG, Jianhong ; ZHANG, Chu: Hyperspectral Imaging with Machine Learning Approaches for Assessing Soluble Solids Content of Tribute Citru. In: *Foods* 12 (2023), Nr. 2, S. 247
- [89] LI, J.; LIU, Z. ; WANG, D.: A Lightweight Algorithm for Recognizing Pear Leaf Diseases in Natural Scenes Based on an Improved YOLOv5 Deep Learning Model. In: *Agriculture* 14 (2024), Nr. 2, S. 273
- [90] LI, Q.; BU, J.; SHU, J.; YU, Z.; TANG, L.; HUANG, S.; GUO, T.; MO, J.; LUO, S.; SOLANGI, G. S. ; HSIANG, T.: Colletotrichum species associated with mango in southern China. In: *Scientific Reports* 9 (2019), S. 18891
- [91] LI, Xiong; LIU, Yande; JIANG, Xiaogang ; WANG, Guantian: Supervised classification of slightly bruised peaches with respect to the time after bruising by using hyperspectral imaging technology. In: *Infrared Physics & Technology* 113 (2021), S. 103557
- [92] LI, Yuhua; LUO, Zhihui; WANG, Fengjie ; WANG, Yingxu: Hyperspectral Leaf Image-Based Cucumber Disease Recognition Using the Extended Collaborative Representation Model. In: *Sensors* 20 (2020), Nr. 14, S. 4045
- [93] LIU, F.; DAMM, U.; CAI, L. ; CROUS, P. W.: Species of the Colletotrichum gloeosporioides complex associated with anthracnose diseases of Proteaceae. In: *Fungal diversity* 61 (2013), S. 89–105
- [94] LIU, L.; TAO, H.; FANG, J.; ZHENG, W.; WANG, L. ; JIN, X.: Identifying Anthracnose and Black Spot of Pear Leaves on Near-infrared Hyperspectroscopy. In: *Nongye Jixie Xuebao/Transactions of the Chinese Society for Agricultural Machinery* 52 (2022), Nr. 2, S. 221–230

- [95] LIU, Nanfeng; TOWNSEND, Philip A.; NABER, Mack R.; BETHKE, Paul C.; HILLS, William B. ; WANG, Yi: Hyperspectral imagery to monitor crop nutrient status within and across growing seasons. In: *Remote Sensing of Environment* 255 (2021), S. 112303
- [96] LIU, Xuan; WANG, Juan; WANG, Hao; HUANG, Yirui ; REN, Zhenhui: Prediction of prunoideae fruit quality characteristics based on machine learning and spectral characteristic acquisition optimization. In: *Food Control* 165 (2024), S. 110627
- [97] LU, G.; FEI, B.: Medical hyperspectral imaging: a review. In: *Journal of biomedical optics* 19(1) (2014)
- [98] LU, Jinzhu; EHSANI, Reza; SHI, Yeyin; ABDULRIDHA, Jaafar; DE CASTRO, Ana I. ; XU, Yunjun: Field detection of anthracnose crown rot in strawberry using spectroscopy technology. In: *Computers and Electronics in Agriculture* 135 (2017), S. 289–299
- [99] M., Jianyou; Z., Guang; L., Qili; S., Ghulam S.; T., Lihua; G., Tangxun; H., Suiping ; H., Tom: Identification and Characterization of Colletotrichum Species Associated with Mango Anthracnose in Guangxi, China. In: *Plant Disease* 102 (2018), Nr. 7, S. 1283–1289
- [100] MADIWALAR, Shriroop C.; WYAWAHARE, Medha V.: Plant disease identification: A comparative study. In: *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 2017, S. 13–18
- [101] MAHLEIN, A.K.; KUSKA, M.T.; THOMAS, S.; BOHNENKAMP, D.; ALISAAC, E.; BEHMANN, J.; WAHABZADA, M. ; KERSTING, K.: Plant disease detection by hyperspectral imaging: from the lab to the field. In: *Advances in Animal Biosciences* 8 (2017), Nr. 2, S. 238–243
- [102] MAKINO, Y.; ISAMI, A.; SUHARA, T.; OSHITA, S.; KAWAGOE, Y.; TSUKADA, M.; ISHIYAMA, R.; SERIZAWA, M.; PURWANTO, Y.A.; AHMAD, U.; MARDJAN, S. ; KUROKI, S.: Non-destructive analysis of internal and external qualities of mango fruits during storage by hyperspectral imaging. In: *Acta Horticulturae* 1011 (2013), S. 443–449
- [103] MALDONADO-CELIS, M. E.; YAHIA, E.M.; BEDOYA, R.; LANDÁZURI, P.; LOANGO, N.; AGUILLÓN, J.; RESTREPO, B. ; OSPINA, J. C. G.: Chemical Composition of Mango (*Mangifera indica* L.) Fruit: Nutritional and Phytochemical Compounds. In: *Frontiers in plant science* 10 (2019), S. 1073
- [104] MISHRA, Puneet; WOLTERING, Ernst ; EL HARCHIOUI, Najim: Improved prediction of ‘Kent’ mango firmness during ripening by near-infrared spectroscopy supported by interval partial least square regression. In: *Infrared Physics & Technology* 110 (2020), S. 103459

- [105] MO, J.; ZHAO, G.; LI, Q.; SOLANGI, G. S.; TANG, L.; GUO, T.; HUANG, S. ; HSIANG, T.: Identification and Characterization of Colletotrichum Species Associated with Mango Anthracnose in Guangxi. In: *Plant Disease* 102 (2018), Nr. 7, S. 1283–1289
- [106] MONTEIRO, S. T.; KOSUGI, Y.; UTO, K. ; WATANABE, E.: Towards applying hyperspectral imagery as an intraoperative visual aid tool. In: *Proceedings of the 4th IASTED international conference on Visualization, Imaging and Image Processing* (2004), S. 483–488
- [107] MONTEON OJEDA, A.; ISLAS, J.S. S.; MORA AGUILERA, J.A.; DE LEON GARCÍA DE ALBA, C.; RODRÍGUEZ, A. P. ; LOPEZ, A. V.: An alternative inoculation technique of Colletotrichum gloeosporioides on mango for early anthracnose tolerance screening. In: *Agronomía Mesoamericana* 29 (2018), Nr. 2, S. 263–274
- [108] MOSHOU, D.; BRAVO, C.; WEST, J.; WAHLEN, S.; MCCARTNEY, A. ; RAMON, H.: Automatic detection of ‘yellow rust’ in wheat using reflectance measurements and neural networks. In: *Computers and electronics in agriculture* 44(3) (2004), S. 173–188
- [109] MUNAWAR, Agus A.; ZULFAHRIZAL; MEILINA, Hesti ; PAWELZIK, Elke: Near infrared spectroscopy as a fast and non-destructive technique for total acidity prediction of intact mango: Comparison among regression approaches. In: *Computers and Electronics in Agriculture* 193 (2022), S. 106657
- [110] MUNERA, Sandra; BESADA, Cristina; ALEIXOS, Nuria; TALENS, Pau; SALVADOR, Alejandra; SUN, Da-Wen; CUBERO, Sergio ; BLASCO, José: Non-destructive assessment of the internal quality of intact persimmon using colour and VIS/NIR hyperspectral imaging. In: *LWT* 77 (2017), S. 241–248
- [111] MUNERA, Sandra; BESADA, Cristina; BLASCO, José; CUBERO, Sergio; SALVADOR, Alejandra; TALENS, Pau ; ALEIXOS, Nuria: Astringency assessment of persimmon by hyperspectral imaging. In: *Postharvest Biology and Technology* 125 (2017), S. 35–41
- [112] NANSEN, C.; ABIDI, N.; SIDUMO, A. J. ; GHARALARI, A. H.: Using Spatial Structure Analysis of Hyperspectral Imaging Data and Fourier Transformed Infrared Analysis to Determine Bioactivity of Surface Pesticide Treatment. In: *Remote Sensing* 2 (2010), Nr. 4, S. 908–925
- [113] NATIONAL MANGO BOARD: *Mango varieties & Availability [online]*. 2024. – Accessed on August 3rd, 2024
- [114] NAVROZIDIS, I.; ALEXANDRIDIS, T.K.; DIMITRAKOS, A.; LAGOPODI, A.L.; MOSHOU, D. ; ZALIDIS, G.: Identification of purple spot disease on asparagus crops across spatial and spectral scales. In: *Computers and Electronics in Agriculture* 148 (2018), S. 322–329

- [115] N'GUETTIA, M. Y.; KOUASSI, N.; DIALLO, H. A. ; KOUAKOU, F. R. Y. Y.: Evaluation of Anthracnose Disease of Mango (*Mangifera indica* L.) Fruits and Characterization of Causal Agent in Côte d'Ivoire. In: *International Journal of Agriculture Innovations and Research* 2 (2014), Nr. 6, S. 1008–1017
- [116] OKYERE, F. G.; CUDJOE, D.; SADEGHI-TEHRAN, P.; VIRLET, N.; RICHE, A. B.; CASTLE, M.; GRECHE, L.; SIMMS, D.; MHADA, M.; MOHAREB, F. ; HAWKESFORD, M. J.: Modeling the spatial-spectral characteristics of plants for nutrient status identification using hyperspectral data and deep learning methods. In: *Frontiers in Plant Science* 14 (2023)
- [117] OMAR, A.F.; ATAN, H. ; MATJAFRI, M.Z.: Peak Response Identification through Near-Infrared Spectroscopy Analysis on Aqueous Sucrose, Glucose, and Fructose Solution. In: *Spectroscopy Letters* 45 (2012), Nr. 3, S. 190–201
- [118] ORTÍZ, M. A.; VARGAS, M. R.; G., R.; MADINAVEITIA, Chew ; VELÁZQUEZ, J. A. M.: Propiedades funcionales de las antocianinas. In: *Biotechnia* 13 (2011), Nr. 2, S. 16–22
- [119] PANCHE, A. N.; DIWAN, A. D. ; CHANDRA1, S. R.: Flavonoids: an overview. In: *Journal of nutritional science* 5 (2016), S. e47
- [120] PANDEY, A.; PANDEY, B.K.; MUTHUKUMA, M.; YADAVA, L.P. ; CHAUHAN, U.K.: Histopathological Study of Infection Process of *Colletotrichum gloeosporioides* Penz and Sacc. on *Mangifera indica* L. In: *Plant Pathology Journal* 11 (2012), Nr. 1, S. 18–24
- [121] PARRA, Francisco: *Estadística y Machine Learning con R [online]*. 2019. – Accessed on August 3rd, 2024
- [122] PARVEZ, GM M.: Pharmacological Activities of Mango (*Mangifera Indica*): A Review. In: *Journal of Pharmacognosy and Phytochemistry* 5 (2016), Nr. 3, S. 1–7
- [123] PATEL, K.K.; KAR, A. ; KHAN, M.A.: Potential of reflected UV imaging technique for detection of defects on the surface area of mango. In: *Journal of food science and technology* 56 (2019), Nr. 3
- [124] PATIL, P.; YALIGAR, N. ; MEENA, S. M.: Comparison of Performance of Classifiers-SVM, RF and ANN in Potato Blight Disease Detection Using Leaf Images. In: *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (2017), S. 1–5
- [125] POLAK, Adam; KELMAN, Timothy; MURRAY, Paul; MARSHALL, Stephen; STOTHARD, David J.; EASTAUGH, Nicholas ; EASTAUGH, Francis: Hyperspectral imaging combined with data classification techniques as an aid for artwork authentication. In: *Journal of Cultural Heritage* 26 (2017), S. 1–11

- [126] PRABU, M.; CHELLIAH, Balika J.: Mango leaf disease identification and classification using a CNN architecture optimized by crossover-based levy flight distribution algorithm. In: *Neural Computing and Applications* 34 (2022), S. 7311–7324
- [127] PÁEZ REDONDO, A. R. *Tecnologías sostenibles para el manejo de la antracnosis (Colletotrichum gloesporioides (Penz.) Penz. y Sacc.) en papaya (Carica papaya L.) y mango (Mangifera indica L.). Boletín técnico No.7.* 2003
- [128] QIN, L. P.; ZHANG, Y.; SU, Q.; CHEN, Y. L.; NONG, Q.; XIE, L.; YU, G. M. ; HUANG, S. L.: First report of anthracnose of *Mangifera indica* caused by *Colletotrichum scovillei* in China. In: *Plant Disease* 103 (2019), Nr. 5
- [129] RAGHAVENDRA, Anitha; GURU, D.S. ; RAO, Mahesh K.: Mango internal defect detection based on optimal wavelength selection method using NIR spectroscopy. In: *Artificial Intelligence in Agriculture* 5 (2021), S. 43–51
- [130] RAI, Nitin; ZHANG, Yu; RAM, Billy G.; SCHUMACHER, Leon; YELLAVAJJALA, Ravi K.; BAJWA, Sreekala ; SUN, Xin: Applications of deep learning in precision weed management: A review. In: *Computers and Electronics in Agriculture* 206 (2023), S. 107698
- [131] RAMOS, P.J.; PRIETO, F.A.; MONTOYA, E.C. ; OLIVEROS, C.E.: Automatic fruit count on coffee branches using computer vision. In: *Computers and Electronics in Agriculture* 137 (2017), S. 9–22. – ISSN 0168–1699
- [132] RAMÍREZ ALBERTO, L.; CABRERA ARDILA, C. E. ; PRIETO ORTIZ, F. A.: A computer vision system for early detection of anthracnose in sugar mango (*Mangifera indica*) based on UV-A illumination. In: *Information Processing in Agriculture* 10 (2023), Nr. 2, S. 204–215
- [133] RIBEIRO, Sônia Machado R.; SCHIEBER, Andreas ; WATSON, Ronald R. (Hrsg.); PREEDY, Victor R. (Hrsg.): *Chapter 34 - Bioactive Compounds in Mango (Mangifera indica L.)*. Academic Press, 2010. – 507–523 S
- [134] RIZVEE, Redwan A.; ORPA, Tasnim H.; AHNAF, Adil; KABIR, Md A.; AHMMAD RASHID, Mohammad R.; ISLAM, Mohammad M.; ISLAM, Maheen; JABID, Taskeed ; ALI, Md S.: LeafNet: A proficient convolutional neural network for detecting seven prominent mango leaf diseases. In: *Journal of Agriculture and Food Research* 14 (2023), S. 100787
- [135] RUNGPICHAYAPICHET, Parika; CHAIYARATTANACHOTE, Nimmitra; KHUWIJITJARU, Pramote; NAKAGAWA, Kyuya; NAGLE, Marcus; MÜLLER, Joachim ; MAHAYOTHEE, Busarakorn: Comparison of near-infrared spectroscopy and hyperspectral imaging for

- internal quality determination of 'Nam Dok Mai' mango during ripening. In: *Journal of Food Measurement and Characterization* 17 (2023), S. 1501–1514
- [136] RUNGPICHAYAPICHET, Parika; MAHAYOTHEE, Busarakorn; NAGLE, Marcus; KHUWIJITJARU, Pramote ; MÜLLER, Joachim: Robust NIRS models for non-destructive prediction of postharvest fruit ripeness and quality in mango. In: *Postharvest Biology and Technology* 111 (2016), S. 31–40
- [137] RUNGPICHAYAPICHET, Parika; NAGLE, Marcus; YUWANBUN, Pasinee; KHUWIJITJARU, Pramote; MAHAYOTHEE, Busarakorn ; MÜLLER, Joachim: Prediction mapping of physicochemical properties in mango by hyperspectral imaging. In: *Biosystems Engineering* 159 (2017), S. 109–120
- [138] SAHA, Dhritiman; MANICKAVASAGAN, Annamalai: Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. In: *Current Research in Food Science* 4 (2021), S. 28–44
- [139] SALAZAR-CASTRO, R.: El cultivo del mango. In: *ASIAVA* (2003), S. 111–119
- [140] SALEEM, Rabia; SHAH, Jamal Hussain; SHARIF, Muhammad; YASMIN, Mussarat; YONG, Hwan-Seung ; CHA, Jaehyuk: Mango Leaf Disease Recognition and Classification Using Novel Segmentation and Vein Pattern Technique. In: *Applied Sciences* 11 (2021), S. 11901
- [141] SANDERS, G.M.; KORSTEN, L.: Comparison of cross inoculation potential of South African avocado and mango isolates of *Colletotrichum gloeosporioides*. In: *Microbiological Research* 158 (2003), Nr. 2, S. 143–150
- [142] SANGEETHA, C. G.; RAWAL, R. D.: Nutritional studies of *Colletotrichum gloeosporioides* (Penz.)Penz. and Sacc. the incitant of mango anthracnose. In: *American-Eurasian Journal of Sustainable Agriculture* 1 (2007), S. 37–41
- [143] SAWANT, S. S.; MANOHARAN, P.: New framework for hyperspectral band selection using modified wind-driven optimization algorithm. In: *International Journal of Remote Sensing* 40 (2019), Nr. 20, S. 7852–7873
- [144] SERNA-ESCOLANO, Vicente; GIMÉNEZ, María J.; ZAPATA, Pedro J.; CUBERO, Sergio; BLASCO, José ; MUNERA, Sandra: Non-destructive assessment of 'Fino' lemon quality through ripening using NIRS and chemometric analysis. In: *Postharvest Biology and Technology* 212 (2024), S. 112870
- [145] SHAHRAYINI, E.; SHAFIZADEH-MOGHADAM, H.; NOROOZI, A.A. ; EGHBAL, M.K.: Multiple-depth modeling of soil organic carbon using visible–near infrared spectroscopy. In: *Geocarto International* 37 (2022), Nr. 5, S. 1393–1407

- [146] SHARMA, Adikshita; SHARMA, Inder M.; SHARMA, Monica; SHARMA, Kishor ; SHARMA, Amit: Effectiveness of fungal, bacterial and yeast antagonists for management of mango anthracnose (*Colletotrichum gloeosporioides*). In: *Egyptian Journal of Biological Pest Control* 31 (2021), Nr. 135
- [147] SHARMA, Sneha; SIRISOMBOON, Panmanas ; PORNCHALOEMPONG, Pimpen: Application of a Vis-NIR Spectroscopic Technique to Measure the Total Soluble Solids Content of Intact Mangoes in Motion on a Belt Conveyor. In: *The Horticulture Journal* 89 (2020), Nr. 5, S. 545–552
- [148] SINGH, U. P.; CHOUHAN, S. S.; JAIN, S. ; JAIN, S.: Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease. In: *IEEE Access* 7 (2019), S. 43721–43729
- [149] SIRIPATRAWAN, U.; MAKINO, Y.: Early Detection of Anthracnose on Mango Fruit Using Hyperspectral Imaging. In: *Rep. Grant. Res., Asahi Glass Foundation* 90 (2021)
- [150] SIRIPATRAWAN, U.; MAKINO, Y.: Hyperspectral imaging coupled with machine learning for classification of anthracnose infection on mango fruit. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 309 (2024), S. 123825
- [151] SIVANKALYANI, Velu; FEYGENBERG, Oleg; DISKIN, Sonia; WRIGHT, Ben ; ALKAN, Noam: Increased anthocyanin and flavonoids in mango fruit peel are associated with cold and pathogen resistance. In: *Postharvest Biology and Technology* 111 (2016), S. 132–139
- [152] SUBEDI, P.P.; WALSH, K.B.: Assessment of sugar and starch in intact banana and mango fruit by SWNIR spectroscopy. In: *Postharvest Biology and Technology* 62 (2011), Nr. 3, S. 238–245
- [153] SUN, Meijun; ZHANG, Dong; LIU, Li ; WANG, Zheng: How to predict the sugariness and hardness of melons: A near-infrared hyperspectral imaging method. In: *Food Chemistry* 218 (2017), S. 413–421
- [154] SUN, W.; DU, Q.: Hyperspectral Band Selection: A Review. In: *IEEE Geoscience and Remote Sensing Magazine* 7 (2019), Nr. 2, S. 118–139
- [155] SWETHA, K.; VENKATARAMAN, Veena; SADHANA, G.D. ; PRIYATHARSHINI, R.: Hybrid approach for anthracnose detection using intensity and size features. In: *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 2016, S. 28–32
- [156] TOVAR-PEDRAZA, J. M.; MORA-AGUILERA, J. A.; NAVA-DÍAZ, C.; LIMA, N. B.; MICHEREFF, S. J.; SANDOVAL-ISLAS, J. S.; CÂMARA, M. P. S.; TÉLIZ-ORTIZ, D. ;

- LEYVA-MIR, S. G.: Distribution and Pathogenicity of Colletotrichum Species Associated With Mango Anthracnose in Mexico. In: *Plant Disease* 104 (2020), Nr. 1, S. 137–142
- [157] TRIPATHI, M. K.; REDDY, P. K.; NEELAKANTAPPA, M.; ANDHARE, C. V. ; SHIVENDRA, S.: Identification of mango variety using near infrared spectroscopy. In: *Indonesian Journal of Electrical Engineering and Computer Science* 31 (2023), Nr. 3, S. 1776–1783
- [158] UDDIN, M. N.; SHEFAT, S. H. T.; AFROZ, M. ; MOON, N. J.: Management of Anthracnose Disease of Mango Caused by Colletotrichum gloeosporioides: A Review. In: *Acta Scientific Agriculture* 2 (2018), Nr. 10, S. 169–177
- [159] ULYA, Millatul; CHAMIDAH, Nur ; SAIFUDIN, Toha: Mango quality prediction based on near-infrared spectroscopy using multi-predictor local polynomial regression modeling. In: *F1000Research* 12 (2024), Nr. 656, S. 1–24
- [160] VALENCIA FRUITS: *El cultivo de mango sigue en auge [online]*. 2022. – Accessed on August 3rd, 2024
- [161] VÉLEZ-RIVERA, Nayeli; BLASCO, José; CHANONA-PÉREZ, Jorge; CALDERÓN-DOMÍNGUEZ, Georgina; DE JESÚS PEREA-FLORES, María; ARZATE-VÁZQUEZ, Israel; CUBERO, Sergio ; FARRERA-REBOLLO, Reynold: Computer Vision System Applied to Classification of “Manila” Mangoes During Ripening Process. In: *Food and Bioprocess Technology* 7 (2014), S. 1183–1194
- [162] VÉLEZ-RIVERA, Nayeli; GÓMEZ-SANCHIS, Juan; CHANONA-PÉREZ, Jorge; CARRASCO, Juan J.; MILLÁN-GIRALDO, Mónica; LORENTE, Delia; CUBERO, Sergio ; BLASCO, José: Early detection of mechanical damage in mango using NIR hyperspectral images and machine learning. In: *Biosystems Engineering* 122 (2014), S. 91–98
- [163] WANG, F.; WANG, F.; ZHANG, Y.; HU, J.; HUANG, J. ; XIE, J.: Rice Yield Estimation Using Parcel-Level Relative Spectral Variables From UAV-Based Hyperspectral Imagery. In: *Frontiers in Plant Science* 10 (2019)
- [164] WEIR, B.S.; JOHNSTON, P.R. ; DAMM, U.: The Colletotrichum gloeosporioides species complex. In: *Studies in Mycology* 73 (2012), S. 115–180
- [165] WONGSILA, S.; CHANTRASRI, P. ; SUREEPHONG, P.: Machine Learning Algorithm Development for detection of Mango infected by Anthracnose Disease. In: *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, 2021, S. 249–252

- [166] YEH, Yu-Hui; CHUNG, Wei-Chang; LIAO, Jui-Yu; CHUNG, Chia-Lin; KUO, Yan-Fu ; LIN, Ta-Te: Strawberry foliar anthracnose assessment by hyperspectral imaging. In: *Computers and Electronics in Agriculture* 122 (2016), S. 1–9
- [167] YEH, Yu-Hui F.; CHUNG, Wei-Chang; LIAO, Jui-Yu; CHUNG, Chia-Lin; KUO, Yan-Fu ; LIN, Ta-Te: A Comparison of Machine Learning Methods on Hyperspectral Plant Disease Assessments. In: *IFAC Proceedings Volumes* 46 (2013), Nr. 4, S. 361–365
- [168] ZAKARIA, L.: Diversity of Colletotrichum species associated with anthracnose disease in tropical fruit crops—A review. In: *Agriculture* 11 (2021), Nr. 4, S. 297
- [169] ZARCO-TEJADA, P. J.; CAMINO, C.; BECK, P. S. A.; CALDERON, R.; HORNERO, A.; HERNÁNDEZ-CLEMENTE, R.; KATTENBORN, T.; MONTES-BORREGO, M.; SUSCA, L.; MORELLI, M.; GONZALEZ-DUGO, V.; NORTH, P. R. J.; LANDA, B. B.; BOSCIA, D.; SAPONARI, M. ; NAVAS-CORTES, J. A.: Previsual symptoms of Xylella fastidiosa infection revealed in spectral plant-trait alterations. In: *Nature Plants* 4 (2018), S. 432–439
- [170] ZAW, M.M.; SETHA, S. ; NARADISORN, M.: Comparison of hot water and UV-C treatments in controlling postharvest disease and maintaining postharvest quality of 'Nam Dok Mai Si Thong' mango. In: *Acta Horticulturae* 1336 (2022), S. 149–158
- [171] ZHANG, Chu; GUO, Chentong; LIU, Fei; KONG, Wenwen; HE, Yong ; LOU, Binggan: Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine. In: *Journal of Food Engineering* 179 (2016), S. 11–18
- [172] ZHANG, Jingcheng; TIAN, Yangyang; YAN, Lijie; WANG, Bin; WANG, Ling; XU, Junfeng ; WU, Kaihua: Diagnosing the symptoms of sheath blight disease on rice stalk with an in-situ hyperspectral imaging technique. In: *Biosystems Engineering* 209 (2021), S. 94–105
- [173] ZHANG, Peipei; WANG, Huaiwen; Ji, Hongwei; LI, Yankun; ZHANG, Xiaochuan ; WANG, Yanan: Hyperspectral imaging-based early damage degree representation of apple: a method of correlation coefficient. In: *Postharvest Biology and Technology* 199 (2023), S. 112309
- [174] ZHANG, Ruihua; WANG, Meng; LIU, Pengfei; ZHU, Tianyu; QU, Xiaotian; CHEN, Xujun ; XIAO, Xinqing: Flexible Vis/NIR sensing system for banana chilling injury. In: *Postharvest Biology and Technology* 207 (2024), S. 112623
- [175] ZHAO, Xiaohu; ZHANG, Jingcheng; HUANG, Yanbo; TIAN, Yangyang ; YUAN, Lin: Detection and discrimination of disease and insect stress of tea plants using hyperspectral imaging combined with wavelet analysis. In: *Computers and Electronics in Agriculture* 193 (2022), S. 106717

-
- [176] ZHENG, Z.; LIU, Y.; HE, M.; CHEN, D.; SUN, L. ; ZHU, F.: Effective band selection of hyperspectral image by an attention mechanism-based convolutional network. In: *RSC Advances* 12 (2022), Nr. 14, S. 8750–8759
- [177] ZHU, Wenjing; CHEN, Hua; CIECHANOWSKA, Izabela ; SPANER, Dean: Application of infrared thermal imaging for the rapid diagnosis of crop disease. In: *IFAC-PapersOnLine* 51 (2018), Nr. 17, S. 424–430
- [178] ZOU, Lin; LI, Huijun; DING, Xuejie; LIU, Zifan; HE, Dongqiong; KOWAH, Jamal A. H.; WANG, Lisheng; YUAN, Mingqing ; LIU, Xu: A Review of The Application of Spectroscopy to Flavonoids from Medicine and Food Homology Materials. In: *Molecules (Basel, Switzerland)* 27 (2022), Nr. 22, S. 7766