

---

A REMOTE SENSING-MACHINE LEARNING FRAMEWORK FOR  
MODELLING FOREST HEALTH

By NITESH KESHAVELAL POONA

---

Dissertation presented for the degree of Doctor of Philosophy (in  
Geoinformatics) in the Faculty of Science at Stellenbosch University.



Supervisor: Prof A van Niekerk

Co-supervisor: Dr R Ismail

December 2020

## **DECLARATION 1: PLAGIARISM**

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 14 January 2020

.....

Copyright © 2020 Stellenbosch University

All rights reserved

## DECLARATION 2: PUBLICATIONS

The research presented in this dissertation includes original work, published in international peer-reviewed journals and international conference proceedings. The preparation of all manuscripts was the principal responsibility of the lead author, Nitesh K. Poona, under the supervision of Dr Riyad Ismail and Prof Adriaan van Niekerk.

1. Poona NK & Ismail R 2019. Developing optimised spectral indices using machine learning to model *Fusarium circinatum* stress in *Pinus radiata* seedlings. *Journal of Applied Remote Sensing* 13, 034515. doi:10.1117/1.JRS.13.034515.
2. Poona NK, Van Niekerk A & Ismail R 2016. Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data. *Sensors* 16, 1918. doi:10.3390/s16111918.
3. Poona NK, Van Niekerk A, Nadel RL & Ismail R 2016. Random forest (RF) wrappers for waveband selection and classification of hyperspectral data. *Applied Spectroscopy* 70, 322-333.
4. Poona NK & Ismail R 2014. Using Boruta-selected spectroscopic bands for the asymptomatic detection of *Fusarium circinatum* stress. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 70, 3764-3772.
5. Poona NK & Ismail R 2013. Reducing hyperspectral data dimensionality using random forest based wrappers. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS2013) held 21-26 July 2013, Melbourne, Australia.
6. Poona NK & Ismail R 2013. Discriminating the occurrence of pitch canker fungus in *Pinus radiata* trees using QuickBird imagery and artificial neural networks. *Southern Forests: a Journal of Forest Science* 75, 29-40.
7. Poona NK & Ismail R 2012. Discriminating the early stages of *Fusarium circinatum* infection of *Pinus radiata* seedlings using high spectral resolution data. Proceedings of the 9<sup>th</sup> International Conference of the African Association of Remote Sensing and the Environment (AARSE2012) held 29 October-2 November 2012, El Jadida, Morocco.
8. Poona NK & Ismail R 2012. Discriminating the occurrence of pitch canker infection in *Pinus radiata* forests using high spatial resolution QuickBird data and artificial neural networks. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS2012) held 22-27 July 2012, Munich, Germany.

## SUMMARY

The utility of remote sensing data, in particular high dimensional spectroscopy data, is now widely used for the detection and monitoring of pest and disease in agriculture and forestry. Coupled with advanced data analytics, spectroscopic data can provide a wealth of information regarding vegetation health, and successfully demonstrates the utility of spectroscopic data and advanced machine learning (ML) algorithms, i.e. tree-based ensemble learners, by developing a remote sensing-machine learning framework for forest health assessment and monitoring. Specifically, the research investigates the use of spectroscopic data for modelling *Fusarium circinatum* stress in *Pinus radiata* and *Pinus patula*.

The research first investigated the utility of novel wrapper feature selection algorithms embedded with the random forest (RF) learner to develop classification models for discriminating healthy, infected, and damaged *P. radiata* and *P. patula* seedlings within a nursery environment. Results showed that reducing data dimensionality results in improved model accuracies. More importantly, the results showed that the RF-Boruta framework yielded the best results.

Two RF variants were subsequently explored, namely oblique random forest (oRF), and rotation forest (rotF). The performances of oRF and rotF were benchmarked against those of traditional RF. All models were evaluated in terms of their ability to discriminate healthy and stressed *Pinus* seedlings. Spectral resampling was employed to reduce data dimensionality. The oRF model yielded the best results, with oRFsvm (oRF employing support vector machine as splitting model) proving to be the most robust.

To extend the utility of model building, the research developed normalised difference two-band spectral indices for real-time *F. circinatum* stress detection. The Boruta algorithm was employed to identify relevant bands, which were used to derive two-band indices. The indices were compared with an extensive list of currently available indices, identified from the literature, to assess the value thereof. Indices were evaluated within univariate and multivariate paradigms, with the latter proving more adept at classifying healthy, damaged, and infected seedlings.

The use of high spatial resolution satellite remote sensing imagery for modelling pitch canker in *P. radiata* trees in a commercial plantation was also evaluated. This exploration served to complement the remote sensing-machine learning framework developed for the nursery environment. In this component of the research, an artificial neural network model was used (whereas tree-based ensemble models were used in the former elements of the research). Results highlight the potential of using high spatial resolution satellite remote sensing for mapping and monitoring of pitch canker infected trees.

Overall, the research successfully demonstrated that high spectral and high spatial resolution remotely sensed data, coupled with advanced data analytics, i.e. tree-based ensemble learners and wrapper algorithms, provides a potentially operational and economically viable framework for *F. circinatum* management within a nursery and plantation environment.

**Keywords:** Hyperspectral remote sensing, spectroscopy, *Pinus*, random forest, Boruta

## OPSOMMING

Afstandwaarnemingsdata, veral hoë-dimensionele spektroskopiedata, word gereeld gebruik vir die opsporing en monitering van plae en siektes in die landbou- en bosbousektor. Tesame met gevorderde data-analise, kan spektroskopiese data 'n magdom inligting verskaf oor die toestand van plantegroei. Spektroskopiese data en gevorderde masjienleer-algoritmes, of altans boom-gebaseerde ensemble-leerders, benut kan word vir die ontwikkeling van 'n raamwerk om die gezondheidstoestand van die bos te assesser en te monitor. Hierdie navorsing ondersoek spesifiek die gebruik van spektroskopiese data vir die modellering van *Fusarium circinatum* in *Pinus radiata* en *Pinus patula*.

Die navorsing het die nut van die seleksie-algoritmes ondersoek deur nuwe wikkelfunksies by die ewekansige woud (RF) algoritme te inkorporeer om klassifikasie Modelle te ontwikkel wat gesonde, besmette en beskadigde *P. radiata*- en *P. patula* saailinge binne 'n kwekeryomgewing onderskei het. Resultate het getoon dat die vermindering van datadimensionaliteit na 'n hoër akkuraatheid van die model toe lei. Die resultate het ook getoon dat die RF-Boruta-raamwerk die beste resultate gelewer het.

Daarna is twee RF-variante ondersoek, naamlik skuins ewekansige woud (oRF), en rotasiewoud (rotF). Die prestasie van oRF en rotF is vergelyk met die van tradisionele RF. Al die Modelle is beoordeel aan die hand van hul vermoë om gesonde en beskadigde Pinus-saailinge te onderskei. Spektrale herversameling is gebruik om die dimensionaliteit van die data te verminder. Die oRF model het die beste resultate gelewer, met oRFsvm (wat ondersteunings-vektor masjien as verdelingsmodel gebruik) wat die sterkste was.

Om die bruikbaarheid van modelbou uit te brei, het die navorsing genormaliseerde verskil tweekbandspektrale indekse ontwikkel om *F. circinatum* stres intyds op te spoor. Die Boruta-algoritme is gebruik om relevante bande te identifiseer en dan om tweekbandindekse af te lei. Die indekse is vergelyk met indekse wat uit die literatuur geïdentifiseer is om die waarde daarvan te beoordeel. Indekse is beoordeel binne eenveranderlike en meerveranderlike paradigmas, en laasgenoemde het gesonde, beskadigde en besmette saailinge beter klassifiseer.

Die gebruik van satellietafstandswaarnemingsbeelde met hoë ruimtelike resolusie vir die modellering van kanker in *P. radiata* bome in 'n kommersiële plantasie is ook ondersoek. Hierdie afdeling van die navorsing het as aanvulling tot die afstandswaarneming-masjienleer raamwerk wat vir die kwekery omgewing ontwikkel is gedien. In hierdie komponent van die navorsing is 'n kunsmatige neurale netwerkmodel gebruik (terwyl boom-gebaseerde ensemble-Modelle in die vorige elemente van die navorsing gebruik is). Resultate beklemtoon die potensiaal van die gebruik van

satellietafstandswaarnemingsdata met 'n hoë ruimtelike resolusie vir die kartering en monitering van besmette bome.

Hierdie navorsing het getoon dat hoë-spektrale en hoë ruimtelike resolusie afstandswaarnemingsdata, tesame met gevorderde data-analise (boomgebaseerde ensemble-leerders en wikkalgoritmes), 'n ekonomiese uitvoerbare bedryfsraamwerk bied vir die bestuur van *F. circinatum* in kwekery- en plantasie-omgewings.

**Sleutelwoorde:** hyperspectral remote sensing, spectroscopy, *Pinus*, random forest, Boruta

## ACKNOWLEDGEMENTS

My heartfelt thanks to:

1. My late parents, for their love;
2. My mentor, the late Professor Doug Rawlings (aka DER), for providing perspective, scientific merit, and guidance;
3. Professor Urmilla Bob for being my mentor, colleague, and friend;
4. Professor Adriaan van Niekerk, for his instruction, guidance, and patience;
5. Dr Riyad Ismail, who has walked the path with me as friend, mentor, instructor, and guide;
6. Professor Altus Viljoen and Dr Diane Mostert for providing and preparing the *F. circinatum* inoculum;
7. Dr Hannel Ham and colleagues for assisting with *P. radiata* seedling inoculation and greenhouse monitoring during the experiments;
8. Institute for Commercial Forestry Research (ICFR) researchers for access to the *P. patula* seedling trials;
9. MTO Forestry and Sappi Forests for access to their plantations and nurseries, and provision of seedlings and enumerator data;
10. The National Research Foundation – German Academic Exchange Service (NRF-DAAD) for financial support;
11. Stellenbosch University for providing the necessary financial and time resources;
12. My sister, Sobhna, for language and grammar editing, and general comments regarding the quality of my manuscripts;
13. Colleagues, friends, and my postgraduate students who have influenced both my academic and personal life over the years.

I have been blessed.



## CONTENTS

DECLARATION 1: PLAGIARISM.....	ii
DECLARATION 2: PUBLICATIONS .....	iii
SUMMARY .....	iv
OPSOMMING .....	vi
ACKNOWLEDGEMENTS .....	viii
CONTENTS.....	ix
TABLES.....	xiii
FIGURES .....	xv
ACRONYMS AND ABBREVIATIONS .....	xviii
CHAPTER 1: GENERAL INTRODUCTION .....	1
1.1 BACKGROUND.....	1
1.1.1 Application.....	1
1.1.2 Remote sensing and machine learning.....	2
1.2 PROBLEM STATEMENT .....	5
1.3 SIGNIFICANCE & RATIONALE FOR UNDERTAKING THE RESEARCH.....	5
1.4 RESEARCH QUESTIONS .....	5
1.5 AIM & OBJECTIVES.....	6
1.6 RESEARCH METHODOLOGY .....	6
1.7 OUTLINE OF DISSERTATION .....	7
CHAPTER 2: RANDOM FOREST CLASSIFICATION OF HYPERSPECTRAL DATA.....	8
2.1 INTRODUCTION.....	9
2.2 RANDOM FOREST .....	11
2.2.1 Feature subset optimisation.....	11
2.2.2 RF sensitivity .....	17
2.2.3 Unsupervised RF.....	19
2.2.4 RF variants .....	20
2.2.5 RF implementations .....	22
2.2.6 Benchmarking RF .....	23
2.3 CONCLUSIONS .....	24
CHAPTER 3: SELECTING THE MOST IMPORTANT BANDS FOR MODELLING <i>FUSARIUM</i> STRESS USING BORUTA .....	26
3.1 INTRODUCTION.....	27

3.2	MATERIALS AND METHODS .....	30
3.2.1	Seedling inoculation.....	30
3.2.2	Spectral data acquisition and pre-processing .....	30
3.2.3	Classification using random forest.....	31
3.2.4	Feature selection with Boruta .....	32
3.2.5	Classification accuracy .....	32
3.3	RESULTS.....	33
3.3.1	Analysis of waveband importance and classification accuracy using random forest ..	33
3.3.2	Waveband selection and classification using the Boruta algorithm .....	34
3.4	DISCUSSION .....	35
3.4.1	Disease progression in <i>P. radiata</i> seedlings .....	38
3.4.2	Waveband selection and classification using the Boruta algorithm .....	38
3.4.3	Asymptomatic detection of <i>F. circinatum</i> infection in <i>P. radiata</i> seedlings.....	40
3.5	CONCLUSIONS .....	42
CHAPTER 4: SELECTING THE MOST IMPORTANT BANDS FOR MODELLING <i>FUSARIUM</i> STRESS: COMPARING RANDOM FOREST WRAPPERS.....		43
4.1	INTRODUCTION.....	44
4.2	MATERIALS AND METHODS .....	46
4.2.1	Symptoms of <i>F. circinatum</i> infection .....	46
4.2.2	Seedling inoculation.....	47
4.2.3	Spectral data acquisition .....	48
4.2.4	Statistical analysis .....	48
4.2.5	Classification accuracy .....	51
4.3	RESULTS.....	52
4.3.1	Waveband selection and classification of <i>Pinus</i> seedlings .....	52
4.3.2	Classification of <i>Pinus</i> seedlings using a combination of bands .....	55
4.4	DISCUSSION .....	59
4.4.1	Hyperspectral dimensionality reduction and classification accuracies.....	60
4.4.2	Waveband selection using the Boruta algorithm .....	61
4.4.3	Classification using a hybrid selection of bands .....	62
4.5	CONCLUSIONS .....	63
CHAPTER 5: OBLIQUE TREE-BASED MODELS FOR DISCRIMINATING <i>FUSARIUM</i> STRESS		64
5.1	INTRODUCTION.....	65

5.2	MATERIALS AND METHODS .....	67
5.2.1	<i>F. circinatum</i> .....	67
5.2.2	Seedling inoculation.....	68
5.2.3	Spectroscopic data acquisition .....	68
5.2.4	Tree-based ensembles .....	70
5.2.5	Spectral resampling.....	72
5.2.6	Classification accuracy .....	72
5.3	RESULTS.....	73
5.4	DISCUSSION .....	79
5.4.1	Classification using all bands.....	79
5.4.2	The effect of spectral resampling on classifier performance .....	80
5.4.3	Robustness of the oblique forest ensembles .....	82
5.5	CONCLUSIONS .....	83
CHAPTER 6: OPTIMISED TWO-BAND NORMALISED DIFFERENCE SPECTRAL INDICES FOR MODELLING <i>FUSARIUM</i> STRESS .....		84
6.1	INTRODUCTION.....	85
6.2	MATERIALS AND METHODS .....	88
6.2.1	Spectral data collection and pre-processing.....	88
6.2.2	Experimental design.....	89
6.2.3	Selection of existing spectral indices .....	90
6.2.4	Development of optimised spectral indices (SIs <sub>opt</sub> ).....	90
6.2.5	Evaluating the existing and optimised indices.....	97
6.3	RESULTS.....	97
6.3.1	Boruta band selection.....	97
6.3.2	Evaluating the indices .....	98
6.3.3	Model validation .....	99
6.4	DISCUSSION .....	100
6.4.1	Boruta band selection.....	100
6.4.2	Performance of existing versus optimised spectral indices .....	101
6.5	CONCLUSIONS .....	102
CHAPTER 7: MODELLING PITCH CANKER USING HIGH SPATIAL RESOLUTION SATELLITE IMAGERY .....		104
7.1	INTRODUCTION.....	105
7.2	MATERIALS AND METHODS .....	108

7.2.1	Site description.....	108
7.2.2	Detecting pine pitch canker in the field .....	108
7.2.3	Field and image data .....	109
7.2.4	Crown-level assessment.....	111
7.2.5	Signature extraction .....	112
7.2.6	Vegetation indices and transformations.....	112
7.2.7	Neural network model.....	113
7.3	RESULTS.....	115
7.3.1	Crown-level assessment.....	115
7.3.2	Neural network model.....	116
7.4	DISCUSSION .....	119
7.4.1	Benefits of crown-level assessments.....	120
7.4.2	Neural network model.....	120
7.5	CONCLUSIONS .....	122
CHAPTER 8: REMOTE SENSING OF FOREST HEALTH: A SYNTHESIS .....		123
8.1	SUMMARY .....	123
8.2	SCIENTIFIC MERITS OF THE RESEARCH.....	123
8.3	REVISITING THE AIM AND OBJECTIVES.....	125
8.4	STRENGTHS AND LIMITATIONS OF THE METHODOLOGY, AND ASUMPTIONS MADE.....	126
8.5	CONCLUSIONS .....	127
8.6	RECOMMENDATIONS AND FUTURE STUDY.....	128
REFERENCES.....		132

## TABLES

Table 1.1: Wavelength ranges typically used in hyperspectral remote sensing studies. ....	2
Table 1.2: Recent applications of hyperspectral remote sensing for forest health assessment. ....	4
Table 2.1: RF variants implemented in R statistical software. ....	23
Table 3.1: Random forest classification results using all bands ( $n = 1769$ ) for the combined classes. The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT. ....	33
Table 3.2: Sensitivity of the Boruta algorithm to increasing number of trees in the forest ( $n_{tree}$ ). The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT. ....	35
Table 3.3: Waveband selection and classification using Boruta-selected bands for the combined classes. The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT. ....	36
Table 3.4: Waveband selection and classification using Boruta-selected bands for the healthy-infected (H-I), and infected-damaged (I-D) class pairs. The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT. ....	37
Table 4.1: Waveband selection and classification using Boruta, recursive feature elimination (RFE), and area under the receiver operating characteristic curve of the random forest (AUC-RF)-selected bands. The three measures of classification accuracy include the out-of-bag (OOB) error, independent test error, and the area under the receiver operating characteristic curve (AUC). ....	56
Table 4.2: Classification accuracies using common subsets of bands selected by combining (i) Boruta and recursive feature elimination (RFE), (ii) Boruta and area under the receiver operating characteristic curve of the random forest (AUC-RF), and (iii) recursive feature elimination (RFE) and area under the receiver operating characteristic curve of the random forest (AUC-RF). The three measures of classification accuracy include the out-of-bag (OOB) error, independent test error, and the area under the receiver operating characteristic curve (AUC). ....	57
Table 5.1: Spectral resampled wavelengths and the associated classification results using the five ensemble classifiers (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model). KHAT values are indicated in parentheses. ....	79
Table 6.1: Existing narrowband spectral indices employed in this study. ....	92
Table 6.2: Resampled Boruta-selected bands for the H-I and I-D class pairs. ....	98
Table 6.3: Results of the univariate and multivariate analyses for the healthy-infected (H-I) and infected-damaged (I-D) class pairs. For the univariate analysis, the best performing index (existing) and bands (optimised) are indicated in parenthesis. For the multivariate analysis, the out-of-bag error is shown, and the number of selected indices indicated in parenthesis. ....	99
Table 6.4: Multivariate results for the healthy-infected (H-I) and infected-damaged (I-D) class pairs for week 1 and week 2, using optimised indices. ....	100

Table 7.1: The two MTO forest compartments selected for the study. ....	109
Table 7.2: The five vegetation indices and tasseled cap transformations used in this study. ....	113
Table 7.3: Classification accuracy for the multilayer perceptron model using the test dataset ( $n = 136$ ). .....	119
Table 8.1: Operational and future spaceborne imaging spectrometer missions. ....	130

## FIGURES

- Figure 1.1: Reflectance spectrum of a healthy *P. radiata* seedling indicating the red-edge region. ....2
- Figure 1.2: Interconnectedness of dissertation chapters (experiments). .....7
- Figure 2.1: Citations of Breiman (2001) from 2001 to 2018. Data source: Scopus Metrics. ....11
- Figure 3.1: Symptom development of *F. circinatum* infection in *P. radiata* seedlings. A healthy seedling (a) is shown as a reference. Symptom expression became more prominent with time, ultimately leading to seedling death. ....27
- Figure 3.2: Waveband importance as determined by the random forest algorithm using optimised *mtry* and *ntree* values for the combined classes. Waveband importance is indicated by the grey bars. The arrow indicates those bands with the highest mean decrease in accuracy. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference. ....36
- Figure 3.3: Boruta-selected bands for the combined classes. The grey bars indicate the most relevant bands selected by Boruta. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference. ....37
- Figure 3.4: Boruta-selected bands for the healthy-infected (H-I) class pair (a) and the infected-damaged (I-D) class pair (b). The grey bars indicate the most relevant bands selected by the Boruta algorithm. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference. ....38
- Figure 3.5: Comparing random forest waveband importance and Boruta waveband selection for the combined classes. The black bars represent the mean decrease in accuracy for random forest, and the grey bars represent bands selected using Boruta. The X-axis is indicative of waveband importance (in the case of RF) / relevance (in the case of Boruta). ....41
- Figure 4.1: Initial symptoms associated with *F. circinatum* infection within a nursery environment. A healthy *P. patula* seedling (a) is shown as a reference. Images courtesy of Institute for Commercial Forestry Research, Pietermaritzburg, South Africa. ....47
- Figure 4.2: Boruta-selected bands, recursive feature elimination-selected bands, and area under the receiver operating characteristic curve of the random forest-selected bands for *P. radiata*. The grey bars indicate the most relevant bands selected by Boruta, recursive feature elimination, and area under the receiver operating characteristic curve of the random forest, respectively. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference. ....54
- Figure 4.3: Boruta-selected bands, recursive feature elimination-selected bands, and area under the receiver operating characteristic curve of the random forest-selected bands for *P. patula*. The grey bars indicate the most relevant bands selected by Boruta, recursive feature elimination, and area under the receiver operating characteristic curve of the random forest respectively. The spectral curve represents the mean signature of a healthy *P. patula* seedling, and is used as a reference. ....54
- Figure 5.1: Experimental setup of the spectroradiometer used for spectral data collection (a) showing the orientation (nadir view) of the pistol relative to the seedling (b). ....69
- Figure 5.2: Mean spectral signature of the healthy ( $n = 50$ ) and infected ( $n = 50$ ) classes. The Healthy (sd) and Infected (sd) signatures represent the 1-sigma standard deviation for the healthy (pink shade) and infected (blue shade) signatures respectively. ....69

Figure 5.3: Visualisation of the decision boundary for (a) RF; (b) oRFridge; (c) oRFpls; and (d) oRFsvm. The margin between the grey and coral areas represents the decision boundary learned. The dots and triangles represent the two classes, i.e., healthy and infected. RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model. .... 73

Figure 5.4: Mean classification accuracies for all tree-based algorithms (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model) considered in this study. The scaled accuracy is the classification accuracy represented on a scale ranging from zero to one. Vertical bars denote 0.95 confidence intervals. .... 74

Figure 5.5: The distribution of the classification accuracy based on the test dataset for all tree-based algorithms (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model) considered in this study. Each boxplot represents the results obtained from 100 repetitions and all bands ( $n = 1769$ ). The scaled accuracy is the classification accuracy represented on a scale ranging from zero to one. .... 75

Figure 5.6: Resampling of the original hyperspectral dataset. Subsets of bands ranged in size from  $n = 884$  (spectral resampling to 2 nm) to  $n = 10$  (spectral resampling to 176 nm). The X-axis represents the wavelength (nm) of the resampled bands whereas the Y-axis represents the reflectance (%). .... 76

Figure 5.7: Comparison of the mean accuracies obtained using all bands and using the resampled bands for the five ensemble classifiers (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model). The red line indicates the mean accuracy obtained using all the original bands ( $n = 1769$ ) whereas the blue bars indicate the mean accuracies for the respective resampled subsets. .... 77

Figure 5.8: Mean classification accuracies using resampled hyperspectral bands ( $n = 800$ ) for each of the tree-based algorithms (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest with ridge regression as splitting model; oRFpls = oblique random forest with PLS as splitting model; oRFsvm = oblique random forest with SVM as splitting model) considered in this study. The scaled accuracy is the classification accuracy represented on a scale ranging from zero to one. Vertical bars denote 0.95 confidence intervals. .... 78

Figure 6.1: Spectral data acquisition using the FieldSpec® Pro Spectroradiometer. .... 89

Figure 6.2: Data analysis workflow employed in this study. .... 90

Figure 6.3: Boruta band selection (grey bars) for the H-I class pair (a) and the I-D class pair (b) for week 3. A mean spectral signature of a healthy (a) and infected (b) is shown for reference. .... 98



Figure 7.1: Tokai Plantation situated in the Western Cape, South Africa. The two compartments (C6c and C17b) selected for the study are indicated. Background image consists of the QuickBird panchromatic band (0.6 m).....	109
Figure 7.2: <i>P. radiata</i> showing signs of pitch canker disease. Infected trees express varied stages of infection, ranging from flagging to extensive canopy dieback, with many of the trees exhibiting advanced stage of infection. ....	110
Figure 7.3: Mean signature values of the infected ( $n = 200$ ), and healthy ( $n = 200$ ) crowns derived from the four QuickBird spectral bands. ....	112
Figure 7.4: Automatically delineated tree crowns in compartment C17b ( $n = 3\ 139$ ) selected on the basis of correspondence between automated isolations (isols) and the manual delineations (greds). Background image consists of the QuickBird panchromatic band. ....	116
Figure 7.5: Neural network topology used in this study. The input layer consisting of the mean crown values extracted from the four QuickBird multispectral bands (B1-B4), five vegetation indices (VI1-VI5), and three tasseled cap components (TC1-TC3) are connected to a single hidden layer with seven hidden layer nodes, which are in turn connected to an output layer with two output nodes representing the crown status.....	117
Figure 7.6: Relationship between learning rate, training time, momentum factor (MF), and prediction accuracy for discriminating healthy and infected crowns. Training time was varied at (a) 500, (b) 1 000, (c) 5 000, (d) 10 000, and (e) 20 000 epochs. ....	118
Figure 7.7: Predicted distribution of healthy and infected crowns for compartment C6c. The background image consists of the QuickBird panchromatic band. ....	120

## ACRONYMS AND ABBREVIATIONS

ACE	artificial contrasts with ensembles
AISA	Airborne Imaging Spectrometer for Applications
ANN	artificial neural network
ANOVA	analysis of variance
ASD	Analytical Spectral Devices, Inc.
AUC-RF	area under the receiver operating characteristic curve of the random forest
AVIRIS	Airborne Visible Infrared Imaging Spectrometer
BRF	boosted random forest
CART	classification and regression trees
CASI	Compact Airborne Spectrographic Imager
CHRIS	Compact High Resolution Imaging Spectrometer
CRF	cascaded random forest
CV	cross-validation
DAIS	Digital Airborne Imaging Spectrometer
DT	decision tree
ELM	extreme learning machine
EMS	electromagnetic spectrum
ERF	enriched random forests
ESA	European Space Agency
FWHM	full width at half maximum
GA	genetic algorithm
GML	Gaussian maximum likelihood
GML-LOOC	Gaussian maximum likelihood with leave-one-out-covariance
GRRF	guided regularised random forest
HyMap	Hyperspectral Mapper
ITC	individual tree crown
J-M	Jeffries-Matusita
k-NN	k-nearest neighbour
LAI	leaf area index
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminant analysis
ML	machine learning

MLP	multilayer perceptron
NASA	National Aeronautics and Space Administration
NB	Naïve Bayes
NIR	near infrared
NN	neural network
OA	overall accuracy
OBIA	object-based image analysis
OOB	out-of-bag
oRF	oblique random forest
PCF	pitch canker fungus
PCA	principal components analysis
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
REP	red-edge position
RF	random forest
RFE	recursive feature elimination
RJ	random jungle
RMSE	root mean square error
rotF	rotation forest
ROSIS	Reflective Optics System Imaging Spectrometer
RRF	regularised random forest
SAM	spectral angle mapper
SFFS	sequential forward feature selection
SI	spectral index / indices
SIMCA	soft independent modelling of class analogy
S/N	signal to noise ratio
SNP	single nucleotide polymorphism
SVM	support vector machine
SWIR	shortwave infrared
TCT	tasseled cap transformation
UAV	unmanned aerial vehicle
URF	unsupervised random forest
VI	vegetation index / indices
VIM	variable importance measure

VIS	visible
VNIR	visible-near infrared
XGBoost	extreme gradient boosting

“There are no secrets to success. It is the result of preparation, hard work, and learning from failure.” Colin Powell

## CHAPTER 1: GENERAL INTRODUCTION

### 1.1 BACKGROUND

#### 1.1.1 Application

The fungus, *Fusarium circinatum* (teleomorph = *Gibberella circinata*) (formerly *Fusarium subglutinans* f. sp. *pini*) is the causal agent of pitch canker of pine trees. The pathogen infects only *Pinus* spp., with more than 30 species of pine being susceptible (EMPPPO 2005). The most susceptible of the *Pinus* spp. is *Pinus radiata* (Monterey pine) (Gordon et al. 2001; Aegerter et al. 2003), a native species to the Monterey region of California, United States of America (Schweisinger 2008). Pitch canker was first described in North Carolina in 1945 (Hepting & Roth 1946), and first detected in the Monterey region in 1992 (Deghi et al. 1994). In South Africa, the fungus was first reported at a Mpumalanga nursery in 1990 (Viljoen et al. 1994) and has since become endemic in nurseries across the country (Porter et al. 2009). Coutinho et al. (2007) and Wingfield et al. (2008) provide a synopsis of the recorded global occurrence of *F. circinatum*, where the pathogen also infects native and commercial stands of several other species of *Pinus*.

Pitch canker is known to spread rapidly and is difficult to control (Schweisinger 2008). Vectors of the disease include birds, wind, rainsplash, and movement of infected plant materials (Viljoen et al. 1997; Gordon et al. 2001; Schweisinger 2008). Additionally, Wingfield et al. (2008) note that insects, in particular *Pissodes nemorensis* (Gebeyehu & Wingfield 2003) is the most significant vector of the disease. Although reported as a vector and wounding agent, *P. nemorensis* is regarded a secondary pest, infesting already stressed trees (Gebeyehu & Wingfield 2003), for example, infesting pine trees stressed due to *F. circinatum* infection.

*F. circinatum* infects the vegetative and reproductive parts of susceptible hosts (Wingfield et al. 2008; Dreaden & Smith 2010). Infection can occur throughout the year, but is limited by the prevailing environmental conditions. Several environmental variables (Gordon 2006; Wingfield et al. 2008) including weather-related injuries, soil and foliar nutrient levels, host susceptibility, planting site characteristics (for example waterlogging and high stand densities), air pollution, temperature (optimum fungal growth at 25°C; optimal spore germination at 20°C), and humidity influence the incidence, establishment, and severity of pitch canker.

Although symptoms of pitch canker are typically expressed on mature trees, the disease affects adolescent trees as well as seedlings (Aegerter et al. 2003; Coutinho et al. 2007; Wingfield et al. 2008). Additionally, infection of branches and stems occur at any age (EPPO 2005). Aegerter et al. (2003) and EPPO (2005) note that symptoms of *F. circinatum* infection are not clearly expressed in

seedlings; this most likely being the cause of pathogenesis in nurseries (Aegerter et al. 2003). Susceptible hosts can, however, become infected during any stage of the tree's life-cycle, as highlighted by Dreaden & Smith (2010).

### 1.1.2 Remote sensing and machine learning

Remote sensing provides an efficient, cost-effective, non-contact, non-destructive, approach to acquiring data relating to a target. Traditional sensors capture data in broad discrete bands across the electromagnetic spectrum (EMS), providing multispectral datasets (Goetz 2009; Mutanga et al. 2009). Advances in sensor design have resulted in the ability to collect data in narrow, discrete contiguous bands across the visible (VIS), near infrared (NIR), and shortwave infrared (SWIR) spectrum, typically from 400 nm to 2 500 nm (Table 1.1). The resulting hyperspectral (i.e. high spectral resolution) dataset is characterised by tens or hundreds of spectral bands that provide a detailed profile, i.e. a spectral signature of the target (Goetz 2009; Figure 1.1). Both multispectral and hyperspectral data are available at varied spatial resolutions (Liang et al. 2012).

Table 1.1: Wavelength ranges typically used in hyperspectral remote sensing studies.

Spectral region	Abbreviation	Wavelength range (nm)
Visible	VIS	400 – 700
Red-edge	RE	680 – 720
Near infrared	NIR	700 – 1 200
Shortwave infrared	SWIR	1200 – 2 500

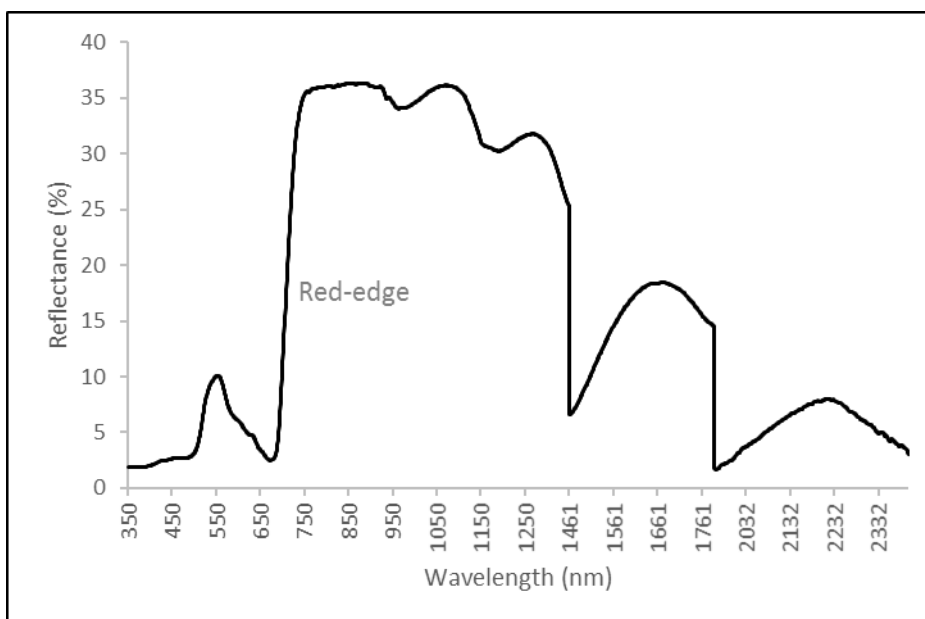


Figure 1.1: Reflectance spectrum of a healthy *P. radiata* seedling indicating the red-edge region.

Hyperspectral data collected using space-borne, air-borne, and field spectrometers have been employed in a wide array of vegetation studies including foliar chemistry modelling (Kokaly & Skidmore 2015; Lepine et al. 2016; Wang et al. 2017), foliar water content estimation (De Jong et al. 2014; Fang et al. 2017), crop phenology modelling (Cole et al. 2014; Lausch et al. 2015), leaf area estimation (Liu et al. 2016; Ali et al. 2017), plant water stress modelling (Loggenberg et al. 2018), species diversity mapping (Ferreira et al. 2016; Laurin et al. 2016; Hakkenberg et al. 2018), and pest and disease modelling in forests (Poona & Ismail 2013; 2014) and agriculture (Abdel-Rahman et al. 2013; Baranowski et al. 2015; Calderón et al. 2015; Adam et al. 2017; Bajwa et al. 2017). The spectral signatures are analysed to extract a wealth of information that is directly linked to the chemical and physical properties of the material under investigation.

Analysing hundreds of spectral bands, is however, not without its challenges. The curse of dimensionality, i.e. large number of predictors ( $p$ ) relative to a limited training set size ( $n$ );  $n \ll p$ , leads to the Hughes phenomenon (Hughes 1968). Additionally, the curse of dimensionality often leads to reduced classification performance (Pal & Foody 2010; Mianji & Zhang 2011). Consequently, researchers have investigated methods that efficiently process high dimensionality data, i.e. improve computational efficiency and classification accuracy.

ML (Goodfellow et al. 2016) algorithms are widely used for processing hyperspectral data. A popular ML algorithm is random forest (RF), proposed by Breiman (2001). RF is a tree-based ensemble that employs bagging (Breiman 1996) and recursive partitioning (Steinberg 2009; Strobl et al. 2009) for model building. RF is often coupled with feature selection algorithms to further improve classification accuracy (Poona & Ismail 2013; 2014; Poona et al. 2016a). Boruta (Kursa et al. 2010; Kursa & Rudnicki 2010) is a wrapper algorithm (Das 2001) that is growing in popularity within the remote sensing community; specifically for hyperspectral data analysis.

Multispectral remotely sensed data has been successfully employed for the detection and mapping of forest health; see for example Ismail et al. (2007). However, there is an increasing trend in the use of hyperspectral remotely sensed data. Hyperspectral remote sensing coupled with ML provides a unique opportunity to acquire timely information on forest health. The increased spectral information content of hyperspectral data, i.e. high dimensional data comprising a large number of contiguous bands, can be employed to detect subtle variations in forest health condition. The studies noted in Table 1.2 confirm (i) the efficacy of employing hyperspectral remotely sensed data and machine learning for modelling forest health, (ii) the utility of vegetation indices (VI) for detecting and monitoring plant stress, and (iii) the need for early stress detection techniques to curb economic losses.



Table 1.2: Recent applications of hyperspectral remote sensing for forest health assessment.

<b>Pest / Pathogen</b>	<b>Host</b>	<b>Sensor</b>	<b>Techniques</b>	<b>Authors</b>
<i>Ips typographus</i>	<i>Picea abies</i>	Fabry-Pérot interferometer (FPI)	Analysis of variance (ANOVA), spectral indices, support vector machine (SVM)	Näsi et al. 2018
<i>Austropuccinia psidii</i>	<i>Melaleuca quinquenervia</i>	Headwall Nano-Hyperspec	Spectral indices, extreme gradient boosting (XGBoost)	Sandino et al. 2018
<i>I. typographus</i>	<i>P. abies</i>	FPI	ANOVA, spectral indices, k-nearest neighbour (k-NN)	Näsi et al. 2015
<i>Dendroctonus ponderosae</i>	<i>Pinus contorta</i>	Airborne Imaging Spectrometer for Applications (AISA)	Continuum removal, ANOVA, similarity analysis (spectral angle mapper (SAM) classification)	Niemann et al. 2015
<i>Sirex noctilio</i>	<i>P. patula</i>	AISA Eagle	RF, SVM	Abdel-Rahman et al. 2014
<i>I. typographus</i>	<i>P. abies</i> , <i>Sorbus aucuparia</i> , <i>Abies alba</i> , <i>Fagus sylvatica</i> , <i>Acer pseudoplatanus</i> , <i>Betula pendula</i> , <i>B. pubescens</i>	HyMap	GA, SVM, Savitzky-Golay filter	Fassnacht et al. 2014
<i>Bursaphelenchus xylophilus</i>	<i>Pinus massoniana</i>	ASD FieldSpec	First order derivatives, spectral indices	Ju et al. 2014
<i>Thaumastocoris peregrinus</i>	<i>Eucalyptus macarthurii</i>	ASD FieldSpec 3 Pro FR	Spectral indices, artificial neural network, sensitivity analysis	Oumar & Mutanga 2014

## 1.2 PROBLEM STATEMENT

No study has investigated the utility of high spectral and high spatial resolution remotely sensed data to characterise *F. circinatum* and pitch canker in South Africa. Research that investigates the potential of remote sensing technologies, coupled with advanced image and signal processing techniques (i.e. ML) to model *F. circinatum* and pitch canker in *P. radiata* and *P. patula*, is urgently needed to manage and control the pathogen. This is especially significant given the established potential of remote sensing to characterise plant stress.

## 1.3 SIGNIFICANCE & RATIONALE FOR UNDERTAKING THE RESEARCH

*F. circinatum* is a pathogen that causes the destruction of pine trees resulting in significant economic losses (Dwinell et al. 1985; Aegerter et al. 2003). *P. radiata* and *P. patula* is recognised as the two most susceptible species in South Africa (Wingfield et al. 1999). *F. circinatum* remains a serious threat to the continued sustainability of commercial pine forests in South Africa (Coutinho et al. 2007; Wingfield et al. 2002, 2008). The favourable climatic conditions along South Africa's coastline increase the susceptibility of pine forests within this region of the country (Coutinho et al. 2007).

The current approach (field surveys and employing visual assessments) to detect and monitor the disease is usually localised to individuals and requires extensive time and labour resources. More significantly, such an approach is impractical given that there is approximately 120 000 ha of *P. radiata* plantations in the Western Cape, and tens of thousands of seedlings in nurseries. There is thus an imminent need to find an efficient way to assess the extent and variability of *F. circinatum* infestations for the effective management of the disease.

It is hypothesised that the damage caused by *F. circinatum* on *P. radiata* can be successfully modelled given the availability of high spatial and high spectral resolution remotely sensed data. It is further envisaged that the findings from this study will contribute to the design of practical operational tools that will help alleviate the incidence of *F. circinatum* infection, and serve as part of an integrated forestry management system.

## 1.4 RESEARCH QUESTIONS

This research was undertaken to answer three principal questions:

1. Can high spectral resolution (hyperspectral) remotely sensed data coupled with RF and Boruta successfully detect asymptomatic (i.e. infected, but no symptom expression) *F. circinatum* stress in *Pinus* seedlings?

2. Can hyperspectral remotely sensed data coupled with RF and Boruta successfully discriminate healthy, damaged, and infected *Pinus* seedlings?
3. Which wavelengths are important / relevant for detecting asymptomatic *F. circinatum* stress in *Pinus* seedlings?
4. Which wavelengths are important / relevant for discriminating healthy, damaged, and infected *Pinus* seedlings?
5. Can high spatial resolution multispectral remotely sensed data successfully model pitch canker disease in a *Pinus* forest?

## 1.5 AIM & OBJECTIVES

The overarching aim of this study was to evaluate the utility of remotely sensed data, specifically high spectral resolution (hyperspectral) data for modelling *F. circinatum* stress in *P. radiata* and *P. patula*.

The specific objectives of this study are to:

1. Assess the utility of the RF ensemble for feature selection and classification of healthy and stressed *Pinus* seedlings using multitemporal hyperspectral data;
2. Evaluate the Boruta feature selection algorithm for identifying the most important spectral bands (wavelengths) for modelling *F. circinatum* stress in *Pinus* seedlings; and
3. Test the viability of high spatial resolution multispectral data for modelling pitch canker stress in *P. radiata* plantations.

## 1.6 RESEARCH METHODOLOGY

The research is quantitative and empirical, employing analytical techniques to build statistical models using multitemporal hyperspectral data. Fundamental to the research is the concept of a spectral signature. Spectral signatures allow for the differentiation in variability of features over varying wavelengths (Figure 1.1). Reflectance spectra at the leaf scale are under the influence of leaf morphology (cell wall thickness, intercellular air spaces) and leaf biochemistry (water content, pigments) (Elvidge 1990); all of which contribute to the reflectance spectra of vegetation (Gates 1965; Knipling 1970). At wavelengths in the visible range (400-700 nm), spectral variability is low due to strong absorption by chlorophyll (Poorter 1995; Cochrane 2000), whereas in the near infrared (NIR; 700-1200 nm), high reflectance is observed resulting from photon scattering attributed to leaf morphology (Woolley 1971; Grant 1987; Asner 1998). Shifts in position of the ‘red-edge’ (Figure 1.1), the abrupt reflectance change in the 680–780nm region of vegetation spectra, can be related to plant health levels (Ghiyammat & Shafri 2010). Pu et al. (2003) recognised that for a healthy plant with

high chlorophyll content and high leaf area index (LAI), the red-edge position (REP) shifts toward the longer wavelengths, whereas for a plant stressed from disease or chlorosis, and consequently low LAI, the REP shifts toward the shorter wavelengths.

Advanced ML algorithms, specifically the tree-based ensemble learner RF (Breiman 2001), is employed for data analysis. RF, coupled with feature selection algorithms, namely wrappers, is employed for model building, i.e. developing statistical models for classification of healthy and stressed *Pinus*. Models are statistically evaluated in terms of their accuracy and robustness using a confusion matrix (Kohavi & Provost 1998), Kappa analysis (or KHAT statistic) (Congalton & Green 2009), and cross-validation (Hastie et al. 2009).

## 1.7 OUTLINE OF DISSERTATION

This dissertation comprises eight chapters. Chapter 1 served to introduce the research by providing a general background to the study, contextualise the significance and rationale of the research, stated the research aim and objectives, and provided an overview of the research methodology. The following chapter, Chapter 2, is presented as a systematic review of the RF algorithm, which features extensively in this research. Chapters 3, 4, 5, 6, and 7 are reformatted manuscripts, published in international peer-reviewed journals. The final chapter (Chapter 8) is a synthesis and contextual analysis of the research, provides concluding remarks, and highlights avenues for future research and development of the methodology presented in this dissertation.

Figure 1.2 illustrates the connection between the various chapters (experiments) undertaken in this research. The results of Chapter 3 directly informs the experimental design of Chapter 4, and indirectly informs the experimental design of Chapters 5 and 6. Similarly, the results of Chapter 4 directly informs Chapter 5, and indirectly Chapter 6. Chapter 6 is thus informed by Chapters 3, 4, and 5. Chapters 3, 4, 5, and 6 form the hyperspectral component of the research. Chapter 7 encapsulates the multispectral component of the research.

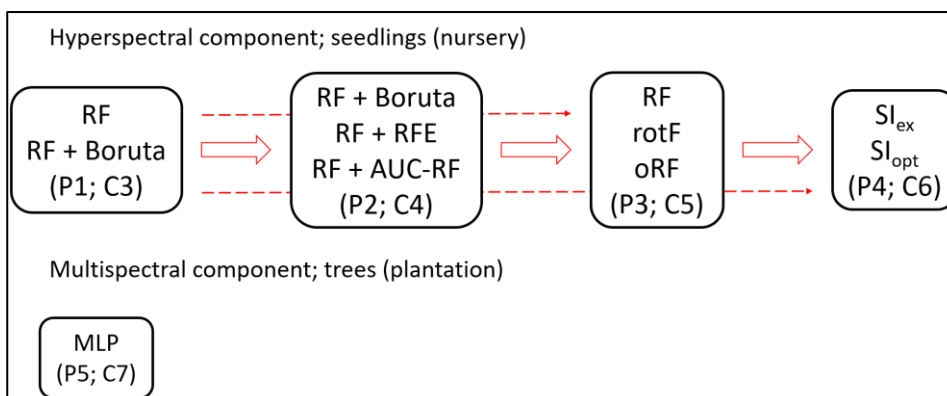


Figure 1.2: Interconnectedness of dissertation chapters (experiments).

## CHAPTER 2: RANDOM FOREST CLASSIFICATION OF HYPERSPECTRAL DATA

Poona NK. Random forest classification of hyperspectral data: A review. In Prep.

### **Abstract**

The random forest (RF) algorithm has seen an exponential increase in use since its introduction in 2001. The RF model has found widespread use in the remote sensing domain, particularly for the analysis of high dimensional spectroscopic (hyperspectral) data. Despite its widespread use in classification and regression tasks, confusion prevails regarding its implementation. This confusion is likely fuelled by the mixed results reported by several researchers. Within the remote sensing community, there is also much debate regarding RF variable importance, feature selection, and hyperparameter tuning. Despite several review papers accounting for the use of RF in various domains, no study to date has provided a meta-analysis and critical evaluation of the RF algorithm in the context of hyperspectral data analysis. Consequently, this review aimed to bridge this gap by deconstructing the key elements of RF and its implementation, critically evaluating the RF algorithm's performance in the presence of noise and imbalanced class distributions, and comparison with other ML algorithms, and finally, reviewing the alternative RF models available. RF was found to be well suited for high dimensional classification tasks, with improved performance achieved using feature selection algorithms prior to classification. Additionally, rotF and oRF show great promise, overcoming the limitations of RF.

**Keywords:** random forest (RF), review, feature subset optimisation, RF sensitivity, RF variants

## 2.1 INTRODUCTION

Traditionally, researchers employed parametric approaches for the classification of remotely sensed data. However, in recent years, attention has shifted toward non-parametric machine learning (ML) approaches given the increased dimensionality of remotely sensed data, and the subsequent sensitivity of parametric classifiers to the Hughes phenomenon (Hughes 1968). Hyperspectral remote sensing data is inherently high dimensional; the number of spectral bands ( $p$ ) is almost always significantly larger than the number of samples ( $n$ ), i.e.  $n \ll p$ . Consequently, the majority of remote sensing studies analysing high dimensional spectral data, have employed ML algorithms in an attempt to mitigate the Hughes effect.

ML (Goodfellow et al. 2016) has been shown to be an effective empirical methodology for the classification of non-linear, high dimensional problems (Lary et al. 2016). A multitude of ML algorithms have been developed over the years; the most popular include Naïve Bayes (NB), SVM, k-NN, ANN, linear discriminant analysis (LDA), and decision tree (DT). For details on these algorithms, and others; see for example Wu & Kumar (2009), Hastie et al. (2009), and James et al. (2013). The need for more advanced learning algorithms is in line with the ever increasing volume and complexity of data coupled with advances in data science and machine intelligence.

Of the numerous ML algorithms available, the most widely used are the ensemble learners (Sheridan et al. 2016; Abellán & Castellano 2017; Loggenberg et al. 2018; Xia et al. 2018). This is likely attributed to their computational efficiency, high classification performance, and robustness when applied to high dimensional, ill-posed problems. Ensemble learners (Rokach 2010, Witten et al. 2011; Zhou 2012; Rokach & Maimon 2015) may be defined as meta-algorithms that combine the predictions of an aggregate of base learners (weak learners) in order to obtain improved prediction accuracy. The principle underpinning ensemble learning is that a composite model, comprising a population of weak learners, has greater generalisation ability compared with a single model.

Aggregating base learners into ensembles is achieved either via voting (nominal outputs; employed for classification tasks) or averaging (numeric outputs; employed for regression tasks) (Zhou 2012). Zhou (2012) defines two ensemble paradigms; sequential methods that exploit the dependence between base learners, as is the case with boosting methods such as adaptive boosting (AdaBoost; Freund & Schapire 1997), and parallel methods that exploit the independence between base learners, as is applied with RF (Breiman 2001). Additionally, ensembles are classified as either homogenous, i.e. base learners are all of the same type, or heterogeneous, i.e. different types of base learners form the ensemble (Zhou 2015).

Tree-based ML approaches, in particular, have found widespread use in hyperspectral studies. A popular tree-based approach is classification and regression trees (CART), proposed by Breiman et al. (1984). The CART model is in principle a binary DT based on recursive partitioning of the feature space (Steinberg 2009; Strobl et al. 2009). For a detailed account of DTs, see Rokach & Maimon (2015). Hastie et al. (2009) and James et al. (2013) highlight several advantages of DT, namely their ease of interpretability, resistance to irrelevant features and outliers, computational efficiency, and parallel nature to human decision-making. The major limitation of DT is their high variance. However, this high variance makes DT ideal base learners for ensembles; the high variability among DT equates to a diverse ensemble of base learners (Blaser & Fryzlewicz 2016). Additionally, Hastie et al. (2009) and Goldstein et al. (2011) indicate that this high variance can be reduced via bagging (bootstrap aggregating).

Bagging or bootstrap aggregating (Breiman 1996) uses an aggregate of DTs as base learners, as opposed to an individual DT, to make a prediction (Hastie et al. 2009; Strobl et al. 2009). The principle underpinning bagging is to grow a committee of DTs from bootstrap samples—without replacement—of the training data, and derive a prediction from each bootstrap sample. Growing multiple DTs equates to simulating multiple training datasets (Goldstein et al. 2011). The final prediction is based on the majority vote averaged over all DTs (Breiman 1996).

An appeal of bagging is the efficiency of computing the model generalisation error using an out-of-bag (OOB) sample (Goldstein et al. 2011). The OOB sample is equivalent to approximately 37% of the original sample, which is not part of the bootstrap sample (Breiman 1996). This OOB sample is subsequently used to compute the OOB error, as an independent test of model accuracy. Hastie et al. (2009) asserts that the OOB error estimate is a close match to the error estimate obtained using cross-validation (CV). The ability to compute generalisation error using the OOB sample is significant given the unfeasibility of analytical methods and computational expense of CV (Goldstein et al. 2011).

A related ensemble method to bagging, is boosting (Freund & Schapire, 1997), based on the work of Kearns (1998), which uses a committee of weak learners to improve prediction accuracy. However, unlike bagging that grows DTs using subsampling, boosting uses all samples and a weight function to manipulate the influence of samples. With each successive iteration, the weight of misclassified samples are increased whereas the weight of correctly classified samples are decreased (Freund & Schapire 1997; Dietterich 2000; Hastie et al. 2009). Like bagging, the final prediction is based on the average prediction, i.e. weighted majority vote over all DTs. Boosting has been shown to reduce both bias and variance, compared with bagging that reduces the variance but has little effect on the bias

(Belgiu & Drăguț 2016). However, boosting may be prone to overfitting (Witten et al. 2011; Zhou 2015).

RF is probably the most widely used ensemble method. The RF algorithm has grown in popularity (Figure 2.1) as the model of choice for complex classification tasks, particularly for the classification of high dimensional data. This popularity is likely attributed to the model's interpretability and performance, when compared with other learning algorithms such as SVM (Abdel-Rahman et al. 2014; Ghamisi et al. 2017; Raczko & Zagajewski 2017; Yuan et al. 2017), artificial neural networks (Ghamisi et al. 2017; Raczko & Zagajewski 2017; Yuan et al. 2017), and boosting trees (Ismail & Mutanga 2011). RF has been shown to be particularly adept in reducing the data dimensionality, and classification of high dimensional problems. Within a remote sensing context, RF has successfully been exploited in a diversity of fields including precision agriculture (Adam et al. 2017; Yuan et al. 2017; Loggenberg et al. 2018), vegetation mapping (Abdel-Rahman et al. 2015; Peerbhay et al. 2015; Agjee et al. 2016), forestry (Abdel-Rahman et al. 2014; Poona & Ismail 2014; Ferreira et al. 2016; Poona et al. 2016a; 2016b; Raczko & Zagajewski 2017), and biodiversity and ecology (Peerbhay et al. 2016; Hakkenberg et al. 2018).

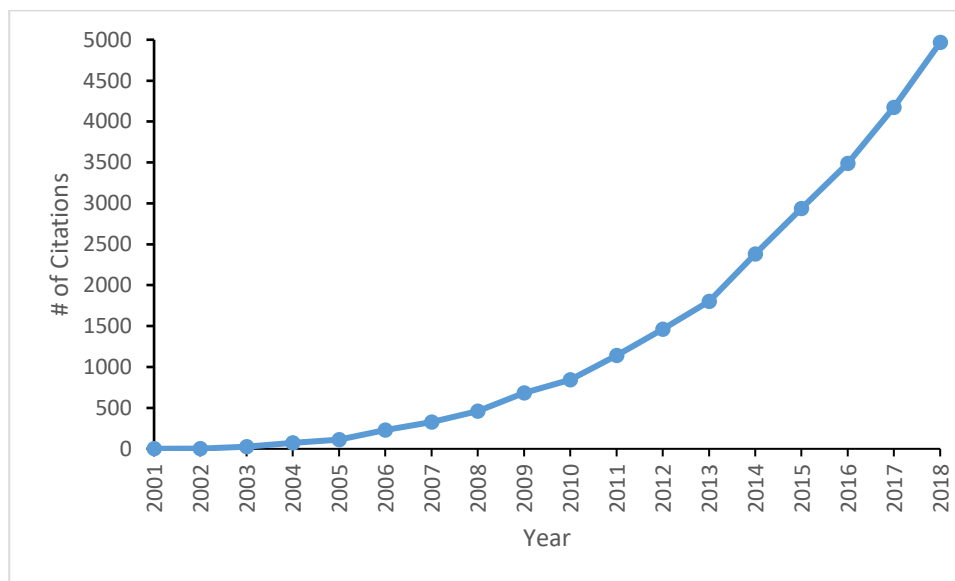


Figure 2.1: Citations of Breiman (2001) from 2001 to 2018. Data source: Scopus Metrics.

## 2.2 RANDOM FOREST

### 2.2.1 Feature subset optimisation

#### 2.2.1.1 Variable (feature) importance

RF calculates two variable importance measures (VIMs) used for feature ranking. Gini importance is computed as the sum of all decreases in Gini impurity of a splitting variable normalised by the number



of trees in the forest. Permutation importance is regarded as a more advanced measure of RF VIM. Permutation importance is calculated as mean decrease in classification accuracy using the OOB observations. The permutation importance is computed by measuring the change in prediction accuracy when the OOB observations are randomly permuted compared with the original observations. The difference in prediction accuracy is then averaged over all trees to compute the permutation importance value (Goldstein et al. 2011; Touw et al. 2013).

Several authors have commented on the biases of the RF importance measures. Breiman et al. (1984), and later Strobl et al. (2007a, 2007b), noted that the Gini importance, and to some degree the permutation importance, is biased in favour of variables comprising a greater number of categories. The bias of the Gini importance measure was echoed by Genuer et al. (2010) and Boulesteix et al. (2012). The authors do, however, point out that the permutation importance measure is more reliable given that it is based on the OOB error measure. In order to overcome the bias of the Gini importance, Strobl et al. (2007a) proposed the use of an alternative RF methodology based on a conditional inference framework developed by Hothorn et al. (2006). Strobl et al. (2007a) suggested using the  $p$ -value (based on the Gini gain and theory of maximally selected statistics). Sandri & Zuccolotto (2008) presented a heuristic loosely based on work by Wu et al. (2007).

In contrast to the biases noted with categorical variables, Strobl et al. (2007b) recorded no biases when using continuous predictors. Hence, for numerical (quantitative) predictor variables, such as high dimensional spectral data, these biases are irrelevant; RF importance measures should be unaffected. No selection bias has been reported in the remote sensing domain. Notably, a limited number of hyperspectral studies have employed the Gini importance; see for example Menze et al. (2009); with only three in the remote sensing domain, namely Poona & Ismail (2013), Poona et al. (2016a) and Adam et al. (2017).

The majority of remote sensing studies have employed the permutation importance, as a measure of variable importance. The widespread use of the permutation importance measure in hyperspectral remote sensing studies (Ismail & Mutanga 2011; Adjorlolo et al. 2013; Poona & Ismail 2014; Abdel-Rahman et al. 2015; Mutanga et al. 2015; Poona et al. 2016a; Agjee et al. 2016) may be attributed to the permutation importance measure being regarded as the more advanced measure of variable importance, and being considered a more reliable indicator of RF variable importance, compared with the Gini importance measure (Genuer et al. 2010).

Despite their extensive use, the most significant limitation of RF importance measures is perhaps that they always output a ranking, irrespective of VIM (Boulesteix et al. 2012). This limitation is significant, as unimportant and redundant features (i.e. features that have no usefulness to the

prediction problem) are all ranked. Consequently, approaches that directly evaluate RF VIM, and undertake feature selection, have been investigated.

#### 2.2.1.2 Feature selection

Several authors (Strobl et al. 2007a, 2007b; Genuer et al. 2010; Goldstein et al. 2011; Boulesteix et al. 2012) note that both Gini importance and permutation importance measures may be biased in favour of highly correlated features when identifying important variables in high dimensional feature space. Additionally, the RF model only provides insight into a feature's importance to a classification problem; the model does not automatically undertake feature selection (Adam et al. 2017). Consequently, much research has been undertaken developing feature selection algorithms that (i) reduce data dimensionality by removing highly correlated / redundant features, (ii) improve computational efficiency through the use of a lower dimensional dataset, and (iii) optimise classification performance by having removed irrelevant / redundant features.

The aim of feature selection is to derive a subset of the most important / relevant features that yield the highest classification accuracies, or put differently, the lowest classification error rates. However, feature selection is not without its limitations. Rokach & Maimon (2015) assert that feature selection may (i) produce a subset that still comprises a relatively large number of variables, (ii) result in reduced classification performance due to the potential loss of important / relevant variables, and (iii) be inefficient in handling high dimensional datasets.

The two primary feature selection techniques are filters and wrappers (Das 2001), of which there are numerous algorithms available. For details, see Chandrashekar & Sahin (2014), Jović et al. (2015), and Ang et al. (2016). Despite their widespread and continued use, filter methods have several limitations to their implementation. Duch (2006) provides a detailed account of these limitations. Consequently, the most widely used algorithms for the classification of high dimensional remotely sensed data include wrappers, which are often embedded with the learning algorithm (classifier). Wrappers use the induction algorithm to score subsets of features based on the feature's predictive power. Feature selection is consequently dependent on the induction algorithm selected, given that the wrapper is often embedded with the induction algorithm (Saeys et al. 2007).

A survey of the published research indicates that the majority of researchers favour the use of embedded algorithms (Blum & Langley 1997; Guyon & Elisseeff 2003). In particular, wrappers embedded with the RF algorithm have gained popularity for the classification of high dimensional remotely sensed data. The embedded approach presents several advantages compared with the standard wrapper approach, namely (i) computational efficiency, (ii) lower risk of overfitting, and (iii) it considers feature dependencies (Ang et al. 2016).

Recursive feature elimination (RFE) has been widely used in high dimensional data analysis studies. A backward elimination search approach based on SVM, aptly named SVM-RFE, was introduced by Guyon et al. (2002) as a wrapper with SVM. The SVM-RFE wrapper found widespread use in microarray analysis (Cannas et al. 2013), metabolomics (Chen et al. 2013), soil analysis (Stevens et al. 2013), text classification (Chapelle & Keerthi 2008), and biomedical data analysis (Huang et al. 2014; Sanz et al. 2018). Studies have also employed SVM-RFE for multi-data analysis, for example Maldonado & Weber (2009) for microarray and credit scoring data analysis, and Dessì & Pes (2015) for internet advertisements, test characterisation, and microarray data analysis. Within a remote sensing domain, Pal (2006), Archibald & Fann (2007), Zhang et al. (2009b) and Pal & Foody (2010) exploited the SVM-RFE algorithm for selecting features from airborne imaging spectrometer data. Díaz-Uriarte & Alvarez de Andrés (2006) proposed a RFE procedure, as a wrapper around the RF algorithm, named RF-RFE. The authors successfully demonstrated the superior performance of RF-RFE in achieving significant data dimensionality reduction coupled with low error rates. In a comparative study of SVM-RFE and RF-RFE, Granitto et al. (2006) demonstrated that RF-RFE outperformed SVM-RFE when applied to proton transfer reaction-mass spectrometry (PTR-MS) analysis of agro-industrial products. RF-RFE has since been widely adopted in an array of applications including quantitative structure-activity relationship (QSAR) modelling (Svetnik et al. 2003), metabolomics studies (Zhou et al. 2012; Degenhardt et al. 2017), ecological modelling (Fox et al. 2017), analysis of satellite remote sensing (Landsat) data (Gregorutti et al. 2017), and microarray analysis (Chen et al. 2018). Coupled with the increased utility of RF for classification, researchers have exploited RF-RFE for the analysis of hyperspectral remote sensing data. For example, Abdel-Rahman et al. (2015) used RF-RFE to select the most important Airborne Imaging Spectrometer for Applications (AISA) Eagle bands for mapping melliferous plants. Sun et al. (2018) employed RF-RFE to extract important bands from hyperspectral imagery for mouldy tea detection.

Building multiple models, as in the case of RFE, can be computationally intensive. Instead of multiple models, Deng & Runger (2012) proposed the regularised framework applied to RF (regularised RF; RRF) that builds a single model for undertaking feature selection. However, in RRF, feature selection is based on only a part of the training data, and may be greedy (Deng & Runger 2013). The regularised framework was subsequently replaced by an enhanced version; guided regularised RF (GRRF) framework (Deng & Runger 2013), in which the RF VIM is used to guide the feature selection process. Additionally, the GRRF framework overcomes the node sparsity issue, which often plagues DT-based models (Deng & Runger 2013). RRF and GRRF have been employed for modelling pest and disease stress in plants (Poona & Ismail 2013; Adam et al. 2017), predicting seabed hardness (Li et al. 2016), and discriminating invasive and indigenous plant species (Murériwa et al. 2016).

Another feature selection algorithm wrapped around RF is the area under the receiver operating characteristic curve of the RF (AUC-RF) (Calle et al. 2011). AUC-RF employs a backward elimination procedure, similar to RFE, based on variable rankings. Only two studies to date have employed AUC-RF. López de Maturana et al. (2013) employed AUC-RF to select an optimal subset of single nucleotide polymorphisms (SNPs) for predicting bladder cancer risk. The authors highlighted the utility of AUC-RF for SNP selection, compared with Bayesian threshold least absolute shrinkage and selection operator (LASSO), and a logistic regression. Within a remote sensing context, Poona et al. (2016a) employed the AUC-RF algorithm for the classification of hyperspectral data. The authors compared the utility of the Boruta, RFE, and AUC-RF for modelling *F. circinatum* stress in *P. radiata* and *P. patula* seedlings successfully demonstrating the robustness of AUC-RF for hyperspectral data analysis.

Boruta (Kursa et al. 2010; Kursa & Rudnicki 2010) is a wrapper embedded with RF that is growing in popularity. The algorithm has been repeatedly shown to outperform most feature selection algorithms in terms of dimensionality reduction and classification performance. For example, Kursa (2014b) investigated the performance of four RF-based feature selection algorithms, namely RF-ACE (artificial contrasts with ensembles) (Tuv et al. 2009), RFE, RRF, and Boruta. Overall, Boruta outperformed the other three algorithms in terms of classification accuracy, feature selection consistency, and computational expense. Li et al. (2016) compared five feature selection algorithms including Boruta and RRF. The authors concluded that Boruta produced some of the most accurate models. In a comparative study of several feature selection algorithms, including RFE and Boruta, Degenhardt et al. (2017) noted Boruta's superior performance compared with the other algorithms tested. Only a limited number of studies to date have exploited Boruta for the analysis of hyperspectral data. In a comparative study employing RRF, RFE, and Boruta, Poona & Ismail (2013) showed that although Boruta produced an equivalent subset ( $n = 17$ ) to RRF, Boruta classification accuracy was higher compared with RRF. Similarly, Poona et al. (2016a) noted Boruta's higher performance compared with RFE and AUC-RF. Agjee et al. (2016) showed that using Boruta resulted in smaller subsets, coupled with lower error rates, compared with using RF-RFE.

### 2.2.1.3 Hyperparameter tuning

The two primary tuning parameters (i.e. hyperparameters) of the RF algorithm are the number of trees grown in the forest (*n<sub>tree</sub>*), and the number of features randomly selected for determining the split at each node in a tree (*m<sub>try</sub>*). For classification tasks, Breiman (2001) proposed that the default number of trees should be 500 (*n<sub>tree</sub>* = 500) and default number of variables selected for node splitting should be the square root of the number of features (*m<sub>try</sub>* =  $p^{1/2}$ , where  $p$  = the number of features). However,

Goldstein et al. (2011) and Boulesteix et al. (2012) assert that the optimal value of *mtry* and *ntree* is dependent on the nature of the feature set, and should subsequently be determined empirically.

According to Boulesteix et al. (2012), the *ntree* value should increase in line with an increase in *p*, thereby giving all features equal opportunity for selection. Hence, for a large feature set, for example  $p = 10\,000$ , the default *ntree* value of 500 should not be used. Ultimately, the *ntree* value should be sufficiently large to yield stable predictions (Duroux & Scornet 2016). A larger *mtry* value results in faster convergence and generation of smaller and more accurate trees, which is particularly desirable when noisy variables are present (Goldstein et al. 2011; Boulesteix et al. 2012). Boulesteix et al. (2012) further advocate that several model iterations be run with varying *mtry* and *ntree* values until stability in the error rate and / or variable importance measure is achieved.

The first studies to suggest optimising *mtry* and *ntree* values is Breiman (2002) and Liaw & Wiener (2002). Probably the first study to empirically examine the influence of the *mtry* and *ntree* value on RF model performance is Díaz-Uriarte & Alvarez de Andrés (2006). Díaz-Uriarte & Alvarez de Andrés (2006) proposed optimising the *ntree* value by using an incremental increasing factor ( $= \{500, 1\,000, 1\,500, 2\,000\}$ ) of the default *ntree* value, and using a multiplicative factor ( $= \{1/3, 1/2, 1, 2, 3\}$ ) of the default *ntree* value. Several authors subsequently adopted this methodology for optimising the *mtry* and *ntree* value. For example, Adam et al. (2012) optimised *mtry* and *ntree* values for the classification of *Cyperus papyrus* L. and its co-existent species in a wetland swamp using resampled in situ spectroscopy data. The authors recorded overall accuracies above 90% with KHAT values above 0.85. Adam et al. (2013) achieved an overall accuracy of 82% (KHAT = 0.63) using optimised RF hyperparameter values to model the susceptibility of *Eucalyptus nitens* to *Coryphodema tritis*. Using optimised *mtry* and *ntree* values, Abdel-Rahman et al. (2014) successfully discriminated *Sirex noctilio* grey-attacked and lightning-struck *Pinus* trees using AISA Eagle imaging spectrometer data, achieving accuracies above 74%. In a study employing RF for the classification of healthy, infected, and damaged *P. radiata* seedlings, Poona & Ismail (2014) showed that using optimised hyperparameter values yielded improved classification accuracies. Abdel-Rahman et al. (2015) used optimised *mtry* and *ntree* values for modelling the flowering of several species of melliferous plants using AISA Eagle imaging spectrometer data. The authors reported accuracies above 80%.

Several studies have demonstrated that optimal results can be obtained using default *ntree* and / or *mtry* values, despite many authors advocating the need to optimise the two RF hyperparameter values. For example, Ismail & Mutanga (2011) optimised the *mtry* and *ntree* values for modelling *Sirex noctilio* infestation levels in a *Pinus patula* plantation using in situ spectral measurements. The best classification accuracies were obtained using an optimised *mtry* value and default *ntree* value. Peerbhay et al. (2015) used default RF hyperparameter values to successfully model *Solanum*

*mauritianum* infestation in a *P. patula* plantation using AISA Eagle imaging spectrometer data. Agjee et al. (2016) modelled the efficacy of *Neochetina* spp. for the biocontrol of *Eichhornia crassipes* in a freshwater ecosystem. After optimising the *mtry* and *ntree* values, the authors found that using default hyperparameter values yielded the best results. Similarly, Chemura et al. (2016) applied default *mtry* and *ntree* values to successfully model *Hemileia vastatrix* infestation on *Coffea arabica* using resampled in situ spectroscopy data.

An assumption in many studies is that the RF ensemble is robust to hyperparameter settings. However, several authors have found that the selection of RF hyperparameter values, can directly influence the VIM. Early work by Díaz-Uriarte & Alvarez de Andrés (2006) found that larger *mtry* values tended to yield more reliable VIMs. Genuer et al. (2010) and Goldstein et al. (2010) also found that *mtry* values greater than the recommended default value ( $p^{1/2}$ ) yielded more reliable VIMs. Additionally, Goldstein et al. (2010) noted that a smaller *ntree* value will likely yield the same prediction accuracies as larger *ntree* values; albeit less reliable trees. Huang & Boutros (2016) reported variable results with respect to *mtry* values and VIM stability. However, the authors noted that larger *ntree* values led to more stable VIMs. Behnamian et al. (2017) noted that both the Gini importance and permutation importance measures were highly variable over  $n$  RF iterations. The authors concluded that a higher *ntree* value should be used to achieve stable VIM rankings. Ultimately, the selection of optimal *mtry* and *ntree* values will be data dependent, and should thus be determined experimentally for each input dataset.

## 2.2.2 RF sensitivity

### 2.2.2.1 Noise

Hyperspectral data, as with all real-world data, is rarely noise-free (Agjee et al. 2018; Jiang et al. 2019). Of primary importance is the negative impact of noise on classifier performance. Noise is managed either through pre-processing the data to remove the noise, i.e. applying de-noising algorithms, or employing algorithms that are robust to noise (Frénay & Verleysen 2014). Several algorithms have been proposed to deal with noise prior to classification. However, these de-noising algorithms are not without their limitations. Consequently, there is growing interest in identifying algorithms that are inherently robust to noise.

Ensemble learners have been widely used owing to their computational accuracy and robustness to noise (Breiman 1996). Breiman (2001) advocated the use of the RF algorithm, due to its computational accuracy, robustness to outliers and noise, and resistance to overfitting. According to Breiman (2001), the robustness of RF to noise may be attributed to (i) using bootstrap sampling (with



replacement) for DT construction, (ii) selecting random coefficients for node splitting in each DT, and (iii) employing majority voting to produce models with low bias and variance. Consequently, several studies have examined the effect of noise on RF classifier performance, including the effect of varying degrees of noise on RF robustness.

In a comparative study of bagging, boosting, and RF (Hamza & Larocque 2005), RF exhibited the greatest robustness to noise. Folleco et al. (2008) showed that RF outperformed both C4.5 and NB under varying noise levels, and proved most robust to data noise. Folleco et al. (2009) investigated the robustness to noise of 11 classifiers, including RF. Results showed that RF yielded the best overall performance, and demonstrated the greatest robustness to noise. Smith & Martinez (2014) compared RF with C4.5, 5-NN, multilayer perceptron (MLP) ANN, and repeated incremental pruning to produce error reduction (RIPPER) to evaluate each algorithm's noise tolerance using 54 datasets. The MLP and RF models yielded the highest classification accuracies. However, their performance deteriorated with increasing noise. Using synthetic and real-world datasets, Ghosh et al. (2016) demonstrated RF's superior performance and robustness to noise, compared with DT and SVM.

Within a remote sensing context, Ismail & Mutanga (2011) assessed the robustness and stability of boosting trees (BT) versus RF to varying levels of noise. The effect of noise was evaluated using handheld field spectrometer shortwave infrared (SWIR) bands, resampled to Hyperspectral Mapper (HyMap) bands. Overall, RF produced lower misclassification error rates compared with BT, demonstrating its robustness and stability in the presence of noise. Pelletier et al. (2017) examined the sensitivity to noise of RF and SVM for land cover mapping using SPOT 4 and Landsat 8 multispectral imagery. The RF model demonstrated greater robustness to noise compared with SVM. Agjee et al. (2018) evaluated the effect of varying levels of simulated noise on RF classification of hyperspectral (spectroscopy) data. RF performance was compared with the oblique RF (oRF) algorithm (Menze et al. 2011) employing ridge regression for multivariate node splitting. The oRF algorithm (discussed in greater detail in section 5.2.4.2) was shown to outperform the traditional RF algorithm, and exhibited greater robustness to increasing levels of noise. Jiang et al. (2019) evaluated the performance of RF, SVM, ANN, and extreme learning machine (ELM) for the classification of hyperspectral imagery, in the presence of noise. The authors concluded that all four algorithms exhibited sensitivity to increasing levels of noise, with classification performance declining with increasing noise. However, RF and ELM were the least affected.

#### 2.2.2.2 Imbalanced class distribution

Imbalanced datasets result from unequal class distributions, which are common in real-world high dimensional datasets, such as hyperspectral data. Training data imbalance negatively influences the

performance of learning algorithms, particularly when applied to high dimensional data (Boulesteix et al. 2012; Lin & Chen 2012; Menardi & Torelli 2014). More importantly, class imbalance generally favours the majority class, leading to models with prediction bias and higher false negative rates (Leevy et al. 2018). In the presence of imbalanced classes, Chen et al. (2004) noted that RF tends to be biased in favour of the majority class.

In an attempt to mitigate the influence of class imbalance on classification accuracy, especially in the case of extremely imbalanced class distributions, Liaw & Wiener (2002) suggested assigning higher probability thresholds to the under-sampled class, instead of applying the default majority voting. Boulesteix et al. (2012) suggested oversampling of the minority class (i.e. class with smaller number of observations) and / or under-sampling of the majority class (i.e. class with larger number of observations) in order to create balanced classes. Janitza et al. (2013) proposed an AUC-based RF permutation VIM to improve RF performance in the presence of imbalanced classes.

Blagus & Lusa (2010) highlighted RF's sensitivity to class imbalance when applied to high dimensional microarray data. In a similar study, Lin & Chen (2012) demonstrated RF's poor performance in the presence of high dimensional microarray data. Del Río et al. (2014) evaluated four strategies with RF classification, including oversampling, under-sampling, the synthetic minority oversampling technique (SMOTE) algorithm (Chawla et al. 2002) that employs an oversampling strategy to generate synthetic samples from the minority class, and cost-sensitive learning (Ling & Sheng 2010) by way of weighted RF (Chen et al. 2004) that considers misclassification cost, in an attempt to reduce total cost. The authors noted that none of the strategies proved valuable in dealing with imbalanced class distributions. Contrary to the above studies, Dittman et al. (2015) found that RF was fairly robust to class imbalance, when applied to bioinformatics data.

Within a remote sensing context, Dalponte et al. (2013) noted RF's sensitivity to class imbalance, when applied to tree species classification using HySpex imaging spectrometry data. Mellor et al. (2015) tested the effect of imbalanced training data on the performance of RF for the classification of various remote sensing and ancillary datasets. Their results showed that using balanced datasets yielded the best classification accuracies. Similarly, Millard & Richardson (2015) tested the effect of sample size and data imbalance on RF performance using light detection and range (LiDAR) data and its derivatives. The authors noted that an increased sample size yielded higher classification accuracy, with balanced training data yielding the highest accuracies.

### **2.2.3 Unsupervised RF**

The majority of RF implementations are supervised, i.e. a set of outcome labels are used for classification. However, many learners—including RF—can be used to undertake unsupervised



learning. The basic premise to unsupervised RF (URF), originally proposed by Breiman & Cutler (2003), is to create a synthetic dataset comprising artificial class labels—randomly selected from the original dataset—thereby creating a binary classification problem (Shi & Horvath 2006; Afanador et al. 2016). For additional detail on URF theory and implementation; see for example Shi & Horvath (2006) and Afanador et al. (2016).

URF has been successfully employed for tumour profiling from microarray data (Shi et al. 2005), spatio-temporal analysis of video data (Pei et al. 2013; Yu et al. 2013), image object detection (Du & Chen 2014, 2015), clinical diagnosis of Alzheimer’s disease using positron emission tomography images (Lu et al. 2015), fault detection in semiconductor manufacturing (Puggini et al. 2015), and population structure analysis in bioinformatics (Alhusain & Hafez 2017).

Within the field of remote sensing data analysis, URF has been used for individual tree crown (ITC) delineation and extraction using LiDAR data (Gupta et al. 2010), and for forest mapping using synthetic aperture radar data (Baron & Erasmi 2017). Only one study to date has employed URF for the classification of hyperspectral data. Peerbhay et al. (2015) used AISA Eagle imaging spectrometry data in conjunction with the RF proximity matrix and Anselin Moran’s I statistic for the detection and mapping of *Solanum mauritianum* in commercial plantations. The authors successfully demonstrated the utility of both the RF outlier detection methodology, as well as the decomposition of the proximity matrix using principal components analysis (PCA) methodology; noting that the latter method yielded superior results.

Liu et al. (2012) proposed Isolation forest (iForest) as an unsupervised non-parametric approach for anomaly detection. The idea underpinning iForest lies in anomaly detection that is independent of distance measures, i.e. there is no need to generate and interpret a proximity matrix. The authors successfully demonstrated iForests’ superior performance compared with four other algorithms including RF. Dalleau et al. (2018) proposed unsupervised extremely randomised trees, that builds on the methodology of URF (Shi & Horvath 2006), and extremely randomised trees (extra trees) (Geurts et al. 2006). No prior study has implemented the unsupervised extra trees methodology.

#### **2.2.4 RF variants**

Since Breiman proposed the RF methodology in 2001, several RF variants have been introduced. Extra trees was proposed by Geurts et al. (2006) as a novel implementation of the traditional RF model. The extra trees model differs from RF by using randomness for selecting node splitting thresholds, instead of node purity, and using the entire set for growing trees, instead of bootstrap samples. Amaratunga et al. (2008) proposed enriched RFs (ERF), a weighted subspace RF method that applies weighted random sampling to select informative features for tree growing. Consequently,

higher weightings are applied to more informative features, and lesser weightings to less informative features. A limitation of ERF, is the algorithm's ability to solve only binary problems. Xu et al. (2012) extended the work of Amaratunga et al. (2008) to solve multiclass problems by using the information gain ratio, instead of the t-test, to calculate feature weights. Random ferns, originally proposed by Özuysal et al. (2007), and later generalised by Kursu (2014a), is based on the NB classifier. A fern is synonymous with a DT, but differs in the computation of posterior probabilities; additively for DT versus multiplicatively for ferns. The Kursu (2014a) implementation additionally introduces bagging in an attempt to improve accuracy. Classification results from varied datasets (Kursu 2014a, 2014b) showed comparable performance to RF. Zhang & Zhang (2008) proposed RotBoost, a novel boosted rotation-based ensemble, which merges the rotation forest (rotF) methodology of Rodríguez et al. (2006) with the boosting methodology of AdaBoost. Classification results from 36 varied datasets demonstrated the high performance of RotBoost compared with rotF. Random jungle (RJ) (Schwarz et al. 2010) was proposed as a fast implementation of RF, specifically for data analysis in genome-wide association studies. The RJ model maintains all the characteristics of Breiman's RF, with the addition of backward elimination for feature selection. Results of RJ were comparable to that of RF, with RJ being less computationally expensive. Boosted RF, proposed by Mishina et al. (2014), integrates boosting with RF. Boosted RF generates complementary learners by constructing successive DTs. Consequently, the number of DTs in the forest is minimised, while maintaining high generality. The authors demonstrated the superior performance of boosted RF using several datasets. Blaser & Fryzlewicz (2016) introduced random rotation ensembles. Instead of using random coefficients to determine the optimal node split, the idea underpinning random rotations is to construct trees, each with a randomly rotated feature space. Such an approach results in a diverse ensemble, with smoother decision boundaries; reminiscent of boundaries generated by rotation forests (Rodríguez et al. 2006). Xia et al. (2018) proposed a boosted rotation-based RF ensemble that uses RF instead of DT as base classifiers, as is the case with RotBoost. The boosted rotation-based RF ensemble was compared with RF as well as four other RF ensembles, namely bagging RF, boosting RF, random subspace RF, rotation RF, for classification of Airborne Visible / Infrared Imaging Spectrometer (AVIRIS) and Reflective Optics System Imaging Spectrometer (ROSIS) imagery. Results showed that the boosted rotation-based RF ensemble generally outperformed all the other models. Deep forest (gcForest) (Zhou & Feng 2017) is a cascaded boosting procedure, analogous to a deep neural network, with hidden layers replaced by an ensemble of DT (such as RF). The deep forest model was shown to yield high accuracies on varied classification tasks (Zhou & Feng 2017; Chen et al. 2019). Zhang et al. (2018) proposed a cascaded RF (CRF) methodology. The CRF approach combines the novelty of ERF (Amaratunga et al. 2008) using a hierarchical random

subspace method for feature selection, with boosted RF (BRF) (Mishina et al. 2014) using the OOB error to update the sample weights. No studies to date have implemented the CRF methodology. The most recent RF variant is dense adaptive cascade forest (daForest) (Wang et al. 2019). The daForest methodology builds on the deep forest methodology by introducing boosting, applying feed-forward connectivity between layers, and incorporating a hyperparameter optimisation layer. The authors showed that daForest outperformed several other models including RF and gcForest.

Two RF variants of growing interest for hyperspectral data classification is the rotF model proposed by Rodríguez et al. (2006), and the oRF model proposed by Menze et al. (2011). The idea of using oblique DTs was initially introduced by Do et al. (2010). Both rotF and oRF employ multivariate splitting for tree construction, compared with orthogonal splitting as employed by RF. Generating multivariate oblique hyperplanes has been shown to be more robust to noise, consequently yielding improved performance. Agjee et al. (2018) demonstrated how oRF and rotF yielded better classification accuracies compared with RF; oRF proving to be most robust.

### 2.2.5 RF implementations

The increasing popularity of RF and the growing interest in the RF variants has led to the development of several implementations. The majority of implementations are available as packages in the R statistical software (R Development Core Team 2019). The *randomForest* package (Liaw & Wiener 2002) in R, is based on the original Fortran 77 code by Breiman & Cutler (2004). The *randomForest* package creates ensembles using default hyperparameter values, i.e.  $n_{tree} = 500$  and  $m_{try} = \sqrt{p}$ , where  $p$  = number of features. Implementation with the *caret* package (Kuhn 2019) allows for hyperparameter value tuning. The *randomForest* package is probably the most widely used; as is evident from the high number of published studies. Table 2.1 summarises the RF model implementations in the R statistical software.

Other implementations include *RandomForestClassifier* and *ExtraTreesClassifier* in Python (Scikit-learn) (Pedregosa et al. 2011) and Cython (Behnel et al. 2011), *RandomForestLearner* and *SimpleRandomForestLearner* in Orange (Demšar et al. 2013; <https://orange.biolab.si>), *CloudForest* (Bressler et al. 2015), a standalone package written in Go (<http://golang.org/>), and *Willows* (Zhang et al. 2009a) that implements CART, RF, and deterministic forest (Zhang & Singer 2003). ALGLIB ([www.alglib.net](http://www.alglib.net)) implements a modified RF algorithm; random decision forest. Additionally, RF is available in several open source and propriety image processing software, e.g. eCognition (Trimble Geospatial 2019), ENVI (Harris Geospatial Solutions, Inc. 2019), Erdas Imagine 2018 (Hexagon AB 2019), Matlab (The Mathworks Inc. 2019), and WEKA (Frank et al. 2016).

Table 2.1: RF variants implemented in R statistical software.

RF model	R package
Extra trees (Geurts et al. 2006)	<i>extraTrees</i> (Simm & Magrans de Abril 2015)
Conditional inference trees (Hothorn et al. 2006)	<i>cforest</i> , included in <i>party</i> (Hothorn et al. 2019)
Rotation forest models (Rodríguez et al. 2006)	<i>rotationForest</i> (Ballings & Van den Poel 2017)
oRF (Menze et al. 2011)	<i>obliqueRF</i> (Menze & Splitthof 2015)
random fern (Kursa 2014a)	<i>rFerns</i> (Kursa 2018)
RRF (Deng & Runger 2012)	<i>RRF</i> (Deng 2019)
weighted subspace RF (Xu et al. 2012)	<i>wsrf</i> (Meng et al. 2017)
RF generator (Wright & Ziegler 2017)	<i>ranger</i> (Wright et al. 2019); includes traditional RF (Breiman 2001), random survival forests (Ishwaran et al. 2008), extra trees (Geurts et al. 2006) and quantile regression forests (Meinshausen 2006)

### 2.2.6 Benchmarking RF

To gauge an algorithm's utility and robustness, its performance (classification accuracies) must be evaluated against that of other widely used algorithms. The performance of RF compared with other ML algorithms including SVM, ANN, extreme gradient boosting (XGBoost), and partial least squares discriminant analysis (PLS-DA), has been evaluated for the classification of hyperspectral data.

Chan et al. (2012) compared RF, RBF-SVM, and boosting (Adaboost) for mapping heath (shrubland) from nadir and off-nadir Compact High Resolution Imaging Spectrometer (CHRIS) imagery. The authors reported mixed results, with no algorithm clearly outperforming the other two, or proving to be most robust. Dalponte et al. (2013) compared RF with a radial basis function SVM (RBF-SVM) and Gaussian maximum likelihood (GML) for tree species mapping in a boreal forest using airborne HySpex imagery. Sequential forward feature selection (SFFS) in combination with the Jeffries-Matusita (J-M) distance was used for feature selection. The RBF-SVM models yielded the highest classification accuracies, outperforming RF and GML. Kong et al. (2013) employed partial least squares discriminant analysis (PLS-DA), soft independent modelling of class analogy (SIMCA), k-NN, RBF-SVM, and RF to develop models for identifying rice seed varieties. The images were captured using a custom-built laboratory-based hyperspectral imaging system. SIMCA, RBF-SVM,

and RF yielded the highest classification accuracies; up to 100%. Employing airborne HyMap and spaceborne Hyperion imagery, combined with vegetation indices (VI) and other ancillary data, Ghosh et al. (2014) evaluated the performance of RF and RBF-SVM to map several tree species in a managed forest. The authors concluded that both RF and RBF-SVM provided equally reliable results. Abdel-Rahman et al. (2014) employed RBF-SVM and RF with AISA hyperspectral imagery to discriminate *Sirex* grey-attacked and lightning-damaged *P. patula* trees. Results of the study showed similar performance by both models. Burai et al. (2015) reported that RBF-SVM yielded the best results, compared with ML and RF, for the classification of herbaceous vegetation from airborne AISA Eagle II imagery. In a comparative study of LDA, L-SVM, RBF-SVM, and RF for tree species classification from airborne ProSpecTIR imagery and using VI, Ferreira et al. (2016) reported that LDA outperformed the other models, with RF yielding the lowest accuracies. Mohite et al. (2017) evaluated the performance of ANN, SVM, RF, and XGBoost for the detection of pesticide residue on grapes from multitemporal handheld spectrometer data. Feature extraction was undertaken using PCA, followed by LASSO and elastic net regularisation for feature selection. XGBoost yielded the best classification accuracy using the top 20 PCs, and RF the lowest accuracy. RF was again outperformed when applied to the LASSO and elastic net subsets. Raczko & Zagajewski (2017) compared RF with RBF-SVM and ANN for classifying five tree species in a natural forest from Airborne Prism Experiment (APEX) imagery. The ANN model produced the best results, with RF producing the lowest classification accuracies. Loggenberg et al. (2018) compared the performance of RF and XGBoost for discriminating healthy and water-stressed vines. RF generally yielded better results for both the full dataset and subset, compared with XGBoost. Sumsion et al. (2019) compared RBF-SVM, ANN (MLP), and RF for classifying tree genus / species using a combination of airborne hyperspectral imagery and airborne LiDAR, from the National Ecological Observatory Network Airborne Observation Platform (NEON AOP). The MLP model outperformed the other two models, yielded more consistent results, and proved to be more robust.

## 2.3 CONCLUSIONS

This review has deconstructed the RF model, and highlighted its utility for the classification of high dimensional hyperspectral data. RF, as an ensemble learner, has been proven to be robust, insensitive to overfitting, and generally yields low error rates in high dimensional classification tasks. RF continually delivers superior results, often outperforming other ML algorithms such as ANN and SVM, as well as ensemble learners such as gradient boosting, and extreme gradient boosting. Additionally, research has demonstrated that the high performance of RF can be improved when coupled with feature selection, e.g. the Boruta wrapper algorithm embedded with RF.

Recent developments in ML have resulted in numerous RF variants. However, two RF variants, namely rotF and oRF have gained interest in the remote sensing community. These two RF variants overcome the inherent limitations, i.e. box-like decision boundary and single feature node splitting of the traditional RF model. Additionally, rotF and oRF have been shown to yield superior classification results, often outperforming RF. However, it must be noted that only a limited number of studies to date have evaluated the utility of rotF and oRF. Additional research is warranted in order to establish their generalised use for the classification of hyperspectral data, in particular, the classification of imaging spectroscopy data.

The traditional RF model continues to be widely used for processing hyperspectral data, and repeatedly yields above satisfactory results. However, RF variants are likely to take centre stage in the future given their better performance compared with the traditional RF model. Future research should thus focus on the RF variants, in particular oRF and rotF. Feature selection algorithms embedded with RF have grown in popularity, given their ability to select an optimum subset of features that are then used for further processing. Algorithms such as Boruta produce smaller feature subsets that ultimately yield high classification accuracies. Thus, RF rankings are less likely to be used in the future, for selecting subsets of important bands. RF's sensitivity to noise and imbalanced class distributions requires further investigation. In particular, class label noise in hyperspectral datasets. The RF proximity matrix provides a powerful means to detecting anomalies in hyperspectral datasets, and for classification where limited or no reference data are available. This is particularly significant in a commercial (operational) environment. Further research is thus required regarding the operational use of URF. The implementation of RF and / or its variants, in fact any machine learning algorithm, depends on several constraints such as input data and computational infrastructure. Ultimately, there is 'no free lunch'!

### CHAPTER 3: SELECTING THE MOST IMPORTANT BANDS FOR MODELLING *FUSARIUM* STRESS USING BORUTA

Poona NK & Ismail R 2014. Using Boruta-selected spectroscopic bands for the asymptomatic detection of *Fusarium circinatum* stress. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 70, 3764-3772.

Poona NK & Ismail R 2012. Discriminating the early stages of *Fusarium circinatum* infection of *Pinus radiata* seedlings using high spectral resolution data. Proceedings of the 9th International Conference of the African Association of Remote Sensing and the Environment (AARSE2012), held 29 October-2 November 2012, El Jadida, Morocco.

#### Abstract

High spectral resolution multitemporal data were used to model asymptomatic stress caused by *F. circinatum* in 3-month old *P. radiata* seedlings. The objectives of the study were to 1) identify an optimal subset of bands that could model asymptomatic stress in *P. radiata* seedlings, and 2) develop a robust classification model for discriminating healthy and stressed seedlings. In order to achieve these objectives, spectral data were collected for healthy, infected, and damaged seedlings using a handheld field spectroradiometer. The data were analysed, first for combined classes (i.e. H-I-D) and then for class pairs (i.e. healthy-infected (H-I) and infected-damaged (I-D)) using the Boruta algorithm. Results indicated that the best discrimination was possible at week three for all classes, with a KHAT value of 0.79 and an out-of-bag error of 14% (CV error = 16%), using a subset of 107 bands. A closer examination of the class pairs, namely H-I and I-D, showed improved discrimination with KHAT values of 0.82 and 0.84, respectively. The H-I class pair was classified using a subset of only 38 bands, whereas the I-D class pair was classified using a subset of 40 bands. Overall, this study demonstrated that it is more difficult to discriminate asymptomatic stress when additional stress related classes are present. Nonetheless, the methodology developed in this study has the potential to be operationalised within a nursery environment for the early detection of *F. circinatum*-induced stress in *P. radiata* seedlings.

**Keywords:** Boruta, feature selection, hyperspectral, random forest (RF), remote sensing.



### 3.1 INTRODUCTION

*F. circinatum* is a highly virulent fungal pathogen (Cram & Fraedrich 2009) causing widespread mortality of *P. radiata* and *P. patula* seedlings (Wingfield et al. 2008). The fungus, which causes pitch canker disease in mature trees, was first reported in 1990 (Morris 2010) and has since become endemic in nurseries across South Africa (Roux et al. 2007; Porter et al. 2009). *F. circinatum* is a serious problem in South African pine nurseries as there is currently no effective method for control (Storer et al. 1998). The problem is compounded by the fact that seedlings provide an important pathway for pathogen propagation (Desperez-Loustau et al. 2006). Symptoms of *F. circinatum* infection may develop in young seedlings, or the fungus may remain latent until stress results in symptom development (Mitchell et al. 2011). The primary symptoms of *F. circinatum* infection include initial tip wilting and discolouration beneath the growing tip (Figure 3.1c). Seedlings later turn brown, with continued needle chlorosis. Severe needle chlorosis results in seedling mortality (Figure 3.1d) (Mitchell et al. 2011).





(a)	(b) Week 1, 2, and 3	(c) Week 4	(d) Week 5
			
A healthy seedling. No visible symptoms of <i>F. circinatum</i> infection.	Seedlings show no visible signs of infection for the first three weeks following inoculation.	Initial symptoms of wilting of needle tips at week four. The ‘falling-off’ of the apical stem is characteristic of <i>F. circinatum</i> infection.	Needle chlorosis five weeks after inoculation, with very few remaining green needles. Several seedlings have already died.

Figure 3.1: Symptom development of *F. circinatum* infection in *P. radiata* seedlings. A healthy seedling (a) is shown as a reference. Symptom expression became more prominent with time, ultimately leading to seedling death.

The inability to readily detect *F. circinatum* infection in young seedlings, and the subsequent mortality of seedlings and young trees in plantations, has dire consequences for the future sustainability of South Africa’s pine trees (Roux et al. 2007). It is hence important to detect the fungus at an early stage of infection in nursery seedlings. The removal of diseased seedlings should alleviate



the development of pitch canker disease at the plantation level. Improved methodologies for early stress detection are thus pertinent, and are a key element in managing the fungal pathogen (Cram & Fraedrich 2009).

Several authors (for example Hoque et al. 1992; Stone et al. 2003; Moshou et al. 2004; Pontius et al. 2005; Jones et al. 2010) have successfully demonstrated the use of spectroscopic data for early stress detection in plants. Early stress detection relies on identifying specific bands (Carter & Miller 1994; Lichtenthaler 1998; West et al. 2003) that correspond to specific physiological responses of the plant in relation to the stress (Chaerle et al. 2007). A specific spectral response can thus be related directly to a specific host-stressor relationship (Mahlein et al. 2012). The ability to relate a spectral response to a specific stressor is significant in the context of this study because the hypothesis that seedlings stressed due to *F. circinatum* infection can be discriminated from seedlings stressed due to physical damage, is tested. Hence, the utility of spectroscopic data offers two key advantages for the application of plant stress detection. Firstly, the high spectral resolution is vital for detecting subtle variations in leaf reflectance (Carter & Knapp 2001; Vigier et al. 2004; Mutanga et al. 2009) which can be associated with a specific stressor. Secondly, repeated spectral measurements allow for the non-destructive collection of multitemporal data (Apan et al. 2005; Sankaran et al. 2010), which may provide useful information with regards to the symptomatic progression of the stressor (Delalieux et al. 2007; Jones et al. 2010).

Although hyperspectral data can provide detailed information on the spectral properties related to plant stress, the inherent high dimensionality of the data makes analysis challenging (Kursa et al. 2010; Pal & Foody 2010). The ‘curse of dimensionality’ often results in reduced classification accuracies arising from the number of samples ( $n$ ) being many times less than the number of features ( $p$ ) (Pal & Foody 2010; Mianji & Zhang 2011). One approach of mitigating the Hughes effect (Hughes 1968) is to reduce the high data dimensionality using feature selection techniques with the aim to improve prediction model accuracy, and to produce an optimal subset of the original feature set in which redundant or irrelevant features have been removed (Hapfelmeier & Ulm 2013). The key advantage of feature selection approaches is that no information relating to individual feature importance is lost (Janecek et al. 2008). Feature selection, or waveband selection in the context of this study, thus involves finding an optimal subset of bands that provides the best classification accuracy.

Recently, several authors have advocated the utility of the random forest (RF) algorithm (Breiman 2001) as a wrapper-based feature selection method. RF is an ensemble of unpruned decision trees (DTs) constructed from bootstrap samples derived from the training data (Breiman 2001). Coupled with a feature selection algorithm, RF provides an efficient means of distinguishing relevant and

irrelevant features (Ismail & Mutanga 2010; Ismail & Mutanga 2011). The algorithm is relatively robust to outliers and noise, and does not over-fit (Biau 2012). Furthermore, it is computationally more efficient than bagging or boosting, is simple to implement, incorporates interactions between predictors, and provides estimates of error, strength, and correlation (Breiman 2001). RF has been successfully applied in a number of hyperspectral studies; see for example Abdel-Rahman et al. (2009), Ismail & Mutanga (2010), Ismail & Mutanga (2011), and Adam et al. (2012). In all of these studies, RF was implemented with a feature selection algorithm, given that RF does not inherently provide for feature selection (Knights et al. 2011), but provides only an importance measure for ranking features (Breiman 2001). For example, Adam et al. (2012) applied RF with a forward variable selection (FVS) technique to identify the important bands for discriminating *Cyperus papyrus* from co-existing species. The FVS technique produced a subset of ten bands from the original 126, equivalent to a 92% reduction in dimensionality. Classification using the subset of ten bands resulted in improved overall classification accuracy. Ismail & Mutanga (2011) used a RF wrapper-based technique on backward variable selection (BVS) (Díaz-Uriarte & Alvarez de Andrés 2006), which yielded a subset of five bands of an original 64 bands. Using only 8% of the original bands resulted in the lowest overall misclassification rate.

A promising wrapper technique embedded with RF is Boruta (Kursa et al. 2010; Kursa & Rudnicki 2010; Kursa & Rudnicki 2011). Unlike the wrapper methods FVS and BVS that aim to find a minimal subset of strongly relevant bands (Blum & Langley 1997; Kohavi & John 1997), Boruta selects bands that are both strongly and weakly relevant (Blum & Langley 1997; Kohavi & John 1997) in terms of providing the best classification accuracy (Kursa & Rudnicki 2011). Kohavi & John (1997) showed that strongly relevant features contribute directly to high model accuracy, whereas weakly relevant features can also contribute to model accuracy. The authors further noted that the selection of relevant features alone does not imply selection of an optimal feature subset. Thus selecting both strongly and weakly relevant bands should provide a model with the highest prediction accuracy.

Several studies have demonstrated the utility of the Boruta algorithm as an efficient technique for feature selection (Kursa et al. 2010; Kursa & Rudnicki 2010; Kursa & Rudnicki 2011) and it has been applied in a number of studies including document indexing (Augereau et al. 2011), microbial metagenomic analyses (Saulnier et al. 2011; Riehle et al. 2012), microarray gene expression studies (Kursa & Rudnicki 2011), and forest biodiversity modelling (Leutner et al. 2012). Leutner et al. (2012) employed Boruta to determine relevant predictive features for modelling forest species richness. This was done in the context of comparing the individual and combined use of LiDAR and hyperspectral data. Boruta reduced the dataset by 72%, from 125 bands to 35 bands. However, the authors assert that using the subset of Boruta-selected bands did not significantly improve the

performance of the final RF model. This is contrary to studies in the other disciplines mentioned above; Augereau et al. (2011), Saulnier et al. (2011), Riehle et al. (2012), Kursa & Rudnicki (2011), and Leutner et al. (2012).

We are not aware of any prior study that has evaluated the efficacy of Boruta for analysing hyperspectral data. In this study, RF was used to model asymptomatic stress in *P. radiata* seedlings infected by *F. circinatum*, using high spectral resolution data. The Boruta algorithm was used to determine the optimal subset of bands, in order to improve RF model classification accuracy. Waveband selection and classification was initially undertaken using the combined classes (i.e. H-I-D), and subsequently undertaken for the class pairs (i.e. H-I and I-D). This study specifically tests the application of the RF and Boruta feature selection algorithm for discriminating healthy and infected *P. radiata* seedlings prior to symptom expression.

## **3.2 MATERIALS AND METHODS**

### **3.2.1 Seedling inoculation**

A total of 150 seedlings were randomly sampled from two trays consisting of three month old *P. radiata* seedlings ( $n = 196$ ). Subsequently, the seedlings were divided into three equal classes ( $n = 50$ ) and labelled as healthy (negative control), artificially damaged (positive control), and infected. A positive control was included to determine if it was possible to discriminate *F. circinatum*-associated stress from stress resulting from physical damage to the seedlings (i.e. artificial wounding). The assumption was that the spectral response of infected seedlings and artificially damaged seedlings was statistically different. For the infected class, seedling inoculation followed the pitch canker fungus (PCF) screening facility best operating practice (Forestry and Agricultural Biotechnology Institute: Pretoria, South Africa) inoculum procedure. Firstly, the apical buds of the seedlings were topped and ten microliters of spore suspension ( $50\,000$  spores. $\text{ml}^{-1}$  in 15% glycerol solution) prepared from the *F. circinatum* isolate (FCC 3579) was then placed onto the topped apical buds. For the artificially damaged class (positive control), only the apical buds of the seedlings were topped.

### **3.2.2 Spectral data acquisition and pre-processing**

Seedlings were monitored for *F. circinatum* symptom development over five weeks following inoculation. Spectroscopic data were collected weekly between 10:00 and 15:00 using an Analytical Spectral Devices (ASD) FieldSpec® Pro FR spectroradiometer. The instrument acquires data in the 350-2500 nm spectral range with a spectral resolution of 3 nm in the 350-1000 nm spectral range and a spectral resolution of 10 nm in the 1000-2500 nm spectral range. The ASD data is then resampled to a spectral resolution of 1 nm (Hatchell 1999). In order to ensure a high signal to noise ratio (S/N),

the instrument was optimised using a Spectralon® white reference Rumpf et al. (2010). The fibre optic cable was attached to a pistol grip mounted onto a tripod and positioned above the sample at the nadir position.

Seedlings were scanned using a 23° field-of-view. Each seedling was rotated through a 360° rotation for five readings (Hatchell 1999) to minimise the effects of shadowing. Spectra were captured weekly per seedling for each of the three classes over a five week period providing a total of 3 750 spectral measurements. The five spectral readings per seedling were then averaged to a single reading per seedling (ASD Inc. 2011) resulting in a total of 750 spectral measurements that were used for analysis. Atmospheric water absorption bands (1350-1460nm and 1790-1960nm) were removed due to the noise resulting from the atmosphere strongly absorbing incident radiation at these wavelengths (Hatchell 1999; Walker 2009).

### 3.2.3 Classification using random forest

RF was developed by Breiman (2001) as an extension of bagging (bootstrap aggregation) trees and is an ensemble of weak unbiased classification or regression trees. Classification using RF is performed by first drawing a bootstrap sample (i.e. with replacement) consisting of approximately two thirds of the original dataset. An unpruned classification tree is then fitted to each bootstrap sample. At each node split, a random subset of possible features ( $mtry$ ) is selected and the final classification is based on a majority vote determined by all trees in the ensemble ( $ntree$ ). Optimisation of the RF algorithm was undertaken following the method proposed by Díaz-Uriarte & Alvarez de Andrés (2006). This method involves using increasing and decreasing factors of the default  $mtry$  value. The values for  $ntree$  were varied from 500 to 2500 by a factor of 500. Within the context of hyperspectral applications this method of optimising the  $mtry$  and  $ntree$  parameters was adopted by Ismail & Mutanga (2010) and Adam et al. (2012).

The remaining one third of the samples (i.e. OOB samples) is used to compute 1) the OOB error, which is an unbiased estimate of the training error and 2) feature importance. The most commonly used measure of RF feature importance is the mean decrease in accuracy (Genuer et al. 2010). Feature importance is calculated by randomly permuting each feature in the OOB sample, and determining the increase in OOB error after each permutation. The higher the increase in OOB error, the more important is the feature (Genuer et al. 2008). The mean decrease in accuracy for a permuted feature  $X^j$  is defined by Equation 3.1 (Genuer et al. 2010):

$$VI(X^j) = \frac{1}{ntree} \sum_t (err_{OOB_{tj}} - err_{OOB_t}) \quad (\text{Equation 3.1})$$

where the summation is over all trees  $t$  of the RF and  $ntree$  is the number of trees, and

$OOB_t$  = features not included in the bootstrap sample used to construct  $t$

$errOOB_t$  = the error / misclassification rate of a single tree  $t$  on the sample  $OOB_t$

$OOB_{tj}$  = permuted sample from randomly permuting the values of  $X^j$  in  $OOB_t$

$errOOB_{tj}$  = the error of the predictor  $t$  on the perturbed sample

RF was implemented using the randomForest library Liaw & Wiener (2002) in the R statistical software (R Development Core Team 2019).

### 3.2.4 Feature selection with Boruta

Waveband selection was undertaken using the Boruta algorithm (Kursa & Rudnicki 2010), to compute an optimal subset of bands for discriminating the three classes (i.e. healthy, infected, and damaged). Boruta is a wrapper embedded with RF that evaluates waveband importance by creating an ensemble of corresponding artificially added ‘shadow’ bands randomly sampled from the original dataset, for each waveband in the dataset. Using this extended dataset, the Boruta algorithm computes and then iteratively compares Z-scores between each waveband and the shadow waveband. Z-scores are based on the mean decrease in accuracy as calculated by the RF algorithm. Waveband importance is assessed by comparing bands in the original dataset with bands in the randomised dataset (Kursa et al. 2010; Kursa & Rudnicki 2010). Many RF models are fit iteratively until either the desired number of iterations is completed (*maxRuns*), or until bands are classified as either *confirmed or rejected*. When the algorithm has stopped due to *maxRuns*, those bands for which importance have not been assigned, are classified as *tentative* (Kursa 2012). In this study, Boruta was run in the default light mode whereby unimportant bands were dropped along with their shadow bands as the algorithm proceeded through the iterations. Using the alternate force mode would result in all shadows being maintained for the entire run of Boruta (Kursa 2012). The force mode is not generally used as it is experimental and has not been fully tested.

### 3.2.5 Classification accuracy

OOB error rates were used to compute the overall classification accuracy using a confusion matrix. A confusion matrix (Kohavi & Provost 1998) shows actual and predicted classifications performed by a classification system. The OOB error rate of a forest is defined by Equation 3.2 (Vincenzi et al. 2011) as:

$$errOOB = \left( \frac{1}{n_{tree}} \right) \sum_{i=1}^{n_{tree}} [y_i - g_{OOB}(X_i)]^2 \quad (\text{Equation 3.2})$$

where  $y_i$  is the  $i$ th element of the training dataset ( $X$ ),  $g_{OOB}$  is the aggregated prediction based on the random trees, and  $(X_i)$  is the bootstrap sample.

Additionally, a discrete multivariate technique called Kappa, or KHAT statistic (Congalton & Green 2009), was used to test whether the values in the confusion matrix are due to true agreement, or chance agreement. In order to test the robustness of the feature selection (i.e. for each iteration) and the classification procedures used in this study, 10-fold CV was performed. CV splits the dataset into equal parts ( $n = 10$ ), with  $n - 1$  parts used as the training dataset and the  $n$ th part used as the test dataset (Hastie et al. 2009).

### 3.3 RESULTS

The results presented in the following sections focus only on week one, week two, and week three given that for these three weeks seedlings were asymptomatic. Week four and week five have been excluded as seedlings showed visual symptoms of *F. circinatum*-induced stress.

#### 3.3.1 Analysis of waveband importance and classification accuracy using random forest

Results for RF with the default *mtry* (the square root of the total number of input bands;  $p = (\sqrt{1769})$ ) and *ntree* (the number of trees to grow in the forest;  $n = 500$ ), and optimised *mtry* and *ntree* values are shown in Table 3.1. The best overall classification results were obtained for week three with an OOB error of 16.67% (CV error = 18.67%) and a KHAT value of 0.75. Optimised *mtry* and *ntree* values provided the best classification results as indicated by both the OOB and CV error rates. RF hyperparameters were subsequently optimised for all subsequent analysis.

Table 3.1: Random forest classification results using all bands ( $n = 1769$ ) for the combined classes. The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT.

Week	Default RF Hyperparameters				Optimised RF Hyperparameters			
	<i>mtry</i>	<i>ntree</i>	OOB error (%)	KHAT	<i>mtry</i>	<i>ntree</i>	OOB error (%)	KHAT
1	42	500	28.67 (28.00)	0.57	210	1000	26.00 (26.67)	0.61
2	42	500	32.00 (35.33)	0.52	210	500	28.67 (32.67)	0.57
<b>3</b>	42	500	<b>17.33 (19.33)</b>	<b>0.74</b>	168	500	<b>16.67 (18.67)</b>	<b>0.75</b>

It is evident from Figure 3.2 that the most important bands for discriminating between the three classes are located in the red-edge region (between 680 and 780 nm), and the NIR region (between 780 and 1200 nm) of the EMS. A limited number of important bands are also located in the SWIR region (between 1200 and 2500 nm) for week three. Bands with the highest mean decrease in accuracy (and therefore the most important bands) were located in the NIR region for week one and week three, whereas for week two, the most important bands are predominately located in the red-edge.



### 3.3.2 Waveband selection and classification using the Boruta algorithm

In order to evaluate the sensitivity of the Boruta algorithm, *n*tree values were varied from 500 to 2500 for each week. It is evident from Table 3.2 that waveband selection is sensitive to changes in the *n*tree value. In total, 15 subsets of bands were evaluated and the most relevant subset of bands were selected based on the classification results as determined by the OOB and CV error rates. Table 3.3 summarises the best subset of bands for each week. Week three provided the best overall classification accuracy with an OOB error of 14% (CV error = 16%) and KHAT of 0.79. These results were achieved using only 107 of the original 1 769 bands. This is equivalent to a 93.95% reduction in dimensionality. The bands selected by the Boruta algorithm are shown in Figure 3.3. Bands in the visible (VIS), red-edge, NIR, and SWIR regions were selected across all three weeks. Table 3.3 provides a detailed summary of the location of these bands. By comparing these results with those achieved for RF waveband importance (Figure 3.2) it is possible to isolate specific regions of the EMS that may be useful to detect asymptomatic stress in *P. radiata* seedlings.

In order to determine if it was possible to discriminate *F. circinatum*-associated stress from stress resulting from physical damage to the seedlings, the following section focusses more specifically on 1) week three since it provided the best classification accuracies, and 2) the H-I, and I-D class pairs. RF with the Boruta feature selection algorithm was implemented for the selected class pairs using the spectral measurements obtained from week three.

The results presented in Table 3.4 show similar classification accuracies for both the H-I (OOB error = 9%, CV error = 11%) and I-D (OOB error = 8%, CV error = 10%) class pairs. These results indicate excellent discrimination potential between the classes. A comparison of the classification results for the combined classes (Table 3.3) and the class pairs (Table 3.4) indicated that better discrimination was possible when evaluating the class pairs as opposed to the combined classes. Improved classification accuracies could thus be obtained for discriminating class pairs as opposed to discriminating the combined classes.

Figure 3.4 shows the location of the relevant bands as selected by the Boruta algorithm for the respective class pairs. More specifically, as shown in Table 3.4, Boruta selected a total of 38 bands that could potentially discriminate the H-I class pair, whereas 40 bands were selected for discriminating the I-D class pair. For the H-I class pair, Boruta selected bands from the VIS ( $n = 1$ ), NIR ( $n = 31$ ), and SWIR ( $n = 6$ ) regions, whereas for the I-D class pair, only bands from the NIR ( $n = 29$ ) and SWIR ( $n = 11$ ) were selected. It is thus evident that discriminating healthy and infected as well as infected and damaged seedlings was predominately dependent on bands in the NIR and SWIR regions.

Table 3.2: Sensitivity of the Boruta algorithm to increasing number of trees in the forest (*ntree*). The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT.

Week	<i>ntree</i>	Number of Boruta bands	OOB error (%)	KHAT
1	500	54	28.00 (26.00)	0.58
	<b>1000</b>	<b>80</b>	<b>24.00 (25.33)</b>	<b>0.64</b>
	1500	95	25.33 (22.67)	0.62
	2000	85	27.33 (25.33)	0.59
	2500	101	24.00 (26.00)	0.64
2	500	39	30.00 (32.00)	0.55
	1000	66	32.00 (28.67)	0.54
	1500	88	29.33 (30.67)	0.56
	<b>2000</b>	<b>82</b>	<b>29.33 (30.67)</b>	<b>0.56</b>
	2500	93	31.33 (32.67)	0.53
3	500	62	14.67 (16.67)	0.78
	1000	78	19.33 (20.00)	0.71
	1500	107	16.00 (17.33)	0.79
	<b>2000</b>	<b>107</b>	<b>14.00 (16.00)</b>	<b>0.79</b>
	2500	125	15.33 (14.00)	0.77

An evaluation of Table 3.3 and Table 3.4 for week three revealed a marked difference in the number of bands required for discriminating *F. circinatum*. For the multiclass setting where all three classes are present, discrimination was possible using 107 bands. However, for the binary setting, where only two classes are present, discrimination was possible using only 38 and 40 bands for the H-I and I-D class pairs respectively. It was thus evident that for a binary classification a reduced subset of bands could be used to discriminate between classes. Furthermore, specific wavelength regions could be identified as relevant for discriminating the class pairs. For example, only 29 bands in the NIR region located between 1118 and 1151 nm, and 11 bands in the SWIR region located between 1337 and 1789 nm are relevant for discriminating infected and damaged seedlings at week three.

### 3.4 DISCUSSION

This study represented a first attempt at selecting an optimal subset of bands from hyperspectral data using the RF and Boruta algorithms. Additionally, the results of this study are significant to the pine forest industry in South Africa given that no such study has previously been undertaken to detect *F. circinatum* infection in *P. radiata* seedlings prior to visual symptom expression. The remote sensing and ML framework presented in this study has potential application for pre-visual seedling screening within a nursery environment. The following sections discuss the results in more detail.



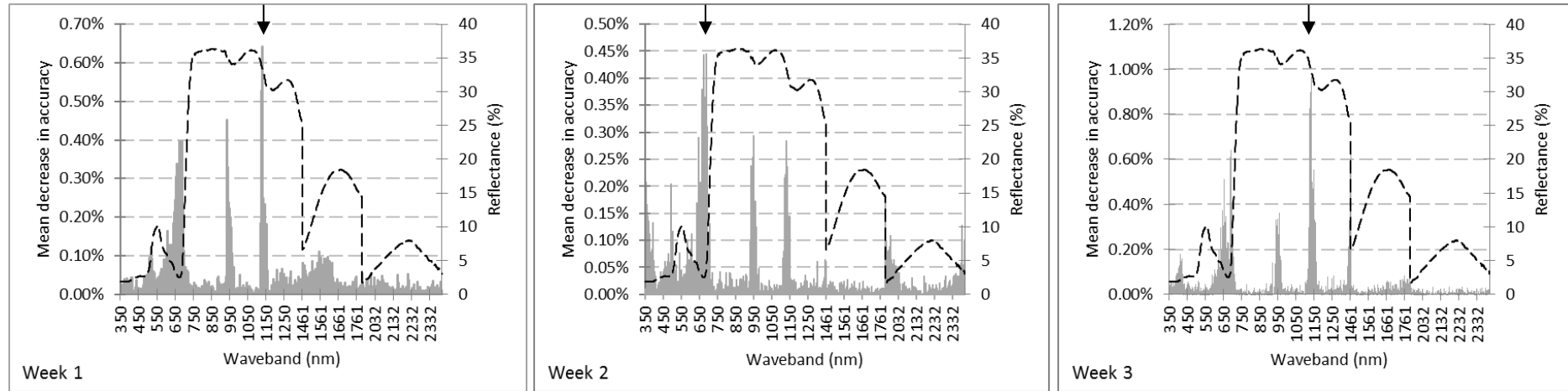


Figure 3.2: Waveband importance as determined by the random forest algorithm using optimised *mtry* and *ntree* values for the combined classes. Waveband importance is indicated by the grey bars. The arrow indicates those bands with the highest mean decrease in accuracy. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference.

Table 3.3: Waveband selection and classification using Boruta-selected bands for the combined classes. The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT.

Week	Number of Boruta bands	% original bands	OOB error (%)	KHAT	VIS (350 nm to 680 nm)		Red-edge (680 nm to 780 nm)		NIR (780 nm to 1200 nm)		SWIR (1200 nm to 2500 nm)	
					Bands	Region	Bands	Region	Bands	Region	Bands	Region
1	80	4.52	24.00 (25.33)	0.64	21	643-680	17	681-699	35	933-1144	7	1498-1588
2	82	4.64	29.33 (30.67)	0.56	30	351-680	15	682-697	34	933-1135	3	1963-1973
3	107	6.05	<b>14.00 (16.00)</b>	<b>0.79</b>	32	400-677	15	683-697	45	933-1153	15	1334-1349

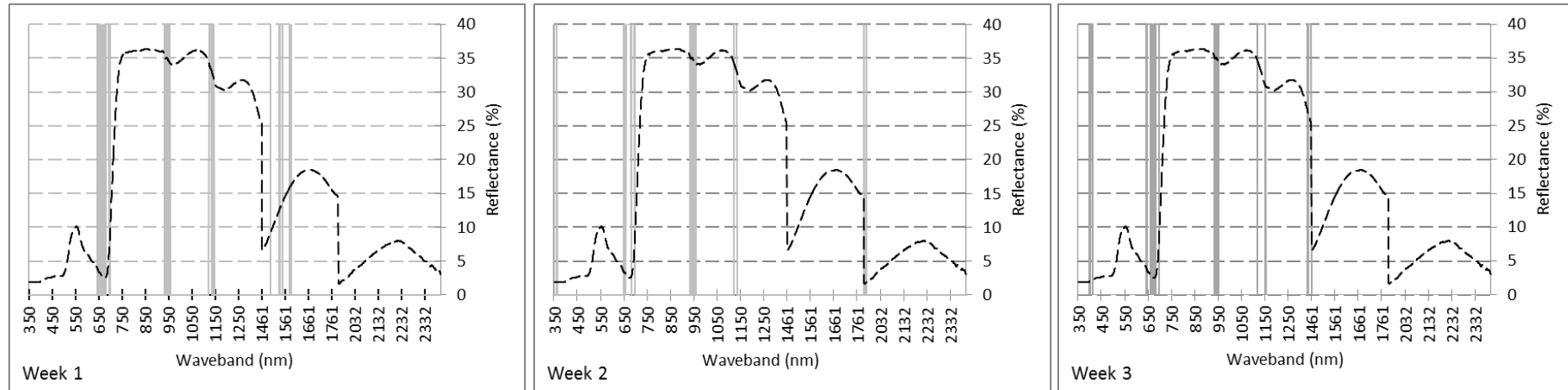


Figure 3.3: Boruta-selected bands for the combined classes. The grey bars indicate the most relevant bands selected by Boruta. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference.

Table 3.4: Waveband selection and classification using Boruta-selected bands for the healthy-infected (H-I), and infected-damaged (I-D) class pairs. The three measures of classification accuracy are the out-of-bag (OOB) error, cross-validation (CV) error (indicated in parentheses), and KHAT.

Class pair	Number of Boruta bands	% original bands	OOB error (%)	KHAT	VIS (350 nm to 680 nm)		Red-edge (680 nm to 780 nm)		NIR (780 nm to 1200 nm)		SWIR (1200 nm to 2500 nm)	
					Bands	Region	Bands	Region	Bands	Region	Bands	Region
H-I	38	2.15	9.00 (11.00)	0.82	1	680	-	-	31	933-1151	6	1344-1349
I-D	40	2.26	8.00 (10.00)	0.84	-	-	-	-	29	1118-1151	11	1337-1789

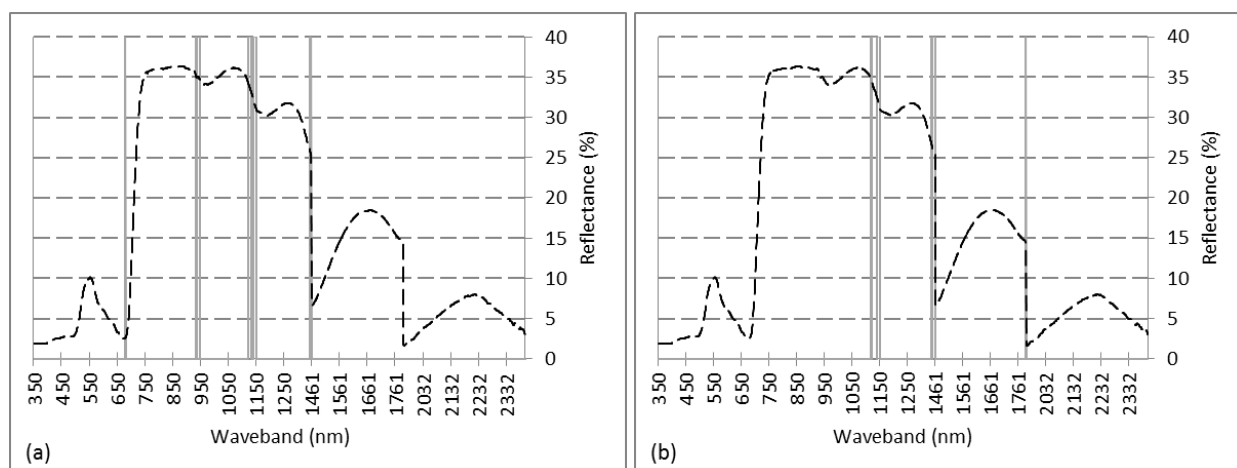


Figure 3.4: Boruta-selected bands for the healthy-infected (H-I) class pair (a) and the infected-damaged (I-D) class pair (b). The grey bars indicate the most relevant bands selected by the Boruta algorithm. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference.

### 3.4.1 Disease progression in *P. radiata* seedlings

An extensive literature search revealed that little is known about the incubation period for *F. circinatum* in *P. radiata* seedlings. However, a study conducted by Gordon et al. (2006) on two year old *P. radiata* tree branches, showed that the incubation period was 31 days, which is the equivalent of approximately four weeks. This incubation period correlates well with the timeframes utilised in this study. No visual symptoms were evident during the first three weeks following inoculation, i.e. week one, two, and three were asymptomatic. At week four, wilting of the growing tip and needle chlorosis were evident (Figure 3.1). By week five, most seedlings had advanced needle chlorosis, with several seedlings having already died.

The earliest removal of infected seedlings from nurseries is critical given that *F. circinatum* is largely responsible for the high mortality of *P. radiata* seedlings in South Africa (Mitchell et al. 2011). High mortality rates reported for *Pinus* species have a major economic impact on the South African forestry industry, with estimated annual losses of 12 million Rand (Morris 2010). Therefore, control of *F. circinatum* within nurseries is the best means of reducing the mortality of field stock (Mitchell et al. 2011). Identifying and immediately removing asymptomatic infected seedlings from nurseries before sowing should help reduce mortality rates and the associated economic losses.

### 3.4.2 Waveband selection and classification using the Boruta algorithm

RF is an efficient technique for analysing hyperspectral data (Ismail et al. 2008; Ham et al. 2011; Ismail & Mutanga 2011). Results of this study confirm the utility of RF as a robust and effective algorithm for the classification of hyperspectral data, where 1)  $n \ll p$  and 2) classes are spectrally similar (Ismail & Mutanga 2011). The results further demonstrate that the use of optimised *mtry* and

*n*tree values produce higher classification accuracies. In this study, the best overall classification results using the RF algorithm were obtained for week three with an OOB error of 16.67% (CV error = 18.67%) and KHAT of 0.75. The higher error rates for week one (OOB error = 26%, CV error = 26.67%) and week two (OOB error = 28.67%, CV error = 32.67%) are likely due to similar spectral reflectance of the classes during the early stages of the experiment. As the disease progressed, biochemical changes within the plant likely resulted in changes in leaf reflectance (Pontius et al. 2005), and subsequently greater spectral dissimilarity. The plant's biochemical response to stress associated with the physical damage likely prompted a different response to that of infection by *F. circinatum*, but also resulting in changes in leaf reflectance. The difference in spectral response would have been more pronounced in seedlings where the fungus remained latent (Mitchell et al. 2011) prior to stress resulting from artificial damage. This could explain the good classification results obtained at week three.

The limitation of the RF algorithm is its inability to select an optimal subset of features for classification. The Boruta algorithm simplified the classification process by reducing the original dataset ( $n = 1\ 769$ ) by as much as 93.95% (i.e. for week 3). The Boruta subset of bands comprised both strongly relevant and weakly relevant bands (Kohavi & John 1997; Kursa & Rudnicki 2011). Strongly relevant bands selected by Boruta correlated well with RF-selected bands, i.e. bands with the highest mean decrease in accuracy (Figure 3.5). Boruta additionally selected weakly relevant bands, i.e. bands that were not strongly relevant, but together with the strongly relevant bands provided the best classification result (Kohavi & John 1997). For week three, strongly relevant bands were located in the red (between 643 and 677 nm), red-edge, NIR, and SWIR regions. Boruta additionally selected weakly relevant bands located in the blue region between 400 and 412 nm, and in the NIR between 933 and 951 nm.

Waveband selection using Boruta showed improved overall classification performance, across all three weeks for both the combined classes and class pairs. Although not previously used for hyperspectral data analysis, the results obtained in this study compare favourably with studies by Kursa et al. (2010) and Kursa & Rudnicki (2010) who demonstrated that the Boruta algorithm provided significant reduction in dimensionality, and often a decrease in error rates.

In this study, the combined classes showed a decrease in error rates for week three, with the OOB error decreasing from 16.67% to 14% and the CV error decreasing from 18.67% to 16% when using RF with the Boruta feature selection algorithm. These classification results were achieved using only 107 bands, equivalent to 6.05% of the original dataset, illustrating the effect of feature selection on classification accuracy. Similar results were obtained for the H-I and I-D class pair analysis with OOB error rates of 9% and 8%, and CV error rates of 11% and 10% achieved using only 38 (2.15%)

and 40 (2.26%) bands respectively. These results also compare favourably to previous forest health studies that have employed wrappers embedded with RF for the classification of hyperspectral data (Ismail & Mutanga 2010; Ismail & Mutanga 2011; Adam et al. 2013).

The hypothesis in this study was that the spectral response of infected seedlings and artificially damaged seedlings was statistically different. The results obtained in this study established that the spectral response of an infected seedling is different from the spectral response of a damaged seedling. Hence it is possible to discriminate *F. circinatum*-associated stress in *P. radiata* seedlings from stress resulting from physical damage to the seedlings. Furthermore, the results confirm that a spectral signature can be related to a specific stressor (i.e. *F. circinatum*) and that individual bands located at specific regions of the EMS have the potential to discriminate healthy and stressed seedlings.

### **3.4.3 Asymptomatic detection of *F. circinatum* infection in *P. radiata* seedlings**

For the combined classes, discrimination of healthy and asymptomatic seedlings was possible using Boruta-selected bands located across the VIS, red-edge, NIR, and SWIR regions for all three weeks (Table 3.2). The bands were located predominately in the VIS and NIR regions for all three weeks. For the H-I class pair, discrimination of healthy and asymptomatic seedlings at week three was possible using Boruta-selected bands across the VIS, NIR, and SWIR regions, with the majority of the bands located in the NIR ( $n = 31$ ) and SWIR ( $n = 6$ ) regions. A single waveband was located in the VIS region at 680 nm. Similar results were obtained for the I-D class pair with Boruta-selected bands dominating the NIR ( $n = 29$ ) and SWIR ( $n = 11$ ) regions. These results compare favourably to previous studies (Naidu et al. 2009; Abdel-Rahman et al. 2010; Grisham et al. 2010) that have attempted to detect asymptomatic stress using high spectral resolution data, and concluded that the most important bands for discrimination were located in the VIS and NIR regions. These results are also similar to Ismail et al. (2008) who found that bands in the VIS and NIR regions showed the highest potential in discriminating the healthy, green, and red, attack stages of *S. noctilio* in *P. patula* trees. Abdel-Rahman et al. (2010) further demonstrated the significance of bands in the red-edge region for discriminating low, medium, and severe sugarcane thrips damage.

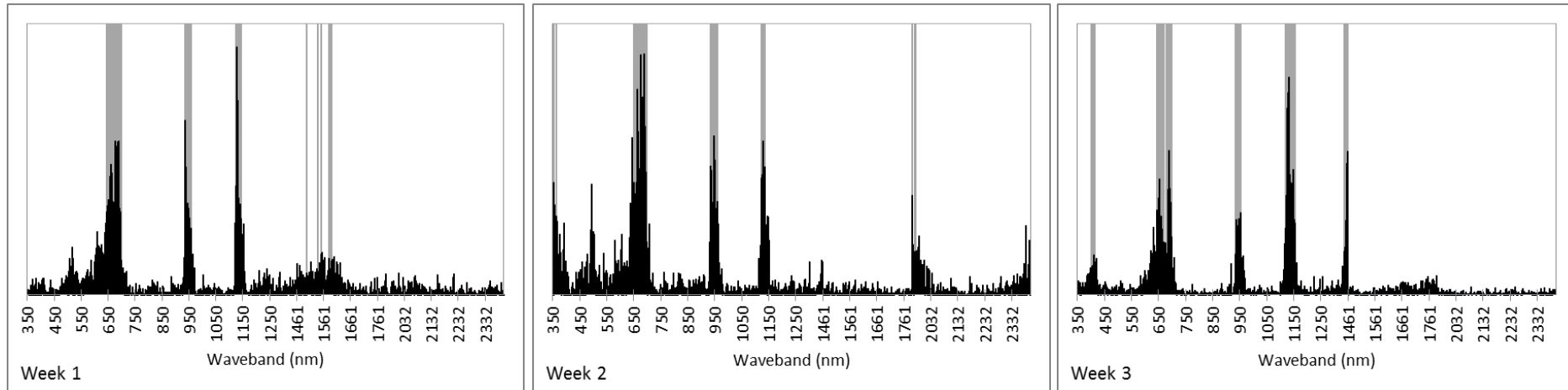


Figure 3.5: Comparing random forest waveband importance and Boruta waveband selection for the combined classes. The black bars represent the mean decrease in accuracy for random forest, and the grey bars represent bands selected using Boruta. The X-axis is indicative of waveband importance (in the case of RF) / relevance (in the case of Boruta).

For the I-D class pair, important bands were located only in the NIR ( $n = 29$ ) and SWIR ( $n = 11$ ) regions (Table 3.3), suggesting that asymptomatic seedlings could be discriminated using bands located in these regions. These results are similar to Ismail & Mutanga (2011) who showed that bands in the SWIR region had the greatest potential for discriminating the early stages of *S. noctilio* infection in *P. patula*. The ability to discriminate *F. circinatum*-induced stress and stress due to damage in *P. radiata* seedlings is significant toward separating damaged and infected seedlings in which the fungus may be latent (Mitchell et al. 2011).

### 3.5 CONCLUSIONS

This study set out to model asymptomatic stress in three month old *P. radiata* seedlings that were inoculated with the fungal pathogen *F. circinatum*. Overall, the results showed that it is possible to detect asymptomatic stress in *P. radiata* seedlings using hyperspectral data. This was possible during the three weeks following inoculation, with week three providing the best discrimination between the classes. Waveband selection using the Boruta algorithm embedded with the RF classification algorithm produced a more accurate model for 1) discriminating healthy and infected seedlings, 2) providing insight into which bands could potentially discriminate healthy and stressed seedlings and 3) discriminating the spectral response of *F. circinatum*-related stress in *P. radiata* seedlings and stress due to physical damage of the seedlings. However, the Boruta-selected bands still need to be validated through further studies. Such validation studies will reinforce the utility of Boruta (compared with other feature selection algorithms) and the identification of selected wavebands for discriminating *F. circinatum*-related stress in *P. radiata* seedlings.

## CHAPTER 4: SELECTING THE MOST IMPORTANT BANDS FOR MODELLING *FUSARIUM* STRESS: COMPARING RANDOM FOREST WRAPPERS

Poona NK, Van Niekerk A, Nadel RL & Ismail R 2016. Random forest (RF) wrappers for waveband selection and classification of hyperspectral data. *Applied Spectroscopy*, 70:2, 322-333.

Poona NK & Ismail R 2013. Reducing hyperspectral data dimensionality using random forest based wrappers. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS2013), held 21-26 July 2013, Melbourne, Australia.

### Abstract

Hyperspectral data collected using a field spectroradiometer was used to model asymptomatic stress in *P. radiata* and *P. patula* seedlings infected with the pathogen *F. circinatum*. Spectral data were analysed using the random forest algorithm. In order to improve the classification accuracy of the model, subsets of bands were selected using three feature selection algorithms: (1) Boruta; (2) recursive feature elimination (RFE); and (3) area under the receiver operating characteristic curve of the random forest (AUC-RF). Results highlighted the robustness of the above feature selection methods when used in conjunction with the random forest algorithm for analysing hyperspectral data. Overall, the Boruta feature selection algorithm provided the best results. When discriminating *F. circinatum* stress in *P. radiata* seedlings, Boruta-selected bands ( $n = 69$ ) yielded the best overall classification accuracies (training error of 17%, independent test error of 17% and an AUC value of 0.91). Classification results were, however, significantly lower for *P. patula* seedlings, with a training error of 24%, independent test error of 38%, and an AUC value of 0.65. A hybrid selection method that utilises combinations of bands selected from the three feature selection algorithms was also tested. The hybrid method showed an improvement in classification accuracies for *P. patula*, but no improvement for *P. radiata*. The results of this study provide impetus towards implementing a hyperspectral framework for detecting stress within nursery environments.

**Keywords:** random forest (RF), *Fusarium circinatum*, feature selection, *Pinus*



## 4.1 INTRODUCTION

*F. circinatum* (syn. *Gibberella circinata*) (Nirenberg & O'Donnell 1998) is a fungal pathogen of significant economic importance to the international forestry industry as the pathogen results in the premature death of seedlings of selected *Pinus* species (Hammerbacher et al. 2009; Mitchell et al. 2011). *F. circinatum*-related seedling mortality severely affect both nursery production and in field performance for up to two years following site establishment, despite attempts to control the pathogen through improved silvicultural practices (Crous 2005; Mitchell et al. 2011). Seedling susceptibility studies have shown that *P. radiata* and *P. patula* are highly susceptible to *F. circinatum* infection (Hodge & Dvorak 2000; Wingfield et al. 2002). The development of an accurate and efficient detection and monitoring methodology is thus essential to reduce the high levels of seedling mortality present within nurseries (Cram & Fraedrich 2009). This is particularly significant as *F. circinatum*-infected seedlings often remain asymptomatic following infection (Poona & Ismail 2014).

The utility of hyperspectral data offers a viable and non-destructive technique for assessing plant health during the asymptomatic stage of infection (Ismail & Mutanga 2011; Poona & Ismail 2014). Exploiting hyperspectral data for asymptomatic infection is not without its difficulties and challenges. The inherent “curse of dimensionality” results in reduced classifier performance, due to the number of bands or features being many times more than the number of samples (Pal & Foody 2010; Mianji & Zhang 2011). Researchers have consequently employed feature selection algorithms, embedded with classification algorithms such as random forest (RF), in an attempt to mitigate the Hughes Effect (Hughes 1968), thereby reducing the number of bands to an optimal subset of bands, and improving overall classification accuracy (Ismail & Mutanga 2011; Adam et al. 2012; Poona & Ismail 2014).

The RF algorithm has been successfully applied in a wide array of studies for the classification of high dimensional data including microarray data analysis (Díaz-Uriarte & Alvarez de Andrés 2006), quantitative structure-activity relationship modelling (Svetnik et al. 2003), genome-wide association studies (Touw et al. 2012), and hyperspectral data analysis (Ismail & Mutanga 2011; Adam et al. 2012; Poona & Ismail 2014). RF is a bagging (i.e. bootstrap aggregation) ensemble process in which classification trees are grown from random samples derived from the training data (Breiman 2001). RF uses bagging as well as random variable selection for constructing individual trees in the ensemble (Díaz-Uriarte & Alvarez de Andrés 2006). The RF algorithm provides several advantages as an ensemble classifier: 1) the algorithm incorporates interactions between features (Hapfelmeier & Ulm 2013), 2) is

computationally more efficient than bagging or boosting, 3) is robust to overfitting, and 4) provides an estimate of variable strength and internal error estimation (Breiman 2001).

Researchers have explored the utility of feature selection algorithms in an attempt to improve RF classification accuracies. Feature selection algorithms are categorised by the manner in which they evaluate feature subsets. Filters are applied as a pre-processing step in which waveband selection is based on the intrinsic properties of the dataset, and therefore independent of the classification algorithm. Unlike filters, wrappers are embedded within the classification algorithm and waveband selection is thus dependent on the selected classification algorithm (Guyon & Elisseeff 2003; Saeys et al. 2007). Ismail & Mutanga (2011) demonstrated that the use of a wrapper produced a greater reduction in dimensionality coupled with improved classification accuracy when compared to using a filter. Additionally, several studies, for example Adam et al. (2012) and Poona & Ismail (2014) have employed wrappers, successfully demonstrating that feature selection within a RF framework results in improved model performance as well as an optimal subset of variables.

Several popular wrapper methodologies that use RF as the base algorithm, have been applied for the analysis of high dimensional data. A backwards elimination procedure termed RFE was introduced by Díaz-Uriarte & Alvarez de Andrés (2006). RFE has been implemented in several hyperspectral studies. For example, Ismail & Mutanga (2010) used RFE to derive a nested subset ( $n = 2$ ) from an original dataset ( $n = 9$ ) of spectral parameters that could best predict *S. noctilio*-induced water stress in *P. patula*. Using RFE also improved the overall RF performance, increasing  $R^2$  values from 0.73 to 0.76. Ismail & Mutanga (2011) later employed RFE to select an optimal subset of bands ( $n = 5$ ) from an original dataset ( $n = 64$ ) for discriminating healthy and asymptomatic *S. noctilio* infestations in *P. patula* trees. Using the RFE selected bands with RF produced the lowest overall misclassification rate of 6.14%.

The Boruta feature selection algorithm (Kursa et al. 2010; Kursa & Rudnicki 2010; Kursa & Rudnicki 2011; Rudnicki et al. 2015) has been used in an array of studies including microbial metagenomic analyses (Riehle et al. 2012), forest biodiversity modelling (Leutner et al. 2012), and boreal forest habitat classification (Räsänen et al. 2014). Within a hyperspectral context, Poona & Ismail (2014) employed the Boruta algorithm to derive an optimal subset of bands ( $n = 38$ ) from an original dataset consisting of 1 769 bands to discriminate healthy and stressed *P. radiata* seedlings infected by the fungal pathogen *F. circinatum*. Their study demonstrated that although stress in *P. radiata* seedlings could be detected one week following artificial inoculation, the best discrimination was possible at three weeks post

inoculation. The study further demonstrated the utility of the Boruta algorithm within a hyperspectral context, which yielded a classification accuracy of 91% using a subset of 38 bands.

Another RF-based wrapper is the area under the receiver operating characteristic curve of the random forest (AUC-RF) (Calle et al. 2011). AUC-RF has been used only in a limited number of studies. Calle et al. (2011) employed the AUC-RF and RFE algorithm to determine an optimal subset of single nucleotide polymorphisms (SNPs) in inflammatory genes associated with bladder cancer risk. In a similar study, López de Maturana et al. (2013) also demonstrated the use of the AUC-RF algorithm, together with the Bayesian threshold Least Absolute Shrinkage and Selection Operator model, to determine SNPs associated with bladder cancer risk. Within a hyperspectral context, this study represents a first attempt at employing the AUC-RF algorithm for feature selection and classification of hyperspectral data.

The aim of this study was to (i) develop empirical models using hyperspectral data and the RF algorithm to discriminate asymptomatic *F. circinatum* infections in *P. radiata* and *P. patula* seedlings within a nursery environment, (ii) test the utility of the Boruta, RFE, and AUC-RF algorithms for selecting an optimal subset of bands that could potentially yield the best classification accuracies, and (iii) test whether combinations of bands selected by the different feature selection algorithms could improve classification accuracies.

## **4.2 MATERIALS AND METHODS**

### **4.2.1 Symptoms of *F. circinatum* infection**

As described by Mitchell et al. (2011), pine seedlings infected with *F. circinatum* in a nursery environment, display symptoms of initial shoot tip wilting and seedling discolouration beneath the growing tip (Figure 4.1(c)). Seedlings also show symptoms of needle chlorosis that are coupled with lesions at the root collar (Figure 4.1 (d)) (Mitchell et al. 2011).

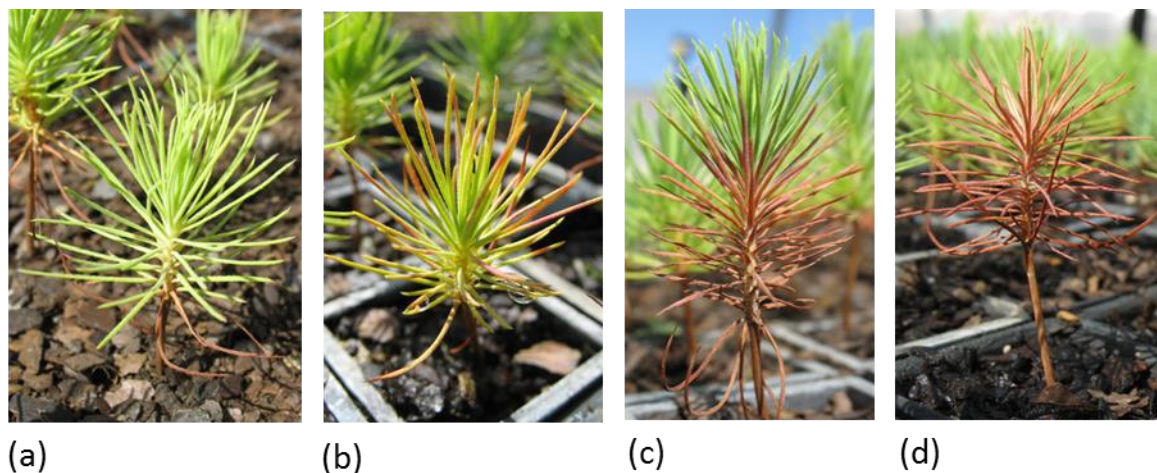


Figure 4.1: Initial symptoms associated with *F. circinatum* infection within a nursery environment. A healthy *P. patula* seedling (a) is shown as a reference. Images courtesy of Institute for Commercial Forestry Research, Pietermaritzburg, South Africa.

#### 4.2.2 Seedling inoculation

Two independent experiments were set up for this study:

##### *Experiment 1:*

A total of 100 seedlings were randomly sampled from two trays of 3-month old *P. radiata* seedlings ( $n = 196$ ). The sampled seedlings were subsequently divided into two equal classes ( $n = 50$ ) and labelled healthy and infected. For the infected class, seedling inoculation followed the PCF screening facility best operating practice inoculum procedure (Forestry and Agricultural Biotechnology Institute: University of Pretoria, Pretoria, South Africa). This procedure involved first topping the apical buds with a sterile razor blade, and then placing a 10  $\mu\text{l}$  spore suspension ( $50\,000$  spores  $\text{ml}^{-1}$ ) of *F. circinatum* isolate (FCC 3579) onto the topped apical buds.

##### *Experiment 2:*

A total of 100 seedlings were randomly sampled from five trays of 2-week old *P. patula* seedlings ( $n = 490$ ). The sampled seedlings were subsequently divided into two equal classes ( $n = 50$ ) and labelled healthy and infected. For the infected class, a 50  $\mu\text{l}$  spore suspension ( $50\,000$  spores  $\text{ml}^{-1}$ ) of *F. circinatum* isolate (FCC 3579) was placed into the soil medium close to the stem of each seedling. Inoculating the soil medium with a higher spore load was undertaken in order to potentially simulate a more ‘natural’ mode of infection within the nursery environment as opposed to artificially wounding the seedlings (i.e. Experiment 1).

### 4.2.3 Spectral data acquisition

For both *P. radiata* and *P. patula* seedlings, hyperspectral data were collected weekly, over a three week period, between 10:00 and 15:00 using a FieldSpec® 3 FR Spectroradiometer (Analytical Spectral Devices, Boulder, CO) following inoculation. For the purpose of this study, only data from week one and week two was used, given that the focus of this study is on modelling asymptomatic stress. The spectroradiometer acquires data in the 350-2500 nm spectral range with a spectral resolution of 3 nm in the VIS-NIR region (350-1000 nm) and 10 nm in the NIR-SWIR region (1000-2500 nm). Reflectance measurements were calibrated using a Spectralon® white reference panel (Curtiss 2013). Five spectral measurements were captured per seedling using the 23° field-of-view (Poona & Ismail 2014). Spectra were later averaged to a single reading per seedling (ASD Inc 2011). The spectral data were then pre-processed to remove atmospheric water absorption bands (1350-1460 nm and 1790-1960 nm) (Hatchell 1999; Walker 2009).

### 4.2.4 Statistical analysis

#### 4.2.4.1 Random forest

RF (Breiman 2001) is an ensemble learner that grows hundreds of unpruned classification trees (*n<sub>tree</sub>*) from bootstrap samples of the original dataset. The bootstrap sample consists of approximately two thirds of the original dataset (known as the “in bag” dataset). A single classification tree is then fitted to each sample. Each classification tree in the ensemble is split into nodes using a random subset of selected features (*m<sub>try</sub>*). From the selected *m<sub>try</sub>* features, a single feature is used to split the nodes. Trees are grown using bagging as well as random variable selection. This process results in trees with low bias and low correlation amongst individual trees (Díaz-Uriarte & Alvarez de Andrés 2006). The final classification is then based on a majority vote of predictions of all trees in the ensemble (Breiman 2001; Goldstein et al. 2011).

Following the recommendation of several hyperspectral studies (Adam et al. 2012; Poona & Ismail 2014), optimised *m<sub>try</sub>* and *n<sub>tree</sub>* values were used for all analyses in this study. The RF hyperparameters were optimised by using increasing and decreasing factors of the default *m<sub>try</sub>* value (the square root of the total number of input bands;  $p = \sqrt{1769}$ ), and varying the *n<sub>tree</sub>* value by a factor of 500 up to a maximum of 2500 (Díaz-Uriarte & Alvarez de Andrés 2006). This approach has been adopted by Ismail & Mutanga (2010), Adam et al. (2012), and Poona & Ismail (2014) for the analysis of hyperspectral data.

The one third of the samples (i.e. out-of-bag (OOB) samples) not used to grow trees is used to compute 1) the OOB error, which provides an internal measure of classification performance, and 2) variable importance, which is based on either the mean decrease in accuracy or the Gini index (Breiman 2001). The RF algorithm was implemented using the randomForest library (Liaw & Wiener 2002) in the R statistical software (R Development Core Team 2019).

#### 4.2.4.2 Random forest waveband importance

A useful by-product of RF is variable importance, which is used for feature ranking. RF variable importance thus provides insight into which bands or set of bands is most relevant for classification (Breiman 2001). The feature selection algorithms used in this study make use of either the Gini importance or the permutation importance as a measure of waveband importance. The Gini importance is computed as the sum of all decreases in Gini impurity of a splitting variable normalised by the number of trees in the forest. Gini impurity ( $\Delta GI(t)$ ) is defined by Equation 4.1 (Kawakubo & Toshida 2012):

$$\Delta GI(t) = P_t GI(t) - P_L GI(t_L) - P_R GI(t_R) \quad (\text{Equation 4.1})$$

where,  $P_t$  is the number of observations before the split,  $P_L$  is the number of observations on the left side of the split,  $P_R$  is the number of observations on the right side of the split,  $GI(t_L)$  is a Gini index on the left side of the node,  $GI(t_R)$  is a Gini index on the right side of the node, and ( $GI(t)$ ) is the Gini index defined by:

$$GI(t) = 1 - \sum_k p(k|t)^2 \quad (\text{Equation 4.2})$$

where,  $p(k|t)$  is the rate at which class  $k$  is correctly classified at node  $t$ .

The permutation importance is calculated as mean decrease in classification accuracy using the OOB observations. It is computed by measuring the change in prediction accuracy when the OOB observations are randomly permuted compared to the original observations. The difference in prediction accuracy is then averaged over all trees to compute the permutation importance value (Goldstein et al. 2011; Touw et al. 2012). The permutation importance for a feature  $X_j$  is defined by Equation 4.3 (Genuer et al. 2010):

$$VI(X_j) = \frac{1}{ntree} \sum_t (errOOB_{tj} - errOOB_t) \quad (\text{Equation 4.3})$$

where,  $ntree$  is the number of trees in the forest,  $OOB_{tj}$  is the permuted sample from randomly permuting the values of  $X_j$  in  $OOB_t$ ,  $errOOB_{tj}$  is the error of the predictor  $t$  on the perturbed sample,  $OOB_t$  is the features not included in the bootstrap sample used to construct  $t$ , and  $errOOB_t$  is the error / misclassification rate of a single tree  $t$  on the sample  $OOB_t$ .



#### 4.2.4.3 Wrapper-based waveband selection within a random forest framework

Díaz-Uriarte & Alvarez de Andrés (2006) introduced the RFE algorithm, which is based on a backward feature selection process. The optimal subset of bands is derived by iteratively building multiple random forests models and successively eliminating the least important bands at each iteration, as determined by the RF permutation importance measure. The optimal subset of bands is then defined by the lowest OOB error rate that is within  $\mu$  standard errors of the best model obtained from fitting all the RF models. Setting  $\mu = 0$  results in a subset of bands with the smallest OOB error, whereas setting  $\mu = 1$  results in the smallest subset of bands whose OOB error falls within the sampling error providing for the best result (Díaz-Uriarte & Alvarez de Andrés 2006). For this study  $\mu = 1$  was used, as this provided a more stable result with smaller subsets of bands (Ismail & Mutanga 2011). The RFE algorithm was implemented using the varSelRF library (Díaz-Uriarte 2012) in the R statistical software (R Development Core Team 2019).

The AUC-RF algorithm (Calle et al. 2011) is an area under the receiver operating characteristic curve based method used in conjunction with RF. AUC-RF first builds a random forest using all bands, with waveband importance determined by the Gini index. The algorithm subsequently incorporates a backwards elimination procedure, similar to the procedure used by Díaz-Uriarte & Alvarez de Andrés (2006). However, rather than simply computing the OOB error, the AUC-RF algorithm computes the area under the ROC curve based on the OOB predictions, also known as the OOB AUC (Calle et al. 2011). At each iteration a new random forest model is built and the least important bands successively eliminated. The process continues until, theoretically, only one waveband remains. The final subset of bands is defined by the highest OOB AUC value and is calculated using Equation 4.4 (Calle et al. 2011):

$$AUC = \frac{1}{n_0} \left( \bar{r}_1 - \frac{n_1}{2} - \frac{1}{2} \right) \quad (\text{Equation 4.4})$$

where  $n_0$  is the number of healthy samples,  $n_1$  is the number of infected samples, and  $\bar{r}_1$  is the mean rank of infected samples.

In order to compute the corrected estimation of classification accuracy of the selected bands and the probability of selection for each waveband, a five-fold CV analysis is performed (repeated 20 times). AUC-RF was implemented using the AUCCRF library (Urrea & Calle 2013) in the R statistical software (R Development Core Team 2019).

The Boruta algorithm (Kursa & Rudnicki 2010) creates an ensemble of shadow bands by randomly sampling the original dataset. Each shadow waveband corresponds to a waveband in the original dataset. In order to evaluate waveband importance, Boruta first computes Z-scores (based on the permutation

importance measure) for each waveband and its corresponding shadow waveband. Using these Z-scores, the algorithm then compares each waveband in the original dataset with bands in the randomised shadow dataset (Kursa et al. 2010; Kursa & Rudnicki 2010). Boruta iteratively fits RF models until 1) the algorithm has completed the specified number of RF iterations (*maxRuns*) or 2) bands are classified as *confirmed*, *rejected*, or *tentative*, based on the waveband's importance (Z-score) (Kursa 2012). Boruta was run in light mode whereby unimportant bands together with their shadow bands were dropped through the iterations of the algorithm (Poona & Ismail 2014). The Boruta algorithm was implemented using the Boruta library (Kursa 2012) in the R statistical software (R Development Core Team 2019).

#### 4.2.4.4 Classification of Pinus seedlings using a combination of bands

It was possible to derive combinations of Boruta, RFE, and AUC-RF selected bands that could be used to classify *P. radiata* and *P. patula*, and potentially improve the overall classification accuracy. The subsets were derived using (i) an intersection, and (ii) a union, of the bands selected by the three feature selection algorithms. An intersection involved combining only those bands selected by both algorithm 1 and algorithm 2, whereas a union involved combining all bands selected by algorithm 1 and algorithm 2. The following combinations of selected bands were considered (i) Boruta and RFE, (ii) Boruta and AUC-RF, and (iii) RFE and AUC-RF.

#### 4.2.5 Classification accuracy

For all feature selection algorithms, week one was used as the training dataset, whereas week two was used as the independent test dataset, in order to provide an independent estimate of model accuracy. However, the classification accuracy was first evaluated using the OOB error rate (or training error) and a confusion matrix. The OOB error rate was computed using Equation 4.5 (Vincenzi et al. 2011):

$$errOOB = \left(\frac{1}{ntree}\right) \sum_{i=1}^{ntree} [y_i - g_{OOB}(X_i)]^2 \quad (\text{Equation 4.5})$$

where  $y_i$  is the  $i$ th element of the training dataset ( $X$ ),  $g_{OOB}$  is the aggregated prediction based on the random trees, and  $(X_i)$  is the bootstrap sample.

Additionally, the area under the receiver operating characteristic curve (AUC) was used to assess classifier performance based on the independent test dataset. The receiver operating characteristic curve is a two dimensional plot of the true positive rate and false positive rate of a classifier, derived from the confusion matrix. The AUC value ranges between zero and one, with a realistic classifier having an AUC value of no less than 0.5 (Fawcett 2003).



## 4.3 RESULTS

### 4.3.1 Waveband selection and classification of *Pinus* seedlings

#### *Experiment 1: P. radiata*

The best subset of bands selected by the Boruta, RFE, and AUC-RF algorithms for the classification of *P. radiata* seedlings is shown in Figure 4.2. Boruta selected bands ( $n = 69$ ) across the VIS (350 nm to 680 nm), red-edge (680 nm to 780 nm), and NIR (780 nm to 1200 nm) regions, whereas RFE selected bands ( $n = 371$ ) across the VIS, red-edge, NIR, and SWIR (1200 nm to 2500 nm) regions. Similar to RFE, AUC-RF selected bands ( $n = 187$ ) across all four regions. It is clear from Figure 4.2 that there is a significant difference in the number and location of bands selected by the Boruta, RFE, and AUC-RF algorithms for the classification of *P. radiata* seedlings. However, there was also significant overlap in the location of certain bands selected by Boruta, RFE, and AUC-RF. For example, in the red-edge region, Boruta selected 17 bands, with RFE and AUC-RF selecting 26 and 20 bands respectively. Table 4.1 provides detail with regards to the specific ranges of the selected bands.

For the classification of *P. radiata* seedlings, using the RF algorithm and all bands ( $n = 1\,769$ ), an optimised *mtry* value of 210, and an optimised *n tree* value of 500 was used. Optimised *mtry* and *n tree* values were also used for Boruta, RFE, and AUC-RF (results not shown). It is evident from Table 4.1 that the best overall classification results were obtained using the Boruta algorithm for the classification of *P. radiata*. A training error of 17%, independent test error of 17%, and an AUC value of 0.91 was obtained using a subset of only 69 bands, that is, 3.90% of the original dataset ( $n = 1\,769$ ). Similar results were achieved using the AUC-RF algorithm, with a training error of 19%, independent test error of 21%, and an AUC value of 0.91. However, AUC-RF selected a total of 187 bands, that is 10.57% of the original dataset ( $n = 1\,769$ ), which is a much larger subset of bands than the Boruta-selected subset. RFE produced the worst results with a training error of 21%, independent test error of 23%, and an AUC value of 0.88. RFE selected a total of 371 bands (i.e. 20.97% of the original dataset), which is significantly larger than the Boruta-selected subset. Overall, Boruta, RFE, and AUC-RF produced a significant reduction in dimensionality. However, only Boruta provided improved classification results compared to using RF and all the bands (training error of 20%, independent test error of 20%, and an AUC value of 0.91).

#### *Experiment 2: P. patula*

Figure 4.3 shows the best subset of bands selected by the Boruta ( $n = 21$ ), RFE ( $n = 62$ ), and AUC-RF ( $n = 23$ ) algorithms for the classification of *P. patula* seedlings. For *P. patula*, all three algorithms

selected bands across the VIS, NIR, and SWIR regions. No bands were selected within the red-edge region. Unlike *P. radiata*, the number and location of bands selected by the Boruta, RFE, and AUC-RF algorithms for the classification of *P. patula* seedlings were very similar. The greatest number of bands selected by Boruta ( $n = 15$ ), RFE ( $n = 43$ ) and AUC-RF ( $n = 14$ ) were located in the SWIR region. Analogous to *P. radiata*, there was overlap in the location of certain bands selected by Boruta, RFE, and AUC-RF (Table 4.1). For example, in the NIR region, Boruta selected four bands with RFE and AUC-RF selecting 11 and 3 bands respectively. Table 4.1 provides detail with regards to the specific ranges of the selected bands.

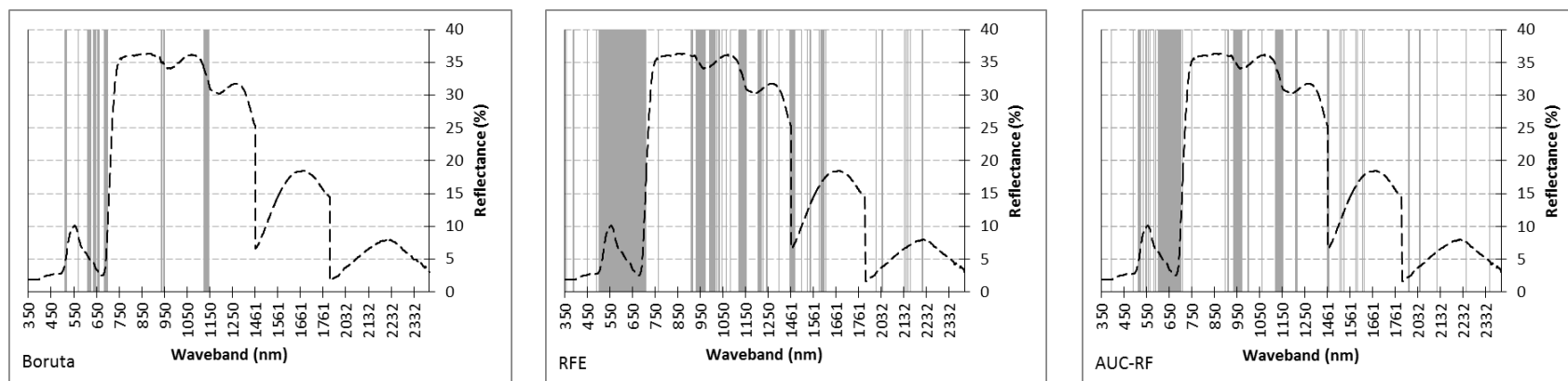


Figure 4.2: Boruta-selected bands, recursive feature elimination-selected bands, and area under the receiver operating characteristic curve of the random forest-selected bands for *P. radiata*. The grey bars indicate the most relevant bands selected by Boruta, recursive feature elimination, and area under the receiver operating characteristic curve of the random forest, respectively. The spectral curve represents the mean signature of a healthy *P. radiata* seedling, and is used as a reference.

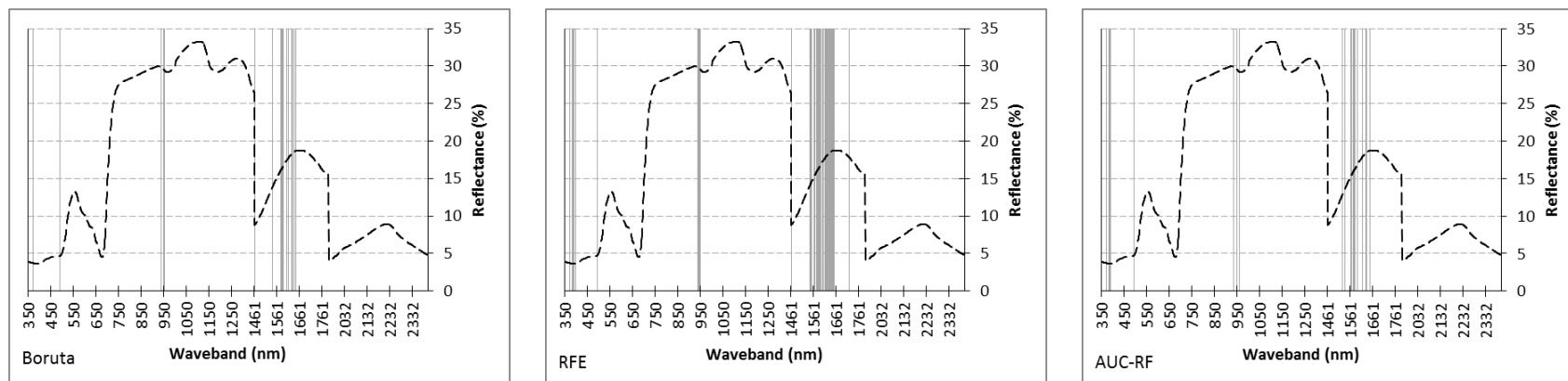


Figure 4.3: Boruta-selected bands, recursive feature elimination-selected bands, and area under the receiver operating characteristic curve of the random forest-selected bands for *P. patula*. The grey bars indicate the most relevant bands selected by Boruta, recursive feature elimination, and area under the receiver operating characteristic curve of the random forest respectively. The spectral curve represents the mean signature of a healthy *P. patula* seedling, and is used as a reference.

For the classification of *P. patula* seedlings, using the RF algorithm and all bands ( $n = 1\ 769$ ), an optimised *mtry* value of 336, and an optimised *nree* value of 500 was used. Optimised *mtry* and *nree* values were also used for Boruta, RFE, and AUC-RF (results not shown). Once again, using the Boruta algorithm produced the best overall results (Table 4.1). A training error of 24%, independent test error of 38%, and an AUC value of 0.65 was obtained using a subset of 21 bands (1.19% of the original dataset). AUC-RF produced similar results with a training error of 24%, independent test error of 40%, and an AUC value of 0.62. RFE again produced the worst results with a training error of 26%, independent test error of 42%, and an AUC value of 0.61. For all three algorithms there was a significant difference between their respective training and independent test error. Similar to *P. radiata*, Boruta, RFE, and AUC-RF produced a significant reduction in dimensionality. However, only Boruta provided improved classification results compared to using RF and all the bands (training error of 27%, independent test error of 39%, and an AUC value of 0.65).

#### 4.3.2 Classification of *Pinus* seedlings using a combination of bands

The hybrid waveband selection approach adopted in this study was implemented as a union as well as an intersection of bands selected by three different feature selection algorithms. For the *P. radiata* dataset, neither the union nor intersection of selected bands resulted in improved classification accuracies. The best overall classification accuracy was obtained using the Boruta subset of selected bands. Similarly, for the *P. patula* dataset, the intersection of selected bands yielded no increase in classification performance. However, a union of the Boruta and RFE selected bands yielded an increase in classification accuracy, when compared to using the Boruta subset. The results using combinations of selected bands for (i) Boruta and RFE, (ii) Boruta and AUC-RF, and (iii) RFE and AUC-RF are shown in Table 4.2.

Table 4.1: Waveband selection and classification using Boruta, recursive feature elimination (RFE), and area under the receiver operating characteristic curve of the random forest (AUC-RF)-selected bands. The three measures of classification accuracy include the out-of-bag (OOB) error, independent test error, and the area under the receiver operating characteristic curve (AUC).

Species	Algorithm	Number of bands	Training (OOB error)	Testing (Independent Error Estimate)	AUC	VIS (350 nm to 680 nm)		Red-edge (680 nm to 780 nm)		NIR (780 nm to 1200 nm)		SWIR (1200 nm to 2500 nm)	
						Bands	Region	Bands	Region	Bands	Region	Bands	Region
<i>P. radiata</i>	RF	1769	20.00	20.00	0.91	331	350-680	100	681-780	420	781-1200	918	1201-2400
	Boruta	<b>69</b>	<b>17.00</b>	<b>17.00</b>	<b>0.91</b>	22	509-661	17	682-699	30	933-1143	-	-
	RFE	371	21.00	23.00	0.88	179	355-680	26	681-761	107	907-1152	59	1203-2214
	AUC-RF	187	19.00	21.00	0.91	71	392-680	20	681-748	75	895-1149	21	1206-2346
<i>P. patula</i>	RF	1769	27.00	39.00	0.65	331	350-680	100	681-780	420	781-1200	918	1201-2400
	Boruta	<b>21</b>	<b>24.00</b>	<b>38.00</b>	<b>0.65</b>	2	369-486	-	-	4	935-951	15	1349-1642
	RFE	62	26.00	42.00	0.61	8	368-492	-	-	11	934-948	43	1349-1717
	AUC-RF	23	24.00	40.00	0.62	6	368-492	-	-	3	933-957	14	1524-1648

Table 4.2: Classification accuracies using common subsets of bands selected by combining (i) Boruta and recursive feature elimination (RFE), (ii) Boruta and area under the receiver operating characteristic curve of the random forest (AUC-RF), and (iii) recursive feature elimination (RFE) and area under the receiver operating characteristic curve of the random forest (AUC-RF). The three measures of classification accuracy include the out-of-bag (OOB) error, independent test error, and the area under the receiver operating characteristic curve (AUC).

Hybrid waveband selection		Species	Number of bands	Training error	Testing (Independent Error Estimate)	AUC	VIS (350 nm to 680 nm)		Red-edge (680 nm to 780 nm)		NIR (780 nm to 1200 nm)		SWIR (1200 nm to 2500 nm)	
							Bands	Region	Bands	Region	Bands	Region	Bands	Region
Intersect	Boruta-RFE	<i>P. radiata</i>	<b>69</b>	<b>17.00</b>	<b>17.00</b>	<b>0.91</b>	22	509-661	17	682-699	30	933-1143	-	-
	Boruta-AUC-RF		63	18.00	21.00	0.88	17	509-660	16	682-699	30	933-1143	-	-
	RFE-AUC-RF		168	17.00	19.00	0.91	69	508-680	19	681-704	71	907-1149	9	1206-1612
	Boruta-RFE	<i>P. patula</i>	<b>7</b>	<b>24.00</b>	<b>37.00</b>	<b>0.65</b>	-	-	-	-	2	935-946	5	1349-1635
	Boruta-AUC-RF		3	30.00	43.00	0.61	-	-	-	-	1	946	2	1576-1628
	RFE-AUC-RF		15	26.00	41.00	0.62	4	368-492	-	-	1	946	10	1562-1648
Union	Boruta-RFE	<i>P. radiata</i>	<b>371</b>	<b>18.00</b>	<b>18.00</b>	<b>0.91</b>	179	355-680	26	681-761	107	907-1152	59	1203-2214
	Boruta-AUC-RF		193	18.00	22.00	0.91	76	392-680	21	981-748	75	895-1149	21	1206-2346
	RFE-AUC-RF		390	17.00	20.00	0.91	181	355-680	27	681-761	111	895-1152	71	1203-2346

	Boruta-RFE	<i>P. patula</i>	<b>76</b>	<b>20.00</b>	<b>35.00</b>	<b>0.65</b>	10	368-492	-	-	13	934-951	53	1349-1717
	Boruta-AUC-RF		41	21.00	40.00	0.64	8	368-492	-	-	6	933-957	27	1349-1648
	RFE-AUC-RF		70	21.00	39.00	0.63	10	368-492	-	-	13	933-957	47	1349-1717

### *Experiment 1: P. radiata*

Using an intersection of the Boruta and RFE selected bands ( $n = 69$ ) produced the same results (training error of 17%, independent test error of 17%, and an AUC value of 0.91) as using the subset of Boruta-selected bands. Using an intersection of RFE and AUC-RF selected bands produced an increase in classification accuracy when compared to using RF with all bands, but no better than using the best feature selection model (i.e. Boruta). The intersection of Boruta and AUC-RF selected bands produced lower accuracies compared to using RF with all bands, as well as using the Boruta subset of bands. Using a union of the Boruta and RFE selected bands produced similar results (training error of 18%, independent test error of 18%, and an AUC value of 0.91) to using the best feature selection model but with a significant increase in dimensionality ( $n = 371$ ). However, the union of Boruta and AUC-RF selected bands, and RFE and AUC-RF selected bands produced no improvement in classification accuracy when compared to using the best feature selection model.

### *Experiment 2: P. patula*

Results indicate that an intersection of the Boruta and RFE selected bands resulted in a marked reduction in dimensionality ( $n = 7$ ) coupled with a marginal (1%) increase in classification accuracy (training error of 24%, independent test error of 37%, and an AUC value of 0.65) when compared to using RF with all bands as well as the best feature selection model (i.e. Boruta). The intersection of Boruta and AUC-RF, and RFE and AUC-RF selected bands both produced lower classification accuracies compared to using RF with all bands as well as using the Boruta subset. Using a union of the Boruta and RFE selected bands resulted in an improvement in classification accuracy (training error of 20%, independent test error of 35%, and an AUC value of 0.65) when compared to using RF with all bands as well as using the Boruta subset. Using a union of Boruta and AUC-RF, and RFE and AUC-RF bands did not result in an improvement in classification accuracy.

## **4.4 DISCUSSION**

This study evaluated the utility of hyperspectral data for the discrimination of *F. circinatum* infections. More specifically, the study evaluated the use of RF, together with the Boruta, RFE, and AUC-RF feature selection algorithms, to derive an optimal subset of bands that could be used to discriminate healthy and infected *P. radiata* and *P. patula* seedlings. Several studies (Ismail & Mutanga 2011; Adam et al. 2012; Riehle et al. 2012; Abdel-Rahman et al. 2014; Poona & Ismail 2014) have successfully illustrated the value of RF wrappers for the analysis of hyperspectral data. In all of these studies, the RF wrapper



framework significantly reduced the dimensionality of the original dataset by identifying an optimal subset of bands thereby simplifying the modelling process (i.e. reducing model complexity), and ultimately improving classification performance. Overall results from this study show that the use of all three feature selection algorithms resulted in a significant decrease in dimensionality, however only Boruta-selected bands produced improve classification accuracies for both experiments.

#### 4.4.1 Hyperspectral dimensionality reduction and classification accuracies

Using the Boruta algorithm reduced the original *P. radiata* dataset by as much as 96.10% and the *P. patula* dataset by 98.81%. These results confirm the findings of Kurska & Rudnicki (2011), Saulnier et al. (2011), Riehle et al. (2012), and Poona & Ismail (2014) who demonstrated the value of the Boruta algorithm with RF to provide significant reduction in dimensionality coupled with increased classification accuracies. Using the Boruta-selected bands with RF resulted in higher classification accuracies for the *P. radiata* dataset (training error of 17%, independent test error of 17%, and an AUC value of 0.91) and for the *P. patula* dataset (training error of 24%, independent test error of 38%, and an AUC value of 0.65) compared to the results obtained when using all the bands with RF. However, classification accuracies were lower than expected for the *P. patula* seedlings. For the *P. radiata* dataset (Experiment 1), it was expected that the fungus would infect the seedlings, followed by rapid progression of the disease, given that (i) the pathogen was introduced directly into the plant, and (ii) *P. radiata* is more susceptible to *F. circinatum* infection (Mitchell et al. 2011). Physiological changes within the seedlings, and concomitant changes in their spectral signature, were thus detectable (Poona & Ismail 2014). However, for the *P. patula* dataset (Experiment 2), seedling infection and subsequent physiological changes were much slower, even though a higher spore load was used for inoculation. The lower level of infection by the fungus can be attributed to (i) the mode of inoculation, i.e. indirect inoculation via the soil medium requiring “natural” wounds for the pathogen to enter the seedling as opposed to direct inoculation via wounding whereby wounds were inoculated with the pathogen, and (ii) *P. patula* being less susceptible than *P. radiata* to *F. circinatum* infection.

The RFE algorithm has been used in several hyperspectral studies (Ismail & Mutanga 2010; Dye et al. 2011), providing for a significant decrease in dimensionality, coupled with high classification performance. In this study, for Experiment 1 and Experiment 2, RFE produced a significantly reduced subset of bands, albeit no improvement in classification accuracies when compared to using RF with all bands as well as using the Boruta subset of bands. Furthermore, the RFE algorithm was significantly outperformed by the Boruta and AUC-RF algorithms, which both produced smaller subsets and better

classification accuracies. For the *P. radiata* dataset, Boruta produced the smallest subset ( $n = 69$ ), whereas RFE produced the largest subset ( $n = 371$ ) of bands. Similarly, for the *P. patula* dataset, Boruta produced the smallest subset ( $n = 21$ ), whereas RFE produced the largest subset ( $n = 62$ ) of bands. Classification results for both the *P. radiata* and *P. patula* dataset were also higher for Boruta, when compared to RFE.

The AUC-RF algorithm provided significant dimensionality reduction and comparative accuracies to using RF with all bands for the classification of *P. radiata* and *P. patula*. For the *P. radiata* dataset (Experiment 1), AUC-RF produced a smaller subset ( $n = 187$ ) as well as comparative accuracies (training error of 19%, independent test error of 21%, and an AUC value of 0.91) when compared to using RF with all bands. For the *P. patula* dataset (Experiment 2), AUC-RF produced a significantly smaller subset ( $n = 23$ ) as well as comparative accuracies (training error of 24%, independent test error of 40%, and an AUC value of 0.62) when compared to using RF with all bands and using the Boruta subset. These results compare favourably to Calle et al. (2011) who demonstrated the utility of the AUC-RF algorithm, and its improved performance compared to using the RFE algorithm. However, the results of Experiment 1 and Experiment 2 indicate that the Boruta algorithm provided the best performing models when compared to using RF with all bands, RFE, as well as AUC-RF. These results confirm the robustness of the Boruta algorithm for waveband selection from hyperspectral data. The following section will focus on the bands selected by the Boruta algorithm.

#### 4.4.2 Waveband selection using the Boruta algorithm

During the early stages of infection, i.e. when the plant is asymptomatic, plants undergo physiological changes such as a decrease in photosynthetic rate, resulting in a change in chlorophyll content (West et al. 2003). This change in chlorophyll content manifests as increased reflectance in the VIS, and a shift in the red-edge toward shorter wavelengths (West et al. 2003). Consequently, the spectral response of a stressed plant would be different from that of a healthy plant. A unique spectral response can thus be derived for stressed plants (Chávez et al. 2012). Furthermore, it may be possible to elucidate the use of specific bands for the detection of a specific stressor, given that physiological changes are often characteristic of a specific pathogen (Mahlein et al. 2012). However, in this study the location of bands selected for *P. radiata* and *P. patula* in response to the same stress agent, differed for the two species.

For the *P. radiata* dataset, Boruta selected bands in the VIS, red-edge, and NIR region, with the most number of bands in the NIR region ( $n = 30$ ). However, no bands were selected in the SWIR region. Bands located in the red-edge region are important, as they provide an indication of plant stress, usually

resulting from changes in leaf chlorophyll concentration (Das et al. 2014). Several studies (Riggins et al. 2011; Buddenbaum et al. 2012; Masaitis et al. 2013) have confirmed the importance of bands in the red-edge as an indicator of plant stress. Contrary to the bands selected for *P. radiata*, Boruta selected bands in the VIS, NIR, and SWIR regions, with no bands selected in the red-edge region for the *P. patula* experiment. Additionally, the greatest number of bands ( $n = 15$ ) were selected in the SWIR region. The SWIR region is an important indicator of leaf moisture content (Zhang et al. 2012), and is sensitive to changes in plant water status (Ismail & Mutanga 2010). Bands in the SWIR region would thus be important toward indicating plant stress (Buddenbaum et al. 2012), due to reduced moisture content of leaves.

These results clearly illustrate the variability in waveband selection for the two *Pinus* species in response to *F. circinatum*. Similarly, Poulos et al. (2012) found that spectral reflectance varied significantly, within the VIS and NIR regions, between four *Pinus* species in response to water stress. This study further demonstrated that the spectral response to the same stressor was different for the two species under investigation. Further physiological studies are required to validate these results.

#### 4.4.3 Classification using a hybrid selection of bands

Several approaches to hybrid feature selection have been noted from the literature (Skurichina & Duin 2005; Tsai & Hsiao 2010; Hsu et al. 2011; Li et al. 2011; Hu et al. 2015). Skurichina & Duin (2005) used forward feature selection, random feature selection, and PCA to select subsets of autofluorescence spectral data, which was later combined using weighted majority voting, the mean rule, and decision templates. Li et al. (2011) adopted a genetic algorithm (GA) support vector machine hybrid feature selection approach to select an optimal subset of HYDICE data. The authors additionally used band grouping based on conditional mutual information to further reduce data dimensionality. There was thus potential for hybrid feature selection methods that combine the results from different feature selection algorithms for the analysis of hyperspectral data.

This study adopted a hybrid approach by combining bands using a union and an intersection of bands selected by the Boruta, RFE, and AUC-RF algorithms. For the *P. radiata* dataset (Experiment 1), using combinations of selected bands resulted in no improvement in classification accuracies. For the *P. patula* dataset (Experiment 2), the intersection of selected bands yielded no increase in classification performance. However, using a union of the Boruta and RFE selected bands yielded a significant increase in classification accuracy, when compared to using the best performing model, i.e. the Boruta subset. Tsai & Hsiao (2010) combined features selected using PCA, GA, and CART by a union and an

intersection method to predict stock prices. Their results showed that the hybrid approach yielded higher prediction accuracy and lower prediction error when compared to using features selected by a single algorithm. Results obtained using a hybrid feature selection approach illustrates the potential for adopting such an approach for future studies. Additionally, other feature selection algorithms need to be tested, and their results combined, for the analysis of hyperspectral data.

## 4.5 CONCLUSIONS

The aim of this study was to evaluate the utility of the Boruta, RFE, and AUC-RF feature selection algorithms, with RF, for the analysis of hyperspectral data. The results show that the Boruta algorithm outperformed both the RFE and AUC-RF algorithms. Additionally, Boruta provided significant reduction in dimensionality coupled with increased classification accuracy when compared to using all the bands with RF. In the context of this study, the three feature selection algorithms selected different bands for the classification of healthy and stressed seedlings. However, for the classification of the *P. patula* dataset, a hybrid selection of bands by the Boruta and RFE algorithms resulted in an increase in classification performance. This demonstrates that features selected by a single algorithm may not necessarily be the optimal subset for classification, and thus requires further investigation.

## CHAPTER 5: OBLIQUE TREE-BASED MODELS FOR DISCRIMINATING *FUSARIUM* STRESS

Poona NK, Van Niekerk A & Ismail R 2016. Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data. *Sensors* 16, 1918. doi:10.3390/s16111918.

### Abstract

Ensemble classifiers are widely used for the classification of spectroscopic data. In this regard, the random forest (RF) ensemble has been successfully applied in an array of applications, and has proven to be robust in handling high dimensional data. More recently, several variants of the traditional RF algorithm including rotation forest (rotF) and oblique random forest (oRF) have been applied to classifying high dimensional data. This study compared the traditional RF, rotF, and oRF (using three different splitting rules, i.e., ridge regression, partial least squares, and support vector machine (SVM) for the classification of healthy and infected *P. radiata* seedlings using high dimensional spectroscopic data. The robustness of these five ensemble classifiers to reduced spectral resolution was further tested, by spectral resampling (binning) of the original spectral bands. The results showed that the three oblique random forest ensembles outperformed both the traditional RF and rotF ensembles. Additionally, the rotF ensemble proved to be the least robust of the five ensembles tested. Spectral resampling of the original bands provided mixed results. Nevertheless, the results demonstrate that using spectral resampled bands is a promising approach to classifying asymptomatic stress in *P. radiata* seedlings

**Keywords:** hyperspectral data, oblique tree-based ensembles, spectral resampling, *Pinus radiata*

## 5.1 INTRODUCTION

Hyperspectral data are characterised by a large number of contiguous bands, ranging from the visible through to the SWIR portion of the EMS (Goetz 2009). For the analysis of plant stress, the high spectral resolution allows for the detection and quantification of a plant's physiological response to stress (Chaerle et al. 2007). This physiological response is exhibited as subtle variations in a plant's spectral response, providing the basis for developing stress detection models (Poona & Ismail 2014; Poona et al. 2016a). Hyperspectral data subsequently provides the opportunity to readily monitor pest and disease stress in agricultural crops and forestry, as demonstrated by Ismail & Mutanga (2011), Abdel-Rahman et al. (2013), Poona & Ismail (2014), Poona et al. (2016a), and others.

The utility of hyperspectral data, especially spectroscopic data, is well established in the remote sensing domain for pest and disease detection. For example, the VNIR spectrum has been particularly useful for the detection of stress in agricultural crops. Chávez et al. (2012) used the 350 nm to 850 nm spectral range to detect bacterial wilt infection caused by *Ralstonia solanacearum* in potato crops. Similarly, Huang et al. (2012) employed leaf and canopy VNIR reflectance data (325 nm to 1075 nm) to detect damage in rice crops caused by *Cnaphalocrocis medinalis*. Within a forestry context, Zhang et al. (2012) used the full spectral range (350 nm to 2500 nm) to model degradation in *Avicennia germinans* and *Rhizophora mangle*. The VNIR and SWIR ranges were also utilised by Poona & Ismail (2014) for modelling asymptomatic *F. circinatum* stress in *P. radiata* seedlings. However, spectroscopic data are highly correlated and there is an a priori assumption that most of the bands will be redundant with only a few key bands producing the best result; see for example Poona & Ismail (2014) and Poona et al. (2016a). Additionally, the limited number of samples ( $n$ ) available coupled with the large number of bands ( $p$ ) presents a statistical challenge (Pal & Foody 2010; Mianji & Zhang 2011).

The random forest (RF) algorithm (Breiman 2001) is particularly well suited for addressing the challenges posed by high dimensional spectral data; see for example Ismail & Mutanga (2011), Adam et al. (2012), Poona & Ismail (2014), and Poona et al. (2016a). Random forest reduces bias (systematic error term independent of the training sample) as well as variance (error due to variability associated with the training sample) by creating unpruned trees thus keeping bias low, and uses randomisation for controlling the diversity between trees in the ensemble (Do et al. 2010). Randomisation is introduced into the ensemble by creating trees using bootstrap aggregation with replacement of samples, as well as for selecting variables that will be used for node splitting (Díaz-Uriarte & Alvarez de Andrés 2006).

However, RF suffers from two primary limitations. First, tree construction is based on a single feature being selected for node splitting. Such trees may be inefficient in dealing with feature dependencies likely inherent in high dimensional spectral data (Do et al. 2010). Second, the majority of current implementations of the RF algorithm utilises orthogonal splits based on univariate decision trees (DTs). According to Menze et al. (2011) the decision boundary generated from orthogonal splits of univariate trees may not be optimal for handling high dimensional spectral data. The argument is that a staircase or box-like decision boundary generated by univariate splits may not be optimal for highly correlated data, such as spectroscopic data, because the data may appear inseparable when their marginal distributions are evaluated (Menze et al. 2011). Building on the initial recommendation of Breiman (2001), Menze et al. (2011) advocated the creation of multivariate DTs by applying a supervised model to learn the splitting rule that results in oblique boundaries rather than the geometrical constrained boundary of orthogonal trees. The only preceding remote sensing study to employ oRF is by Bassa et al. (2016) for land cover and land use mapping.

Research by Do et al. (2010) on 15 high dimensional datasets showed that oRF using a support vector machine (SVM) as the node splitting model (oRFsvm) produced higher classification accuracies compared with using the traditional RF and SVM. Overall findings showed that using the oRFsvm model resulted in an improvement in the mean classification accuracy of 3.57% and 6.35% when compared with the traditional RF and SVM classifiers respectively. Similarly, Menze et al. (2011) compared the oblique version of RF together with seven other classifiers, including RF and SVM, for the classification of high dimensional spectral data. Overall results showed that oRF outperformed all classifiers, with oRF using ridge regression providing the best results.

A related oblique tree-based ensemble approach is rotF (Rodríguez et al. 2006). Unlike oRF that uses supervised models to determine the optimal split direction, rotF applies PCA on bootstrap samples to derive the optimal rotation of the axes for node splitting. rotF encourages diversity in the model through random subset selection and using PCA for feature selection. High accuracy is sought through preserving the discriminatory information of the training data by retaining all the principal components (Rodríguez et al. 2006). Within a remote sensing context, Kavzoglu et al. (2015) applied rotF for the classification of multispectral WorldView-2 data highlighting its superior performance over the RF, SVM, and nearest neighbour algorithms. Du et al. (2015) also found that rotF outperformed RF and SVM when applied to the classification of fully polarimetric synthetic aperture radar imagery. Two studies (Xia et al. 2014;



Xia et al. 2015) applied rotF for the classification of AVIRIS, ROSIS, and Digital Airborne Imaging Spectrometer (DAIS) data. Results showed that rotF outperformed all classifiers including RF and SVM. Several studies have successfully applied hyperspectral data for asymptomatic stress detection. For example, Grisham et al. (2010) used multitemporal spectroscopic data with LDA to detect sugarcane yellow leaf in sugarcane plantations, caused by *Polerovirus*. Two studies (Calderón et al. 2013; Calderón et al. 2015) applied high resolution hyperspectral imagery with LDA and SVM classifiers to discriminate verticillium wilt severity in olive plantations, caused by *Verticillium dahliae*. De Castro et al. (2015) used spectroscopic data with ANOVA and neural network classifiers to model laurel wilt severity in avocado crops caused by *Raffaelea lauricola*. Only two studies (Poona & Ismail 2014; Poona et al. 2016a) have previously investigated the use of hyperspectral data for modelling *F. circinatum* stress in *P. radiata*, and discriminating healthy and stressed seedlings.

Poona & Ismail (2014) successfully demonstrated the use of the RF ensemble for modelling asymptomatic stress in *P. radiata* seedlings. The authors applied RF with the Boruta algorithm (Kursa et al. 2010; Kursa & Rudnicki 2010) for waveband selection and classification of healthy, infected, and damaged *P. radiata* seedlings. Results of their study indicated that hyperspectral data can successfully discriminate *F. circinatum* stress (discrimination of healthy and infected seedlings was achieved with accuracies above 80%). The authors further demonstrated that selected bands can potentially be used to discriminate stress with improved accuracy. Poona et al. (2016a) confirmed the findings of Poona & Ismail (2014) and additionally showed that a combination of selected bands could be used for modelling *F. circinatum* stress in *P. radiata* and *P. patula* seedlings.

It is within this context that the utility of the RF, oRF, and rotF ensembles for the classification of hyperspectral data were evaluated. The study was undertaken as a series of experiments. First, the five ensemble classifiers, i.e., RF, rotF, and oRF (with ridge regression, partial least squares, and SVM as the node splitting models) was tested using all hyperspectral bands ( $n = 1769$ ). Then, the effect of decreasing the spectral resolution on the classification performance of the five ensemble classifiers was evaluated. More specifically, the RF, rotF, and oRF ensemble classifiers were applied to modelling asymptomatic stress in *P. radiata* seedlings associated with *F. circinatum* infection.

## 5.2 MATERIALS AND METHODS

### 5.2.1 *F. circinatum*



*F. circinatum* (synonym *Gibberella circinata*) (Nirenberg & O'Donnell 1998) is a fungal plant pathogen that is now endemic in South African nurseries (Porter et al. 2009). It is one of the most significant pathogens to infect *Pinus* seedlings worldwide (Coutinho et al. 2007), with *P. radiata* being highly susceptible (Wingfield et al. 2008). Within the nursery environment, *Pinus* seedlings often succumb to *F. circinatum* infection. Initial symptoms include wilting and discolouration of the growing tip, with death of the root tips and collar rot observed in later stages of infection. Fungal growth on the seedling stem may be visible at an advanced stage of infection (Mitchell et al. 2011). Britz et al. (2005) note that *F. circinatum* is the most significant of pathogens infecting *Pinus*, with the fungus now prevalent in *P. radiata* plantations across the Western Cape Province of South Africa (Coutinho et al. 2007).

### 5.2.2 Seedling inoculation

A total of 100 seedlings were randomly sampled from two trays of 3-month old *P. radiata* seedlings ( $n = 196$ ). The seedlings were subsequently divided into two equal classes ( $n = 50$ ) labelled healthy and infected. For the infected class, seedling inoculation followed the PCF Screening Facility Best Operating Practice (Forestry and Agricultural Biotechnology Institute: Pretoria, South Africa) inoculum procedure. This procedure involved first topping the apical buds, followed by placing a 10  $\mu\text{L}$  spore suspension (50 000 spores  $\text{mL}^{-1}$ ) of *F. circinatum* isolate (FCC 3579) onto the topped apical buds. Seedlings were kept in a greenhouse for the duration of the study.

### 5.2.3 Spectroscopic data acquisition

Spectral data were collected weekly between 10:00 and 15:00 using a FieldSpec<sup>®</sup> Pro FR Spectroradiometer (Analytical Spectral Devices, Boulder, CO, USA) over a three week period following inoculation. The instrument acquires data in the 350–2500 nm spectral range with a spectral resolution of 3 nm in the VIS-NIR region (350 nm to 1000 nm) and 10 nm in the NIR-SWIR region (1000 nm to 2500 nm). Reflectance measurements were calibrated using a Spectralon<sup>®</sup> white reference panel (Curtiss 2013). Five spectral measurements were captured per seedling using the 23° field-of-view (Poona & Ismail 2014; Poona et al. 2016a). The experimental setup of the spectroradiometer for all data collection is shown in Figure 5.1. The setup provided for consistent (controlled) background reflectance for all measurements.

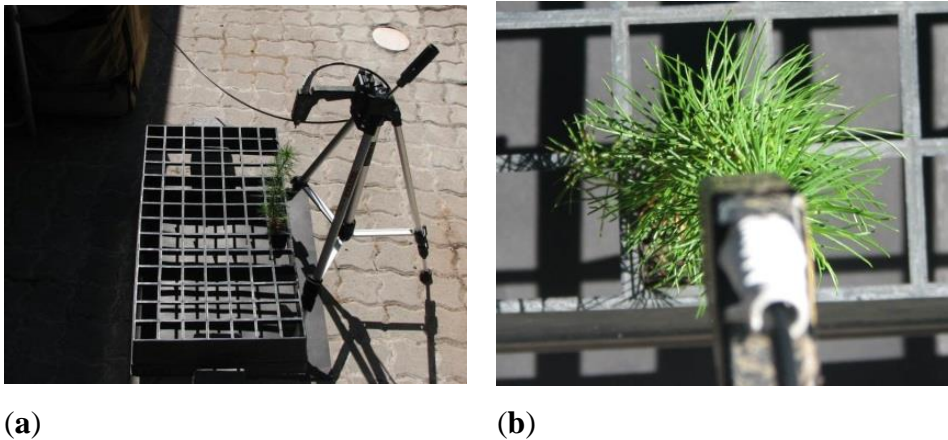


Figure 5.1: Experimental setup of the spectroradiometer used for spectral data collection (a) showing the orientation (nadir view) of the pistol relative to the seedling (b).

Spectra were later averaged to a single reading per seedling (ASD Inc. 2011). The spectral data were then pre-processed to remove atmospheric water absorption bands (1350–1460 nm and 1790–1960 nm) (Hatchell 1999; Walker 2009), and noisy bands (2401–2500 nm). Figure 5.2 illustrates the mean spectral signature of the healthy and infected seedlings captured at week one.

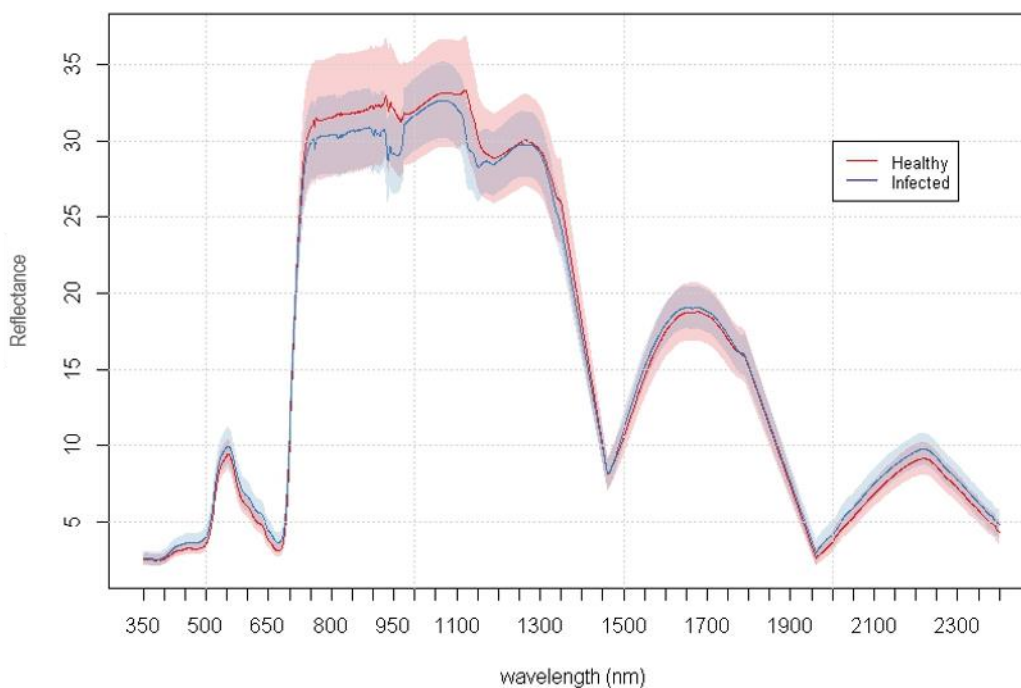


Figure 5.2: Mean spectral signature of the healthy ( $n = 50$ ) and infected ( $n = 50$ ) classes. The Healthy (sd) and Infected (sd) signatures represent the 1-sigma standard deviation for the healthy (pink shade) and infected (blue shade) signatures respectively.

## 5.2.4 Tree-based ensembles

### 5.2.4.1 Random forest

The RF algorithm (Breiman 2001) is an extension of bootstrap aggregation of CART. The RF algorithm builds models by aggregating large numbers of trees (*ntree*) on bootstrap samples of the original dataset. Trees are maximally grown, i.e. trees are not pruned. RF randomly selects a subset of bands (*mtry*) to create the node splits for individual trees in the ensemble, thereby reducing the correlation between trees in the ensemble. The *mtry* hyperparameter value is equal to the number of bands randomly sampled as candidates for node splitting in each tree. The *mtry* hyperparameter controls the bias variance trade-off since using fewer bands per node will produce less correlated trees, thereby reducing the overall variance but increasing the bias, as individual trees are now less accurate (Díaz-Uriarte & Alvarez de Andrés 2006). The default *mtry* value is equal to the square root of the total number of bands ( $p$ ). The final classification is based on a majority vote of predictions of all trees in the ensemble (Goldstein et al. 2011). RF was implemented using the randomForest library (Liaw & Wiener 2002) in the R statistical software (R Development Core Team 2019). The default *mtry* hyperparameter value ( $mtry = p^{1/2}$ ) and an *ntree* value of 500 was used for model building (Liaw & Wiener 2002).

### 5.2.4.2 Oblique Random Forest

The oRF model shares the same ensemble creating process (i.e., bootstrap aggregation and the selection of random variables for node splitting) as RF, but differs in the manner in which the optimal split direction at each node of the tree is created. The original RF implementation uses random coefficients to create optimal splits using a single variable selected from the user-defined *mtry* variables whereas oRF uses all the selected *mtry* variables to learn the optimal split direction using a supervised model. Additionally, unlike the original RF implementation, oRF scales (zero mean and unit variance) the variables to enhance model stability (Menze et al. 2011). According to Menze et al. (2011) models for the node splits may consider (i) class label information only (for example logistic regression and linear discriminant analysis (LDA)), (ii) data variation (for example PCA), or (iii) an optimum between class label correlation and data (for example ridge regression, partial least squares (PLS), and SVM).

This study considered (i) ridge regression, (ii) PLS, and (iii) SVM for multivariate node splitting. Ridge regression aims to improve determination of the regression coefficients and reduce the variance among highly correlated bands by imposing a penalty on the coefficients (Addink et al. 2007):

$$RSS(\lambda) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (\text{Equation 5.1})$$

where  $\lambda$  controls the shrinkage of the regression coefficients,  $n$  is the number of samples,  $y$  is class label,  $\hat{y}$  is the regression prediction,  $p$  is the number of bands, and  $\beta_j$  is the  $j$ th regression coefficient. PLS computes a set of weights and loadings for a set of factors that is used to model the variance among the bands and the classes. These weights and loadings are further used to compute the cumulative importance ( $B$ -value) of each band (Equation 5.2); the higher the  $B$ -value, the higher the band importance (Jones et al. 2010):

$$B = w(p'w)^{-1}q' \quad (\text{Equation 5.2})$$

where  $B$  is the cumulative wavelength importance,  $w$  is the band weight,  $p$  is the band loading, and  $q$  is the class weight.

For a training dataset of  $k$  classes represented by  $\{x_i, y_i\}$ ,  $i = 1, \dots, k$ , where  $x \in \mathbf{R}^N$  is an  $N$ -dimensional space and  $y \in \{-1, +1\}$  is the class label, SVM seeks to find a separating hyperplane that maximizes the perpendicular distance between the healthy and infected classes by solving the constrained optimisation problem (Pal & Foody 2010):

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (\text{Equation 5.3})$$

where  $w$  is a vector that determines the orientation of the separating hyperplane, and  $b$  is a scalar that determines the offset of the hyperplane from the origin.

For all models, the regularisation parameters were optimised using the OOB samples at each node (Menze & Splitthoff 2015). Oblique RF was implemented using the `obliqueRF` library (Menze & Splitthoff 2015) in the R statistical software (R Development Core Team 2019). Default hyperparameter values of `mtry` (i.e., the square root of the total number of bands), and a `ntree` value of 300 were used for model building (Menze et al. 2011).

#### 5.2.4.3 Rotation Forest

rotF (Rodríguez et al. 2011) is a tree-based ensemble approach that uses DT as the base learner. It is similar to RF with respect to training independent trees, but differs by using a different subset of extracted features to train each tree. The key principle underpinning rotF is the use of PCA to first transform the original feature space to a new rotated feature space and subsequently undertake feature extraction for each base classifier (Rodríguez et al. 2006). Feature extraction is applied to subsets of bands, with all principal components then used for training each DT. Random partitioning of the feature set leads to greater diversity of the bootstrap samples. Similar to RF, the final classification result is based on a majority vote of the combined DT (Rodríguez & Alonso 2011). rotF was implemented in the R statistical

software (R Development Core Team 2019), using  $n_{tree} = 100$  and the default hyperparameter values of  $mtry$  (i.e., the square root of the total number for bands) for model building. A  $n_{tree}$  value of 100 was used, given that using  $n_{tree} = 10$  (Rodríguez et al. 2006) did not provide valuable results (not shown).

### 5.2.5 Spectral resampling

This study employed spectral resampling to reduce data dimensionality, and subsequently test the effect of a reduced dimensionality on classification accuracy. Several approaches to spectral resampling have been found in the literature. For example, Franke et al. (2009) and Mewes et al. (2011) used a stepwise merging approach, which involved summation of the full width at half maximum (FWHM) values of adjacent bands, to resample HyMap spectra. Adam et al. (2012) and Adjorlolo et al. (2013) applied user-defined bandwidths (equivalent to FWHM) fit to a Gaussian (normal distribution) model to resample spectral measurements to HyMap spectra. Dalponte et al. (2009) used the mean of contiguous spectral bands to spectral resample AISA Eagle bands ranging from 4.6 nm to 36.8 nm in increments of 4.6 nm. The original bands ( $n = 1769$ ) were incrementally resampled using user-defined waveband centres, based on the mean of adjacent bands. Subsets of bands were created by binning (resampling) bands into specified wavelength ranges, i.e., from 2 nm to 176 nm. Resampling of the hyperspectral bands was performed using the pavo library (Maia et al. 2015) in the R statistical software (R Development Core Team 2019). The resulting eight subsets ranged in size from  $n = 884$  to  $n = 10$  bands that were then used to test the robustness of the ensemble classifiers used in this study.

### 5.2.6 Classification accuracy

An independent test dataset (i.e., captured during week two) was used for assessing classification accuracy. This provided an independent estimate of model accuracy. All algorithms were trained using the spectral measurements obtained during week one and subsequently tested using the spectral measurements collected during week two of the experiment. Classification accuracy was then evaluated using overall accuracy derived from a confusion matrix (Kohavi & Provost 1998). Additionally, a discrete multivariate technique called Kappa analysis was used to assess classification accuracy. A KHAT statistic (Congalton & Green 2009) provides a measure of agreement between actual (“observed”) agreement and chance (“expected”) agreement:

$$\widehat{K} = \frac{p_o - p_c}{1 - p_c} \quad (\text{Equation 5.4})$$

where  $p_o$  is the actual agreement and  $p_c$  is the expected agreement. Models were replicated ( $n = 100$ ) (Ismail & Mutanga 2010) to provide a more robust measure of model generalisation, and descriptive statistics (mean accuracy and standard deviation) computed.

### 5.3 RESULTS

To better understand the difference in behaviour of the RF and oRF models, the topology of the decision boundary learned by each ensemble classifier was examined (Figure 5.3). The decision boundary was modelled using the first two principal components extracted from a PCA of the original hyperspectral dataset ( $n = 1769$ ). Figure 5.3a clearly illustrates the staircase or box-like decision boundary generated by univariate orthogonal splits, as used by RF (Menze et al. 2011; Blazer & Fryzlewicz 2015). For the oRF ensembles (Figure 5.3b–d) however, the smoother decision boundary is reminiscent of multiple rotated trees using random multivariate splits (Menze et al. 2011).

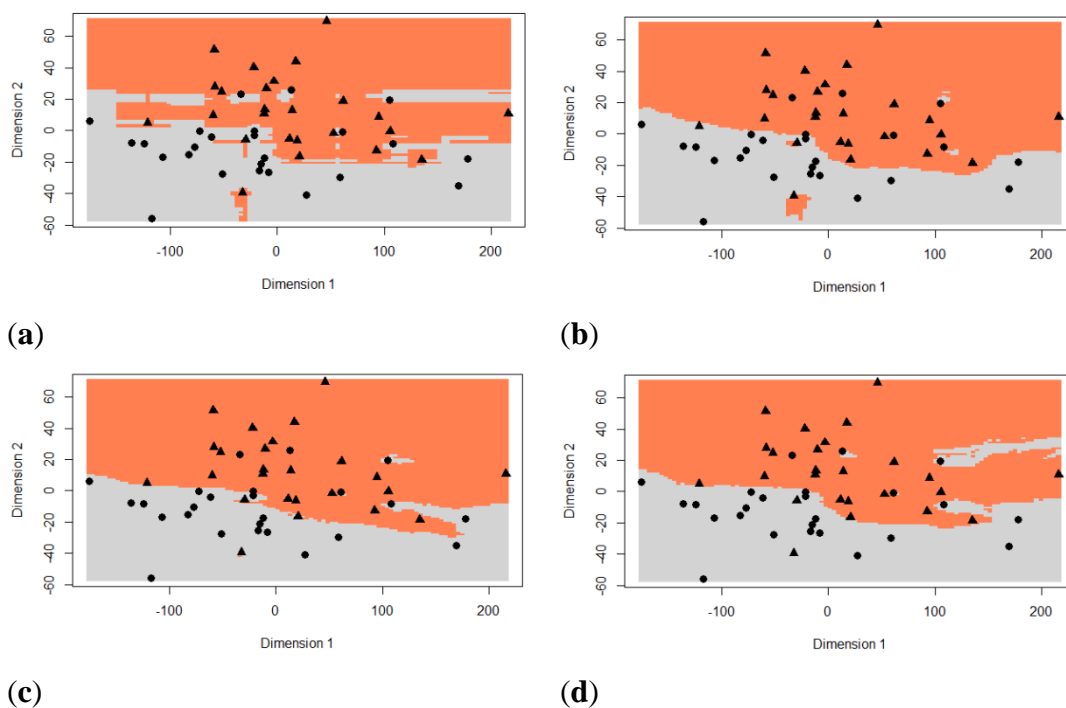


Figure 5.3: Visualisation of the decision boundary for (a) RF; (b) oRFridge; (c) oRFpls; and (d) oRFsvm. The margin between the grey and coral areas represents the decision boundary learned. The dots and triangles represent the two classes, i.e., healthy and infected. RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model.

Figure 5.4 shows the resulting mean classification accuracies obtained for the five ensemble classifiers using all bands ( $n = 1769$ ) based on 100 model runs. For all ensembles, the mean model accuracy was above 80% (KHAT values ranged from  $0.61 \pm 0.16$  to  $0.87 \pm 0.02$ ). The oRFsvm model produced the highest mean classification accuracy of  $93.59\% \pm 0.85\%$ . In comparison, the traditional RF model yielded the lowest mean classification accuracy of  $81.8\% \pm 1.82\%$ . rotF yielded a similar accuracy of  $82.73\% \pm 3.06\%$  when compared with RF, but has a higher variability of accuracy values denoted by the wider confidence interval.

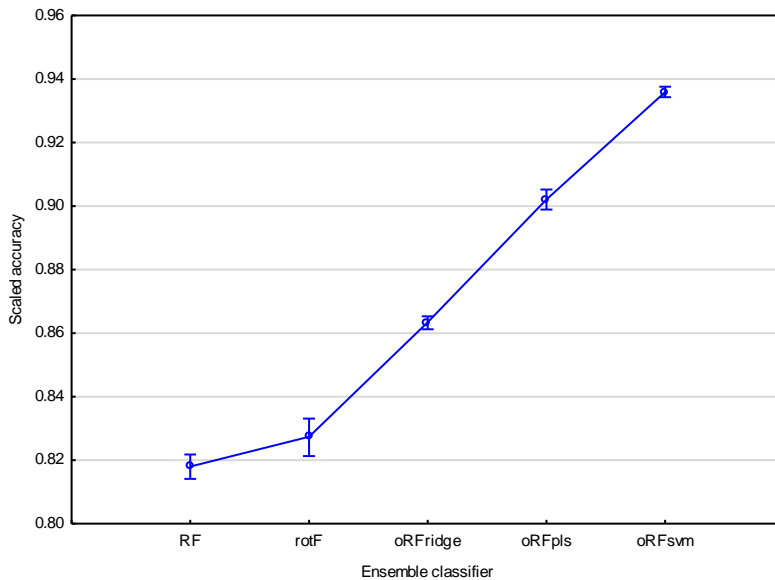


Figure 5.4: Mean classification accuracies for all tree-based algorithms (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model) considered in this study. The scaled accuracy is the classification accuracy represented on a scale ranging from zero to one. Vertical bars denote 0.95 confidence intervals.

It is evident from Figure 5.5 that the oRFsvm ensemble also has the smallest range of accuracy values between the upper and lower quartiles. This indicates higher classification results and better generalisation ability when compared with the other ensembles. Conversely, the rotF model has the largest range of accuracy values between the upper and lower quartiles. This indicates lower generalisation ability.

To determine if the classification accuracies obtained using the five tree-based ensemble classifiers were statistically different, a one-way ANOVA was performed followed by Fisher's least significance difference (LSD) test (Del Fiore et al. 2010) with bootstrapping (Efron & Tibshirani 1993). The results showed that there was no significant difference between the accuracies obtained for the RF and rotF



models at  $p = 0.05$ . However, there was a significant difference between the accuracies obtained for the three oRF models, i.e., oRFridge, oRFpls, and oRFsvm. Additionally, there was a significant difference between the RF model accuracy and the oRFridge, oRFpls, and oRFsvm model accuracy, as well as between the rotF model accuracy and the oRFridge, oRFpls, and oRFsvm model accuracy. Figure 5.5 indicates that the oRFridge, oRFpls, and oRFsvm models produced significantly higher mean accuracies (ranging between 86% and 94%) compared with RF and rotF models that produced significantly lower, and statistically similar, accuracies (ranging between 80% and 84%).

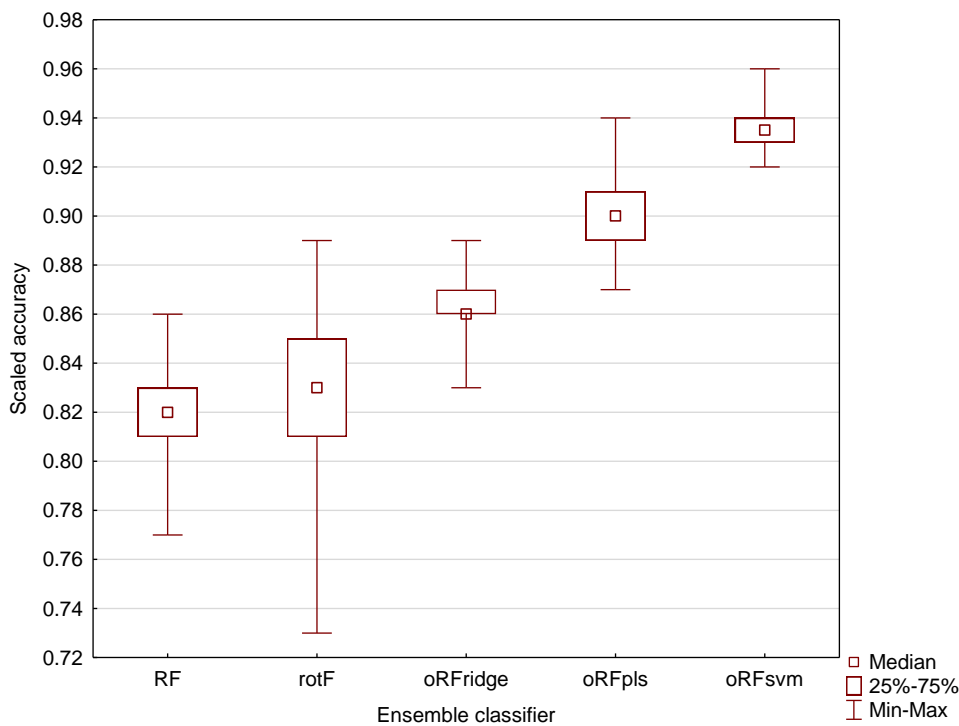


Figure 5.5: The distribution of the classification accuracy based on the test dataset for all tree-based algorithms (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model) considered in this study. Each boxplot represents the results obtained from 100 repetitions and all bands ( $n = 1769$ ). The scaled accuracy is the classification accuracy represented on a scale ranging from zero to one.

Figure 5.6 shows the result of spectral resampling of the original hyperspectral dataset ( $n = 1769$ ). Resampling of the hyperspectral bands resulted in subsets of bands ranging in size from  $n = 884$  (resampled to 2 nm) to  $n = 10$  (resampled to 176 nm). These subsets were used to generate models using each of the five ensemble classifiers. The results illustrated in Figure 5.7 show that for all ensembles, except oRFridge, the mean classification accuracy remained stable when using bands resampled to 2 nm



ranging up to 63 nm. However, bands resampled to 126 nm and 176 nm show a significant decrease in mean classification accuracy for all ensembles considered in this study. The oRFsvm ensemble provided the most consistent accuracies across all resampled bands and is thus shown to be the most robust of all the ensembles considered in this study.

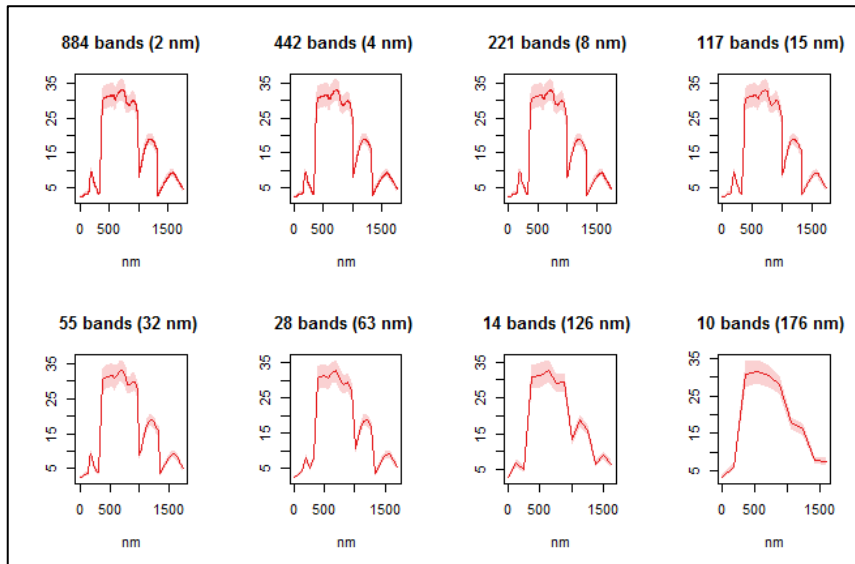


Figure 5.6: Resampling of the original hyperspectral dataset. Subsets of bands ranged in size from  $n = 884$  (spectral resampling to 2 nm) to  $n = 10$  (spectral resampling to 176 nm). The X-axis represents the wavelength (nm) of the resampled bands whereas the Y-axis represents the reflectance (%).

A one-way ANOVA was again performed, followed by Fisher's LSD test (Del Fiore et al. 2010) with bootstrapping (Efron & Tibshirani 1993) to determine if the classification accuracies of all the ensemble models obtained using the spectral resampled bands were statistically different. The results show that there was no significant difference in accuracy between the three oRF models, i.e., oRFridge, oRFpls, and oRFsvm, at  $p = 0.05$ . This is contrary to the results obtained when using all hyperspectral bands. The results also indicated that the RF and rotF model accuracies were significantly different from each other as well as from the oRFridge, oRFpls, and oRFsvm model accuracies. It is clear from Figure 5.8 that the oRFridge, oRFpls, and oRFsvm models produced similar accuracies (ranging between 90% and 92%) compared with the RF and rotF models which have significantly lower mean accuracies.

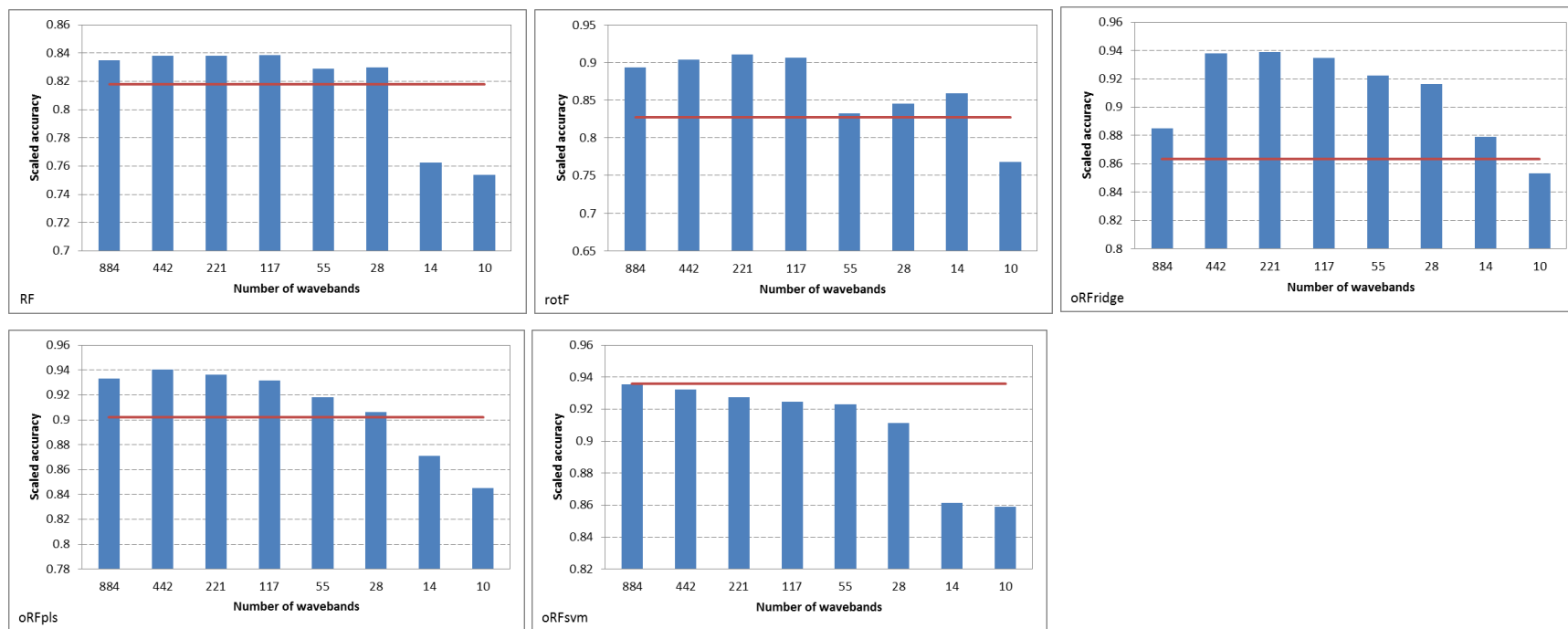


Figure 5.7: Comparison of the mean accuracies obtained using all bands and using the resampled bands for the five ensemble classifiers (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model). The red line indicates the mean accuracy obtained using all the original bands ( $n = 1769$ ) whereas the blue bars indicate the mean accuracies for the respective resampled subsets.

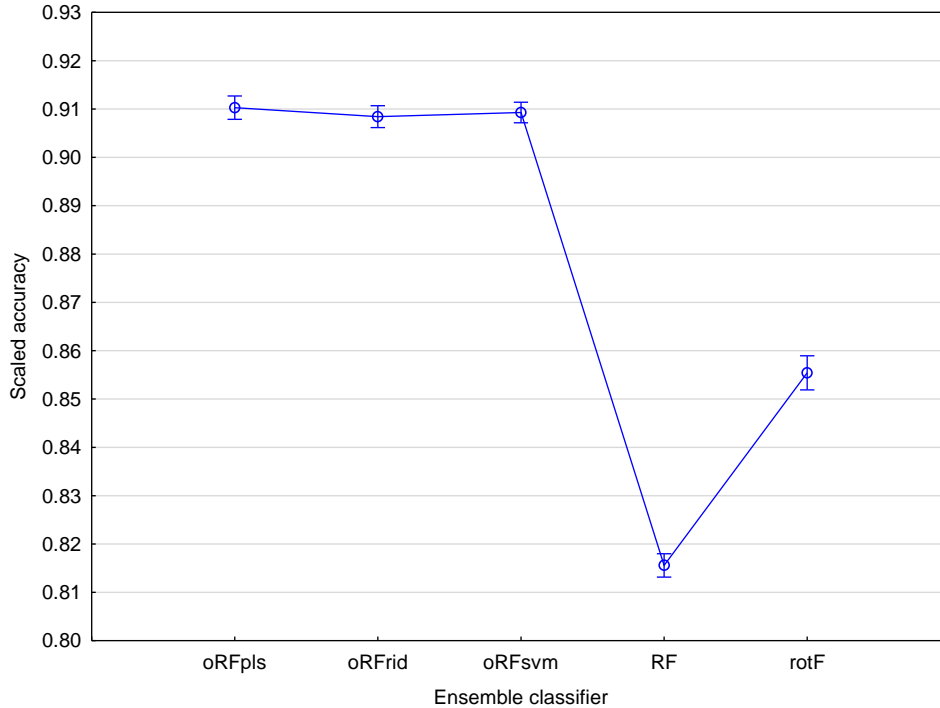


Figure 5.8: Mean classification accuracies using resampled hyperspectral bands ( $n = 800$ ) for each of the tree-based algorithms (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest with ridge regression as splitting model; oRFpls = oblique random forest with PLS as splitting model; oRFsvm = oblique random forest with SVM as splitting model) considered in this study. The scaled accuracy is the classification accuracy represented on a scale ranging from zero to one. Vertical bars denote 0.95 confidence intervals.

Table 5.1 summarises the highest and lowest mean classification accuracies (and associated spectral resampled bands) for all the ensemble classifiers considered in this study. Overall results indicate that the three oRF ensembles, i.e., oRFridge, oRFpls, and oRFsvm, produced the highest mean classification accuracies. Additionally, the oRFridge model had the lowest standard deviation of 0.48 when using bands ( $n = 221$ ) resampled to 8 nm. In comparison, RF produced a highest mean classification accuracy of only  $84\% \pm 0.60\%$  using bands ( $n = 117$ ) resampled to 15 nm. For all ensembles, classification using a very coarse spectral resolution, that is spectral resampling to 176 nm ( $n = 10$ ), yielded the lowest mean classification accuracy.

Comparing the results in Table 5.1 with the mean classification accuracies obtained using all bands ( $n = 1769$ ), it is evident that spectral resampling resulted in an overall increase in classification accuracy. For example, for rotF, the highest mean classification accuracy achieved was  $91\% \pm 0.85\%$ , using bands ( $n = 221$ ) resampled to 8 nm compared with  $83\% \pm 3.06\%$  using all bands. This is equivalent to an increase of more than 8% in classification accuracy. The only exception, in which there was no change in

classification accuracy, was for oRFsvm with a highest mean classification accuracy of  $94\% \pm 0.77\%$  using the resampled bands compared with  $94\% \pm 0.85\%$  using all bands.

Table 5.1: Spectral resampled wavelengths and the associated classification results using the five ensemble classifiers (RF = random forest; rotF = rotation forest; oRFridge = oblique random forest using ridge regression as splitting model; oRFpls = oblique random forest using PLS as splitting model; oRFsvm = oblique random forest using SVM as splitting model). KHAT values are indicated in parentheses.

Ensemble Classifier	Highest Accuracy (%)	Resampled bands(nm)	Resampled bands( <i>n</i> )	Lowest Accuracy (%)	Resampled bands(nm)	Resampled bands( <i>n</i> )
RF	$84 \pm 0.60$ ( $0.68 \pm 0.01$ )	15	117	$75 \pm 1.35$ ( $0.51 \pm 0.03$ )	176	10
rotF	$91 \pm 0.85$ ( $0.80 \pm 0.04$ )	8	221	$77 \pm 1.24$ ( $0.55 \pm 0.07$ )	176	10
oRFridge	$94 \pm 0.48$ ( $0.88 \pm 0.01$ )	8	221	$85 \pm 1.75$ ( $0.71 \pm 0.03$ )	176	10
oRFpls	$94 \pm 0.75$ ( $0.88 \pm 0.02$ )	4	442	$85 \pm 1.28$ ( $0.69 \pm 0.03$ )	176	10
oRFsvm	$94 \pm 0.77$ ( $0.87 \pm 0.02$ )	2	884	$86 \pm 1.20$ ( $0.72 \pm 0.03$ )	176	10

## 5.4 DISCUSSION

Tree-based ensemble classifiers are widely used for the classification of high dimensional data; see for example Ismail & Mutanga (2011), Poona & Ismail (2014), and Poona et al. (2016a). Their popularity is driven by the basic premise that using many weak classifiers should yield better classification accuracy than a single classifier (Rokach 2010). In this study, five tree-based ensemble classifiers i.e., random forest (RF), rotation forest (rotF), oRF using ridge regression as the splitting model (oRFridge), oRF using PLS as the splitting model (oRFpls), and oRF using SVM as the splitting model (oRFsvm) were compared. Additionally, this study specifically examined the effect of spectral resolution on the ensemble's ability to classify healthy and infected *P. radiata* seedlings using high dimensional spectral data. The following sections discuss the experimental results in more detail.

### 5.4.1 Classification using all bands

RF has become a popular ensemble classifier for the analysis of hyperspectral data, given that it is relatively robust to outliers and noise and is not prone to overfitting (Biau 2012). The analysis shows that RF was generally outperformed by the other tree-based ensembles considered in this study. This indicates

that RF may not be the optimal ensemble classifier for the classification of spectroscopic data. When using all bands ( $n = 1769$ ) the RF ensemble only marginally outperformed rotF with a mean classification accuracy of  $82\% \pm 1.82\%$  for RF compared with  $79\% \pm 3.06\%$  for rotF. More importantly, RF was significantly outperformed by oRFridge ( $86\% \pm 1.06\%$ ), oRFpls ( $90\% \pm 1.66\%$ ), and oRFsvm ( $94\% \pm 0.85\%$ ).

Contrary to previous studies, for example Rodríguez et al. (2006), Stiglic & Kokol (2007), and Xia et al. (2014) that have demonstrated the superior performance of rotF compared with RF, this study shows that rotF produced the lowest overall classification accuracies. rotF was the least robust of all the ensemble classifiers, yielding variable classification accuracies ranging from a minimum of 73% to a maximum of 89% with a standard deviation of 3.06%.

However, the results of this study compare favourably with those of Do et al. (2010) and Menze et al. (2011). For example, Do et al. (2010) tested the performance of RF, SVM, and oRF. The key finding of their study was that oRF outperformed both RF and SVM by an average of 3.57% and 6.35% respectively. The results of this study show that oRF (using SVM as the splitting rule) outperformed RF by an average of 12%. Menze et al. (2011) also showed that for the classification of high dimensional spectral data, the RF ensemble was outperformed by the oRF ensembles, with oRFridge yielding the best classification result. The analyses further indicate that although oRFridge outperformed RF, oRFridge was outperformed by both oRFpls and oRFsvm, with the oRFsvm ensemble providing the best classification accuracy when using the entire hyperspectral dataset. The stable results of oRFsvm may be attributed to the ability of SVM to effectively handle ill-posed problems, i.e. classification of a high dimensional feature space with limited training samples, coupled with its higher generalisation ability (Abdel-Rahman et al. 2014). Of note is that only a limited number of studies have investigated the use of oRF for the analysis of high dimensional data; see for example Do et al. (2010), Menze et al. (2011), and Do et al. (2015). Additionally, the results of this study highlight the potential to use oRF in a binary application.

#### **5.4.2 The effect of spectral resampling on classifier performance**

In this study, a total of 100 samples was used, i.e., healthy ( $n = 50$ ) and infected ( $n = 50$ ). All models were constructed using a decreasing number of bands ( $p$ ) while maintaining the number of samples ( $n$ ) constant. Models constructed from a larger number of samples compared with the number of bands ( $n < p$ ) generally achieved the highest accuracy. This is evident from Figure 5.7, where the highest accuracies are obtained using bands spectrally resampled to 2 nm, 4 nm, 8 nm, and 15 nm. A similar result is

observed for models constructed with an equivalent number of samples and bands ( $n \approx p$ ); evident using bands spectrally resampled to 32 nm and 63 nm. However, models constructed with a lower number of bands compared with the number of samples ( $n > p$ ) showed the lowest classification performance. These results are evident using bands spectrally resampled to 126 nm and 176 nm. This trend was also observed by Dalponte et al. (2009) and Mewes et al. (2011) who found that models constructed from a lower number of bands yielded the lowest accuracies.

Spectral resampling of the hyperspectral bands produced mixed results with respect to the ensemble model employed. For example, from an evaluation of the mean classification accuracy obtained for RF, rotF, and oRFpls using the original bands compared with using the spectral resampled bands, it is evident that improved classification performance was achieved using the spectrally resampled bands. For oRFridge and oRFpls, the spectrally resampled bands yielded a significant increase in the classification performance. However, for oRFsvm, the spectrally resampled bands did not yield any significant improvement in the mean classification accuracy. Several authors; see for example Mladenović et al. (2004), Pal & Mather (2005), and Pal (2006) found that the performance of the linear SVM is not significantly influenced by a reduced dimensionality. The robustness of SVM has been illustrated using oRFsvm for the classification using all bands (Section 5.4.1). Similar results were demonstrated by Dalponte et al. (2009) using the SVM, GML with leave-one-out-covariance estimator (GML-LOOC), and LDA classifiers. The authors noted that the SVM classifier yielded the highest KHAT values, and remained stable across all spectral resampled subsets. KHAT values were generally lower for the GM-LOOC and LDA classifiers.

Overall, the results reaffirm the findings of previous research (Poona & Ismail 2014; Poona et al. 2016a), demonstrating that decreasing the data dimensionality leads to improved overall classification accuracy, and that a lower dimensional dataset can be used to efficiently discriminate healthy and infected seedlings. In this study, all ensemble classifiers displayed a similar trend in classification performance with the resampled datasets, i.e., classification accuracy remained stable at lower FWHM values and decreased at higher FWHM values. A similar trend was observed by Dalponte et al. (2009) and Mewes et al. (2011). Although lower accuracies were obtained at a spectral resolution of 126 nm and 176 nm, the results indicate that it is still possible to discriminate the two classes (healthy and infected). For example, for both RF and rotF, classification accuracy was above 75% using bands resampled to 176 nm. In the case of oRFridge, oRFpls, and oRFsvm, classification accuracy was above 84% using bands resampled to 176 nm.

### 5.4.3 Robustness of the oblique forest ensembles

This study evaluated the use of RF ensembles, including rotF and oRF, to model asymptomatic stress in *P. radiata* seedlings associated with *F. circinatum* infection. Previous studies, for example Do et al. (2010) and Menze et al. (2011) have demonstrated the superior performance of oblique forest ensembles compared with other classifiers such as RF, CART, and SVM. The use of oblique RF was found to be particularly suitable for the processing of high dimensional spectral data.

As previously indicated, the staircase or box-like decision boundary generated by univariate splits, as is the case with CART and RF, may not be optimal for the classification of highly correlated data, such as high dimensional spectroscopic data (Menze et al. 2011). Consequently, learners that comprise multivariate DT via generation of oblique decision boundaries would be more suited to analysing high dimensional, highly correlated hyperspectral data. The results obtained in this study clearly confirm this notion. In this study, the traditional RF ensemble constructed from univariate DT was outperformed by all three oRF ensembles as well as the rotF ensemble, which are constructed from multivariate DT. Additionally, the use of an algorithm to estimate the splitting rule for the oRF ensembles—ridge regression, PLS, and SVM was used in this study—likely contributed to the improved performance of the oRF ensemble and consequently the high classification accuracies. Friedl & Brodley (1997) showed that multivariate DTs incorporating splitting rules produced significantly higher classification accuracies compared with univariate DTs and Bayesian classifiers. Similarly, Pal & Mather (2003) showed that multivariate DTs produced comparatively high classification accuracies compared with univariate DTs, artificial neural networks, and Bayesian classifiers.

The classification results further indicate that the performance of the oRF ensembles is not significantly affected by the multicollinearity, albeit the fact that higher classification accuracies were obtained when a lower dimensionality, i.e., spectral resolution was used. In this study the dataset size was systematically reduced by spectral resampling (binning) of the original dataset ( $n = 1769$ ) into discrete subsets of bands. The results of Dalponte et al. (2009), Franke et al. (2009), Mewes et al. (2011), Adam et al. (2012), and Adjorlolo et al. (2013) illustrate that reducing the input data dimensionality results in improved classification performance. This notion is reinforced by the results achieved in this study using the oRF ensembles to classify high dimensional spectroscopic data. The results show that a subset of bands, generated by spectral resampling of the original dataset ( $n = 1769$ ), achieves accuracies above 90%, when an oblique node splitting model is used. The results of this study thus demonstrate the potential for

operationalisation of the oblique ensemble model for the asymptomatic detection of *F. circinatum* infection in *P. radiata* seedlings within a nursery environment.

## 5.5 CONCLUSIONS

This study aimed to evaluate the performance of various ensemble classifiers for the analysis of high dimensional spectral data. Additionally, the study tested the robustness of these ensembles to reduced data dimensionality and sample size. Some important conclusions from this study are, firstly, that rotF and oRF may be more suitable than RF for the analysis of high dimensional spectral data. Secondly, rotF is sensitive to both dimensionality and sample size, and produces less robust models compared with RF and oRF. Thirdly, the oRF ensemble using varied splitting models is more robust and yields better classification results compared with rotF and RF. Finally, the methods employed in this study require further investigation to evaluate their operational potential.



## CHAPTER 6: OPTIMISED TWO-BAND NORMALISED DIFFERENCE SPECTRAL INDICES FOR MODELLING *FUSARIUM* STRESS

Poona NK & Ismail R 2019. Developing optimised spectral indices using machine learning to model *Fusarium circinatum* stress in *Pinus radiata* seedlings. *Journal of Applied Remote Sensing* 13, 034515. doi:10.1117/1.JRS.13.034515.

### Abstract

Narrowband normalised difference spectral indices have found wide application in vegetation studies. Consequently, several studies have investigated the utility of optimised spectral indices for targeted applications. The objective of this study was to statistically develop optimised two-band normalised difference spectral indices from a subset of hyperspectral bands derived using the Boruta wrapper algorithm. These indices were applied to model *F. circinatum* stress in *P. radiata* seedlings. The performance of the developed optimised indices were compared with a selection of widely used existing spectral indices ( $n = 111$ ) noted in the literature. Analyses were undertaken within a univariate (using the Jeffries-Matusita distance) and a multivariate (using the random forest algorithm) framework. The results clearly demonstrate the improved accuracies using optimised spectral indices (overall accuracy ranged from 76-96%) compared with using existing indices (overall accuracy ranged from 83-90%). Additionally, the results show that a multivariate approach yields superior results compared with a univariate approach. Overall, the results demonstrate the operational potential of optimised two-band normalised difference spectral indices within a multivariate paradigm.

**Keywords:** spectral indices, Boruta, random forest (RF), Jeffries-Matusita distance.

## 6.1 INTRODUCTION

Several authors have successfully demonstrated the use of narrowband hyperspectral data for modelling pest and disease stress in vegetation (Abdel-Rahman et al. 2013; Calderón et al. 2013; Oumar et al. 2013; Ashourloo et al. 2014a; Poona & Ismail 2014; Poona et al. 2016a). High dimensional narrowband hyperspectral (spectroscopic) data are characterised by many—often several hundred—narrow contiguous bands across the visible (VIS) and infrared portions of the EMS. These bands typically range from 400 nm in the VIS region, to 2500 nm in the SWIR region. In the VIS spectrum (ranging from approximately 400 nm to 700 nm), leaf reflectance is influenced by changes in plant biochemistry; in the NIR spectrum (ranging from approximately 700 nm to 1200 nm) by leaf structural attributes; whereas in the SWIR region (ranging from approximately 1200 nm to 2500 nm), leaf reflectance is primarily a function of leaf water content (Jacquemoud & Ustin 2001). Thus, when a plant is under stress, subtle variations in leaf reflectance in the VIS, NIR, and SWIR regions can be detected, quantified, and modelled (Poona & Ismail 2014).

Spectral indices (SIs), derived from narrowband hyperspectral measurements, are commonly used in vegetation studies to assess plant physiology and chemistry (Mahlein et al. 2013). Using the spectral information contained in only a few spectral channels, SI combine two or more narrow spectral bands as linear combinations through ratioing, summing, or differencing (Jackson & Huete 1991). For example, the blue index (BI) (Calderón et al. 2013) is a ratio of the reflectance at 450 nm and 490 nm (i.e.  $R_{570}/R_{670}$ ), whereas the health index (HI) (Mahlein et al. 2013) is based on the normalised difference of 534 nm and 698 nm, and the reflectance at 704 nm (i.e.  $[(R_{534}-R_{698})/(R_{534}+R_{698})]-1/2(R_{704})$ ). A large number of SIs have been developed and extensively applied in remote sensing studies of vegetation; several specifically developed for the detection and monitoring of plant disease and stress. For example, Calderón et al. (2013) and Calderón et al. (2015) employed SIs ( $n = 31$ ) and ( $n = 21$ ) respectively, and airborne hyperspectral imagery (wavelength range from 400 nm to 885 nm) for modelling *Verticillium* wilt of *Olea europea* L. caused by *Verticillium dahlia* Kleb. Overall results showed that SIs could successfully be used for the early detection of *V. dahlia*-induced stress and discrimination of *Verticillium* wilt disease severity. Ashourloo et al. (2014a) used SIs ( $n = 22$ ) and leaf spectral measurements (ranging from 450 nm to 1000 nm) to model *Puccinia triticina* disease severity in *Triticum dicoccum*. Results showed the potential of SIs for the classification of *P. triticina* disease severity. The authors further noted the value of using SIs in reducing both data dimensionality and computational expense. Feng et al. (2016) employed SIs ( $n = 28$ ) and canopy spectral measurements

(ranging from 400 nm to 1000 nm) to model *Blumeria graminis* f. sp. *tritici* disease severity in four winter wheat cultivars. The results highlighted the potential of using SIs for modelling *B. graminis* disease severity. More recently, Shi et al. (2017) successfully demonstrated the use of SIs ( $n = 14$ ) for the detection and discrimination of yellow rust, aphid infection, and powdery mildew in *Triticum aestivum* L. leaves.

According to Mahlein et al. (2013) and Huang et al. (2014), the use of narrowband SIs is, however, limited in their application to detecting and linking a specific stressor to a specific stress / disease. Numerous studies have shown that only certain portions of the EMS, located at discrete wavelengths, are of benefit to detecting plant stress (Abdel-Rahman et al. 2013; Calderón et al. 2013; Oumar et al. 2013; Ashourloo et al. 2014a; Poona & Ismail 2014; Poona et al. 2016a; 2016b). Identifying and isolating these discrete wavelengths (bands) can thus inherently improve the ability to model plant stress using spectroscopic data. Consequently, the development of optimised spectral indices ( $SI_{opt}$ ), statistically derived from these discrete bands, i.e. the most important bands, may prove indispensable in detecting stress caused by a specific stress agent (Mahlein et al. 2013; Huang et al. 2014). Additionally, the identification of the most important bands can provide insight to correlating a specific stressor to a specific stress (Mahlein et al. 2013).

Several studies have highlighted the utility of  $SI_{opt}$  for modelling plant stress. More significantly, these studies demonstrated the superior performance of  $SI_{opt}$  compared with the performance of SIs noted in the literature; henceforth referred to as existing SIs ( $SI_{ex}$ ). For example, Prabhakar et al. (2011) applied  $SI_{ex}$  ( $n = 20$ ) and  $SI_{opt}$  ( $n = 4$ ) for modelling stress in *Gossypium hirsutum* L. caused by *Empoasca devastans* (Ishida). The authors used the Pearson correlation coefficient ( $r$ ) to determine the most important bands, i.e. those bands with the highest absolute correlation coefficients, for developing the four  $SI_{opt}$ . Results showed that only 14 of the 20  $SI_{ex}$  were statistically significant, whereas all four developed  $SI_{opt}$  were statistically significant in discriminating stress severity. Overall, the four developed  $SI_{opt}$  outperformed the commonly used  $SI_{ex}$ . Mahlein et al. (2013) applied  $SI_{ex}$  ( $n = 12$ ) and  $SI_{opt}$  ( $n = 4$ ) for detecting and identifying *Cercospora* leaf spot, sugar beet rust, and powdery mildew in sugar beet plants. The most important bands for developing the four  $SI_{opt}$  were derived using the RELIEF-F algorithm (Robnik-Šikonja & Kononenko 2003). Overall results showed that the developed  $SI_{opt}$  outperformed all the  $SI_{ex}$  in detecting *Cercospora* leaf spot as well as classifying healthy and diseased leaves. Similarly, Huang et al. (2014) employed  $SI_{ex}$  ( $n = 10$ ) and  $SI_{opt}$  ( $n = 4$ ) for identifying and monitoring disease stress in *Triticum aestivum* L. caused by powdery mildew, yellow rust, and

aphids. The authors also employed the RELIEF-F algorithm to determine the most important bands for developing the four  $SI_{opt}$ . The developed  $SI_{opt}$  produced high classification accuracies (>85%) for discriminating between the diseases, provided the best discrimination of healthy and diseased winter wheat, and overall outperformed the  $SI_{ex}$ .

Within a forestry context, Fassnacht et al. (2012) employed visible-near infrared (VNIR) HyMap wavelengths ( $n = 39$ ), ranging from 455 nm to 986 nm, to model *Ips typographus* L. infestation in a mixed species natural forest. The 39 wavelengths were used to develop three-angle  $SI_{opt}$ , and compared with  $SI_{ex}$  ( $n = 82$ ) via a genetic algorithm coupled with a nearest centroid classifier, and a support vector machine classification. Overall results illustrated the superior performance of using  $SI_{opt}$  compared with using  $SI_{ex}$ . Oumar et al. (2013) employed  $SI_{ex}$  ( $n = 23$ ) and  $SI_{opt}$  ( $n = 20$ ) with leaf spectral measurements (ranging from 426.82 nm to 2395.50 nm), for predicting *T. peregrinus* infestation in *E. macarthurii*. The original bands were first resampled to the Hyperion sensor calibrated bands ( $n = 198$ ), and subsequently used to compute two-band normalised indices ( $n = 1\ 081$ ). The top 20 indices were selected based on their highest linear regression coefficient ( $R^2$ ) values between the computed indices and visual damage. Overall, the  $SI_{opt}$  dataset ( $n = 20$ ) provided better predictive power of *T. peregrinus* damage compared with the  $SI_{ex}$  dataset ( $n = 23$ ).

The random forest (RF) algorithm (Breiman 2001) is an ensemble of CART that has found widespread use in remote sensing for both classification and regression tasks. Of particular significance is the importance rankings assigned to features. Feature importance is computed using an out-of-bag (OOB) sample, equivalent to approximately 34% of the input data. This OOB sample is additionally used to compute an unbiased estimate of the model training error, known as the OOB error. Researchers have exploited the RF feature importance as a precursor to undertaking feature selection for development of  $SI_{opt}$ . For example, Abdel-Rahman et al (2013) employed RF to select the ten most important bands for generating  $SI_{opt}$  for estimating leaf nitrogen concentration in sugarcane. The authors successfully demonstrated the use of optimised spectral indices for predicting sugarcane leaf nitrogen concentration ( $R^2 = 0.67$ ). More recently, Chemura et al. (2016) employed the RF model to select the most important  $SI_{ex}$  ( $n = 7$ ) for discriminating *Hemileia vastarix* infection levels on *Coffea arabica*, obtaining an overall classification accuracy of 82.5%.

Boruta (Kursa et al. 2010; Kursa & Rudnicki 2010) is a feature selection algorithm wrapped around the RF ensemble. Poona & Ismail (2014) and Poona et al. (2016a) successfully demonstrated that a subset of bands, derived using the Boruta algorithm, could best discriminate healthy and stressed *P. radiata*

seedlings infected with the fungal pathogen *F. circinatum*. Poona & Ismail (2014) further demonstrated that the spectral response of healthy, infected, and damaged seedlings was significantly different and that using both strongly and weakly relevant Boruta-selected bands provided the best classification accuracies.

No study to date has employed Boruta-selected bands for developing  $SI_{opt}$ . Additionally, no study to date has compared the implementation of  $SI_{ex}$  and  $SI_{opt}$  within a univariate and multivariate framework. In this study, the Boruta algorithm was employed to first derive a subset of the most important bands for discriminating healthy, infected, and damaged *P. radiata* seedlings. The Boruta-selected bands were subsequently used to statistically derive two-band normalised difference  $SI_{opt}$ , to model *F. circinatum* stress in *P. radiata* seedlings. The  $SI_{ex}$ , selected from the literature, was then compared with the indices statistically derived in this study ( $SI_{opt}$ ), for the asymptomatic detection of *F. circinatum* stress in *P. radiata* seedlings. The hypothesis was that the  $SI_{opt}$  developed from Boruta-selected bands will provide better classification accuracies compared with the subset of Boruta bands (Poona & Ismail 2014) as well as existing indices widely used in the literature ( $SI_{ex}$ ).

*Application:* *F. circinatum* is a significant pathogen of *Pinus* spp. globally; the most susceptible species being *P. radiata* (EPPO 2005). The fungus has a global presence, infecting several species of native and commercial *Pinus* stands (Coutinho et al. 2007; Wingfield et al. 2008), and is endemic in nurseries across South Africa (Porter et al. 2009). The fungus spreads rapidly and is difficult to control (Schweisinger 2008), with infection often leading to seedling mortality (Mitchell et al. 2011). Consequently, an accurate and efficient operational methodology to discriminate healthy and stressed seedlings in a nursery environment is paramount in order to reduce high seedling mortality.

## 6.2 MATERIALS AND METHODS

### 6.2.1 Spectral data collection and pre-processing

Narrowband hyperspectral data collected from healthy, damaged, and infected *P. radiata* seedlings inoculated with the pathogen *F. circinatum* formed the basis for this study. A total of 100 seedlings were randomly sampled from two trays of three month old *P. radiata* seedlings ( $n = 196$ ). The sampled seedlings were subsequently divided into three equal classes ( $n = 50$ ) and labelled healthy, infected, and damaged. The infected class was inoculated with a 10 mL spore suspension ( $50\,000$  spores  $mL^{-1}$ ) of *F. circinatum* isolate (FCC 3579) following the PCF screening facility best operating practice inoculum procedure (Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South

Africa). The inoculation procedure involved first topping the apical buds with a sterile razor blade, and subsequently placing the inoculum onto the topped apical buds. The damaged class only had the apical buds topped.

Spectral measurements were captured weekly (for a period of three weeks) between 10:00 and 15:00 using a FieldSpec® Pro FR Spectroradiometer (Analytical Spectral Devices, Boulder, CO) following inoculation. The FieldSpec® Pro acquires data in the 350–2500 nm spectral range with a spectral resolution of 3 nm @ 700 nm, 10 nm @ 1400 nm, and 12 nm @ 2100 nm. Reflectance measurements were calibrated in field using a Spectralon white reference panel (Poona & Ismail 2014; Poona et al. 2016a; 2016b). Figure 6.1 illustrates the setup used for spectral data acquisition.



Figure 6.1: Spectral data acquisition using the FieldSpec® Pro Spectroradiometer.

Five spectral measurements were captured per seedling using the 23° field-of-view, which were later averaged to a single reading per seedling. A total of 450 spectra (50 spectra per class); each spectrum comprising 1769 bands following removal of the atmospheric water absorption bands (1350-1460 nm and 1790-1960 nm) (Hatchell 1999; Walker 2009) and noisy bands (2401-2500 nm), were used for further processing.

### 6.2.2 Experimental design

To evaluate the ability of spectral indices to discriminate healthy, infected, and damaged seedlings, spectral data were combined into two class pairs, namely H-I and I-D. All spectral indices, i.e.  $SI_{s_{ex}}$  and  $SI_{s_{opt}}$  were developed using spectral data collected during week 3. The spectral data collected during week 1 and week 2 were subsequently used as independent validation data. Figure 6.2 provides an



overview of the data analysis workflow adopted in this study. The following sections provide details regarding the methods adopted in this study.

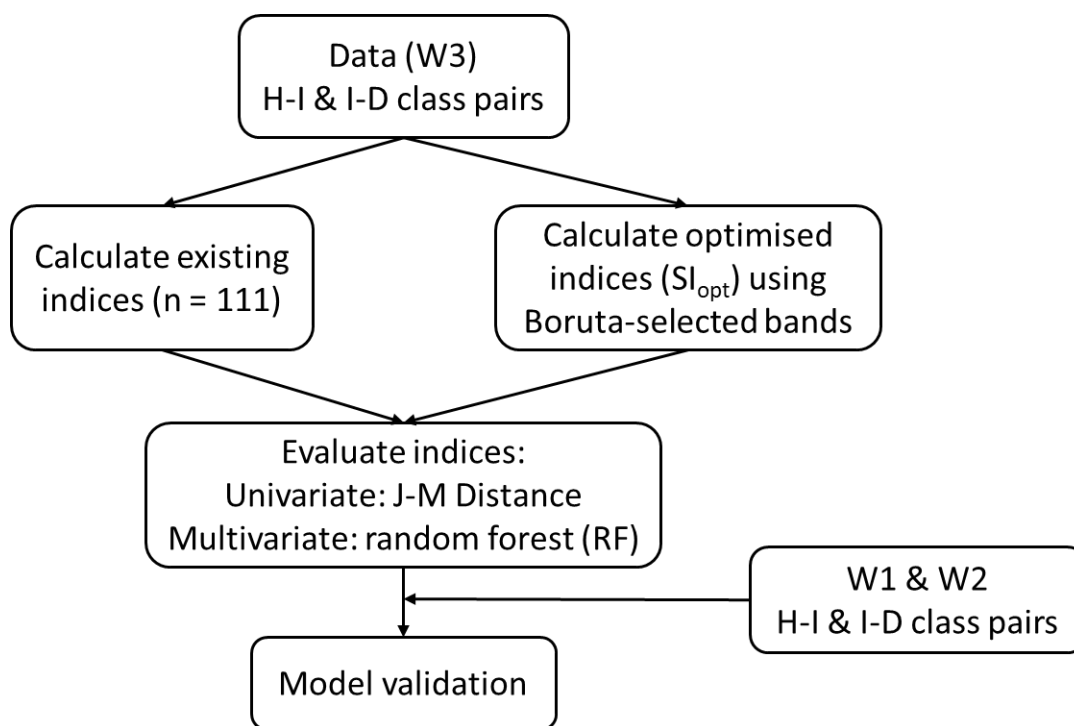


Figure 6.2: Data analysis workflow employed in this study.

### 6.2.3 Selection of existing spectral indices

Following a review of the literature (Zarco-Tejada et al. 2005; Prabhakar et al. 2011; Agapiou et al. 2012; Hernández-Clemente et al. 2012; Calderón et al. 2013; Oumar et al. 2013; Ashourloo et al. 2014a, 2014b; Feng et al. 2016), a set of narrowband spectral indices ( $n = 111$ ) was selected; henceforth referred to as existing indices ( $SI_{ex}$ ). The set comprised two-band ( $n = 74$ ), three-band ( $n = 29$ ), and four-band ( $n = 8$ ) spectral indices. The  $SI_{ex}$  included six thematic types namely; greenness, light use efficiency, senescent carbon, leaf pigment, canopy water content, and plant stress indices (Thenkabail et al. 2013; Table 6.1). Indices were subsequently computed using the spectral measurements from week 3.

### 6.2.4 Development of optimised spectral indices ( $SI_{opt}$ )

Developing and optimising a ratio-based index from all possible two-band combinations of  $n$  wavelengths can yield a possible  $n \times (n - 1)$  spectral indices, and from all possible three-band combinations, a possible  $n \times (n - 1) \times (n - 2)$  spectral indices (Wang et al. 2016). Consequently, using all wavelengths within the 350 nm to 2400 nm wavelength range ( $n = 1769$ ; following removal of noisy

bands) would yield 3 127 592 possible two-band indices, and more than  $5.5 \times 10^9$  possible three-band combinations. The decision was made to focus on developing two-band indices. It was necessary to employ a band selection method, such as a filter or a wrapper (Kohavi & John 1997), to reduce the data dimensionality, and select the most important bands for  $SI_{opt}$  development.

Spectra representing each of the three classes, i.e. healthy ( $n = 50$ ), infected ( $n = 50$ ), and damaged ( $n = 50$ ) were processed using the Boruta wrapper algorithm (Kursa et al. 2010; Kursa & Rudnicki 2010; 2011). Boruta was used to select an optimal subset of bands that could best discriminate the three classes, i.e. healthy, infected, and damaged (Poona & Ismail 2014). Boruta evaluates band importance by creating an ensemble of “shadow” bands that are randomly sampled from the original dataset and correspond to each of the original bands in the dataset. Boruta initially computes Z-scores (based on the mean decrease in accuracy score) for each band and its corresponding shadow band. The algorithm then iteratively compares Z-scores between each band and its corresponding shadow band in order to evaluate band importance (Kursa et al. 2010; Kursa & Rudnicki 2010). Boruta proceeds to iteratively fit RF models until either the specified number of runs is completed (maxRuns) or all bands are classified as either confirmed or rejected (Kursa & Rudnicki 2018). Boruta was run in light mode, i.e. both unimportant and corresponding shadow bands were dropped through iterations of the algorithm (Poona & Ismail 2014). RF models were built using default hyperparameter values, i.e.  $mtry =$  square root of the total number of input bands, and  $n tree = 500$ . The Boruta algorithm was implemented using the Boruta library (Kursa & Rudnicki 2018) in the R statistical software (R Development Core Team 2019).

To compute the normalised difference indices, Mahlein et al. (2013) and Huang et al. (2014) employed a “normalised wavelength difference” approach using the RELIEF-F algorithm (Robnik-Šikonja & Kononenko 2003). The normalised wavelength difference requires two wavelengths, one selected from the best weighted single band, and one from the worst weighted single band. Similarly, in this study, a normalised wavelength difference approach was adopted by selecting the strongly relevant and weakly relevant Boruta bands based on the mean decrease in classification accuracy score as determined by the RF algorithm (Poona & Ismail; 2014). The selected Boruta bands were used to compute all possible combinations of two-band SIs:

$$SI = \frac{R_i - R_j}{R_i + R_j} \quad (\text{Equation 6.1})$$

where  $R_i$  and  $R_j$  are the reflectance values of relevant Boruta-selected bands. Computation of the SIs was undertaken using the R statistical software (R Development Core Team 2019).



Table 6.1: Existing narrowband spectral indices employed in this study.

Index	Acronym	Equation	Reference
Anthocyanin reflectance index	ARI	$(1/R_{550}) - (1/R_{700})$	Gitelson et al. 2006
Blue index	BI	$R_{450}/R_{490}$	Calderón et al. 2013
Blue/green index 1	BGI 1	$R_{400}/R_{550}$	Zarco-Tejada et al. 2005
Blue/green index 2	BGI 2	$R_{450}/R_{550}$	Zarco-Tejada et al. 2005
Blue/red index 1	BRI 1	$R_{400}/R_{690}$	Zarco-Tejada et al. 2005
Blue/red index 2	BRI 2	$R_{450}/R_{690}$	Zarco-Tejada et al. 2005
Carter stress index 1	CSI 1	$R_{605}/R_{760}$	Carter 1994
Carter stress index 2	CSI 2	$R_{695}/R_{760}$	Carter 1994
Carter stress index 3	CSI 3	$R_{710}/R_{760}$	Carter 1994
Carter stress index 4	CSI 4	$R_{695}/R_{420}$	Carter 1994
Carter stress index 5	CSI 5	$R_{695}/R_{670}$	Carter 1994
Carotenoid/chlorophyll ratio index	CCRI	$(R_{680} - R_{500})/R_{750}$	Garrity et al. 2011
Carotenoid reflectance index 1	$CRI_{550}$	$(1/R_{510}) - (1/R_{550})$	Gitelson et al. 2001
Carotenoid reflectance index 2	$CRI_{700}$	$(1/R_{510}) - (1/R_{700})$	Gitelson et al. 2001
Carotenoid reflectance index 3	$RNIR * CRI_{550}$	$(1/R_{510}) - (1/R_{550}) * R_{770}$	Gitelson et al. 2006
Carotenoid reflectance index 4	$RNIR * CRI_{700}$	$(1/R_{510}) - (1/R_{700}) * R_{770}$	Gitelson et al. 2006
Chlorophyll index	CI	$(R_{415} - R_{435}) / (R_{415} + R_{435})$	Barnes 1992
Chlorophyll stress index 1	ChSI 1	$R_{415}/R_{695}$	Read et al. 2002
Chlorophyll stress index 2	ChSI 2	$R_{708}/R_{915}$	Zhao et al. 2005
Chlorophyll stress index 3	ChSI 3	$R_{551}/R_{915}$	Zhao et al. 2005
Curvature index	CI	$R_{675} * R_{690} / (R_{683})^2$	Zarco-Tejada et al. 2000
Damage sensitive spectral index	DSSI	$(R_{747} - R_{901} - R_{537} - R_{572}) / [(R_{747} - R_{901}) + (R_{537} - R_{572})]$	Mirik et al. 2006
Datt index 1	DI 1	$(R_{850} - R_{710}) / (R_{850} - R_{680})$	Datt 1999a, 1999b
Datt 2	DI 2	$R_{672} / (R_{550} * R_{708})$	Datt 1998
Datt 3	DI 3	$R_{860} / (R_{550} * R_{708})$	Datt 1999a, 1999b
Datt 4	DI 4	$R_{850}/R_{710}$	Datt 1999a, 1999b
Datt 5	Di 5	$R_{672}/R_{550}$	Datt 1998

Disease water stress index 1	DWSI 1	$R800/R1660$	Apan et al. 2004
Disease water stress index 2	DWSI 2	$R1660/R550$	Apan et al. 2004
Disease water stress index 3	DWSI 3	$R1660/R680$	Apan et al. 2004
Disease water stress index 4	DWSI 4	$R550/R680$	Apan et al. 2004
Disease water stress index 5	DWSI 5	$(R800+R550)/(R1660+R680)$	Apan et al. 2004
Double difference index	DDI	$(R749-R720)-(R701-R672)$	Le Maire et al. 2004
Double difference index (new)	DDIn	$2(R710-R(710-50)-R(710+50))$	Le Maire et al. 2008
Gitelson index 1	GI 1	$1/R500$	Gitelson et al. 2001
Gitelson index 2	GI 2	$1/R700$	Gitelson et al. 2001
Gitelson index 3	GI 3	$(R750-R800/R695-R740)-1$	Gitelson et al. 2003
Gitelson and Merzlyak 1	GM 1	$R750/R550$	Gitelson and Merzlyak 1994
Gitelson and Merzlyak 2	GM 2	$R750/R700$	Gitelson and Merzlyak 1994
Green normalised difference vegetation index	GNDVI	$(R750-R550)/(R750+R550)$	Gitelson et al. 1996
Healthy index	HI	$(R534-R698/R534+R698)-1/2(R704)$	Mahlein et al. 2013
Lichtenthaler index	LI	$R440/R740$	Lichtenthaler et al. 1996
Maccioni index	MI	$(R780-R710)/(R780-R680)$	Maccioni et al. 2001
Modified anthocyanin content index	MACI	$R940/R530$	Steele et al. 2009
Modified anthocyanin reflectance index	MARI	$R800[(1/R550)-(1/R700)]$	Gitelson et al. 2006
Modified chlorophyll-absorption-integral	MCAI	$[(R545+R752)/2](R752-R545)-\sum_{R545}^{R752}(1.158R)$	Laudien et al. 2003
Modified chlorophyll absorption reflectance index 1	MCARI 1	$[(R700-R670)-0.2(R700-R550)](R700/R670)$	Daughtry et al. 2000
Modified chlorophyll absorption reflectance index 2	MCARI 2	$[(R750-R705)-0.2(R750-R550)](R750/R705)$	Wu et al. 2008
Modified chlorophyll absorption reflectance index 3	MCARI 3	$1.2[2.5(R800-R670)-1.3(R800-R550)]$	Haboudane et al. 2004
Modified chlorophyll absorption reflectance index 4	MCARI 4	$1.5[2.5(R800-R670)-1.3(R800-R550)]/[(2R800+1)2-(6R800-5(R670)0.5)-0.5]0.5$	Haboudane et al. 2004

Modified chlorophyll absorption reflectance index/ Optimised soil-adjusted vegetation index	MCARI/OSAVI 1	$[(R700-R670)-0.2(R700-R550)](R700/R670)/[1+0.16(R800-R670)/(R800+R670+0.16)]$	Rondeaux et al. 1996
Modified chlorophyll absorption reflectance index/ Optimised soil-adjusted vegetation index	MCARI/OSAVI 2	$[(R750-R705)-0.2(R750-R550)](R750/R705)/[1+0.16(R750-R705)/(R750+R705+0.16)]$	Wu et al. 2008
Modified normalised difference vegetation index	MNDVI	$(R800-R680)/(R800+R680-2R445)$	Sims & Gamon 2002
Modified normalised difference vegetation index 705	MNDVI <sub>705</sub>	$(R750-R705)/(R750+R705-2R445)$	Sims & Gamon 2002
Modified photochemical reflectance index	MPRI	$(R515-R530)/(R515+R530)$	Hernández-Clemente et al. 2011
Modified soil-adjusted vegetation index 1	MSAVI 1	$0.5[2R800+1-\{(2R800+1)^2-8(R800-R670)\}^{0.5}]$	Qi et al. 1994
Modified soil-adjusted vegetation index 1	MSAVI 2	$0.5[2R750+1-\{(2R750+1)^2-8(R750-R705)\}^{0.5}]$	Qi et al. 1994
Modified triangular vegetation index 1	MTVI1	$1.2[1.2(R800-R550)-2.5(R670-R550)]$	Haboudane et al. 2004
Modified triangular vegetation index 2	MTVI2	$1.5[1.2(R800-R550)-2.5(R670-R550)]/[(2R800+1)^2-(6R800-5(R670)0.5)-0.5]^{0.5}$	Haboudane et al. 2004
Moisture stress index	MSI	$R1600/R820$	Hunt & Rock 1989
Nitrogen stress index 1	NSI 1	$R415/R710$	Read et al. 2002
Nitrogen stress index 2	NSI 2	$R517/R413$	Zhao et al. 2005
Normalised difference lignin index	NDLI	$[\log(1/R1754)-\log(1/R1680)]/[\log(1/R1754)+\log(1/R1680)]$	Serrano et al. 2002
Normalised difference vegetation index 1	NDVI1	$(R800-R670)/(R800+R670)$	Rouse et al. 1974
Normalised difference vegetation index 2	NDVI2	$(R750-R705)/(R750+R705)$	Gitelson & Merzlyak 1994
Normalised difference vegetation index 3	NDVI 3	$(R800-R680)/(R800+R680)$	Lichtenthaler et al. 1996

Normalised difference vegetation index 4	NDVI 4	$(R682-R553)/(R682+R553)$	Gandia et al. 2004
Normalised pigment chlorophyll index	NPCI	$(R680-R430)/(R680+R430)$	Peñuelas et al. 1993, 1994
Optimised soil-adjusted vegetation index 1	OSAVI 1	$(1+0.16)(R800-R670)/(R800+R670+0.16)$	Rondeaux et al. 1996
Optimised soil-adjusted vegetation index 2	OSAVI 2	$(1+0.16)(R750-R705)/(R750+R705+0.16)$	Wu et al. 2008
Photochemical reflectance index 1	PRI 1	$(R570-R531)/(R570+R531)$	Gamon et al. 1992
Photochemical reflectance index 2	PRI 2	$(R512-R531)/(R512+R531)$	Hernández-Clemente et al. 2011
Photochemical reflectance index 3	PRI 3	$(R570-R531-R670)/(R571+R531+R670)$	Hernández-Clemente et al. 2011
Pigment specific normalised difference	PSND	$(R800-R470)/(R800+R470)$	Blackburn 1998
Pigment specific simple ratio a	PSSRa	$R800/R675$	Blackburn 1998
Pigment specific simple ratio b	PSSRb	$R800/R650$	Blackburn 1998
Pigment specific simple ratio c	PSSRc	$R800/R500$	Blackburn 1998
Pigment specific normalised difference a	PSND a	$(R800-R675)/(R800+R675)$	Blackburn 1998
Pigment specific normalised difference b	PSND b	$(R800-R650)/(R800+R650)$	Blackburn 1998
Pigment specific normalised difference c	PSND c	$(R800-R500)/(R800+R500)$	Blackburn 1998
Plant senescence reflectance index	PSRI	$(R678-R500)/R750$	Merzlyak et al. 1999
R520/R500	ZT 1	$R520/R500$	Zarco-Tejada et al. 2012
R515/R570	ZT 2	$R515/R570$	Zarco-Tejada et al. 2012
R515/R670	ZT 3	$R515/R670$	Zarco-Tejada et al. 2012
Ratio analysis of reflectance spectra	RARS	$R746/R513$	Chappelle et al. 1992
Red-edge	RE	$R750/R710$	Zarco-Tejada et al. 2001
Red-edge chlorophyll index	CI <sub>red-edge</sub>	$[R800/R700]-1$	Gitelson et al. 2006
Red-edge inflection	REI	$700+40[\{((R670+R780)/2)-R700\}/(R740-R700)]$	Clevers et al. 2002
Redness index	RI	$R700/R670$	Gitelson et al. 2000

Renormalised difference vegetation index	RDVI	$(R800-R670)/(R800+R670)0.5$	Rougean & Breon 1995
Simple ratio 1	SR 1	$R750/R700$	Gitelson & Merzlyak 1994
Simple ratio 2	SR 2	$R750/R705$	Gitelson & Merzlyak 1996
Simple ratio 3	SR 3	$R752/R690$	Gitelson & Merzlyak 1996
Simple ratio 4	SR 4	$R750/R550$	Gitelson & Merzlyak 1996
Simple ratio 5	SR 5	$R700/R670$	McMurtrey et al. 1994
Simple ratio 6	SR 6	$R675/R700$	Chappelle et al. 1992
Simple ratio 7	SR 7	$R750/R710$	Zarco-Tejada et al. 2001
Simple ratio 9	SR 9	$R800/R680$	Blackburn 1998
Simple ratio 10	SR 10	$R440/R690$	Lichtenthaler et al. 1996
Simple ratio pigment index	SRPI	$R430/R680$	Peñuelas et al. 1995
Spectral polygon vegetation index	SPVI	$0.4[3.7(R800-R670)-1.2(R550-R670)]$	Vincini et al. 2006
Structure insensitive pigment index 1	SIPI 1	$(R800-R445)/(R800-R680)$	Peñuelas et al. 1995
Structure insensitive pigment index 2	SIPI 2	$(R800-R505)/(R800-R690)$	Blackburn 1998
Structure insensitive pigment index 3	SIPI 3	$(R800-R470)/(R800-R680)$	Blackburn 1998
Transformed chlorophyll absorption reflectance index	TCARI 1	$3[(R700-R670)-0.2(R700-R550)(R700/R670)]$	Haboudane et al. 2002
Transformed chlorophyll absorption reflectance index	TCARI 2	$3[(R750-R705)-0.2(R750-R550)(R750/R705)]$	Wu et al. 2008
Transformed chlorophyll absorption reflectance index/Optimised soil-adjusted vegetation index 1	TCARI/OSAVI 1	$3[(R700-R670)-0.2(R700-R550)(R700/R670)]/[1+0.16(R800-R670)/(R800+R670+0.16)]$	Haboudane et al. 2002
Transformed chlorophyll absorption reflectance index/Optimised soil-adjusted vegetation index 2	TCARI/OSAVI 2	$3[(R750-R705)-0.2(R750-R550)(R750/R705)]/[1+0.16(R750-R705)/(R750+R705+0.16)]$	Wu et al. 2008
Triangular chlorophyll index	TCI	$0.5(R700-R550)-1.5(R670-R550)*(R700/R670)0.5$	Haboudane et al. 2008
Vogelmann red-edge index 1	VRI 1	$R740/R720$	Vogelmann et al. 1993
Vogelmann red-edge index 2	VRI 2	$(R734-R747)/(R715+R726)$	Vogelmann et al. 1993

### 6.2.5 Evaluating the existing and optimised indices

In order to assess the utility of the existing and optimised indices to discriminate the H-I and I-D class pairs, analyses were undertaken as two experiments. Experiment 1 adopted an univariate approach, which employed the Jeffries-Matusita (J-M) distance (Richards & Jia 2006) to determine the discriminatory power of each index. The J-M distance ranges from zero to two (scaled values); the higher the J-M distance, the better the discriminatory power of the index. Consequently, the index with the highest J-M distance should be best at discriminating the H-I and I-D class pairs respectively. For normally distributed classes, the J-M distance is defined by Equation 6.2 (Richards & Jia 2006):

$$J_{ij} = 2(1 - e^{-B}) \quad (\text{Equation 6.2})$$

where  $B$  is the Bhattacharyya distance and is defined as (Kailath 1967):

$$B = \frac{1}{8} (m_i - m_j)^t \left\{ \frac{C_i + C_j}{2} \right\}^{-1} (m_i - m_j) + \frac{1}{2} \ln \left\{ \frac{|(C_i + C_j)/2|}{\sqrt{|C_i|}|C_j|}} \right\} \quad (\text{Equation 6.3})$$

where  $i$  and  $j$  represent the two classes being compared,  $C_i$  is the covariance matrix of  $i$ ,  $m_i$  is the mean vector of  $i$ , and  $|C_i|$  is the determinant of  $C_i$ .

Experiment 2 employed the RF algorithm as a multivariate approach to determine the best combination of indices providing the highest discriminatory power. RF (Breiman 2001) grows hundreds of unpruned classification trees (*n<sub>tree</sub>*) using bootstrap samples of the original data set. A bootstrap sample consists of approximately two thirds of the original data set. A single classification tree is then fit to each bootstrap sample. A single feature selected from a random subset of selected features (*m<sub>try</sub>*) is selected for node splitting. Trees are maximally grown, i.e. without pruning, using bagging as well as random variable selection. The final classification is based on a majority vote of predictions of all trees in the ensemble (Breiman 2001). RF was implemented using default hyperparameter values (Poona et al. 2016b). RF models were built using the randomForest library (Liaw & Wiener 2002) in the R statistical software (R Development Core Team 2019).

For Experiment 1 and Experiment 2, spectral data at week 3 were employed to determine the discriminatory ability of the existing and optimised indices. Consequently, the data at week 1 and week 2 were employed as independent datasets to validate the models.

## 6.3 RESULTS

### 6.3.1 Boruta band selection

Boruta feature selection resulted in a reduction in data dimensionality of more than 97.50% for both the H-I ( $n = 38$ ) and I-D ( $n = 40$ ) class pairs (Figure 6.3). For the H-I class pair, selected bands were

located in the VIS region ( $n = 1$ ) at 680 nm, NIR region ( $n = 31$ ) from 933 nm to 1151 nm, and SWIR region ( $n = 6$ ) from 1344 nm to 1349 nm. However, for the I-D class pair, bands were located only in the NIR region ( $n = 29$ ) from 1118 nm to 1151 nm, and SWIR region ( $n = 11$ ) from 1337 nm to 1789 nm. The greater number of bands located in the NIR region for both the H-I and I-D class pairs, indicate the significance of the NIR region in discriminating healthy, infected, and damaged seedlings.

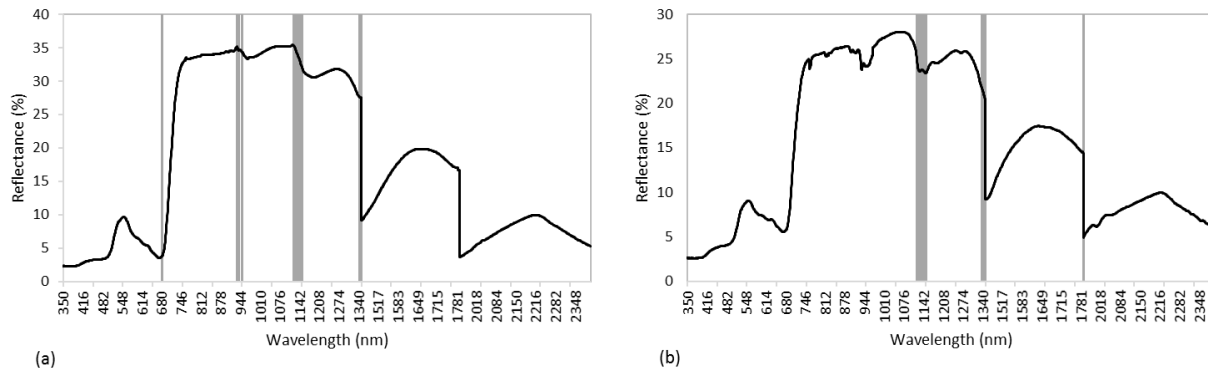


Figure 6.3: Boruta band selection (grey bars) for the H-I class pair (a) and the I-D class pair (b) for week 3. A mean spectral signature of a healthy (a) and infected (b) is shown for reference.

It is evident from Figure 6.3 that the resulting Boruta subset comprised contiguous bands. In order to further reduce the contiguity of the Boruta-selected bands, spectral resampling was employed. Several spectral resampling methodologies are available, including stepwise merging (Franke et al. 2009; Mewes et al. 2011), applying user-defined bandwidths fit to a Gaussian model (Adam et al. 2012; Adjorlolo et al. 2013), incremental resampling using user-defined band centres (Poona et al. 2016b), and incremental resampling using the mean of contiguous bands (Dalponte et al. 2009). In this study, the mean of contiguous bands was computed, resulting in a total of eleven datasets (Table 6.2). The datasets for the H-I class pair ( $n = 5$ ) and the I-D class pair ( $n = 6$ ), were subsequently employed to develop  $SIs_{opt}$  for discriminating the H-I and I-D class pairs using Equation 6.1.

Table 6.2: Resampled Boruta-selected bands for the H-I and I-D class pairs.

Class pair	Number of bands( $n$ )	Wavelengths
H-I	5	680 nm, 935 nm, 1128 nm, 1146 nm, 1347 nm
I-D	6	1118 nm, 1131 nm, 1146 nm, 1337 nm, 1345 nm, 1789 nm

### 6.3.2 Evaluating the indices

Table 6.3 shows the results of the univariate analysis using the J-M distance measure (experiment 1) and multivariate analysis using Boruta (experiment 2) for the H-I and I-D class pairs. Within a

univariate framework, the modified anthocyanin content index (mACI) (Steele et al. 2009), a two-band index, demonstrated the best discriminatory ability for the H-I class pair, yielding a J-M distance of 0.62. The new double difference index (DDIn) (Le Maire et al. 2008), which is also a two-band index, yielded a J-M distance of 0.33, proving best for discriminating the I-D class pair. The optimised indices yielded only marginally better J-M distances compared with the existing indices. For example 0.62 (existing) versus 0.71 (optimised) for the H-I class pair. Overall, the existing and optimised indices yielded low separability measures (0.33-0.71) for both the H-I and I-D class pairs. The low J-M distances indicate the difficulty in applying a single index to discriminate healthy, infected, and damaged seedlings. Worth noting is that the optimised indices for discriminating both the H-I and I-D class pairs were based on bands selected within the NIR region. The selection of NIR bands reinforces the significance of the NIR region for modelling vegetation stress.

Table 6.3: Results of the univariate and multivariate analyses for the healthy-infected (H-I) and infected-damaged (I-D) class pairs. For the univariate analysis, the best performing index (existing) and bands (optimised) are indicated in parenthesis. For the multivariate analysis, the out-of-bag error is shown, and the number of selected indices indicated in parenthesis.

Analysis Type (Experiment)	H-I		I-D	
	Existing	Optimised	Existing	Optimised
Univariate (J-M distance)	0.62 (mACI)	0.71 (1128; 1146)	0.33 (DDIn)	0.48 (1118; 1131)
Multivariate (Boruta)	90 ( $n = 111$ )	96 ( $n = 20$ )	83 ( $n = 111$ )	93 ( $n = 30$ )

Results of the multivariate analysis, using Boruta, showed a marked improvement in the ability to discriminate the H-I and I-D class pairs, compared with the univariate approach. Existing indices showed greater potential for discrimination, yielding an overall accuracy of 90% and 83% for the H-I and I-D class pairs respectively. However, such accuracies were achieved using the full set of indices ( $n = 111$ ). Conversely, the highest accuracies were obtained using a combination of only  $n = 20$  and  $n = 30$  indices for discriminating the H-I (OA = 96%) and I-D (OA = 93%) class pairs. The results indicate that (i) optimised indices outperformed the existing indices, (ii) a lesser number of optimised indices yielded higher accuracies compared with a larger number of existing indices, and (iii) a combination of indices (i.e. a multivariate framework) can better discriminate healthy, infected, and damaged seedlings compared with using a single index.

### 6.3.3 Model validation

The spectral data collected for week 1 and week 2 served as independent validation datasets. The study focussed on the optimised indices within a multivariate framework, given (i) the higher accuracies obtained using  $SIs_{opt}$  compared with using  $SIs_{ex}$ , (ii) the lesser number of  $SIs_{opt}$  used (Table



6.3), and (iii) the overarching aim of the study, which is to develop optimised indices using ML. The validation results using the data from week 1 and week 2 are presented in Table 6.4.

Table 6.4: Multivariate results for the healthy-infected (H-I) and infected-damaged (I-D) class pairs for week 1 and week 2, using optimised indices.

Dataset	H-I ( $n = 20$ )				I-D ( $n = 30$ )			
	OA (%)	KHAT	Sensitivity	Specificity	OA (%)	KHAT	Sensitivity	Specificity
Week 1	81	0.62	1	0.62	76	0.52	0.94	0.58
Week 2	82	0.64	1	0.62	91	0.82	1	0.82

The results in Table 6.4 indicate that the H-I class pair can readily be discriminated at week 1 (OA = 81%; KHAT = 0.62) using a combination of  $SI_{opt}$  ( $n = 20$ ). Conversely, the I-D class pair is more difficult to discriminate at week 1 (OA = 76%; KHAT = 0.52). However, discrimination of the I-D class pair using a combination of  $SI_{opt}$  ( $n = 30$ ) is significantly improved at week 2 (OA = 91%; KHAT = 0.82). An evaluation of the overall results presented in Table 6.3 and Table 6.4 demonstrate the ability to discriminate healthy, infected, and damaged seedlings, with high accuracy, from within one week of infection.

## 6.4 DISCUSSION

In this research, the Boruta algorithm was employed to first select the most important bands, and subsequently statistically derive and evaluate optimised two-band normalised difference spectral indices ( $SI_{opt}$ ) for the classification of healthy, infected, and damaged *P. radiata* seedlings. The discriminatory power of the  $SI_{opt}$  were compared with existing indices ( $SI_{sex}$ ;  $n = 111$ ) within a univariate as well as a multivariate framework. The following sections discuss the results in further detail.

### 6.4.1 Boruta band selection

The Boruta algorithm embedded with RF has proven to be an effective wrapper approach for feature selection and classification of hyperspectral data (Leutner et al. 2012; Poona & Ismail 2014; Poona et al. 2016a). For example, Poona & Ismail (2014) showed that using Boruta resulted in improved model performance for the classification of healthy, infected, and damaged seedlings. Using Boruta-selected wavelengths ( $n = 38$  for the H-I class pair, and  $n = 40$  for the I-D class pair), the authors achieved classification accuracies of 91% (KHAT = 0.82) and 92% (KHAT = 0.84) for discriminating the H-I and I-D class pairs respectively. Probably the most significant advantage to using Boruta is the selection of strongly relevant and weakly relevant bands (Kursa & Rudnicki 2011). Poona & Ismail (2014) and Poona et al. (2016a) showed that the selection of all relevant bands leads to superior

model development, compared with using RF importance bands as well as feature selection algorithms that select only important bands. Consequently, using Boruta-selected bands to develop two-band normalised difference spectral indices, should yield optimal results.

Boruta selected bands primarily in the NIR region for discriminating both the H-I and I-D class pair. For the H-I class pair, additional bands were located in the red and SWIR regions, whereas additional bands were only located in the SWIR region for the I-D class pair. These results are consistent with previous studies illustrating the significance of the red, NIR, and SWIR regions for modelling pest and disease stress in vegetation (Poona & Ismail 2014; Baranowski et al. 2015).

In this study, Boruta bands were selected and subsequently employed to develop two-band normalised difference indices ( $SI_{opt}$ ) to discriminate H-I and I-D class pairs. Using  $SI_{opt}$  ( $n = 20$  for the H-I class pair, and  $n = 30$  for the I-D class pair) within a multivariate framework yielded accuracies of 96% and 93% for the H-I and I-D class pairs respectively. These results represent an overall improvement on the results obtained by Poona & Ismail (2014) using subsets of discrete bands.

#### **6.4.2 Performance of existing versus optimised spectral indices**

Within a univariate framework, the two-band indices yielded the best results. Of all the existing indices evaluated ( $n = 111$ ), the modified anthocyanin content index (mACI) and new double difference index (DDIn) were the best performing  $SI_{ex}$ . The mACI is based on the ratio of reflectance in the NIR (940 nm) and green (530 nm) regions, and are used to estimate leaf anthocyanin concentration (Steele et al. 2009). The DDIn is used to estimate leaf chlorophyll content, and is based on reflectance within and proximal to the red-edge region, namely 660 nm, 710 nm, and 760 nm (Le Maire et al. 2008). Changes in chlorophyll and anthocyanin concentration have been linked to plant stress (Chalker-Scott 1999; Hernández-Clemente et al. 2017; Trojak & Skowron 2017). However, the results show that a single index demonstrated limited potential to discriminate the H-I and I-D class pairs. This was evident from a J-M distance of 0.62 (mACI), and J-M distance of 0.33 (DDIn) for the H-I and I-D class pairs respectively. Similar results were obtained using an optimised index for both the H-I and I-D class pair (Table 6.3).

The multivariate results obtained using the  $SI_{opt}$  (Table 6.4) show high sensitivity for both the H-I and I-D class pairs, at week 1 and week 2. This is indicative of the model's ability to correctly classify healthy seedlings, in the case of the H-I class pair, and infected seedlings, in the case of the I-D class pair. The lower specificity results illustrate the difficulty in correctly classifying infected seedlings, in the case of the H-I class pair, and damaged seedlings, in the case of the I-D class pair at week 1 and week 2. However, the higher specificity (0.82) for the I-D class pair illustrates improved classification of damaged seedlings at week 2. These results are likely due to the similarity in spectral

response between the respective classes during the early stages of *F. circinatum* infection (Poona & Ismail 2014).

Our results support the premise that  $SI_{opt}$  improves the modelling of asymptomatic plant stress using hyperspectral data. Improved results have since been successfully demonstrated by Prabhakar et al. (2011) and others. Additionally, applying  $SI_{opt}$  within a multivariate framework yields optimal results, thereby demonstrating operational potential. Only one study (Oumar et al. 2013) to date employed spectral indices within a multivariate framework for predicting *T. peregrinus* damage in an *E. macarthurii* plantation. The authors successfully illustrated that a combination of  $SI_{opt}$  ( $n = 20$ ) provided better predictive power compared with  $SI_{ex}$  ( $n = 23$ ).

Overall, multivariate analysis produced the best predictions, using both existing and optimised indices. No study to date has compared the utility of spectral indices within a univariate and multivariate framework. The novelty of this study is thus in the development of  $SI_{opt}$  using Boruta embedded with RF, and compare their performance against  $SI_{ex}$  within a univariate and multivariate framework. Additionally, the study successfully demonstrated the utility of  $SI_{opt}$  developed from Boruta bands for classification of healthy, infected, and damaged *P. radiata* seedlings infected with the fungal pathogen *F. circinatum*.

## 6.5 CONCLUSIONS

The overall aim of this study was to develop optimised spectral indices ( $SI_{opt}$ ) for the classification of healthy, infected, and damaged *P. radiata* seedlings. In order to achieve this aim, two-band normalised difference spectral indices were developed using Boruta-selected bands. The results of this study demonstrated several important findings:

1. The Boruta wrapper embedded with the RF ensemble provides for efficient feature selection and classification;
2. Optimised spectral indices developed using Boruta-selected bands, provide better discrimination of *Fusarium circinatum* stress in *P. radiata* compared with using existing indices, as well as using a subset of discrete spectral bands; and
3. A multivariate framework is more conducive for modelling disease stress, compared with using a univariate framework. However, this needs to be further evaluated across other stressors.

Overall, the results demonstrate the operational potential for employing optimised two-band normalised difference spectral indices developed using Boruta-selected bands for classifying healthy, damaged, and stressed *P. radiata* seedlings within a nursery environment. Ultimately, the methodology developed in this study could readily be applied to varied applications in agriculture,

forestry, ecology, and earth sciences, employing optimised normalised difference two-band spectral indices for classification.

## CHAPTER 7: MODELLING PITCH CANKER USING HIGH SPATIAL RESOLUTION SATELLITE IMAGERY

Poona NK & Ismail R 2013. Discriminating the occurrence of pitch canker fungus in *Pinus radiata* trees using QuickBird imagery and artificial neural networks. *Southern Forests: a Journal of Forest Science*, 75:1, 29-40.

Poona NK & Ismail R 2012. Discriminating the occurrence of pitch canker infection in *Pinus radiata* forests using high spatial resolution QuickBird data and artificial neural networks. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS2012), held 22-27 July 2012, Munich, Germany.

### Abstract

Pathogenic fungi, such as *F. circinatum*, present a serious threat to *P. radiata* plantations. The effective management of infected trees is thus paramount. Coupled with advanced techniques, high spatial resolution remotely sensed data provides the necessary tools to effectively identify and map infected trees. This paper explores the utility of transformed high spatial resolution QuickBird imagery and artificial neural networks for the detection and mapping of pitch canker disease. Individual tree crowns (ITCs) were delineated using an automated segmentation and classification approach within an object-based image analysis environment. Subsequently, several vegetation indices including the tasseled cap transformation were calculated and incorporated into a neural network model. The feed-forward neural network showed high discriminatory power with an overall accuracy of 82.15% and KHAT of 0.65. The results of this study show great potential for the future application of crown-level mapping of pitch canker disease at a landscape scale.

**Keywords:** *Fusarium circinatum*, artificial neural networks, QuickBird, tree crown

## 7.1 INTRODUCTION

Pitch canker is an episodic disease caused by the fungus *F. circinatum* (teleomorph = *Gibberella circinata*), and infects only *Pinus* forests. In South Africa, *P. radiata* and *P. patula* are recognised as being the two most susceptible forest species (Wingfield et al. 1999). Of an estimated 1.3 million hectares of commercial forest occurring in South Africa, 51% of these forests comprise of *Pinus* plantations with 58 000 hectares of *P. radiata* trees planted almost exclusively in the Western Cape (DAFF 2010). The first outbreak of pitch canker in South Africa was observed at Sappi's Ngodwana nursery in 1990 (Morris 2010). Fifteen years later, the first outbreak occurred in a forest stand of *P. radiata* in Tokai Plantation (Coutinho et al. 2007).

The accurate detection and monitoring of pitch canker at a wide spatial coverage is required for an effective management programme. Although no effective mechanisms are available for treating pitch canker (Wingfield et al. 2008), detection and monitoring of diseased trees remain an essential component of an effective management programme. Storer et al. (2002) noted the tendency of a number of individuals within a stand to sustain a few infections prior to any individual becoming severely infected. The detection of infected individuals could thus provide insight into the spatial patterns of disease occurrence, and the opportunity to develop models of disease spread. Remote sensing can provide timely information on forest health status (Coops et al. 2009; Woodall et al. 2010), and at a lower cost than traditional field sampling (Pouliot et al. 2002). Additionally, remotely sensed data provide the necessary spatial and spectral information that can be directly related to the biochemical and biophysical properties of forests.

The utility of high spatial resolution (1-4 m) multispectral imagery has been extensively investigated in forest health studies. The finer spatial resolution allows for the detection of stress and disease at the crown-level (Lee & Cho 2006), thereby allowing for discrimination of individual healthy and diseased trees (Coops et al. 2006a). The impact of stress in forest trees is often first manifested in the crowns. Tree crowns are thus a good indicator of general tree health and vigour (Zarnoc et al. 2004; Goodwin et al. 2005). Subsequently, Leckie et al. (2005) noted that tree crown damage can be accurately classified using automated ITC delineation, or once the pixels defining a crown have been identified.

Several studies (for example Bunting & Lucas 2006; Whiteside & Boggs 2009; Boggs 2010) have demonstrated that object-based image analysis (OBIA) can accurately delineate crowns. OBIA may be defined as the segmentation (partitioning) of image data into meaningful objects (segments) that can then be analysed (Hay & Castilla 2006). OBIA builds on the concepts of segmentation, edge detection, feature extraction, and classification (Blaschke 2010). The OBIA approach is a rule-based

system that uses rule-sets (a set of threshold rules) to define how information is used to assign classes (for example crown, soil, shadow). These rules are used to define the segmentation, delineation, and subsequent classification of objects.

Crown delineation from high spatial resolution multispectral data has been adopted in several forest health assessment studies. Leckie et al. (2004) mapped *Pseudotsuga menziesii* infected by *Phellinus weirii* using 0.6 m CASI airborne imagery, with classification accuracies ranging from 55 to 82%. White et al. (2005) used IKONOS spaceborne imagery to map *P. contorta* infected by *D. ponderosae*, and achieved classification accuracies ranging from 71 to 92%. Using digital multispectral airborne imagery Sims et al. (2007) achieved classification accuracies of 92% and 68% when mapping *P. radiata* trees infected by *Diplodia pinea* and *Essigella californica* respectively.

Spectral information from remote sensing imagery has been successfully used as indicators of forest health (for example Malenovsky et al. 2009; Ghiyamat & Shafri 2010). Leaf reflectance is largely dependent on concentrations of chlorophylls and carotenoids (Lichtenthaler et al. 1998). Changes in chlorophyll concentration as a result of stress, is thus directly correlated to changes in leaf reflectance (Lichtenthaler et al. 1998; Carter & Knapp 2001). Healthy trees thus exhibit high reflectance in the NIR region with corresponding low reflectance in the red region (Lichtenthaler et al. 1998; Wang et al. 2010) due to high leaf chlorophyll content. Conversely, stressed trees exhibit increased reflectance in the far red region and negligible change in reflectance in the NIR region (Lichtenthaler et al. 1998; Carter & Knapp 2001), unless stress is coupled with changes in leaf cellular structure or water content (Govender et al. 2009). Consequently, vegetation indices (VI), derived from these spectral bands, are widely used in remote sensing studies to characterise forest health and vigour (Jackson & Huete 1991; Wang et al. 2010).

Probably the most widely used VI is the normalised difference vegetation index (NDVI), which has been extensively used in forest health assessment studies (for example Coops et al. 2006a; Ismail et al. 2008). Wang et al. (2010) provide a good review of the application of VIs and spectral transformations in forest health studies. A commonly used spectral transformation that provides information with regards to forest health and vitality is the tasseled cap transformation (TCT). TCT re-projects the spectral bands along the principal directions of brightness, greenness, and yellowness (Skakun et al. 2003; Yarbrough et al. 2005). The brightness axis is associated with overall background reflectance, greenness axis with variations in vegetation vigour, and yellowness related to variations in senescence (Mather 1999; Karl & Maurer 2010). Research has shown that chlorosis associated with damaged trees are organised along the directions of the new transformed bands (Skakun et al. 2003). Consequently, Lee & Cho (2006) applied the TCT to detect damaged *Pinus densiflora* and *Pinus thunbergii* infested by *Bursaphelenchus xylophilus*. Similarly, Coops et al. (2006a, 2006b), and



Wulder et al. (2006a, 2006b) used the TCT to successfully distinguish *P. contorta* stands infected by *D. ponderosae*.

Several approaches are available for the classification of remotely sensed data. Classification approaches based on the Bayesian decision rule, such as the GML classifier, have been extensively used in the remote sensing of forest health while employing high spatial resolution multispectral imagery. For example, using QuickBird imagery, Hick & Logan (2009) employed the ML classifier to map red-attack *Pinus albicaulis* trees infected by *D. ponderosae*. However, the utility of the ML classifier is limited due to the (i) assumption that the data fits a normal (Gaussian) distribution, and (ii) low classification accuracies achieved when the data does not fit an adequate multivariate statistical model, and the number of training samples are limited (Atkinson & Tatnall 1997; Waske et al. 2009). ML approaches such as support vector machines (SVM) and artificial neural networks (ANN) have been increasingly used for processing and interpreting remote sensing imagery. SVM perform better than or at least equally well as most state of the art classifiers (Burgess 1998; Waske et al. 2009). Unfortunately, SVM are limited due to the (i) challenge in selection of the appropriate kernel and appropriate kernel function parameters (Burgess 1998; Frohlich & Zell 2005), and (ii) resulting models' susceptibility to overfitting (Burgess 1998).

Modelled on the efficiency of the human brain to process and interpret large amounts of data (Atkinson & Tatnall 1997), ANNs are highly suited for modelling ecological data, which often display non-linear relationships and rarely follow a parametric statistical distribution (Jensen et al. 1999; Cunningham et al. 2009). The key advantage to using ANNs with remotely sensed data is that the algorithm makes no underlying assumption on the data distribution in feature space (Atkinson & Tatnall 1997; Kavzoglu 2009). Within forest environments, ANNs have been employed in a wide array of applications including forest change detection (Mas et al. 2004; Kehl et al. 2012), forest structure mapping (Ingram et al. 2005), forest growth modelling (Huang et al. 2012), and forest health modelling (Klobučar et al. 2010). The most widely used ANN is the multilayer perceptron (MLP) ANN trained with an error backpropagation algorithm (Huang 2009; Waske et al. 2009; Günther & Fritsch 2010).

To summarise, although extensive research with regards to the application of high spatial resolution remote sensing to forest health is on-going, only a limited number of studies have investigated the application to pathogenic fungal damage of forests (for example, Kelly 2002; Leckie et al. 2004), and investigated the potential of high spatial resolution imagery for automated ITC delineation and subsequent crown-level tree health assessment (for example, Leckie et al. 2004; White et al. 2005; Sims et al. 2007). This research focuses on testing the utility of high spatial resolution imagery for quantifying forest health. Specifically, this paper explores the utility of QuickBird imagery and the



MLP neural network (NN), to identify and discriminate healthy and diseased *P. radiata* trees infected by *F. circinatum*.

## 7.2 MATERIALS AND METHODS

### 7.2.1 Site description

The study was undertaken at Tokai Plantation, situated on the eastern slopes of the Peninsula mountains, neighbouring the Cape Peninsula National Park, between latitudes 34°1'43" and 34°4'50" S and longitudes 18°23'25" and 18°26'37" E (Figure 7.1). Tokai is one of the oldest plantations in South Africa, with a total plantation area of 756.4 hectares. The topography ranges from very steep slopes to level sand flats. The geology is mainly of the Peninsula formation with resultant soils being mainly erodible clay to loam on the upper slopes, and sandy soils lower down. Annual precipitation ranges from 996 mm to 1270 mm, with rainfall mainly during the winter months, May to August. Temperatures range from 1°C to 38°C, with a mean annual temperature of 15°C. The local geology and climate provides ideal site qualities for the production of pine veneer and sawlogs (Kirkman 2009).

### 7.2.2 Detecting pine pitch canker in the field

*F. circinatum* infects the vegetative and reproductive parts of susceptible trees (Wingfield et al. 2008; Dreaden & Smith 2010). Infection can occur throughout the year (Wingfield et al. 2008), but is limited by the prevailing environmental conditions (Gordon 2006). The first symptom of pitch canker is the wilting and discolouration of needles (known as flagging) (Figure 7.2a), which eventually fall off, resulting in branch dieback (Gordon et al. 2001; Aegerter et al. 2003). Dieback occurs from the branch tips to the infection sites, as a result of girdling cankers obstructing water flow (Gordon et al. 2001).

Larger-diameter branches including the main stem (trunk) eventually also become infected (Aegerter et al. 2003). These infections are characterised by bleeding resinous cankers (Figure 7.2b) (Coutinho et al. 2007; Wingfield et al. 2008). Repeated infections result in increased disease severity and consequent extensive tip dieback (Figure 7.2c) in the canopy (Aegerter et al. 2003; EMPPO 2005), which is the primary and most noticeable symptom of the disease. Many of the stressed trees may also become infested by *Pissodes nemorensis* (Figure 7.2d) (Gebeyehu & Wingfield 2003). Infection of the main stem and larger branches causes accelerated tree decline, with girdling of the main stem resulting in tree death (EMPPO 2005).

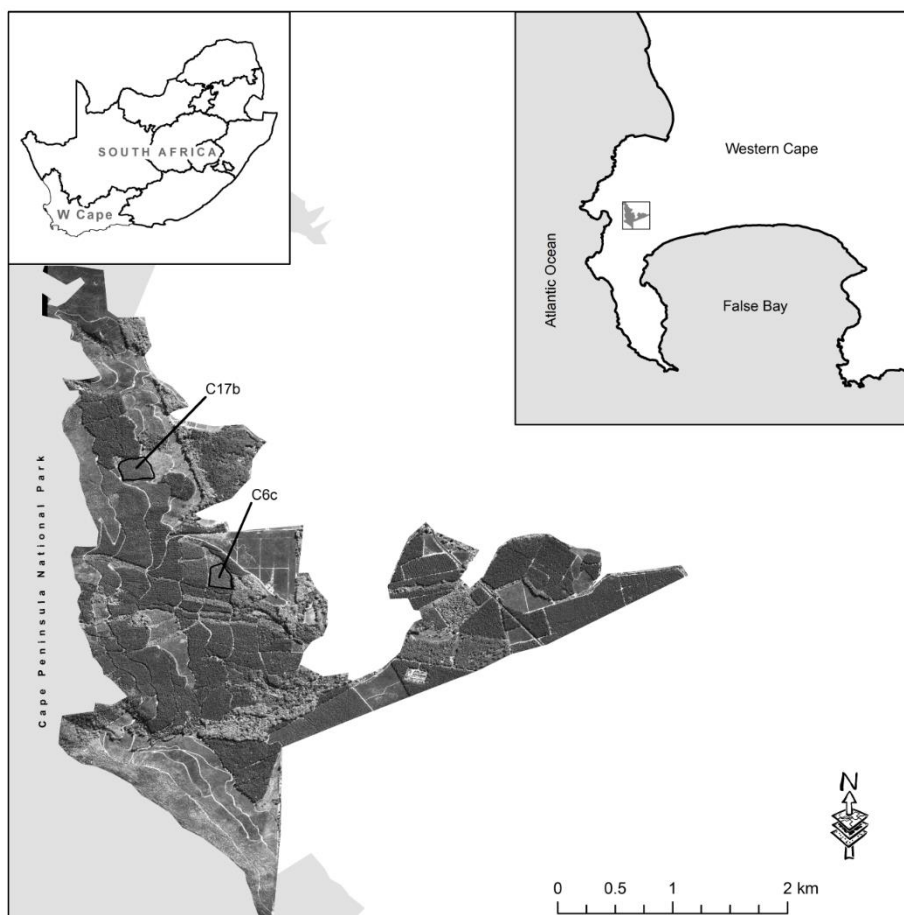


Figure 7.1: Tokai Plantation situated in the Western Cape, South Africa. The two compartments (C6c and C17b) selected for the study are indicated. Background image consists of the QuickBird panchromatic band (0.6 m).

### 7.2.3 Field and image data

MTO Forestry (Pty) Ltd. forest inventory data were used to identify prospective sites (compartments) based on tree age, area, number of stems per hectare (based on survival rate at four years), and the health status of compartments (Figure 7.1; Table 7.1). MTO Forestry inventory data (captured June 2008), combined with Forestry and Agricultural Biotechnology Institute data (captured August 2006 and April 2007) were used to identify and confirm the health status of trees. A follow-up field visit was conducted in August 2010. Healthy and infected trees were identified using a visual assessment procedure. A healthy tree was defined as a tree showing no signs of PCF infection, whereas an infected tree was defined as one showing signs of flagging, branch dieback, and/or bleeding resinous cankers (Figure 7.2).

Table 7.1: The two MTO forest compartments selected for the study.

Site	Compartment	Age (months)	Espacement	Area (ha)	Stems per ha	Health status
1	C6c	81	3 x 3 m	2.741	989	Infected
2	C17b	81	3 x 3 m	3.898	989	Healthy





Figure	Photograph	Symptom
(a)		<p>Wilting and chlorosis of pine needles (flagging). Evidence of the early signs of pitch canker fungus infection.</p>
(b)		<p>Bleeding resinous canker on the main stem. Characteristic of the advanced stage of pitch canker fungus infection.</p>
(c)		<p>Extensive canopy dieback resulting from multiple branch tip infections.</p>
(d)		<p>Secondary infection by the pine weevil <i>Pissodes nemorensis</i>. The weevil has been recorded as both a vector and wounding agent related to <i>F. circinatum</i> infection.</p>

Figure 7.2: *P. radiata* showing signs of pitch canker disease. Infected trees express varied stages of infection, ranging from flagging to extensive canopy dieback, with many of the trees exhibiting advanced stage of infection.

A total of 400 trees (200 per compartment) were randomly sampled. Trees that were open-grown, at the edge of an opening, or at the periphery of the forest compartment were not sampled as it was found by Leckie et al. (2004) that these crowns exhibited markedly different spectral signatures from the rest of the compartment. The health status of each sampled tree was confirmed and noted. Each sampled tree was subsequently identified and marked on a hardcopy print of crowns delineated from a high resolution orthorectified aerial photograph (12.5 cm). The dataset ( $n = 400$ ) was randomly split into a training dataset (66%) and a test dataset (34%).

A single QuickBird image (ortho ready standard, 25 km<sup>2</sup> subset) was acquired on 5 March 2008. QuickBird acquires 11-bit data in four multispectral bands with a 2.44 m spatial resolution. The spectral resolution of the multispectral bands is as follows: 0.447–0.512  $\mu\text{m}$  (blue); 0.499–0.594  $\mu\text{m}$  (green); 0.620–0.688  $\mu\text{m}$  (red) and 0.755–0.874  $\mu\text{m}$  (NIR). Additionally a single panchromatic band (0.525–0.924  $\mu\text{m}$ ), with a 0.6 m spatial resolution was also acquired (Krause 2005). The image was orthorectified and subsequently atmospherically corrected using the QUick Atmospheric Correction (QUAC) module in ENVI 4.8 (ITTVIS 2010).

#### **7.2.4 Crown-level assessment**

This study employed an automated segmentation and classification that incorporated a local maxima detection and region growing approach (Ke & Quackenbush 2007; Li et al. 2008) to delineate tree crowns. Ke & Quackenbush (2011) provide a good review of the various approaches to automatic individual crown detection and delineation. The analysis was undertaken within an OBIA environment using eCognition Developer 8 (Trimble 2011).

The following steps describe the procedure employed in eCognition:

1. A rule-set was developed using the four QuickBird multispectral bands and the single panchromatic band.
2. A multi-threshold segmentation was used on the panchromatic band to mask out shadow and bare soil as 'background'.
3. An NDVI layer was generated from the multispectral bands and used in another multi-threshold segmentation to mask out non-vegetation as 'background'.
4. Following a chess-board segmentation (using the resolution of the panchromatic band), pixels (not classified as 'background') with the local maximum value (in a 3x3 filter) was delineated as seeds.
5. Seeds were grown to the size of 13 pixels (4.68 m<sup>2</sup>) and the objects merged, resulting in the final delineated tree crowns.

The accuracy and geo-location of the crown delineation process was assessed utilising individual trees (Wang et al. 2004) and an aggregated approach (Pouliot et al. 2002). Individual assessments were based on evaluating the correspondence between the OBIA crown delineations and manually delineated crowns (Leckie et al. 2005). Tree crowns were manually delineated on a very high spatial resolution (12.5 cm) orthorectified colour aerial photograph as undertaken in previous studies (for example Ke & Quackenbush 2009; Bunting et al. 2010). The aggregated assessment was based on comparing the OBIA crown delineations with MTO field inventory data (enumerated as stems per hectare).

### 7.2.5 Signature extraction

Reflectance data from the QuickBird spectral bands were used to derive a spectral signature for each OBIA-delineated crown ( $n = 400$ ). The signature for each crown thus comprised the average value of pixels (representing each crown) extracted from each of the four spectral bands. Figure 7.3 shows the average spectral values of healthy and infected crowns for the blue, green, red, and NIR bands. A one-way analysis of variance (ANOVA) was then used to determine if there was a statistically significant difference between the two crown classes, i.e. healthy and infected, based on the spectral information extracted from the four QuickBird multispectral bands.

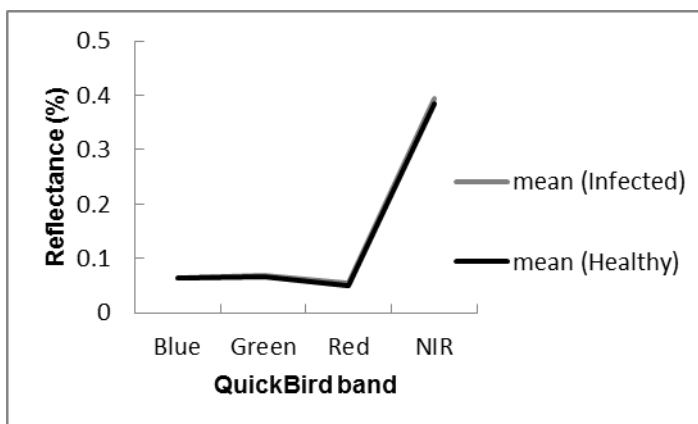


Figure 7.3: Mean signature values of the infected ( $n = 200$ ), and healthy ( $n = 200$ ) crowns derived from the four QuickBird spectral bands.

### 7.2.6 Vegetation indices and transformations

VIs are calculated as either the ratio of two or  $n$  spectral bands, or the linear combination of  $n$  spectral bands (Jackson & Huete 1991). Selection of the VIs used in this study was based on a review of satellite remote sensing in forest health studies by Wang et al. (2010). The selected VIs have been used in several forest health studies (Table 7.2). Previously only available for Landsat MSS (Kauth & Thomas 1976), Landsat TM (Crist & Cicone 1984), and IKONOS (Horne 2003) imagery,

Yarbrough et al. (2005) developed TCT coefficients for QuickBird imagery using the Gram-Schmidt Orthogonalisation method. The TCT is calculated as the linear combination of the four QuickBird spectral bands. The brightness, greenness, and yellowness components for the QuickBird imagery were calculated based on the TCT coefficients (Table 7.2) provided by Yarbrough et al. (2005). VI and TCT were subsequently computed using the crown signatures extracted from the four QuickBird spectral bands, and formed input values to the ANN model.

Table 7.2: The five vegetation indices and tasseled cap transformations used in this study.

Index	Estimation	Forest health Reference
Modified Soil-Adjusted Vegetation Index (MSAVI <sub>2</sub> )	$\frac{2R_{NIR} + 1 - \sqrt{(2R_{NIR} + 1)^2 - 8(R_{NIR} - R_{RED})}}{2}$	Bonneau et al. (1999)
Normalised Difference Vegetation Index (NDVI)	$\frac{R_{NIR} - R_{RED}}{R_{NIR} + R_{RED}}$	Coops et al. (2006a)
Red-Green Index (RGI)	$\frac{R_{RED}}{R_{GREEN}}$	Coops et al. (2006a) Wulder et al. (2008)
Simple Ratio (SR)	$\frac{R_{NIR}}{R_{RED}}$	Royle & Lathrop (1997)
Wide Dynamic Range Vegetation Index (WDRVI)	$\frac{(\alpha + 1)NDVI + (\alpha - 1)}{(\alpha - 1)NDVI + (\alpha + 1)}$	Eklundh et al. (2009)
Tasseled Cap - Brightness	$(0.319R_{BLUE}) + (0.542R_{GREEN}) + (0.49R_{RED}) + (0.604R_{NIR})$	
Tasseled Cap - Greenness	$(-0.121R_{BLUE}) + (-0.331R_{GREEN}) + (-0.517R_{RED}) + (0.78R_{NIR})$	
Tasseled Cap - Wetness	$(0.652R_{BLUE}) + (0.375R_{GREEN}) + (-0.639R_{RED}) + (-0.163R_{NIR})$	

### 7.2.7 Neural network model

A multilayer feed-forward ANN (Skidmore et al. 1997) was used to discriminate between infected and healthy trees at a crown-level using remotely sensed data. The network was implemented in Waikato Environment for Knowledge Analysis 3.7.5 (WEKA) open source data mining software (Hall et al. 2009). According to Skidmore et al. (1997) the learning process of a feed-forward network consists of two phases; a feed-forward phase and a backpropagation phase. In the first phase (forward phase), the input values  $o_i$  (spectral bands, VI, TCT) are multiplied with the weight value  $w_{ij}$  for each



node in the hidden layer (Equation 7.1). The resulting value  $Z_j$  is then transformed by a sigmoidal activation function (Equation 7.2), adding non-linearity to the network. The output value  $o_k$  is calculated by multiplying the values for each node  $o_j$  in the hidden layer by the weight value  $w_{jk}$ . For a three-layered ANN,  $z_k$  can be similarly calculated as in Equation 7.1. The feed-forward phase stops after the output value  $o_k$  has been calculated.

$$Z_j = \sum_j w_{ij} o_i \quad (\text{Equation 7.1})$$

where  $Z_j$  is the sum of the products at the hidden nodes,  $o_i$  is the input data, and  $w_{ij}$  is the weight.

$$O_j = \frac{1}{1 + e^{-(z_j + \theta)/\theta_o}} \quad (\text{Equation 7.2})$$

where  $\theta$  is a threshold or bias, and  $\theta_o$  is a constant.

The second phase (backpropagation phase) involves calculation of the root mean square error (RMSE) of the predicted value, that is, the difference between the input and output value. This information is passed backwards through the network, coupled with a weight adjustment. This cycle represents a single epoch. A number of epochs are repeated iteratively until either a local RMSE is attained, or the number of specified epochs is completed. Learning is thus achieved *via* an iterative process (Kavzoglu 2009).

#### *Preparing the input data*

The input data to the ANN model comprised the crown-level spectral data extracted from the four QuickBird multispectral bands, the five VIs calculated from the crown-level spectral data, and the three TCT components calculated using the coefficients in Table 7.2. Skidmore et al. (1997) demonstrated that the training speed of ANN can be improved by normalising the input data between 0 and 1. This was achieved using Equation 7.3:

$$X_{input} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (\text{Equation 7.3})$$

where  $X_{input}$  is the normalised input parameters,  $X_i$  is the parameter value,  $X_{min}$  is the minimum value, and  $X_{max}$  is the maximum value in the training dataset.

#### *Optimising the neural network model*

Network training involved performing a number of runs while varying the learning rate and momentum factor. The learning rate reflects the amount the weights are updated during each cycle, whereas the momentum factor is the momentum applied to the weights during updating. The learning

rate, momentum factor, and number of epochs were optimised using a training dataset ( $n = 66\% \times 400 = 264$ ). Model performance was assessed using a test dataset ( $n = 34\% \times 400 = 136$ ) that was not used during the network training process. Implementation of the MLP ANN followed the procedure used by Skidmore et al. (1997), as explained in the section ‘Neural network model’. For a more detailed description of ANN, please see for example Huang et al. (2009) and Waske et al. (2009). The ability of the model to correctly discriminate healthy and infected tree crowns based on the test data (34%) was evaluated using a contingency matrix. The contingency matrix provides descriptive statistics in terms of overall accuracy, and user’s and producer’s accuracy. Additionally, a discrete multivariate technique called Kappa was used, which provided a KHAT statistic as a measure of agreement between the training and test data (Congalton & Green 2009).

## 7.3 RESULTS

### 7.3.1 Crown-level assessment

To ensure high spatial accuracy of the automated tree crown delineation, high correspondence between the OBIA crown delineations and the manually delineated crowns was required. Based on the recommendations of Leckie et al. (2005), a good match was defined as (i) a greater than 50% overlap between the OBIA and manually delineated crowns, and (ii) a 1:1 correspondence, that is, only one OBIA-delineated crown associated with only one manually delineated crown. Based on these recommendations, a total of 2 264 and 3 139 crowns in compartments C6c and C17b were respectively identified as having good correspondence. Figure 7.4 shows the automatically delineated tree crowns selected on the basis of the good correspondence achieved between the OBIA and the manual crown delineations. The OBIA tree crown delineations also compared well with the MTO inventory data. MTO inventory data consisted of 989 stems per ha and was based on survival rate at four years. More specifically, the aggregated assessment showed that the OBIA tree crown delineations accounted for more than 80% of trees in both compartments.

It should be noted that from the 5 403 OBIA-delineated crowns, the sample used in this study consisted of 400 crowns that were field verified and used for model development and testing. Subsequently, the model was used to predict the condition of the remaining crowns in an effort to produce an operational product for the detection and monitoring of the pitch canker disease.



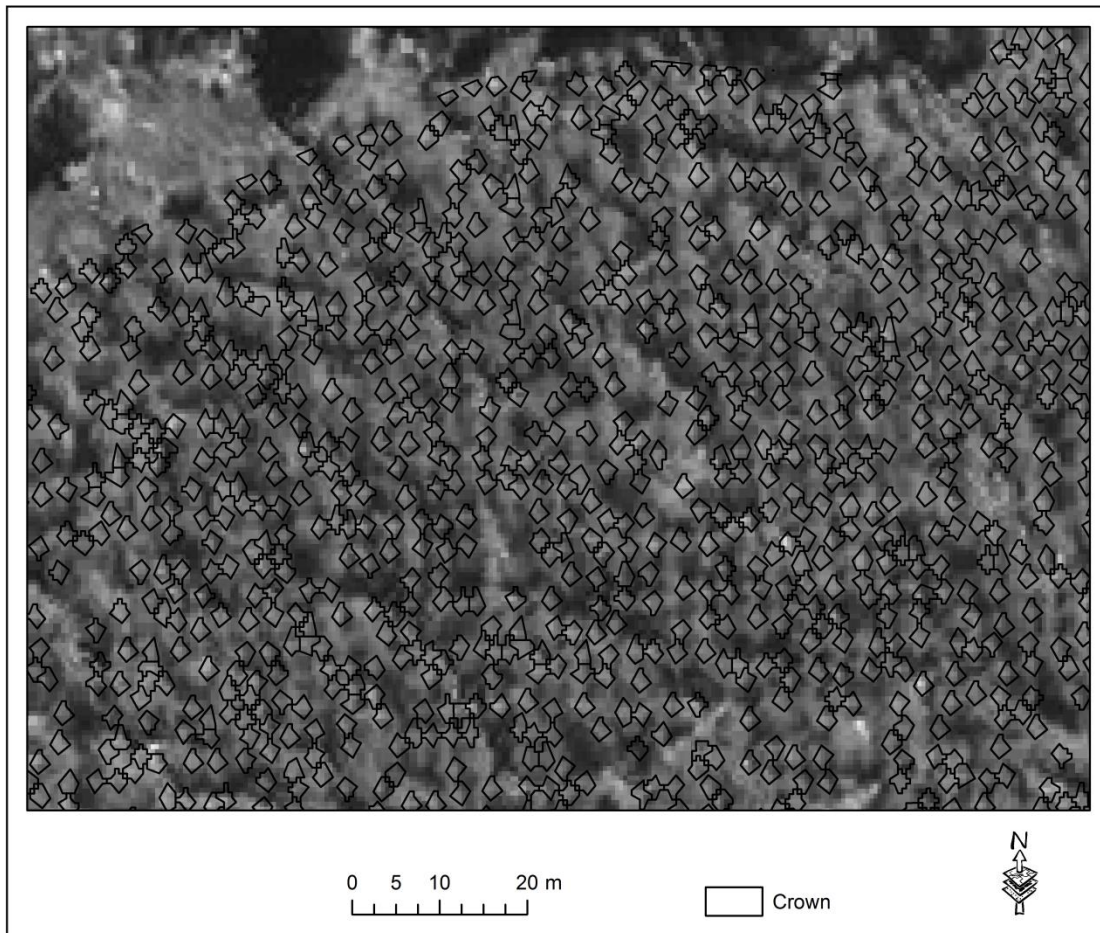


Figure 7.4: Automatically delineated tree crowns in compartment C17b ( $n = 3\ 139$ ) selected on the basis of correspondence between automated isolations (isols) and the manual delineations (greds). Background image consists of the QuickBird panchromatic band.

### 7.3.2 Neural network model

Results of the ANOVA indicated that there was a significant difference ( $p < 0.001$ ) between the QuickBird bands and the crown classes. These signatures subsequently formed the basis for calculating the VI and TCT, which were inputs to the MLP ANN model.

Network topology was defined by a 12:7:2 architecture: twelve input nodes, a single hidden layer with seven hidden layer nodes, and two output nodes (Figure 7.5). The network was defined by a single hidden layer, based on the recommendations by Mills et al. (2006) and Mas & Flores (2008). Additionally, the number of hidden layer nodes was calculated by averaging the sum of the number of attributes and classes (Witten et al. 2011). The network thus comprised three layers; an (i) input layer  $i$ , (ii) internal (hidden) layer  $j$ , and (iii) output layer  $k$ .

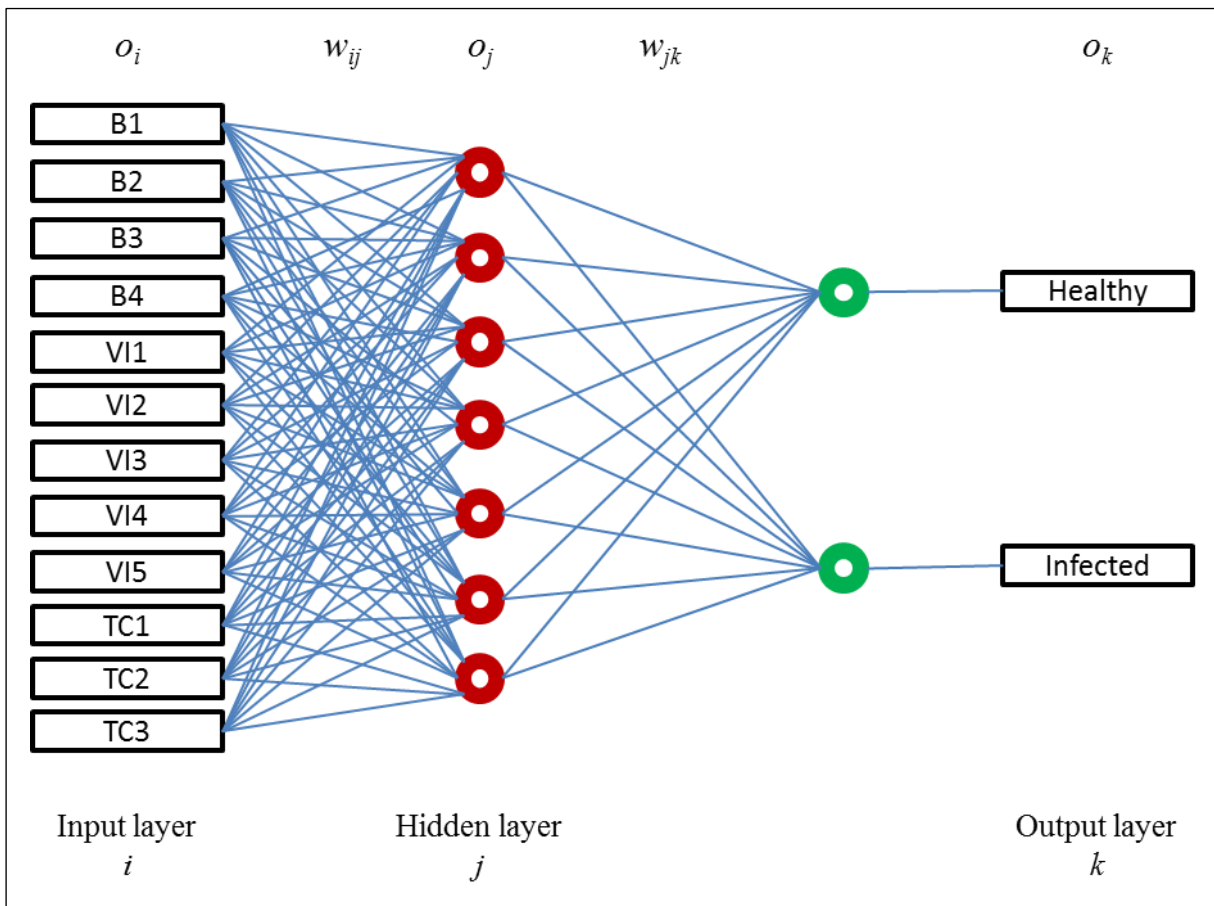


Figure 7.5: Neural network topology used in this study. The input layer consisting of the mean crown values extracted from the four QuickBird multispectral bands (B1-B4), five vegetation indices (VI1-VI5), and three tasseled cap components (TC1-TC3) are connected to a single hidden layer with seven hidden layer nodes, which are in turn connected to an output layer with two output nodes representing the crown status.

Figure 7.6 shows the overall results of varying the learning rate, momentum factor (MF), and number of epochs on model performance using a 66-34 percentage split for training ( $n = 264$ ) and testing ( $n = 136$ ) the ANN model. The learning rate was varied from 0.01 to 0.3 in increments of 0.01, and the MF varied from 0.1 to 0.5 in increments of 0.1. The training times (number of epochs to train) used were 500, 1 000, 5 000, 10 000, and 20 000 epochs. As the number of epochs increased from 500 (Figure 7.6a) to 20 000 (Figure 7.6e), the overall average model accuracy decreased from 78.54% to 76.68%. The decrease in model accuracy was coupled to a decrease in the average KHAT from 0.57 to 0.53, and an increase in the average RMSE from 0.40 to 0.45. Furthermore, Figure 7.6 shows that the overall accuracy decreased with increasing learning rate as well as with increasing MF.

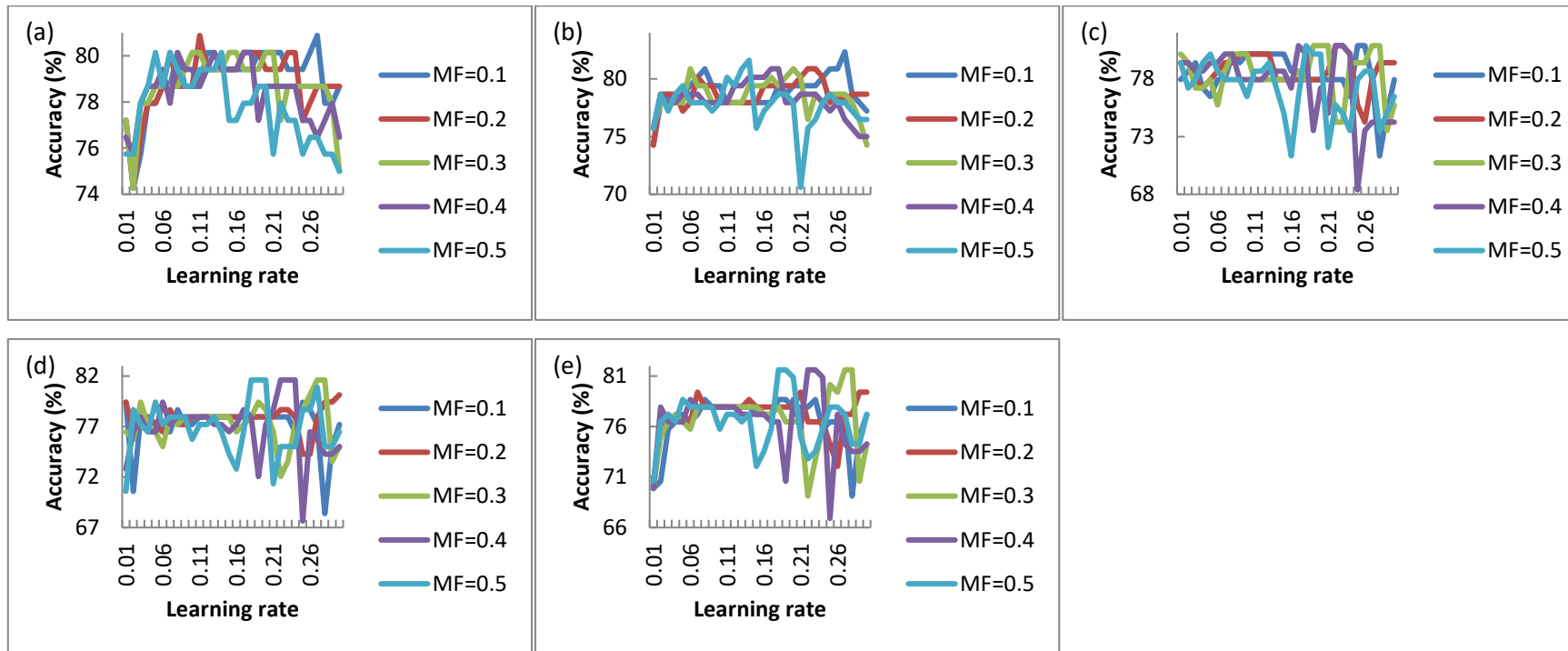


Figure 7.6: Relationship between learning rate, training time, momentum factor (MF), and prediction accuracy for discriminating healthy and infected crowns. Training time was varied at (a) 500, (b) 1 000, (c) 5 000, (d) 10 000, and (e) 20 000 epochs.

Selection of the optimal model was based on the network that yielded the highest accuracy and KHAT statistic, coupled with the lowest RMSE. The results of the accuracy assessments are presented in Table 7.3. The optimised model parameters consisted of a learning rate of 0.27, a momentum factor of 0.1, a training time of 1 000 epochs (Figure 7.6b) and RMSE of 0.40. An overall accuracy of 82.35% coupled with high user's and producer's accuracies above 80%, indicated that the MLP model was able to adequately and accurately classify healthy and infected crowns while a KHAT of 0.65 indicated good agreement between the train and test data.

Table 7.3: Classification accuracy for the multilayer perceptron model using the test dataset ( $n = 136$ ).

	<b>Healthy</b>	<b>Infected</b>
<b>User's accuracy (%)</b>	80.60	84.06
<b>Producer's accuracy (%)</b>	83.08	81.69
<b>Overall accuracy</b>	82.35	
<b>KHAT</b>	0.65	

The model was subsequently used to predict the condition of the remaining delineated crowns ( $n = 5\ 003$ ) in an effort to produce an operational product for the detection and monitoring of the pitch canker disease. The model developed on the training dataset assigned each crown from the test dataset a probability from one to 100% with respect to belonging to either the healthy or the infected class. A probability greater than or equal to 50% was assigned as true for that class, while a probability of less than 50% was assigned as false for that class. For example, for the infected class, if a crown had a probability score of 51%, the final class assignment was infected. The result of the prediction on the delineated crowns for compartment C6c ( $n = 2\ 264$ ) is illustrated in Figure 7.7. Of the 2 264 crowns, 1 517 (67%) were scored as being infected, while 747 (33%) were scored as being healthy.

## **7.4 DISCUSSION**

This study represented a first attempt at the application of high spatial resolution remotely sensed data and data mining techniques, specifically a backpropagation ANN, for detecting and mapping trees infected by pine pitch canker. The remote sensing framework developed in this study provides the impetus for the implementation of an efficient forest health monitoring system. The utility of the proposed framework could facilitate the improved management of infected compartments, and provide significant information with regards to pitch canker infection rates and distributions. The sections below discuss the results in more detail.



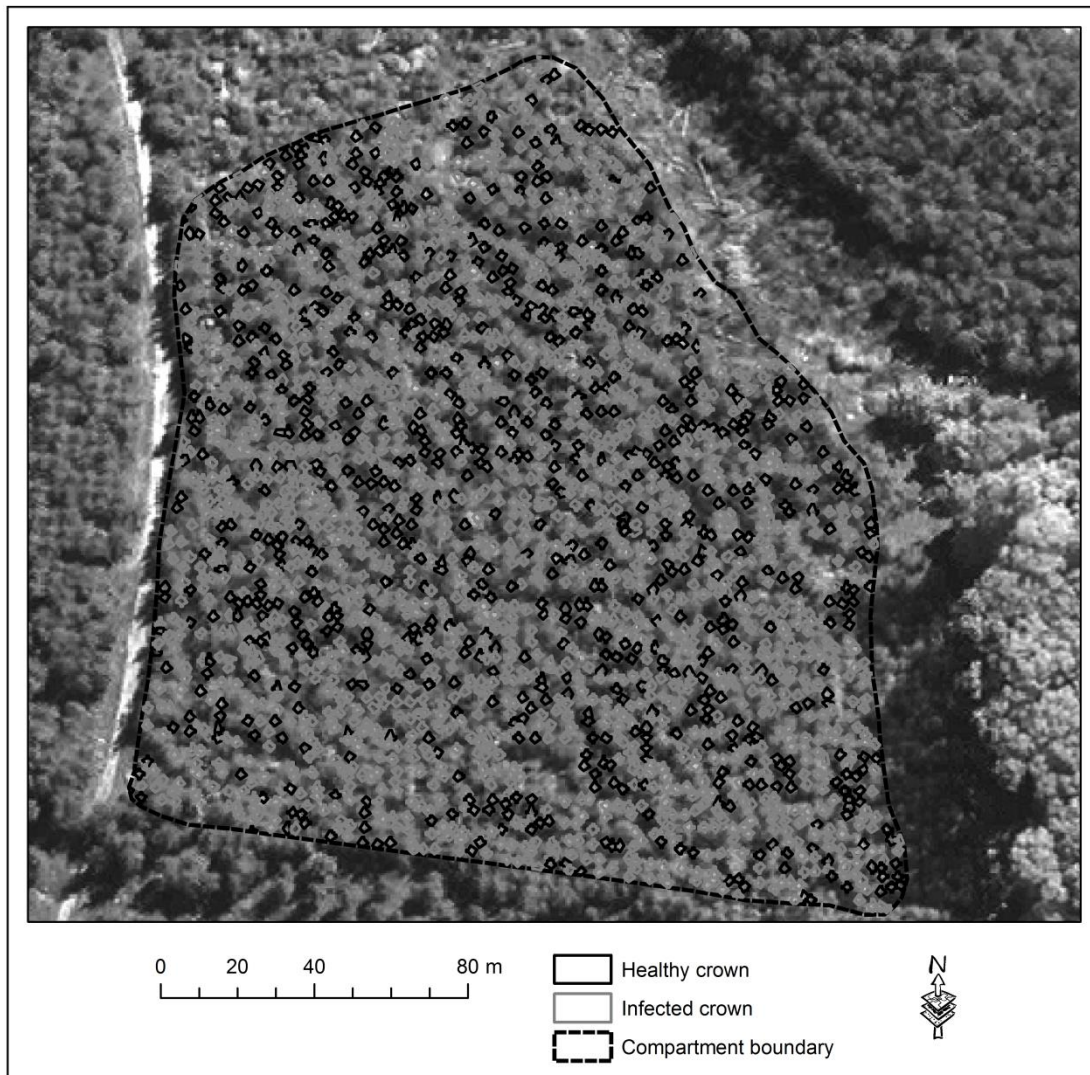


Figure 7.7: Predicted distribution of healthy and infected crowns for compartment C6c. The background image consists of the QuickBird panchromatic band.

#### 7.4.1 Benefits of crown-level assessments

Automated tree crown delineation within an OBIA environment was achieved with relatively high efficiency and accuracy. The 0.6 m (panchromatic) spatial resolution of QuickBird imagery was sufficiently high for detecting and delineating *P. radiata* crowns. The accuracies attained for automated crown delineation compared favourably to studies by Wang et al. (2004) who achieved 83% correspondence between manually and automatically delineated crowns and Bunting & Lucas (2006) who achieved delineation accuracies ranging from ~19-67%.

#### 7.4.2 Neural network model

ML algorithms such as ANN represent an alternative approach to exploiting remotely sensed data for assessing forest health (Klobučar et al. 2010). The performance of the ANN model is ultimately determined by the network topology, for which there is no standard operating rules, and subsequently

varies with dataset and application. ANN, in particular the MLP model, have been extensively applied in a multitude of remote sensing studies, with authors demonstrating their robustness, efficiency to handle large datasets, and high predictive power.

Within this context, a MLP ANN was built and trained to discriminate healthy and infected *P. radiata* crowns ( $n = 400$ ), using spectral information from high spatial resolution QuickBird imagery, and VIs and TCT components derived from the four QuickBird spectral bands. Network topology was defined using the literature as a guide. The learning rate, MF, and training time was varied, such that an optimal model could be acquired that could best discriminate healthy and infected *P. radiata* crowns. An overall accuracy of 82.35%, user's and producer's accuracies of above 80%, and a KHAT of 0.65 indicated that the model developed in this study expressed high discriminatory power in classifying healthy and infected crowns, based on a sample of 400 crowns.

The application of ANN is however not without limitations. The number of hidden layer nodes, training set size, and number of epochs was of considerable importance during model development, directly impacting on model performance. These factors affect the model's ability to generalise, that is, interpolate and extrapolate test data (Atkinson & Tatnall 1997; Huang 2009). Several authors (for example Atkinson & Tatnall 1997; Skidmore et al. 1997; Del Frate et al. 2007) have shown that increased training time results in overtraining (overfitting the data), that is, when the network learns from the training data but is unable to generalise to the new (test) data. Additionally, the training of an ANN is computationally expensive (Huang 2009; Waske et al. 2009). Consequently, ANN models must be optimised with several (10s, often 100s) runs in order for the model to attain good generalisation capabilities.

Despite these limitations, the results of this study indicate that high spatial resolution imagery and ANN (i.e. MLP model) provide a relevant framework for the effective management of pitch canker at the crown-level. Automated crown delineation for large compartments across extensive landscapes can be readily achieved using the OBIA approach proposed in this study. This would provide full coverage of the study area at a crown-level. Infected individual trees can thus be isolated, followed by proactive prevention of further spread. Further research is however required towards understanding the mechanisms of infection, disease spread, host susceptibility, and symptom expression in individual *P. radiata* trees. It is worth investigating the potential for discriminating pine trees at varied levels of infection intensity. Several authors, for example Wulder et al. (2008) and Ismail et al. (2008) having undertaken studies on *D. ponderosae* and *S. noctilio* respectively, discriminated between a green attack, red-attack, and grey-attack stage. Discriminating the stages of attack provides for estimation of insect populations, and thus has implications for developing

effective disease management protocols. Although these stages have only been defined for insect attack of forests, this principle could be extrapolated and applied to fungal infection and damage.

## 7.5 CONCLUSIONS

This paper explored the utility of QuickBird imagery and the MLP ANN, to identify and discriminate healthy and diseased *P. radiata* trees infected by *F. circinatum*. Major findings of this study include:

1. The results of this research indicate that high spatial resolution QuickBird imagery can successfully identify and discriminate healthy and infected *P. radiata* trees at a crown-level;
2. An overall accuracy of 82.35%, user's and producer's accuracies above 80%, and a KHAT of 0.65 indicate that the model expressed high discriminatory power in classifying healthy and infected crowns, based on a sample of 400 crowns; and
3. The remote sensing and ML framework proposed in this study can be operationalised to identify individual infected crowns and estimate density of infected trees as part of an efficient management routine.

## CHAPTER 8: REMOTE SENSING OF FOREST HEALTH: A SYNTHESIS

### 8.1 SUMMARY

Remote sensing technologies are frequently employed as a non-contact, non-destructive, indirect assessment method to collect data on vegetation health. Such data are often collected (i) in situ using handheld and / or terrestrial sensors (Ju et al. 2014; Oumar & Mutanga 2014; Loggenberg et al. 2018); often referred to as proximal sensing, (ii) by employing airborne platforms such as unmanned aerial vehicles (Näsi et al. 2015; Dash et al. 2017; Näsi et al. 2018; Sandino et al. 2018) and aircraft (Fassnacht et al. 2014; Näsi et al. 2018), and (iii) from sensors on-board spaceborne (satellite) platforms (Kumaresan 2018; Kayet et al. 2019). The sensors employed are either imaging or non-imaging. Spectral signals acquired using non-imaging sensors, such as the handheld FieldSpec spectroradiometer, are directly processed to extract information on vegetation status and plant health (Poona & Ismail 2012b; 2014). Spectral data from imaging sensors, such as QuickBird, are processed to model spatial variability in vegetation condition (Poona & Ismail 2012a; 2013).

ML algorithms are most often employed for processing high dimensional data. Popular learning algorithms include LDA, DTs, ANNs, NB classifier, k-NN, and SVMs (Zhou 2012). However, ensemble methods such as boosting and bagging are widely used. RF (Breiman 2001), an extension of bagging, is a popular choice for hyperspectral data analysis. Additionally, algorithms that reduce data dimensionality are often employed. Embedded with a wrapper, RF provides an efficient framework for feature selection and classification of hyperspectral data (Poona & Ismail 2013; 2014; Poona et al. 2016a).

In the context of this study, in situ remotely sensed data were employed to model plant health / forest condition. Spectral measurements collected using a handheld spectroradiometer were used to develop statistical models for discriminating healthy and stressed *Pinus* seedlings infected with the *F. circinatum* pathogen. Additionally, imaging data were analysed to detect and spatially model healthy and stressed *P. radiata* trees within a forest plantation. The following sections provide a synthesis of the research in the context of its scientific merit, aim and objectives, methodology, gaps and assumptions, and future opportunities.

### 8.2 SCIENTIFIC MERITS OF THE RESEARCH

Infected seedlings can remain asymptomatic for extended periods, making visual stress detection infeasible. Additionally, laboratory analysis is often required to confirm cause of mortality (Gordon et al. 2015). Thus, the ability to detect asymptomatic stress within the nursery environment is



paramount. According to Vainio et al. (2019), early detection coupled with effective monitoring and control, can lead to eradication of *F. circinatum*. The availability of hyperspectral remote sensing technology and ML techniques, presented a unique opportunity to develop a framework for in situ early detection of stressed (infected and damaged) seedlings. This research evaluated and successfully demonstrated the utility of hyperspectral remote sensing and ML for asymptomatic stress detection of *F. circinatum*-infected *P. radiata* and *P. patula* seedlings.

This research (Chapters 3, 4, and 5) illustrated the successful implementation of Boruta with RF for feature selection and classification of hyperspectral data. Deriving a subset of the most important / relevant bands is important for two primary reasons; developing customised sensors comprising specific bands, and using existing multispectral sensors employing narrowband interference filters (Stone & Mohammed 2017). In both cases, the targeted wavelengths are those sensitive to *F. circinatum* infection. Such sensing setups could provide an operational, cost-effective, real-time detection tool for industry.

The ability to identify stress / disease related to a specific stressor is significant to industry. In this research, artificially damaged seedlings served as a positive control to determine whether physical damage and *F. circinatum* infection could be discriminated. The results of the study presented in Chapter 3, confirms that stress due to damage and stress due to *F. circinatum* infection present unique spectral responses, and can thus be discriminated. Healthy and infected seedlings were discriminated with 91% accuracy (KHAT = 0.82), and infected and damaged seedlings with 92% accuracy (KHAT = 0.84). Extrapolating the success of this study, it is likely that individual biotic and abiotic stressors will present with unique spectral responses that can be discriminated using hyperspectral data and ML. This notion is valuable within a nursery environment, as pathogenic stress, e.g. due to *F. circinatum* infection, which almost always results in seedling mortality, must be separable from other stresses such as drought, which can be reversed.

This research (Chapter 6) further illustrated the successful development and implementation of optimised two-band normalised difference spectral indices for *F. circinatum* detection. The impetus for developing such indices is the need for *F. circinatum*-specific indices, given that  $SI_{\text{ex}}$  are not stressor / disease-specific (Mahlein et al. 2013). Healthy and infected seedlings were discriminated with 96% accuracy using  $SI_{\text{ex}}$  ( $n = 20$ ), whereas infected and damaged seedlings were discriminated with 93% accuracy using  $SI_{\text{ex}}$  ( $n = 30$ ). The specific bands used to derive the  $SI_{\text{opt}}$  can be used to develop customised sensors that are optimised for detecting *F. circinatum* infection. Such a system could prove invaluable within a nursery environment.

Despite the high number of studies exploiting hyperspectral data and ML for plant stress detection, this research is novel in investigating *F. circinatum* in *Pinus* using hyperspectral data and ML. The

remote sensing–machine learning framework proposed in this research illustrates a viable and cost-effective approach to managing the health of commercial forestry. The proposed framework could provide for rapid detection and identification of asymptomatic infected and / or damaged seedlings within a nursery environment. The infected and / or damaged seedlings can be removed from the nursery and destroyed, thereby limiting the transfer of infected seedlings to the field, and reducing post-planting mortality. Ultimately, such a framework would form part of an integrated management solution, i.e. eradicating *F. circinatum* from nurseries.

### 8.3 REVISITING THE AIM AND OBJECTIVES

This research set out to develop a framework employing remotely sensed data and ML technologies and techniques, for near real-time cost-effective forest health monitoring. Specifically, this research aimed to evaluate the utility of high spectral and high spatial resolution remotely sensed data for modelling *F. circinatum*-induced stress in *P. radiata* and *P. patula*. The ability to successfully detect *F. circinatum*-induced stress would form part of a monitoring framework that could be operationalised within a commercial forestry environment. In order to achieve this aim, three objectives were set.

The first two objectives, i.e. explore the utility of RF for feature selection and classification of healthy and stressed *Pinus* seedlings using multitemporal hyperspectral data, and evaluate the utility of Boruta for identifying the most important spectral bands for modelling *F. circinatum* stress in *Pinus* seedlings, are inextricably linked, as illustrated in Chapters 3, 4, and 5. This research supports the utility of RF for hyperspectral data analysis. More importantly, this research establishes the utility of RF and multitemporal spectroscopy data for developing asymptomatic stress detection models. Additionally, Boruta (Kursa & Rudnicki 2010) was shown to be an efficient wrapper for data dimensionality reduction and selection of the most relevant bands for model development. Results presented in Chapter 3 demonstrate the success achieved using the RF-Boruta framework.

Despite the potential demonstrated by the RF-Boruta framework for asymptomatic stress detection, it was imperative to explore other frameworks. Consequently, Chapter 4 investigated the efficiency of Boruta compared with RFE (Díaz-Uriarte & Alvarez de Andrés 2006) and AUC-RF (Calle et al. 2011), within the RF framework. The three feature selection algorithms were evaluated according to their ability to discriminate healthy, infected, and damaged *P. radiata* and *P. patula* seedlings. The RF-Boruta framework proved to be most efficient, outperforming the RF-RFE and AUC-RF models. The promising results yielded using the RF-Boruta framework provided the impetus to extend the analyses to develop spectral indices. Chapter 6 explored the utility of Boruta-selected bands for developing optimised two-band normalised difference spectral indices (SIs<sub>opt</sub>) for discriminating

healthy, infected, and damaged *Pinus* seedlings. Statistically deriving an index from a linear combination of only two spectral bands could improve classification accuracy.  $SI_{opt}$  were benchmarked against existing indices ( $SI_{ex}$ ), within a univariate and multivariate framework, to gauge their performance and value as an operational tool. Overall, using  $SI_{opt}$  yielded improved classification accuracies, illustrating their value for classification.

The third objective, i.e. to test the viability of high spatial resolution multispectral data for modelling pitch canker stress in *P. radiata*, was addressed by employing high spatial resolution QuickBird satellite imagery to identify healthy and pitch canker infected *P. radiata* trees within a commercial stand. Classification models were developed using the feed-forward back propagation MLP ANN. Despite the challenges with implementation of ANN, the MLP ANN model yielded promising results. This study (Chapter 7) is the first of its kind. High spatial resolution multispectral data were applied to model *F. circinatum* damage in a *P. radiata* forest, employing an automated ITC delineation approach (using OBIA) for crown-level tree health assessment.

#### **8.4 STRENGTHS AND LIMITATIONS OF THE METHODOLOGY, AND ASUMPTIONS MADE**

The methodology adopted in this study is rooted in the ability to statistically detect change (i.e. deviations) in a spectral signature between healthy and stressed plants. At leaf scale, the spectral signature of vegetation is influenced by leaf surface properties, internal structure, and biochemistry. Stress in plants induces defence responses, resulting in altered biochemical and physiological mechanisms that are exhibited as deviations in the spectral signature (Martinelli et al. 2015). Thus, when comparing signatures of healthy and stressed plants, the deviations in the signatures, e.g. blue shift can be statistically computed, and used as an indicator of asymptomatic stress.

The empirical nature of the individual studies; experimental by design, provides a basis for determining causality. The basic premise is that healthy seedlings express a “healthy signature”, infected seedlings express an “infected signature”, and damaged seedlings express a “damaged signature”. The experiments and concomitant results are based on controlled experiments, i.e. simulations, which are common in empirical studies. However, these experiments need to be replicated within a nursery environment, i.e. in the real-world, and on various sets of data to determine their generalisation and true operational ability. The use of field spectroscopy data allows for the experiments to be replicated on a larger scale.

The implementation of ML algorithms for modelling is a key strength of this research. ML is a proven methodology for analysis of high dimensional data (Lary et al. 2016). Ensemble learners, in particular, have grown in popularity for classification of hyperspectral data. This research employed RF as a

tree-based ensemble learner. RF's strengths were exploited for classification as well as feature selection. The efficiency of feature selection was improved using wrappers, in particular, Boruta. Identifying and isolating the most relevant bands, compared with the most important bands, yielded more accurate models for the given classification problem.

The primary datasets used in this study were acquired using a handheld spectroradiometer. Despite being an in situ (proximal) mode of data collection, signals may nonetheless be affected by atmospheric water vapour, scattering, background radiance, and wind (McCoy 2005). Additionally, when collecting multitemporal data it is impractical to maintain a consistent viewing geometry for every target at every time (t). In order to minimise these environmental effects, as well as the effect of different viewing geometries, (i) data collection was scheduled between 10:00 and 15:00 each week, (ii) the spectroradiometer was systematically optimised using the Spectralon® white reference panel (Poona & Ismail 2014), (iii) data collection was paused during cloud overpass and / or changes in air movement, and (iv) seedlings were scanned five times through a 360° rotation (Hatchell 1999). This study focussed on modelling a single stress factor, *F. circinatum*, in two *Pinus* species, *P. radiata* and *P. patula*. Seedlings were sown from uninfected seed, and manually inoculated; via topping (in the case of *P. radiata*) and via the soil medium (in the case of *P. patula*). Seedlings were maintained in a semi-controlled environment; in a greenhouse (in the case of *P. radiata*) and under shade cloth (in the case of *P. patula*) for the duration of the experiments. Several key assumptions are that (i) all seedlings were healthy prior to sampling, and subsequent damage and inoculation, (ii) there was no transfer of infection to healthy seedlings, (iii) no cross infection occurred while seedlings were housed in the nursery, (iv) no external stress was present; stress expression was attributed to *F. circinatum* infection in the case of infected seedlings, and topping in the case of damaged seedlings.

The remote sensing–machine learning framework proposed in this study is not a solution to eradicating *F. circinatum* from nurseries. *F. circinatum* in nurseries is a complex problem, influenced by environmental stress, host physiology, and forest management (Gordon et al. 2015). Rather, the framework should complement current nursery management protocols and assist in curbing (i) the nursery to field infection transfer rates, (ii) seedling mortality in field, and ultimately (iii) economic losses.

## 8.5 CONCLUSIONS

This research has investigated the utility of both in situ hyperspectral and spaceborne multispectral remotely sensed data, coupled with ML techniques to develop a framework for the cost-effective and near real-time detection and identification of *F. circinatum* and pitch canker in *Pinus* seedlings and

trees. The overall results show that the individual studies are successful in developing such a framework. From these results, the following conclusions are drawn:

1. Remote sensing is an invaluable technology for the forestry sector. The utility of in situ spectroscopy for developing asymptomatic stress detection models shows much promise, and requires further investigation;
2. Ensemble learners such as RF, are a viable approach to developing accurate and robust classification models;
3. Boruta embedded with RF provides for efficient feature selection and classification of hyperspectral (high dimensional) data. Despite the assumptions and limitations of this study, the successes achieved with the RF-Boruta framework highlight the potential of these algorithms for developing of an operational tool for implementation within a nursery environment;
4. Non-imaging spectroscopy can readily be employed within a nursery environment, for detecting, identifying, and quantifying stress / disease prior to symptom expression; and
5. The methodology successfully adopted and demonstrated in this research can be extrapolated to other domains such as agriculture for crop health and yield estimation, geology for mineral mapping, and the environment for modelling contaminants and vegetation stress.

## 8.6 RECOMMENDATIONS AND FUTURE STUDY

A fully operational remote sensing framework for forest health monitoring should ideally include detection (changes / deviations in spectral signature of healthy seedling), identification (discriminating *F. circinatum* stress from other stressors; linking specific bands to a specific stressor), and quantification (pitch canker disease severity; level / stages of infection) protocols (Mahlein et al. 2012; Martinelli et al. 2015). This study focussed on detection, and in part, identification of *F. circinatum* stress, as part of a monitoring framework within a commercial environment. The ability to detect multiple stressors, at varying severity levels, as well as identification (diagnosis) of stressor-specific symptoms, is yet to be assessed. A future study should make use of multi-stressor samples, for example *F. circinatum* (Poona & Ismail 2014), *S. noctilio* (Ismail & Mutanga 2011), physical damage (Poona & Ismail 2014), and drought (Xulu et al. 2019). Such a study could form the foundation for developing models to discriminate multiple stress agents, as well as identify and isolate the stress / disease-specific bands (Mahlein et al. 2013).

Unsupervised tree-based clustering (Peerbhay et al. 2015; Afanador et al. 2016; Dalleau et al. 2018; Xulu et al. 2019) represents a promising approach to asymptomatic detection of stressed seedlings in a nursery, as well as detection and quantification of pitch canker in forest stands. An unsupervised

approach eliminates the need for a priori knowledge regarding the health status of the sampled seedlings / trees used for model building. Stressed seedlings / trees are detected as anomalies, and clustered accordingly; clustering is based on the proximity matrix (Afanador et al. 2016). A future study could implement an URF methodology using the Boruta-selected bands identified in this research. If successful, such an approach that combines RF-Boruta and URF could facilitate / enhance the framework proposed in this research.

This research has identified several variants of the traditional RF model that requires investigation, specifically within a hyperspectral remote context. Of the particular interest are the oRF and rotF models that yielded better classification results and robustness compared with RF. The oRF and rotF models certainly require further investigation for the analysis of hyperspectral as well as other high dimensional remote sensing datasets.

Further research is required to investigate the utility of imaging spectrometry to detect, identify, and quantify pitch canker, both at crown and stand level. Such a study could make use of the myriad of manned airborne hyperspectral systems available, such as FPI (Näsi et al. 2018), AISA (Niemann et al. 2015), and HyMap (Fassnacht et al. 2014), as well as numerous unmanned systems (Adão et al. 2017). Unfortunately, sensor availability and accessibility, pilot licensing, and logistical costs, are limiting factors to deploying airborne systems, particularly in South Africa. A viable alternative to airborne systems is spaceborne systems. Operational and future hyperspectral missions (Table 8.1) represent opportunities for spaceborne multitemporal monitoring of forest health. However, it is unlikely that all missions will provide global coverage. Additionally, data policies may restrict “full, free, and open” data access.

Table 8.1: Operational and future spaceborne imaging spectrometer missions.

Status	Sensor	Agency	Reference / URL [accessed 30 October 2019]
Operational	Hyperspectral Precursor and Application Mission (PRISMA)	Italy	<a href="https://earth.esa.int/web/eoportal/satellite-missions/p/prisma-hyperspectral">https://earth.esa.int/web/eoportal/satellite-missions/p/prisma-hyperspectral</a>
	Compact High Resolution Imaging Spectrometer (CHRIS)	European Space Agency (ESA)	<a href="https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/proba/instruments/chris">https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/proba/instruments/chris</a>
	HyperSpectral Imaging Satellite (HysIS)	India	<a href="https://directory.eoportal.org/web/eoportal/satellite-missions/h/hysis">https://directory.eoportal.org/web/eoportal/satellite-missions/h/hysis</a>
	HyperSCOUT	ESA	<a href="https://hyperscout.nl/">https://hyperscout.nl/</a> <a href="https://directory.eoportal.org/web/eoportal/satellite-missions/g/gomx-4">https://directory.eoportal.org/web/eoportal/satellite-missions/g/gomx-4</a>
	DLR's Earth Sensing Imaging Spectrometer (DESI)	Germany (German Aerospace Centre)	<a href="https://www.dlr.de/content/en/articles/news/2018/2/20180629_hyperspectral-earth-observation-instrument-desis-sets-off-for-the-iss_28665.html">https://www.dlr.de/content/en/articles/news/2018/2/20180629_hyperspectral-earth-observation-instrument-desis-sets-off-for-the-iss_28665.html</a>
Future	Environmental Mapping and Analysis Program (EnMAP)	Germany	<a href="https://earth.esa.int/web/eoportal/satellite-missions/e/enmap">https://earth.esa.int/web/eoportal/satellite-missions/e/enmap</a>
	Spaceborne Hyperspectral Applicative Land and Ocean Mission (SHALOM)	Israel; Italy	Feingersh & Ben-Dor (2015)
	Surface Biology and Geology (SBG)	National Aeronautics and Space Administration (NASA)	<a href="https://science.nasa.gov/earth-science/decadal-sbg">https://science.nasa.gov/earth-science/decadal-sbg</a>
	Hyperspectral Infrared Imager (HyspIRI)	NASA	<a href="https://hyspiri.jpl.nasa.gov/">https://hyspiri.jpl.nasa.gov/</a>
	Hyperspectral X IMagery (HypXIM)	France	Michel et al. (2011)
	Copernicus Hyperspectral Imaging Mission (CHIME)	ESA	Nieke & Rast (2018)



	FLEX	ESA	<a href="https://eoportal.org/web/eoportal/satellite-missions/f/flex">https://eoportal.org/web/eoportal/satellite-missions/f/flex</a>
	Hyperspectral Imager Suite (HISUI)	Japan	<a href="https://eoportal.org/web/eoportal/satellite-missions/content/-/article/iss-utilization-hisui-hyperspectral-imager-suite-">https://eoportal.org/web/eoportal/satellite-missions/content/-/article/iss-utilization-hisui-hyperspectral-imager-suite-</a>

## REFERENCES

- Abdel-Rahman EM, Ahmed FB, Van den Berg M & Way MJ 2010. Potential of spectroscopic data sets for sugarcane thrips (*Fulmekiola serrata* Kobus) damage detection. *International Journal of Remote Sensing* 31, 4199-4216.
- Abdel-Rahman EM, Makovi DM, Landmann T, Piironen R, Gasim S, Pellikka P & Raina SK 2015. The utility of AISA Eagle hyperspectral data and random forest classifier for flower mapping. *Remote Sensing* 7, 13298-13318.
- Abdel-Rahman EM, Mutanga O, Adam E & Ismail R 2014. Detecting *Sirex noctilio* grey-attacked and lightning-struck pine trees using airborne hyperspectral data, random forest and support vector machines classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing* 88, 48-59.
- Abdel-Rahman EM, Van den Berg M, Way MJ & Ahmed FB 2009. Hand-held spectrometry for estimating thrips (*Fulmekiola serrata*) incidence in sugarcane. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 12–17 July, Cape Town, South Africa. doi:10.1109/IGARSS.2009.5417322.
- Abdel-Rahman E, Way M, Ahmed F, Ismail R & Adam E 2013. Estimation of thrips (*Fulmekiola serrata* Kobus) density in sugarcane using leaf-level hyperspectral data. *South African Journal of Plant and Soil* 30, 91-96.
- Abellán J & Castellano JG 2017. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications* 73, 1-10.
- Adam E, Deng H, Odindi J, Abdel-Rahman EM & Mutanga O 2017. Detecting the early stage of Phaeosphaeria leaf spot infestations in maize crop using in situ hyperspectral data and guided regularized random forest algorithm. *Journal of Spectroscopy* 2017. doi:10.1155/2017/6961387.
- Adam E, Mutanga O & Ismail R 2013. Determining the susceptibility of *Eucalyptus nitens* forests to *Coryphodema tristis* (cossid moth) occurrence in Mpumalanga, South Africa. *International Journal of Geographical Information Science* 27, 1-15.
- Adam EM, Mutanga O, Rugege D & Ismail R 2012. Discriminating the papyrus vegetation (*Cyperus papyrus* L.) and its co-existent species using random forest and hyperspectral data resampled to HyMap. *International Journal of Remote Sensing* 33, 552-569.
- Adão T, Hruška J, Pádua L, Bessa J, Peres E, Morais R & Sousa JJ 2017. Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing* 9, 1110. doi: 10.3390/rs9111110.

- Addink EA, de Jong SM & Pebesma EJ 2007. The importance of scale in object-based mapping of vegetation parameters with hyperspectral imagery. *Photogrammetric Engineering and Remote Sensing* 73, 905-912.
- Adjorlolo C, Mutanga O, Cho MA & Ismail R 2013. Spectral resampling based on user-defined inter-band correlation filter: C3 and C4 grass species classification. *International Journal of Applied Earth Observations and Geoinformation* 21, 535-544.
- Afanador NL, Smolinska A, Tran TN & Blanchet L 2016. Unsupervised random forest: a tutorial with case studies. *Journal of Chemometrics* 30, 232-241.
- Aegerter BJ, Gordon TR, Storer AJ & Wood DL 2003. Pitch canker: A technical review. Turf Image, Inc., Monterey, California.
- Agapiou A, Hadjimitsis DG & Alexakis DD 2012. Evaluation of broadband and narrowband vegetation indices for the identification of archaeological crop marks. *Remote Sensing* 4, 3892-3919.
- Agjee NH, Ismail R & Mutanga O 2016. Identifying relevant hyperspectral bands using Boruta: a temporal analysis of water hyacinth biocontrol. *Journal of Applied Remote Sensing* 10, 042002. doi:10.1117/1.JRS.10.042002.
- Agjee NH, Mutanga O, Peerbhay K & Ismail R 2018. The impact of simulated spectral noise on random forest and oblique random forest classification performance. *Journal of Spectroscopy*. doi:10.1155/2018/8316918.
- Alexandridis TK, Katagis T, Gitas IZ, Silleos NG, Eskridge KM & Gritzas G 2010. Investigation of aggregation effects in vegetation condition monitoring at a national scale. *International Journal of Remote Sensing* 24, 507-521.
- Alhusain L & Hafez AM 2017. Cluster ensemble based on random forests for genetic data. *BioData Mining* 10. doi:10.1186/s13040-017-0156-2.
- Ali AM, Darvishzadeh R, Skidmore AK & Van Duren I 2017. Specific leaf area estimation from leaf and canopy reflectance through optimization and validation of vegetation indices. *Agricultural and Forest Meteorology* 236, 162-174.
- Amaratunga D, Cabrera J & Lee Y-S 2008. Enriched random forests. *Bioinformatics* 24, 2010-2014.
- Ang JC, Mirzal A, Haron H & Hamed HNA 2016. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13, 971-989.
- Apan A, Held A, Phinn S & Markley J 2004. Detecting sugarcane 'orange rust' disease using EO-1 Hyperion hyperspectral imagery. *International Journal of Remote Sensing* 25, 489-498.

- Apan A, Datt B & Kelly R 2005. Detection of pests and diseases in vegetable crops using hyperspectral sensing: a comparison of reflectance data for different sets of symptoms. Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial conference of the Spatial Sciences Institute, 14-16 September, Melbourne, Australia.
- Archibald R & Fann G 2007. Feature selection and classification of hyperspectral images with support vector machines. *IEEE Geoscience and Remote Sensing Letters* 4, 674-677.
- ASD Inc. 2011. ViewSpec Pro V6.0.11. Analytical Spectral Devices Inc. (ASD), Boulder, CO, USA.
- Ashourloo D, Mobasheri MR & Huete A 2014a. Developing two spectral disease indices for detection of wheat leaf rust (*Puccinia triticina*). *Remote Sensing* 6, 4723-4740.
- Ashourloo D, Mobasheri MR & Huete A 2014b. Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. *Remote Sensing* 4, 5107-5123.
- Asner GP 1998. Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sensing of Environment* 64, 234-253.
- Atkinson PM & Tatnall ARL 1997. Introduction: Neural networks in remote sensing. *International Journal of Remote Sensing* 18, 699-709.
- Augereau O, Journet N & Domenger J-P 2011. Document images indexing with relevance feedback: an application to industrial context. Proceedings of IEEE International Conference on Document Analysis and Recognition (ICDAR), 18-21 September, Beijing, China. doi:10.1109/ICDAR.2011.240.
- Bajwa SG, Rupe JC & Mason J 2017. Soybean disease monitoring with leaf reflectance. *Remote Sensing* 9, 127. doi:10.3390/rs9020127.
- Ballings M & Van den Poel D 2017. Fit and deploy rotation forest models. In Package 'rotationForest'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/rotationForest/index.html> [last accessed 10 September 2019].
- Baron D & Erasmi S 2017. High resolution forest maps from interferometric TanDEM-X and multitemporal Sentinel-1 SAR data. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 85, 389-405.
- Baranowski P, Jedryczka M, Mazurek W, Babula-Skowronska D, Siedliska A & Kaczmarek J 2015. Hyperspectral and thermal imaging of oilseed rape (*Brassica napus*) response to fungal species of the genus *Alternaria*. *PLoS ONE* 10. doi:10.1371/journal.pone.0122913.

- Barnes JD, Balaguer L, Manrique E, Elvira S & Davison AW 1992. A reappraisal of the use of DMSO for the extraction and determination of chlorophylls a and b in lichens and higher plants. *Environmental and Experimental Botany* 32, 85-100.
- Bassa Z, Bob U, Szantoi Z & Ismail R 2016. Land cover and land use mapping of the iSimangaliso Wetland Park, South Africa: Comparison of oblique and orthogonal random forest algorithms. *Journal of Applied Remote Sensing* 10. doi:0.1117/1.JRS.10.015017.
- Behnamian A, Millard K, Banks SN, White L, Richardson M & Pasher J 2017. A systematic approach for variable selection with random forests: achieving stable variable importance values. *IEEE Geoscience and Remote Sensing Letters* 14, 1988-1992.
- Behnel S, Bradshaw R, Citro C, Dalcin L, Seljebotn DS & Smith K 2011. Cython: the best of both worlds. *Computing in Science & Engineering* 13, 31–39.
- Belgiu M & Drăguț L 2016. Random forest in remote sensing - A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114, 24-31.
- Biau G 2012. Analysis of a random forests model. *Journal of Machine Learning Research* 13, 1063-1095.
- Blackburn 1998. Spectral indices for estimating photosynthetic pigment concentrations: A test using senescent tree leaves. *International Journal of Remote Sensing* 19, 657-675.
- Blagus R & Lusa L 2010. Class prediction for high dimensional class-imbalanced data. *BMC Bioinformatics* 11. doi:10.1186/1471-2105-11-523.
- Blaschke T 2010. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* 65, 2-16.
- Blaser R & Fryzlewicz P 2015. Random rotation ensembles. *Journal of Machine Learning Research* 2, 1-15.
- Blum AL & Langley P 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 245-271.
- Boggs GS 2010. Assessment of SPOT 5 and QuickBird remotely sensed imagery for mapping tree cover in savannas. *International Journal of Applied Earth Observation and Geoinformation* 12, 217-224.
- Bonneau LR, Shields KS & Civco DL 1999a. A technique to identify changes in hemlock forest health over space and time using satellite image data. *Biological Invasions* 1, 269-279.
- Boulesteix A-L, Janitza S, Kruppa J & König IR 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining and Knowledge Discovery* 2, 493-507.

- Boyd DS & Danson FM 2005. Satellite remote sensing of forest resources: three decades of research development. *Progress in Physical Geography* 29, 1-26.
- Breiman L 1996. Bagging predictors. *Machine Learning* 24, 123-140.
- Breiman L 2001. Random forests. *Machine Learning* 45, 5-32.
- Breiman & Cutler A 2003. Manual on setting up, using, and understanding random forests V4.0. [https://www.stat.berkeley.edu/~breiman/Using\\_random\\_forests\\_v4.0.pdf](https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf) [last accessed 22 October 2019].
- Breiman L and Cutler A 2004. Random Forests. Department of Statistics, University of California, Berkeley. [http://www.math.usu.edu/~adele/forests/cc\\_software.htm](http://www.math.usu.edu/~adele/forests/cc_software.htm) [last accessed 22 October 2019].
- Breiman L, Friedman J, Olshen R & Stone C 1984. Classification and Regression Trees. Chapman & Hall, New York, NY, USA. pp 368.
- Bressler R, Kreisberg RB, Bernard B, Niederhuber JE, Vockley JG, Schmulevich I & Knijnenburg TA 2015. CloudForest: a scalable and efficient random forest implementation for biological data. *PLoS ONE* 10(12): e0144820. doi:10.1371/journal.pone.0144820.
- Britz H, Coutinho TA, Wingfield BD, Marasas WFO & Wingfield MJ 2005. Diversity and differentiation in two populations of *Gibberella circinata* in South Africa. *Plant Pathology* 54, 46-52.
- Buddenbaum H, Stern O, Stellmes M, Stoffels J, Poeschel P, Hill J & Werner W 2012. Field imaging spectroscopy of beech seedlings under dryness stress. *Remote Sensing* 4, 3721-3740.
- Bulman LS, Dunningham AG, Sims NC, Culvenor DS & Brownlie RK 2006. Evaluation of remote sensing technologies for forest health. Ensis Report No. 38641. pp 39.
- Bunting P & Lucas L 2006. The delineation of tree crowns in Australian mixed species forests using hyperspectral Compact Spectrographic Imager (CASI) data. *Remote Sensing of Environment* 101, 230-248.
- Bunting P, Lucas RM, Jones K & Bean AR 2010. Characterisation and mapping of forest communities by clustering individual tree crowns. *Remote Sensing of Environment* 114, 2536-2547.
- Burai P, Deák B, Valkó O & Tomor T 2015. Classification of herbaceous vegetation using airborne hyperspectral imagery. *Remote Sensing* 7, 2046-2066. doi:10.3390/rs70202046.
- Burges CJC 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121-167.
- Calderón R, Navas-Cortés JA, Lucena C & Zarco-Tejada PJ 2013. High-resolution airborne hyperspectral and thermal imagery for early detection of *Verticillium* wilt of olive using

- fluorescence, temperature and narrowband spectral indices. *Remote Sensing of Environment* 139, 231-245.
- Calderón R, Navas-Cortés JA & Zarco-Tejada PJ 2015. Early detection and quantification of *Verticillium* wilt in olive using hyperspectral and thermal imagery over large areas. *Remote Sensing* 7, 5584-5610.
- Calle ML, Urrea V, Boulesteix A-L & Malats N 2011. AUC-RF: a new strategy for genomic profiling with random forest. *Human Heredity* 72: 121-132.
- Cannas LM, Dessì N & Pes B 2013. Assessing similarity of feature selection techniques in high dimensional domains. *Pattern Recognition Letters* 34, 1446-1453.
- Carter GA 1994. Ratios of leaf reflectances in narrow wavebands as indicators of plant stress. *International Journal of Remote Sensing* 15, 697-703.
- Carter GA & Knapp AK 2001. Leaf optical properties in higher plants: linking spectral characteristics to stress and chlorophyll concentration. *American Journal of Botany* 88, 677-684.
- Carter GA & Miller RL 1994. Early detection of plant stress by digital imaging within narrow stress-sensitive wavebands. *Remote Sensing of Environment* 50, 295-302.
- Chaerle L, Leinonen I, Jones HG & Van Der Straeten D 2007. Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging. *Journal of Experimental Botany* 58, 773-784.
- Chalker-Scott L 1999. Environmental significance of anthocyanins in plant stress responses. *Photochemistry and Photobiology* 70, 1-9.
- Chan JC-W, Beckers P, Spanhove T & Borre JV 2012. An evaluation of ensemble classifiers for mapping Natura 2000 heathland in Belgium using spaceborne angular hyperspectral (CHRIS/Proba) imagery. *International Journal of Applied Earth Observation and Geoinformation* 18, 13-22.
- Chandrashekar G & Sahin F 2014. A survey on feature selection methods. *Computers and Electrical Engineering* 40, 16-28.
- Chapelle C & Keerthi SS 2008. Multi-class feature selection with support vector machines. Proceedings of the American Statistical Association, 3–7 August, Denver, USA.
- Chapman L 2008. An introduction to ‘upside-down’ remote sensing. *Progress in Physical Geography* 32, 529-542.
- Chappelle EW, Kim MS & McMurtrey III JE 1992. Ratio analysis of reflectance spectra (RARS): an algorithm for the remote estimation of the concentrations of chlorophyll A, chlorophyll B, and carotenoids in soybean leaves. *Remote Sensing of Environment* 39, 239-247.



- Chávez P, Yarlequé C, Loayza H, Mares V, Hanco P, Priou S, Del Pilar Márquez M, Posadas A, Zorogastúa P, Flexas J & Quiroz R 2012. Detection of bacterial wilt infection caused by *Ralstonia solanacearum* in potato (*Solanum Tuberosum* L.) through multifractal analysis applied to remotely sensed data. *Precision Agriculture* 13, 236-255.
- Chawla NV, Bowyer KW, Hall LO & Kegelmeyer WP 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Applied Intelligence Research* 16, 321-357.
- Chemura A, Mutanga O & Dube T 2016. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. *Precision Agriculture* 18, 859-881.
- Chen C, Liaw A & Breiman L 2004. Using random forest to learn imbalanced data. Technical report 666, Statistics Department, University of California Berkeley. pp 12.
- Chen Q, Meng Z, Liu X, Jin Q & Su R 2018. Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes* 9. doi:10.3390/genes9060301.
- Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, Jia W & Zhao A 2013. Random forest in Clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-Based Complementary and Alternative Medicine*. doi:10.1155/2013/298183.
- Chen Z-H, Li L-P, He Z, Zhou J-R, Li Y & Wong L 2019. An improved deep forest model for predicting self-interacting proteins from protein sequence using wavelet transformation. *Frontiers in Genetics* 10. doi:10.3389/fgene.2019.00090.
- Clevers JGPW, De Jong SM, Epema GF, Van der Meer FD, Bakker WH, Skidmore AK & Scholte KH 2002. Derivation of the red edge index using the MERIS standard band setting. *International Journal of Remote Sensing* 23, 3169-3184.
- Cochrane MA 2000. Using vegetation reflectance variability for species level classification of hyperspectral data. *International Journal of Remote Sensing* 21, 2075-2087.
- Cole B, McMorrow J & Evans M 2014. *ISPRS Journal of Photogrammetry and Remote Sensing* 90, 49-58.
- Congalton RG & Green K 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, 2nd ed. Chapman & Hall/CRC: Boca Raton, Florida, USA, pp. 105-120.
- Coops NC, Goodwin N, Stone C & Sims N 2006. Application of narrowband digital camera imagery to plantation canopy condition assessment. *Canadian Journal of Remote Sensing* 32, 1-14.
- Coops NC, Johnson M, Wulder MA & White JC 2006a. Assessment of QuickBird high spatial resolution imagery to detect red attack damage due to mountain pine beetle infestation. *Remote Sensing of Environment* 103, 67-80.

- Coops N, Stanford M, Old K, Dudzinski M, Culvenor D & Stone C 2003. Assessment of Dothistroma needle blight of *Pinus radiata* using airborne hyperspectral imagery. *Phytopathology* 93, 1524-1532.
- Coops NC, Waring RH, Wulder MA & White JC 2009. Prediction and assessment of bark beetle-induced mortality of lodgepole pine using estimates of stand vigour derived from remotely sensed data. *Remote Sensing of Environment* 113, 1058-1066.
- Coops NC, Wulder MA & White JC 2006b. Integrating remotely sensed and ancillary data sources to characterize a mountain pine beetle infestation. *Remote Sensing of Environment* 105, 83-97.
- Coutinho TA, Steenkamp ET, Mangwaketsi K, Wilmot M & Wingfield MJ 2007. First outbreak of pitch canker in a South African pine plantation. *Australasian Plant Pathology* 36, 256-261.
- Cram MM & Fraedrich SW 2009. Seed diseases and seedborne pathogens of North America. *Tree Planters' Notes* 53, 35-44.
- Crist EP & Cicone RC 1984. Application of the tasseled cap concept to simulated Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing* 50, 343-352.
- Crous JW 2005. Post establishment survival of *Pinus patula* in Mpumalanga, one year after planting. *South African Forestry Journal* 205, 3-8.
- Cunningham SC, Nally RM, Read J, Baker PJ, White M, Thomson JR & Griffioen P 2009. A robust technique for mapping vegetation condition across a major river system. *Ecosystems* 12, 207-219.
- Curtiss B 2013. Reference Measurements: The What, Why, When, and How\|. Analytical Spectral Devices Inc. (ASD): Boulder, CO, USA.
- DAFF (Department of Agriculture, Forestry and Fisheries – Forestry Branch: Republic of South Africa) 2010. Report on commercial timber resources and primary roundwood processing in South Africa: 2008/2009. Directorate: Forestry Technical and Information Services, Pretoria. pp 137.
- Dalleau K, Couceiro M & Smail-Tabbone M 2018. Unsupervised extremely randomized trees. In Phung D, Tseng V, Webb G, Ho B, Ganji M & Rashidi L (Eds), *Advances in Knowledge Discovery and Data Mining (PAKDD): Lecture Notes in Computer Science*, 10939. Springer: Cham.
- Dalponte M, Bruzzone L, Vescovo L & Gianelle D 2009. The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. *Remote Sensing of Environment* 113, 2345–2355.
- Dalponte M, Ørka HO, Gobakken T, Gianelle D & Næsset E 2013. Tree species classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* 51, 2632-2645.

- Das PK, Choudhary KK, Laxman B, Rao SVCK & Seshasai MVR 2014. A modified linear extrapolation approach towards red edge position detection and stress monitoring of wheat crop using hyperspectral data. *International Journal of Remote Sensing* 35, 1432-1449.
- Das S 2001. Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the Eighteenth International Conference on Machine Learning, 28 June-1 July, Williamstown, USA.
- Dash JP, Watt MS, Pearse GD, Heaphy M & Dungey HS 2017. Assessing very high resolution UAV imagery for monitoring forest health during a simulated disease outbreak. *ISPRS Journal of Photogrammetry and Remote Sensing* 131, 1-14.
- Datt 1998. Remote sensing of chlorophyll a, chlorophyll b, chlorophyll a+b, and total carotenoid content in eucalyptus leaves. *Remote Sensing of Environment* 66, 111-121.
- Datt 1999a. A new reflectance index for remote sensing of chlorophyll content in higher plants: tests using eucalyptus leaves. *Journal of Plant Physiology* 154, 30-36.
- Datt 1999b. Visible / near infrared reflectance and chlorophyll content in eucalyptus leaves. *International Journal of Remote Sensing* 20, 2741-2759.
- Daughtry CST, Walthall CL, Kim MS, De Colstoun EB & McMurtrey JE 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment* 74, 229-239.
- De Castro AI, Ehsani R, Ploetz R, Crane JH & Abdulridha J 2015. Optimum spectral and geometric parameters for early detection of laurel wilt disease in avocado. *Remote Sensing of Environment* 171, 33-44.
- De Jong SM, Addink EA & Doelman JC 2014. Detecting leaf-water content in Mediterranean trees using high-resolution spectrometry. *International Journal of Applied Earth Observation and Geoinformation* 27, 128-136.
- Degenhardt F, Seifert S & Szymczak S 2017. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics* 20, 492-503.
- Deghi GS, Huffman T & Culver JW 1994. California's native Monterey pine populations: Potential for sustainability. *Fremontia* 23, 14-23.
- Del Fiore A, Reverberi M, Ricelli A, Pinzari F, Serranti S, Fabbri AA, Bonifazi G & Fanelli C 2010. Early detection of toxigenic fungi on maize by hyperspectral imaging analysis. *International Journal of Food Microbiology* 144, 64-71.
- Del Frate FD, Pacifici F, Schiavon G & Solimini C 2007. Use of neural networks for automatic classification from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing* 45, 800-809.

- Del Río S, López V, Benítez JM & Herrera F 2014. On the use of MapReduce for imbalanced big data using random forest. *Information Sciences* 285, 112-137.
- Delalieux S, Van Aardt J, Keulemnas W, Schrevens E & Coppin P 2007. Detection of biotic stress (*Venturia inaequalis*) in apple trees using hyperspectral data: non-parametric statistical approaches and physiological implications. *European Journal of Agronomy* 27, 130-143.
- Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Štajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M, Zupan B 2013. Orange: data mining toolbox in Python. *Journal of Machine Learning Research* 14, 2349-2353.
- Deng H 2019. Regularized random forest. In Package ‘RRF’. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/RRF/index.html> [last accessed 11 September 2019].
- Deng H & Runger G 2012. Feature selection via regularized trees. Proceedings of IEEE World Congress on Computational Intelligence, 10-15 June, Brisbane, Australia. arXiv:1201.1587v3.
- Deng H & Runger G 2013. Gene selection with guided regularized random forest. *Pattern Recognition* 46, 3483-3489.
- Desperez-Loustau M-L, Marçais B, Nageleisen L-M, Piou D & Vannini A 2006. Interactive effects of drought and pathogens in forest trees. *Annals of Forest Science* 63, 597-612.
- Dessi N & Pes B 2015. Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications* 42, 4632-4642.
- Díaz-Uriarte R 2012. Variable selection from random forests using OOB error. In: Package ‘varSelRF’. The Comprehensive R Archive Network: Vienna, Austria. Available online: <http://cran.r-project.org/web/packages/varSelRF/index.html> [last accessed 1 November 2019].
- Díaz-Uriarte R & Alvarez de Andrés S 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 1-13.
- Dietterich TG 2000. An experimental comparison of three methods for constructing ensembles of decision trees - bagging, boosting, and randomisation. *Machine Learning* 40, 139-157.
- Dittman DJ, Khoshgoftaar TM & Napolitano A 2015. The effect of data sampling when using RF on imbalanced bioinformatics data. Proceedings of IEEE International Conference on Information Reuse and Integration, 13-15 August, San Francisco, USA. doi: 10.1109/IRI.2015.76.
- Do T-N, Lenca P & Lallich S 2015. Classifying many-class high dimensional fingerprint datasets using random forest of oblique decision trees. *Vietnam Journal of Computer Science* 2, 3-12.
- Do T-N, Lenca P, Lallich S & Pham N-K 2010. Classifying very-high dimensional data with random forests of oblique decision trees. In Guillet F, Ritschard G, Zighed D & Briand H (Eds), *Advances in Knowledge Discovery and Management*, 292, 39-55. Springer: Berlin/Heidelberg, Germany.

- Dreaden T & Smith J 2010. Pitch canker disease of pines. IFAS Report FOR236. University of Florida, School of Forest Resources and Conservation, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences. pp 4.
- Du P, Samat A, Waske B, Liu S & Li Z 2015. Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing* 105, 38-53.
- Du S & Chen S 2014. Salient object detection via random forest. *IEEE Signal Processing Letters* 21, 51-54.
- Du S & Chen S 2015. Detecting co-salient objects in large image sets. *IEEE Signal Processing Letters* 21, 145-148.
- Duch W 2006. Filter methods. In Guyon I, Nikravesh M, Gunn S & Zadeh LA (Eds), *Feature Extraction: Studies in Fuzziness and Soft Computing* 207, 89-117. Springer: Berlin, Heidelberg. doi:10.1007/978-3-540-35488-8\_4.
- Duroux R & Scornet E 2016. Impact of subsampling and pruning on random forests. arXiv:1603.04261v1.
- Dwinell LD, Barrows-Broadus J & Kuhlman EG 1985. Pitch canker: a disease complex of southern pines. *Plant Disease* 69, 270-276.
- Dye M, Mutanga O & Ismail R 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict *Pinus patula* age in KwaZulu-Natal, South Africa. *Geocarto International* 26, 275-289.
- Efron B & Tibshirani RJ 1993. An Introduction to the Bootstrap. Chapman & Hall: New York, NY, USA. pp 436.
- Eklundh L, Johansson T & Solberg S 2009. Mapping insect defoliation in Scots pine with MODIS time-series data. *Remote Sensing of Environment* 113, 1566-1573.
- Elvidge DE 1990. Visible and near infrared reflectance characteristics of dry plant materials. *Remote Sensing of Environment* 11, 1775-1795.
- EPPO (European and Mediterranean Plant Protection Organisation) 2005. Data sheets on quarantine pests: *Gibberella circinata*. *EPPO Bulletin* 35, 383-386.
- Fang M, Ju W, Zhan W, Cheng T, Qui F & Wang J 2017. *Remote Sensing of Environment* 196, 13-27.
- Fassnacht FE, Latifi H, Ghosh A, Joshi PK & Koch B 2014. Assessing the potential of hyperspectral imagery to map bark beetle-induced tree mortality. *Remote Sensing of Environment* 140, 533-548.

- Fassnacht FE, Latifi H & Koch B 2012. An angular vegetation index for imaging spectroscopy data—preliminary results on forest damage detection in the Bavarian National Park, Germany. *International Journal of Applied Earth Observation and Geoinformation* 19, 308-321.
- Fawcett T 2003. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861-874.
- Feingersh T & Ben-Dor E 2015. SHALOM – A Commercial Hyperspectral Space Mission. In Qian S-E (Ed) *Optical Payloads for Space Missions*, 247-263. doi:10.1002/9781118945179.ch11.
- Feng W, Shen W, He L, Duan J, Guo B, Li Y, Wang C & Guo T 2016. Improved remote sensing detection of wheat powdery mildew using dual-green vegetation indices. *Precision Agriculture* 17, 608-627.
- Ferreira MP, Zortea M, Zanotta DC, Shimabukuro YE & De Souza Filho CR 2016. Mapping tree species in tropical seasonal semi-deciduous forests with hyperspectral and multispectral data. *Remote Sensing of Environment* 179, 66-78.
- Folleco A, Khoshgoftar TM, Van Hulse J & Bullard L 2008. Software quality modelling: the impact of class noise on the random forest classifier. Proceedings of IEEE Congress on Evolutionary Computation, 1-6 June, Hong Kong, China. doi:10.1109/CEC.2008.4631321.
- Folleco AA, Khoshgoftar TM, Van Hulse J & Napolitano A 2009. Identifying learners robust to low quality data. *Informatica* 33, 245-259.
- Fox EW, Hill RA, Leibowitz SG, Olsen AR, Thornbrugh DJ & Weber MH 2017. Assessing the accuracy and stability of variable selection methods for random forest modelling in ecology. *Environmental Monitoring and Assessment* 189, 316. doi:10.1007/s10661-017-6025-0.
- Frank E, Hall MA & Witten IH 2016. The WEKA Workbench. Online Appendix for *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition. <https://www.cs.waikato.ac.nz/ml/weka/index.html> [last accessed 22 October 2019].
- Franke J, Mewes T & Menz G 2009. Requirements on spectral resolution of remote sensing data for crop stress detection. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 12–17 July, Cape Town, South Africa. doi:10.1109/IGARSS.2009.5416884.
- Frénay B & Verleysen M 2014. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 845-869.
- Freund Y & Schapire RE 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 119-139.
- Friedl MA & Brodley CE 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment* 51, 399-409.



- Frohlich H & Zell A 2005. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN), 31 July-4 August, Tubingen, Germany. doi:10.1109/IJCNN.2005.1556085.
- Gamon JA, Peñuelas J & Field CB 1992. A narrowwaveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sensing of Environment* 41, 35-44.
- Gandia S, Fernandez JC & Moreno J 2004. Retrieval of vegetation biophysical variables from CHRIS/PROBA data in the SPARC campaign. Proceedings of 2nd CHRIS/Proba workshop, 28-30 April, Frascati, Italy.
- Garrity SR, Eitel JUH & Vierling LA 2011. Disentangling the relationships between plant pigments and the photochemical reflectance index reveals a new approach for remote estimation of carotenoid content. *Remote Sensing of Environment* 115, 628-635.
- Gates DM, Keegan HJ, Schleter JC & Weidner VR 1965. Spectral properties of plants. *Applied Optics* 4, 11-20.
- Gebeyehu S & Wingfield MJ 2003. Pine weevil *Pissodes nemorensis*: threat to South African pine plantations and options for control. *South African Journal of Science* 99, 531-536.
- Genuer R, Poggi J-M & Tuleau-Malot C 2010. "Variable selection using random forests". *Pattern Recognition Letters* 31, 2225-2236.
- Genuer R, Poggi J-M & Tuleau C 2008. Random forests: some methodological insights. arXiv:0811.3619v1.
- Geurts P, Ernst D & Wehenkel L 2006. Extremely randomized trees. *Machine Learning* 63, 3-42.
- Ghamisi P, Plaza J, Chen Y, Li J & Plaza A 2017. Advanced spectral classifiers for hyperspectral images: A review. *IEEE Geoscience and Remote Sensing Magazine* 5, 8-32.
- Ghiyammat A & Shafri HZM 2010. A review on hyperspectral remote sensing for homogeneous and heterogeneous forest biodiversity assessment. *International Journal of Remote Sensing* 31, 1837-1856.
- Ghosh A, Fassnacht FE, Joshi PK & Koch B 2014. A framework for mapping tree species combining hyperspectral and LiDAR data: role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation* 26, 49-63.
- Ghosh A, Manwani N & Sastry PS 2016. On the robustness of decision tree learning under label noise. *JMLR: Workshop and conference proceedings* 1-17. arXiv:1605.06296v2.
- Gitelson AA, Gritz Y & Merzlyak MN 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology* 160, 271-282.



- Gitelson AA, Kaufman YJ & Merzlyak MN 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment* 58, 289-298.
- Gitelson AA, Keydan GP & Merzlyak MN 2006. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. *Geophysical Research Letters* 33, L11402. doi:10.1029/2006GL026457.
- Gitelson AA & Merzlyak MN 1994. Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. leaves. Spectral features and relation to chlorophyll estimation. *Journal of Plant Physiology* 143, 286-292.
- Gitelson AA & Merzlyak MN 1996. Signature analysis of leaf reflectance spectra - algorithm development for remote sensing of chlorophyll. *Journal of Plant Physiology* 148, 494-500.
- Gitelson AA, Merzlyak MN & Chivkunova OB 2001. Optical properties and non-destructive estimation of anthocyanin content in plant leaves. *Photochemistry and Photobiology* 74, 38-45.
- Gitelson AA, Yacobi YZ, Schalles JF, Rundquist DC, Han L, Stark R & Etzion D 2000. Remote estimation of phytoplankton density in productive waters. *Advances in Limnology* 55, 121-136.
- Goetz AFH 2009. Three decades of hyperspectral remote sensing of the Earth: A personal view. *Remote Sensing of Environment* 113, S5-S16.
- Goldstein BA, Hubbard AE, Cutler A & Barcellos LF 2010. An application of random forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics* 11. doi: 10.1186/1471-2156-11-49.
- Goldstein BA, Polley EC & Briggs FBS 2011. Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology* 10, 1-34.
- Goodenough DG, Pearlman J, Chen H, Dyk A, Han T, Li J, Miller J & Niemann KO 2004. Forest information from hyperspectral sensing. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 20-24 September, Anchorage, Alaska, USA. doi:10.1109/IGARSS.2004.1369826.
- Goodfellow I, Bengio Y & Courville A 2016. Deep Learning. MIT Press. <http://www.deeplearningbook.org/> [last accessed 12 December 2019]
- Goodwin N, Coops NC & Stone C 2005. Assessing plantation canopy condition from airborne imagery using spectral mixture analysis and fractional abundances. *International Journal of Applied Earth Observation and Geoinformation* 7, 11-28.
- Gordon TR 2006. Pitch canker disease of pines. *Phytopathology* 96, 657-659.
- Gordon TR, Kirkpatrick SC, Aegerter BJ, Wood DL & Storer AJ 2006. Susceptibility of Douglas fir (*Pseudotsuga menziesii*) to pitch canker, caused by *Gibberella circinata* (anamorph = *Fusarium circinatum*). *Plant Pathology* 55, 231-237.

- Gordon TR, Storer AJ & Wood DL 2001. The pitch canker epidemic in California. *Plant Disease* 85, 1128-1139.
- Gordon TR, Swett CL & Wingfield MJ 2015. Management of *Fusarium* diseases affecting conifers. *Crop Protection* 73, 28-39.
- Govender M, Dye PJ, Weiersbye IM, Witkowski ETF & Ahmed F 2009. Review of commonly used remote sensing and ground-based technologies to measure plant water stress. *Water SA* 35, 741-752.
- Granitto PM, Furlanello C, Biasioli F & Gasperi F 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Systems* 83, 83-90.
- Grant L 1987. Diffuse and specular characteristics of leaf reflectance. *Remote Sensing of Environment* 22, 309-322.
- Green RO, Eastwood ML, Sarture CM, Chrien TG, Aronsson M, Chippendale BJ, Faust JA, Pavri BE, Chovit CJ, Solis M, Olah MR & Williams O 1998. Imaging spectroscopy and the airborne visible / infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment* 65, 227-248.
- Gregorutti B, Michel B & Saint-Pierre P 2017. Correlation and variable importance in random forests. *Statistics and Computing* 27, 659-678.
- Grisham MP, Johnson RM & Zimba PV 2010. Detecting sugarcane yellow leaf virus infection in asymptomatic leaves with hyperspectral remote sensing and associated leaf pigment changes. *Journal of Virological Methods* 167, 140-145.
- Gupta S, Weinacker H & Koch B 2010. Comparative analysis of clustering-based approaches for 3-D single tree detection using airborne fullwave LiDAR data. *Remote Sensing* 2, 968-989.
- Günther F & Fritsch S 2010. Neuralnet: training of neural networks. *The R Journal* 2, 30-38.
- Guyon I & Elisseeff A 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157-1182.
- Guyon I, Weston J, Barnhill S & Vapnik V 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46, 389-422.
- Haboudane D, Miller JR, Tremblay N, Zarco-Tejada PJ & Dextraze L 2002. Integrated narrowband vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment* 81, 416-426.
- Haboudane D, Miller JR, Pattey E, Zarco-Tejada PJ & Strachan IB 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment* 90, 337-352.

- Haboudane D, Tremblay N, Miller JR & Vigneault P 2008. Remote estimation of crop chlorophyll content using spectral indices derived from hyperspectral data. *IEEE Transaction on Geoscience and Remote Sensing* 46, 423-437.
- Hakkenberg CR, Peet RK, Urban DL & Song C 2018. Modelling plant composition as community continua in a forest landscape with LiDAR and hyperspectral remote sensing. *Ecological Applications* 28, 177-190.
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P & Witten IH 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11, 10-18.
- Ham J, Chen Y, Crawford MM & Ghosh J 2011. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* 43, 492-501.
- Hammerbacher A, Wright LP, Wingfield BD, Wingfield MJ & Coutinho TA 2009. Factors affecting pine pitch canker modelled on Michaelis-Menten kinetics. *Botany* 87, 36-42.
- Hamza M & Larocque D 2005. An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation* 75, 629-643.
- Hapfelmeier A & Ulm K 2013. A new variable selection approach using random forests. *Computational Statistics & Data Analysis* 60, 50-69.
- Harris Geospatial Solutions Inc. 2019. ENVI. <https://www.harrisgeospatial.com/Software-Technology/ENVI> [last accessed 22 October 2019].
- Hastie T, Tibshirani R & Friedman J 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. New York: Springer, 2009, pp. 763.
- Hatchell DC, Ed., "ASD technical guide 3rd ed. Boulder, CO: Analytical Spectral Devices, Inc. (ASD), 1999, pp. 140.
- Hatala JA, Crabtree RL, Halligan KQ & Moorcroft PR 2010. Landscape-scale patterns of forest pest and pathogen damage in the Greater Yellowstone Ecosystem. *Remote Sensing of Environment* 114, 375-384.
- Hay GJ & Castilla G 2006. Object-based image analysis: strengths, weaknesses, opportunities and threats (SWOT). Proceedings of 1st International Conference on Object-based Image Analysis (OBIA), 4-5 July, Salzburg, Austria.
- Hepting GH & Roth ER 1946. Pitch canker, a new disease of some southern pines. *Journal of Forestry* 44, 742-744.
- Hernández-Clemente R, Navarro-Cerrillo RM, Suárez L, Morales F & Zarco-Tejada PJ 2011. Assessing structural effects on PRI for stress detection in conifer forests. *Remote Sensing of Environment* 115, 2360-2375.

- Hernández-Clemente R, North PRJ, Hornero A & Zarco-Tejada PJ 2017. Assessing the effects of forest health on sun-induced chlorophyll fluorescence using the FluorFLIGHT 3-D radiative transfer model to account for forest structure. *Remote Sensing of Environment* 193, 165-179.
- Hexagon AB 2019. ERDAS IMAGINE: World-class remote sensing software. <https://www.hexagongeospatial.com/products/power-portfolio/erdas-imagine/erdas-imagine-remote-sensing-software-package> [last accessed 22 October 2019].
- Hicke JA & Logan J 2009. Mapping whitebark pine mortality caused by a mountain pine beetle outbreak with high spatial resolution satellite imagery. *International Journal of Remote Sensing* 30i 4427-4441.
- Hodge GR & Dvorak WS 2000. Differential responses of Central American and Mexican pine species and *Pinus radiata* to infection by the pitch canker fungus. *New Forests* 19, 241-258.
- Hoque E, Hutzler PJS & Hiendl H 1992. Reflectance, colour, and histological features as parameters for the early assessment of forest damages. *Canadian Journal of Remote Sensing* 18, 104-110.
- Horne JH 2003. A tasseled cap transformation for Ikonos images. Proceedings of American Society for Photogrammetry & Remote Sensing (ASPRS), 5-9 May, Anchorage, Alaska.
- Hothorn T, Hornik K, Strobl C & Zeileis A 2019. A laboratory for recursive partitioning. In Package 'party'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/party/index.html> [last accessed 11 September 2019].
- Hothorn T, Hornik K & Zeileis A 2006. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics* 15, 651-674.
- Houborg R & Boegh E 2008. Mapping leaf chlorophyll and leaf area index using inverse and forward canopy reflectance modelling and SPOT reflectance data. *Remote Sensing of Environment* 112, 186-202.
- Hsu H-H, Hsieh C-W & Lu M-D 2011. Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications* 38, 8144-8150.
- Hu Z, Bao Y, Xiong T & Chiong R 2015. Hybrid filter-wrapper feature selection for short-term load forecasting. *Engineering Applications of Artificial Intelligence* 40, 17-27.
- Huang BFF & Boutros PC 2016. The parameter sensitivity of random forests. *BMC Bioinformatics* 17:331. doi 10.1186/s12859-016-1228-x.
- Huang J, Gao G, Guo F 2012. Forest growth simulation based on artificial neural network. In Qian Z, Cao L, Su W, Wang T & Yang H (Eds), *Lecture Notes in Electrical Engineering: Recent Advances in Computer Science and Information Engineering* 1, 657-663.

- Huang M-L, Hung Y-H, Lee WM, Li RK & Jiang B-R 2014. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. *The Scientific World Journal*. doi: 10.1155/2014/795624.
- Huang J, Liao H, Zhu Y, Sun J, Sun Q & Liu X 2012. Hyperspectral detection of rice damaged by rice leaf folder (*Cnaphalocrocis medinalis*). *Computers and Electronics in Agriculture* 82, 100-107.
- Huang W, Guan Q, Luo J, Zhang J, Zhao J, Liang D, Huang L & Zhang D 2014. New optimized spectral indices for identifying and monitoring winter wheat diseases. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7, 2516-2524.
- Huang Y 2009. Advances in artificial neural networks – methodological development and application. *Algorithms* 2, 973-1007.
- Huete AR 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* 25, 295-309.
- Hughes GF 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* 14, 55-63.
- Hunt ER & Rock BN 1989. Detection of changes in leaf water content using Near- and Middle-Infrared reflectances. *Remote Sensing of Environment* 30, 43-54.
- Hurley BP, Govender P, Coutinho TA, Wingfield BD & Wingfield MJ 2007. Fungus gnats and other Diptera in South African forestry nurseries and their possible association with the pitch canker fungus. *South African Journal of Science* 103, 43-46.
- Ingram JC, Dawson TP & Whittaker RJ 2005. Mapping tropical forest structure in south eastern Madagascar using remote sensing and artificial neural networks. *Remote Sensing of Environment* 94, 491-507.
- Ishwaran H, Kogalur UB, Blackstone EH & Laue MS 2008. Random survival forests. *The Annals of Applied Statistics* 2, 841-860.
- Ismail R & Mutanga O 2010. A comparison of regression tree ensembles: Predicting *Sirex noctilio* induced water stress in *Pinus patula* forests of KwaZulu-Natal, South Africa. *International Journal of Applied Earth Observation and Geoinformation* 12S, S45-S51.
- Ismail R & Mutanga O 2011. Discriminating the early stages of *Sirex noctilio* infestation using classification tree ensembles and shortwave infrared bands. *International Journal of Remote Sensing* 32, 4249-4266.
- Ismail R, Mutanga O & Ahmed F 2008. Discriminating *Sirex noctilio* attack in pine forest plantations in South Africa using high spectral resolution data. In Kalacska M & Sánchez-Azofeifa A (Eds),

*Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests*, 161-174. Chapman & Hall/CRC: Boca Raton, Florida, USA.

Ismail R, Mutanga O & Bob U 2007. Forest health and vitality: the detection and monitoring of *Pinus patula* trees infected by *Sirex noctilio* using digital multispectral imagery. *Southern Hemisphere Forestry Journal* 69, 39-47.

Ismail R, Mutanga O, Kumar L & Bob U 2008. Determining the optimal spatial resolution of remotely sensed data for the detection of *Sirex noctilio* infestations in pine plantations in KwaZulu-Natal, South Africa. *South African Geographical Journal* 90, 22-30.

ITTVIS (ITT Visual Information Solutions) 2010. ENVI 4.8, Boulder, Colorado.

Jackson RD & Huete AR 1991. Interpreting vegetation indices. *Preventive Veterinary Medicine* 11, 185-200.

Jacquemoud S & Ustin SL 2001. Leaf optical properties: a state of the art. Proceedings of 8th International Symposium of Physical Measurements & Signatures in Remote Sensing, 8-12 January, Aussois, France.

James G, Witten D, Hastie T & Tibshirani R 2013. *An introduction to statistical learning – with applications in R*. Springer, New York. pp 426.

Janecek AGK, Gansterer WN, Demel MA & Ecker GF 2008. On the relationship between feature selection and classification accuracy. *Journal of Machine Learning Research: Workshop Conference Proceedings* 4, 90-105.

Janitza S, Strobl C & Boulesteix A-L 2013. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-119.

Jensen JR, Qiu F & Ji M 1999. Predictive modelling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensor data. *International Journal of Remote Sensing* 20, 2805-2822.

Jiang J, Ma J, Wang Z, Chen C & Liu X 2019. Hyperspectral image classification in the presence of noisy labels. *IEEE Transactions on Geoscience and Remote Sensing* 57, 851-865.

Jones CD, Jones JB & Lee WS 2010. Diagnosis of bacterial spot of tomato using spectral signatures. *Computers and Electronics in Agriculture* 74, 329-335.

Jović A, Brkić K & Bogunović N 2015. A review of feature selection methods with applications. Proceedings of International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 25-29 May, Opatija, Croatia. doi:10.1109/MIPRO.2015.7160458.

Ju Y, Pan J, Wang X & Zhang H 2014. Detection of *Bursaphelenchus xylophilus* infection in *Pinus massoniana* from hyperspectral data. *Nematology* 16, 1197-1207.



- Kailath T 1967. The Divergence and Bhattacharyya Distance measures in signal selection. *IEEE Transactions on Communication Theory* 15, 52-60.
- Karl JW & Maurer BA 2010. Multivariate correlations between imagery and field measurements across scales: comparing pixel aggregation and image segmentation. *Landscape Ecology* 25, 591-605.
- Kauth RJ & Thomas GS 1976. The tasseled cap - a graphical description of the spectral-temporal development of agricultural crops as seen by Landsat. Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, 29 June-1 July, Purdue University, West Lafayette, Indiana.
- Kavzoglu T 2009. Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software* 24, 850-858.
- Kavzoglu T, Colkesen I & Yomralioglu T 2015. Object-based classification with rotation forest ensemble learning algorithm using very-high-resolution WorldView-2 image. *Remote Sensing Letters* 6, 834-843.
- Kawakubo H & Yoshida H 2012. Rapid feature selection based on random forests for high dimensional data. *Information Processing Society of Japan: SIG Notes* 89, 1-7.
- Kayet N, Pathak K, Chakrabarty A, Singh CP, Chowdary VM, Kumar S & Sahoo S 2019. Forest health assessment for geo-environmental planning and management in hilltop mining areas using Hyperion and Landsat data. *Ecological Indicators* 106, 105471. doi:10.1016/j.ecolind.2019.105471.
- Ke Y & Quackenbush LJ 2007. Forest species classification and tree crown delineation using QuickBird imagery. Proceedings of American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Conference, 7-11 May, Tampa, Florida.
- Ke Y & Quackenbush LJ 2009. Individual tree crown detection and delineation from high spatial resolution imagery using active contour and hill-climbing methods. Proceedings of American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Conference, 9-13 March, Baltimore, Maryland.
- Ke Y & Quackenbush LJ 2011. A review of methods for automatic individual tree-crown detection and delineation from passive remote sensing. *International Journal of Remote Sensing* 32, 4725-4747.
- Kearns M 1988. Thoughts on hypothesis boosting. Machine learning class project, unpublished.
- Kehl TN, Todt V, Veronez MR & Cazella SC 2012. Amazonian forest deforestation detection tool in real time using artificial neural networks and satellite images. *Sustainability* 4, 2566-2573.
- Kelly NM 2002. Monitoring sudden oak death in California using high-resolution imagery. General Technical Report PSW-GTR-184. USDA Forest Service, California.



- Kirkman K 2009. MTO conservation management plan. Steytlerville: MTO Forestry.
- Klobučar D, Pernar R, Lončarić S, Subašić M, Seletković A & Ančić M 2010. Detecting forest damage in CIR aerial photographs using a neural network. *Croatian Journal of Forest Engineering* 31, 157-163.
- Knights D, Costello EK & Knight R 2011. Supervised classification of human microbiota. *FEMS Microbiology Reviews* 35, 343-359.
- Knipling EB 1970. Physical and physiological basis for the reflectance of visible and near infrared radiation from vegetation. *Remote Sensing of Environment* 1, 155-159.
- Kohavi R & John GH 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97, 273-324.
- Kohavi R & Provost F 1998. Glossary of terms. *Machine Learning* 30, 271-274.
- Kokaly RF & Skidmore AK 2015. Plant phenolics and absorption features in vegetation reflectancespectra near 1.66  $\mu\text{m}$ . *International Journal of Applied Earth Observation and Geoinformation* 43, 55-83.
- Kokaly RF, Despain DG, Clark RN & Livo KE 2003. Mapping vegetation in Yellowstone National Park using spectral feature analysis of AVIRIS data. *Remote Sensing of Environment* 84, 437-456.
- Kong W, Zhang C, Liu F, Nie P & He Y 2013. Rice seed cultivar identification using near infrared hyperspectral imaging and multivariate data analysis. *Sensors* 13, 8916-8927. doi:10.3390/s130708916.
- Krause K 2005. Radiometric use of QuickBird imagery. Technical Note. Longmont, Colorado: DigitalGlobe, Inc.
- Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C & Hunt T 2019. Classification and regression training. In Package 'caret'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/obliqueRF/index.html> [last accessed 1 November 2019].
- Kumaresan PR 2018. Spectral based vegetation discrimination and forest health assessment using Hyperion (EO-1) in Yelagiri Hills, Tamil Nadu. *International Journal of Applied Engineering Research* 13, 13826-13832.
- Kursa MB 2012. Important attribute search using Boruta algorithm. In: Package 'Boruta'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <http://cran.r-project.org/web/packages/randomForest/index.html> [last accessed 1 November 2019].

- Kursa MB 2014a. rFerns: an implementation of the random ferns method for general-purpose machine learning. *Journal of Statistical Software* 61. doi:10.18637/jss.v061.i10.
- Kursa MB 2018. Random ferns classifier. In Package ‘rFerns’. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/rFerns/index.html> [last accessed 7 September 2019].
- Kursa MB 2014b. Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics* 15(1). doi: 10.1186/1471-2105-15-8.
- Kursa MB, Jankowski A & Rudnicki WR 2010. Boruta – A system for feature selection. *Fundamenta Informaticae* 101, 271-285.
- Kursa MB & Rudnicki WR 2010. Feature selection with the Boruta package. *Journal of Statistical Software* 36, 1-13.
- Kursa MB & Rudnicki WR 2011. A deceiving charm of feature selection: the microarray case study. In Czachorski T, Kozielski S & Stanczyk U (Eds), *Man-Machine Interactions* 2, 103, 145-152. Springer-Verlag: Heidelberg, Berlin.
- Kursa MB & Rudnicki WR 2011. The all relevant feature selection using random forest. arXiv preprint arXiv:1106.5112.
- Kursa MB & Rudnicki WR 2018. Feature selection with the Boruta algorithm. In: Package ‘Boruta’. The Comprehensive R Archive Network: Vienna, Austria. <http://cran.r-project.org/web/packages/Boruta/index.html> [last accessed 1 November 2019].
- Kvas M, Marasas WFO, Wingfield BD, Wingfield MJ & Steenkamp ET 2009. Diversity and evolution of *Fusarium* species in the *Gibberella fujikuroi* complex. *Fungal Diversity* 34, 1-31.
- Lary DJ, Alavi AH, Gandomi AH & Walker AL 2016. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7, 3-10.
- Laudien R, Bareth G & Doluschitz R 2003. Analysis of hyperspectral field data for detection of sugar beet diseases. Proceedings of the European Federation for Information Technology in Agriculture, Food and the Environment (EFITA), 5-9 July, Debrecen, Hungary.
- Laurin GV, Puletti N, Hawthorne W, Liesenberg V, Corona P, Papale D, Chen Q & Valentini R 2016. Discrimination of tropical forest types, dominant species, and mapping of functional guilds by hyperspectral and simulated multispectral Sentinel-2 data. *Remote Sensing of Environment* 176, 163-176.
- Lausch A, Salbach C, Schmidt A, Doktor D & Merbach I 2015. Deriving phenology of barley with imaging hyperspectral remote sensing. *Ecological Modelling* 295, 123-135.

- Le Maire G, François C & Dufrêne E 2004. Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements. *Remote Sensing of Environment* 89, 1-28.
- Le Maire G, François C, Soudani K, Berveiller D, Pontauiller J-Y, Bréda N, Genet H, Davi H & Dufrêne E 2008. Calibration and validation of hyperspectral indices for the estimation of broadleaved forest leaf chlorophyll content, leaf mass per area, leaf area index and leaf canopy biomass. *Remote Sensing of Environment* 112, 3846-3864.
- Leckie DG, Jay C, Gougeon FA, Sturrock RN & Paradine D 2004. Detection and assessment of trees with *Phellinus weirii* (laminated root rot) using high resolution multi-spectral imagery. *International Journal of Remote Sensing* 25, 793-818.
- Leckie DG, Gougeon FA, Tinis S, Nelson T, Burnett CN & Paradine D 2005. Automated tree recognition in old growth conifer stands with high resolution digital imagery. *Remote Sensing of Environment* 94, 311-326.
- Lee SH & Cho HK 2006. Detection of the pine trees damaged by pine wilt disease using high spatial remote sensing data. In Kerle N & Skidmore A (Eds), Proceedings of the ISPRS Commission VII Symposium 'Remote Sensing: From Pixels to Processes', 8-11 May, Enschede, The Netherlands.
- Levy JL, Khoshgoftaar TM, Bauder RA & Seliya N 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5, 42. doi: 10.1186/s40537-018-0151-6.
- Lepine LC, Ollinger SV, Ouimette AP & Martin ME 2016. Examining spectral reflectance features related to foliar nitrogen in forests: implications for broad-scale nitrogen mapping. *Remote Sensing of Environment* 173, 174-186.
- Leutner BF, Reineking B, Müller J, Bachmann M, Beierkuhnlein C, Dech S & Wegmann M 2012. Modelling forest  $\alpha$ -diversity and floristic composition – on the added value if LiDAR plus hyperspectral remote sensing. *Remote Sensing* 4, 2818-2845.
- Li J, Tran M & Siwabessy J 2016. Selecting optimal random forest predictive models: a case study on predicting the spatial distribution of seabed hardness. *PLoS ONE* 11. doi:10.1371/journal.pone.0149089.
- Li S, Wu H, Wan D & Zhu J 2011. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Systems* 24, 40-48.
- Li Z, Hayward RF, Zhang J & Liu Y 2008. Individual tree crown delineation techniques for vegetation management in power line corridor. Proceedings of the 10th International Conference on Digital Image Computing: Techniques and Applications (DICTA), 1-3 December, Canberra, Australia. doi:10.1109/DICTA.2008.21.

- Liang S, Li X & Wang J (Eds) 2012. *Advanced Remote Sensing*. Academic Press: Oxford, UK. pp. 800.
- Liaw A & Wiener M 2002. Classification and Regression by RandomForest. *R News* 2, 18-22.
- Lichtenthaler HK 1998. The stress concept in plants: an introduction. *Annals of the New York Academy of Sciences* 851, 187-198.
- Lichtenthaler HK, Lang M, Sowinska M, Heisel F & Miehé, JA 1996. Detection of vegetation stress via a new high resolution fluorescence imaging system. *Journal of Plant Physiology* 148, 599-612.
- Lichtenthaler HK, Wenzel O, Buschmann C & Gitelson A 1998. Plant stress detection by reflectance and fluorescence. *Annals of the New York Academy of Sciences* 851, 271-285.
- Lin W-J & Chen JJ 2012. Class-imbalanced classifiers for high dimensional data. *Briefings in Bioinformatics* 14, 13-26.
- Ling CX & Sheng VS 2010. Cost-Sensitive Learning. In Sammut C & Webb GI (Eds), *Encyclopaedia of Machine Learning*. Springer: Boston, MA. doi:10.1007/978-0-387-30164-8\_181.
- Liu K, Zhou Q, Wu W, Xia T & Tang H 2016. Estimating the crop leaf area index using hyperspectral remote sensing. *Journal of Integrative Agriculture* 15, 475-491.
- Liu FT, Ting KM & Zhou Z-H 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data* 6, 3:1-3:39. doi:10.1145/2133360.2133363.
- Leevy JL, Khoshgoftaar TM, Bauder RA & Seliya N 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5. doi.org/10.1186/s40537-018-0151-6.
- Loggenberg K, Strever A, Greyling B & Poona N 2018. Modelling water stress in a Shiraz vineyard using hyperspectral imaging and machine learning. *Remote Sensing* 10. doi:10.3390/rs10020202.
- López de Maturana E, Ye Y, Calle ML, Rothman N, Urrea V, Kogevinas M, Petrus S, Chanock SJ, Tardón A, García-Closas M, González-Neira A, Vellalta G, Carrato A, Navarro A, Lorente-Galdós B, Silverman DT, Real FX, Wu X & Malats N 2013. Application of multi-SNP approaches Bayesian LASSO and AUC-RF to detect main effects of inflammatory-gene variants associated with bladder cancer risk. *PLoS ONE* 8. doi:10.1371/journal.pone.0083745.
- Lu S, Xia Y, Cai TW & Feng DD 2015. Semi-supervised manifold learning with affinity regularization for Alzheimer's disease identification using positron emission tomography imaging. In 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 25-29 August 2015, Milano, Italy. 2251-2254.
- Maccioni A, Agati G & Mazzinghi P 2001. New vegetation indices for remote measurement of chlorophylls based on leaf directional reflectance spectra. *Journal of Photochemistry and Photobiology B: Biology* 61, 52-61.

- Maia R, Eliason C & Bitton P-P 2015. A Cohesive Framework for Parsing, Analyzing and Organizing Color from Spectral Data. In Package 'Pavo'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/pavo/index.html> [last accessed 1 November 2019].
- Mahlein A-K, Oerke E-C, Steiner U & Dehne H-W 2012. Recent advances in sensing plant diseases for precision crop protection. *European Journal of Plant Pathology* 133, 197-209.
- Mahlein A-K, Rumpf T, Welke P, Dehne H-W, Plümer L, Steiner U & Oerke E-C 2013. Development of spectral indices for detecting and identifying plant diseases. *Remote Sensing of Environment* 128, 21-30.
- Maldonado S & Weber R 2009. A wrapper method for feature selection using support vector machines. *Information Sciences* 179, 2208-2217.
- Malenovsky Z, Mishra KB, Zeemk F, Rascher U & Nedbal L 2009. Scientific and technical challenges in remote sensing of plant canopy reflectance and fluorescence. *Journal of Experimental Botany* 60, 2987-3004.
- Mas JF & Flores JJ 2008. The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing* 29, 617-663.
- Mas JF, Puig H, Palacio JL & Sosa-López A 2004. Modelling deforestation using GIS and artificial neural networks. *Environmental Modelling & Software* 19, 461-471.
- Martin ME, Newman SD, Aber JD & Congalton RG, 1998. Determining forest species composition using high spectral resolution remote sensing data. *Remote Sensing of Environment* 65, 249-254.
- Martinelli F, Scalenghe R, Davino S, Panno S, Scuderi G, Ruisi P, Villa P, Stroppiana D, Boschetti M, Goulart LR, Davis CE & Dandekar AM 2015. Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development* 35, 1-25.
- Masaitis G, Mozgeris G & Augustaitis A 2013. Spectral reflectance properties of healthy and stressed coniferous trees. *iForest* 6, 30-36.
- Mather PM 1999. *Computer processing of remotely-sensed images* (2<sup>nd</sup> Ed). Chichester: John Wiley & Sons Ltd.
- McCoy RM 2005. *Field methods in remote sensing*. The Guildford Press: New York.
- McMurtrey JE, Chappelle EW, Kim MS, Meisinger JJ & Corp LA 1994. Distinguishing nitrogen fertilization levels in field corn (*Zea mays* L.) with actively induced fluorescence and passive reflectance measurements. *Remote Sensing of Environment* 47, 36-44.
- Meinshausen N 2006. Quantile regression forests. *Journal of Machine Learning Research* 7, 983-999.

- Mellor A, Boukir S, Haywood A & Jones S 2015. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing* 105, 155-168.
- Menardi G & Torelli N 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery* 28, 92-122.
- Meng Q, Zhao H, Williams GJ, Lv J, Xu B & Huang JZ 2017. Weighted subspace random forest for classification. In Package 'wsrf'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/wsrif/index.html> [last accessed on 7 September 2019].
- Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W & Hamprecht FA 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 10. doi: 10.1186/1471-2105-10-213.
- Menze BH, Kelm BM, Splitthoff DN, Koethe U & Hamprecht FA 2011. On oblique random forests. In Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M (Eds), *Machine Learning and Knowledge Discovery in Databases*, 453-469. Springer: Berlin, Heidelberg. doi:10.1007/978-3-642-23783-6\_29.
- Menze B & Splitthoff N 2015. Oblique Random Forests from Recursive Linear Model Splits. In Package 'obliqueRF'; The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/obliqueRF/index.html> [last accessed 22 October 2019].
- Merton R 1998. Monitoring community hysteresis using spectral shift analysis and the red-edge vegetation stress index. Proceedings of the seventh annual Jet Propulsion Laboratory (JPL) Airborne Earth Science workshop, 12-16 January, NASA, Pasadena, California, USA.
- Merzlyak MN, Gitelson AA, Chivkunova OB & Rakitin VY 1999. Non-destructive optical detection of leaf senescence and fruit ripening. *Physiologia Plantarum* 106, 135-141.
- Mewes T, Franke J & Menz G 2011. Spectral requirements on airborne hyperspectral remote sensing data for wheat disease detection. *Precision Agriculture* 12, 795-812.
- Mianji FA & Zhang Y 2011. Robust hyperspectral classification using relevance vector machine. *IEEE Transactions in Geoscience and Remote Sensing* 49, 2100-2112.
- Michel S, Gamet P & Lefevre-Fonollosa M-J 2011. HYPXIM — A hyperspectral satellite defined for science, security and defence users. Proceedings of 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 6-9 June, Lisbon, Portugal. doi: 10.1109/WHISPERS.2011.6080864.



- Millard K & Richardson M 2015. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. *Remote Sensing* 7, 8489-8515.
- Mills H, Cutler MEJ & Fairbairn D 2006. Artificial neural networks for mapping regional-scale upland vegetation from high spatial resolution imagery. *International Journal of Remote Sensing* 27, 2177-2195.
- Mirik M, Michels Jr GJ, Kassymzhanova-Mirik S, Elliott NC, Catana V, Jones DB & Bowling R 2006. Using digital image analysis and spectral reflectance data to quantify damage by greenbug (Hemitera: Aphididae) in winter wheat. *Computers and Electronics in Agriculture* 51, 86-98.
- Mishina Y, Tsuchiya M & Fujiyoshi H 2014. Boosted random forest. Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP), 5-8 January, Lisbon, Portugal.
- Mitchell RG, Steenkamp ET, Coutinho TA & Wingfield MJ 2011. The pitch canker fungus, *Fusarium circinatum*: implications for South African forestry. *Southern Forests: A Journal of Forest Science* 73, 1-13.
- Mladeníć D, Brank J, Grobelnik M & Milic-Frayling N 2004. Feature selection using linear classifier weights: Interaction with classification models. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 25-29 July, Sheffield, UK. doi: 10.1145/1008992.1009034.
- Mohite J, Karale Y, Pappula S, Shabeer TPA, Sawant SD & Hingmire S 2017. Detection of pesticide (Cyantraniliprole) residue on grapes using hyperspectral sensing. Proceedings of Sensing for Agriculture and Food Quality and Safety IX, 9-13 April, Anaheim USA. doi:10.1117/12.2261797.
- Morris A 2010. A review of pitch canker (*Fusarium circinatum*) as it relates to plantation forestry in South Africa. Howick: Shaw Research Centre, Sappi Forests (Pty) Ltd., Research Document #/2010, 2010.
- Moshou D, Bravo C, West J, Wahlen S, McCartney A & Ramon H 2004. Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. *Computers and Electronics in Agriculture* 44, 173-188.
- Mureriwa N, Adam E, Sahu A & Tesfamichael S 2016. Examining the spectral separability of *Prosopis glandulosa* from co-existent species using field spectral measurement and guided regularized random forest. *Remote Sensing* 8. doi:10.3390/rs8020144.
- Mutanga O, Adam E, Adjorlolo C & Abdel-Rahman EM 2015. Evaluating the robustness of models developed from field spectral data in predicting African grass foliar nitrogen concentration



- using WorldView-2 image as an independent test dataset. *International Journal of Applied Earth Observation and Geoinformation* 34, 178-187.
- Mutanga O, Van Aardt J & Kumar L 2009. Imaging spectroscopy (hyperspectral remote sensing) in southern Africa: an overview. *South African Journal of Science* 105, 193-198.
- Naidu RA, Perry EM, Pierce FJ & Mekuria T 2009. The potential of spectral reflectance technique for the detection of *Grapevine leafroll-associated virus-3* in two red-berried wine grape cultivars. *Computers and Electronics in Agriculture* 66, 38-45.
- Näsi R, Honkavaara E, Blomqvist M, Lyytikäinen-Saarenmaa P, Hakala T, Viljanen N, Kantola T & Holopainen M 2018. Remote sensing of bark beetle damage in urban forests at individual tree level using a novel hyperspectral camera from UAV and aircraft. *Urban Forestry & Urban Greening* 30, 72-83.
- Näsi R, Honkavaara E, Lyytikäinen-Saarenmaa P, Blomqvist M, Litkey P, Hakala T, Viljanen N, Kantola T, Tanhuanpää T & Holopainen M 2015. Using UAV-Based Photogrammetry and Hyperspectral Imaging for Mapping Bark Beetle Damage at Tree-Level. *Remote Sensing* 7, 15467-15493.
- Nieke J & Rast M 2018. Towards the Copernicus hyperspectral imaging mission for the environment (CHIME). Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 22-27 July, Valencia, Spain. doi:10.1109/IGARSS.2018.8518384.
- Niemann KO, Quinn G, Stephen R, Visintini F & Parton D 2015. Hyperspectral remote sensing of mountain pine beetle with an emphasis on previsual assessment. *Canadian Journal of Remote Sensing* 41, 191-202.
- Nirenberg HI & O'Donnell K 1998. New *Fusarium* species and combinations within the *Gibberella fujikuroi* species complex. *Mycologia* 90, 434-458.
- Oumar Z & Mutanga O 2014. Predicting water stress induced by *Thaumastocoris peregrinus* infestations in plantation forests using field spectroscopy and neural networks. *Journal of Spatial Science* 59, 79-90.
- Oumar Z, Mutanga O & Ismail R 2013. Hyperspectral indices using field spectra resampled to the Hyperion. *International Journal of Applied Earth Observation and Geoinformation* 21, 113-121.
- Özuysal M, Fua P & Lepetit V 2007. Fast keypoint recognition in ten lines of code. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 17-22 June, Minneapolis, USA. doi:10.1109/CVPR.2007.383123.
- Pal M 2006. Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. *International Journal of Remote Sensing* 27, 2877-2894.

- Pal M & Foody GM 2010. Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing* 48, 2297-2307.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, David Cournapeau D, Brucher M, Perrot M & Duchesnay É 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.
- Peerbhay KY, Mutanga O & Ismail R 2015. Random forests unsupervised classification: the detection and mapping of *Solanum mauritianum* infestations in plantation forestry using hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, 3107-3122.
- Peerbhay KY, Mutanga O, Lottering R & Ismail R 2016. Mapping *Solanum mauritianum* plant invasions using WorldView-2 imagery and unsupervised random forests. *Remote Sensing of Environment* 182, 39-48.
- Pei Y, Kim T-K & Zha H 2013. Unsupervised random forest manifold alignment for lipreading. Proceedings of IEEE International Conference on Computer Vision, 1-8 December, Sydney, Australia. doi:10.1109/ICCV.2013.23.
- Pelletier C, Valero S, Inglada J, Champion N, Sicre CM & Dedieu G 2017. Effect of class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing* 9. doi:10.3390/rs9020173.
- Peñuelas J, Baret F & Filella I 1995. Semi-empirical indices to assess carotenoids / chlorophyll a ratio from leaf spectral reflectance. *Photosynthetica* 31, 221-230.
- Peñuelas J, Gamon JA, Fredeen AL, Merino J & Field CB 1994. Reflectance indices associated with physiological changes in nitrogen- and water-limited sunflower leaves. *Remote Sensing of Environment* 48, 135-146.
- Peñuelas J, Gamon JA, Griffin KL & Field CB 1993. Assessing community type, plant biomass, pigment composition, and photosynthetic efficiency of aquatic vegetation from spectral reflectance. *Remote Sensing of Environment* 46, 110-118.
- Poona NK & Ismail R 2012a. Discriminating the occurrence of pitch canker infection in *Pinus radiata* forests using high spatial resolution QuickBird data and artificial neural networks. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 22-27 July, Munich, Germany. doi: 10.1109/IGARSS.2012.6350698.
- Poona NK & Ismail R 2012b. Discriminating the early stages of *Fusarium circinatum* infection of *Pinus radiata* seedlings using high spectral resolution data. Proceedings of the 9<sup>th</sup> International

Conference of the African Association of Remote Sensing and the Environment (AARSE), 29 October-2 November, El Jadida, Morocco.

- Poona NK & Ismail R 2013. Discriminating the occurrence of pitch canker fungus in *Pinus radiata* trees using QuickBird imagery and artificial neural networks. *Southern Forests: a Journal of Forest Science*, 75:1, 29-40.
- Poona NK & Ismail R 2013. Reducing hyperspectral data dimensionality using random forest based wrappers. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 21-26 July, Melbourne, Australia. doi: 10.1109/IGARSS.2013.6723063.
- Poona NK & Ismail R 2014. Using Boruta-selected spectroscopic wavebands for the asymptomatic detection of *Fusarium circinatum* stress. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7, 3764-3772.
- Poona NK, Van Niekerk A, Nadel RL & Ismail R 2016a. Random forest (RF) wrappers for waveband selection and classification of hyperspectral data. *Applied Spectroscopy* 70, 322-333.
- Poona N, Van Niekerk A & Ismail R 2016b. Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data. *Sensors* 16, 1918. doi: 10.3390/s16111918.
- Poorter L, Oberbauer SF & Clark DB 1995. Leaf optical properties along a vertical gradient in a tropical rain forest canopy in Costa Rica. *American Journal of Botany* 82, 1257-1263.
- Pontius J, Hallett R & Martin M 2005. Assessing hemlock decline using visible and near-infrared spectroscopy: indices comparison and algorithm development. *Applied Spectroscopy* 59, 836-843.
- Porter B, Wingfield MJ & Coutinho TA 2009. Susceptibility of South African native conifers to the pitch canker pathogen, *Fusarium circinatum*. *South African Journal of Botany* 75, 380-382.
- Pouliot DA, King DJ, Bell FW & Pitt DG 2002. Automated tree crown detection and delineation in high-resolution digital camera imagery of coniferous forest regeneration. *Remote Sensing of Environment* 82, 322-334.
- Poulos HM, Berlyn GP & Mills SA 2012. Differential stress tolerance of four pines (Pinaceae) across the elevation gradient of the San Bernardino Mountains, Southern California, USA. *Journal of the Torrey Botanical Society* 139, 96-108.
- Prabhakar M, Prasad YG, Thirupathi M, Sreedevi G, Dharajothi B & Venkateswarlu B 2011. Use of ground based hyperspectral remote sensing for detection of stress in cotton caused by leafhopper. *Computers and Electronics in Agriculture* 79, 189-198.
- Pu R, Gong P, Biging GS & Larrieu MR 2003. Extraction of red edge optical parameters from Hyperion data for estimation of forest leaf area index. *IEEE Transactions on Geoscience and Remote Sensing* 41, 916-621.

- Puggini L, Doyle J & McLoone S 2015. Fault detection using random forest similarity distance. *IFAC-PapersOnLine* 48, 583-588.
- Qi J, Chehbouni A, Huete AR, Kerr YH & Sorooshian S 1994. A modified soil adjusted vegetation index. *Remote Sensing of Environment* 48, 119-126.
- R Development Core Team 2019. R: A language and environment for statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.r-project.org/> [last accessed September 17 2019].
- Raczko E & Zagajewski B 2017. Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sensing* 50, 144-154.
- Räsänen A, Kuitunen M, Tomppo E & Lensu A 2014. Coupling high-resolution satellite imagery with ALS-based canopy height model and digital elevation model in object-based boreal forest habitat type classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 94, 169-182.
- Read JJ, Tarpley L, McKinion JM & Reddy KR 2002. Narrow waveband reflectance ratios for remote estimation of nitrogen status in cotton. *Journal of Environmental Quality* 31, 1442-1452.
- Richards JA & Jia X 2006. *Remote sensing digital image analysis: An introduction*. 4<sup>th</sup> Ed. Springer: Germany. p 439.
- Riehle K, Coarfa C, Jackson A, Ma J, Tandon A, Paithankar S, Raghuraman S, Mistretta T-A, Saulnier D, Raza S, Diaz M-A, Shulman R, Aagaard K, Versalovic J & Milosavljevic A 2012. The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences. *BMC Bioinformatics* 13, 1-13.
- Riggins JJ, Defibaugh y Chávez JM, Tullis JA & Stephen FM 2011. Spectral identification of previsual northern red oak (*Quercus rubra* L.) foliar symptoms related to oak decline and red oak borer (Coleoptera: Cerambycidae) attack. *Southern Journal of Applied Forestry* 35, 18-25.
- Robnik-Šikonja M & Kononenko I 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* 53, 23-69.
- Rodríguez JJ & Alonso CJ 2011. Rotation-based ensembles. In Conejo R, Urretavizcaya M, Pérez-de-la-Cruz J-L (Eds), *Current Topics in Artificial Intelligence*, 3040, 498-506 Springer: Berlin/Heidelberg, Germany.
- Rodríguez JJ, Kuncheva LI & Alonso CJ 2006. Rotation forest: A new classifier ensemble method. *IEEE Transactions Pattern Analysis and Machine Intelligence* 28, 1619-1630.
- Rokach L 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33, 1-39.
- Rokach L & Maimon O 2015. *Data mining with decision trees: Theory and applications*. 2<sup>nd</sup> Ed. World Scientific Publishing Co. Pte. Ltd. Singapore.

- Rodríguez JJ, Kuncheva LI & Alonso CJ 2006. Rotation forest: A new classifier ensemble method. *IEEE Transactions Pattern Analysis and Machine Intelligence* 28, 1619-1630.
- Rondeaux G, Steven M & Baret F 1996. Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment* 55, 95-107.
- Rougean J-L & Breon FM 1995. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment* 51, 375-384.
- Rouse JW, Haas RH, Schell JA, Deering DW & Harlan JC 1974. Monitoring the vernal advancement and retrogradation (greenwave effect) of natural vegetation (Type III Final Report). NASA Goddard Space Flight Center, Greenbelt, MD.
- Roux J, Eisenberg B, Kanzler A, Nel A, Coetzee V, Kietzka E & Wingfield MJ 2007. Testing of selected South African *Pinus* hybrids and families for tolerance to the pitch canker pathogen, *Fusarium circinatum*. *New Forests* 33, 109-123.
- Royle DD & Lathrop RG 1997. Monitoring hemlock forest health in New Jersey using Landsat TM data and change detection techniques. *Forest Science* 43, 327-335.
- Rudnicki WR, Wrzesień M & Paja W 2015. All relevant feature selection methods and applications. In: Stańczyk U & Jain LC (Eds), *Feature selection for data and pattern recognition, Studies in computational intelligence* 584. pp. 11-28.
- Rumpf T, Mahlein A-K, Steiner U, Eorke E-C, Dehne H-W & Plümer L 2010. Early detection and classification of plant disease with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture* 74, 91-99.
- Saeys Y, Inza I & Larrañaga P 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507-2517.
- Sampson PH, Zarco-Tajada PJ, Mohammed GH, Miller JR & Noland TL 2003. Hyperspectral remote sensing of forest condition: estimating chlorophyll content in tolerant hardwoods. *Forest Science* 49, 381-391.
- Sandino J, Pegg G, Gonzalez F & Smith G 2018. Aerial mapping of forests affected by pathogens using UAVs, hyperspectral sensors, and artificial intelligence. *Sensors* 18, 944. doi:10.3390/s18040944.
- Sandri M & Zuccolotto P 2008. A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics* 17, 611-628.
- Sankaran S, Mishra A, Ehsani R & Davis C 2010. A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture* 72, 1-13.

- Sanz H, Valim C, Vegas E, Oller JM & Reverter F 2018. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics* 19. doi:10.1186/s12859-018-2451-4.
- Saulnier DM, Riehle K, Mistretta T-A, Diaz M-A, Mandal D, Raza S, EM Weidler EM, X. Qin X, Coafra C, Milosavljevic A, Petrosino JF, Highlander S, Gibbs R, Lynch SV, Shulman RJ & Versalovic J 2011. Gastrointestinal microbiome signatures of paediatric patients with irritable bowel syndrome. *Gastroenterology* 141, 1782-1791.
- Schwarz DF, König IR & Ziegler A 2010. On safari to Random Jungle – a fast implementation of random forests for high dimensional data. *Bioinformatics* 26, 1752-1758.
- Schweisinger JJ 2008. Mapping the effects of the pine disease pitch canker using satellite imagery and geospatial technology. Monterey Pine Forest Habitat Monitoring Project. Turf Image, Inc.: Monterey, CA.
- Serrano L, Peñuelas J & Ustin SL 2002. Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: Decomposing biochemical from structural signals. *Remote Sensing of Environment* 81, 355-364.
- Sheridan RP, Wang WM, Liaw A, Ma J & Gifford EM 2016. Extreme gradient boosting as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modelling* 56, 2353-2360.
- Shi T, Seligson D, Beldegrun AS, Palotie A & Horvath S 2005. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology* 18, 547-557.
- Shi T & Horvath S 2006. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics* 15, 118-138.
- Shi Y, Huang W, Luo J, Huang L & Zhou X 2017. Detection and discrimination of pests and diseases in winter wheat based on spectral indices and kernel discriminant analysis. *Computers and Electronics in Agriculture* 141, 171-180.
- Simm J & Magrans de Abril I 2015. Extremely randomized trees (ExtraTrees) method for classification and regression. In Package ‘extraTrees’. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/extraTrees/index.html> [last accessed 10 September 2019].
- Sims DA & Gamon JA 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment* 81, 337-354.



- Sims NC, Stone C, Coops NC & Ryan P 2007. Assessing the health of *Pinus radiata* plantations using remote sensing data and decision tree analysis. *New Zealand Journal of Forestry Science* 37, 57-80.
- Skakun RS, Wulder MA & Franklin SE 2003. Sensitivity of the thematic mapper enhanced wetness difference index to detect mountain pine beetle red-attack damage. *Remote Sensing of Environment* 86, 433-443.
- Skidmore AK, Turner BJ, Brinkhof W & Knowles E 1997. Performance of a neural network: mapping forests using GIS and remotely sensed data. *Photogrammetric Engineering & Remote Sensing* 63, 501-514.
- Skurichina M & Duin RPW 2005. Combining Feature Subsets in Feature Selection. In Oza NC, Polikar R, Kittler J & Roli F (Eds), *Multiple Classifier Systems*. Berlin, Germany: Springer, 2005. Pp 165-175.
- Smith MR & Martinez T 2014. Becoming more robust to label noise with classifier diversity. arXiv:1403.1893.
- Steele MR, Gitelson AA, Rundquist DC & Merzlyak MN 2009. Nondestructive estimation of anthocyanin content in grapevine leaves. *American Journal of Enology and Viticulture* 60, 87-92.
- Steinberg 2009. CART: Classification and regression trees. In Wu X & Kumar V (Eds), *The top ten algorithms in data mining*. Chapman & Hall/CRC: Boca Raton, Florida, USA. pp 201.
- Stevens A, Nocita M, Tóth G, Montanarella L & Van Wesemael B 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PLoS ONE* 8. doi:10.1371/journal.pone.0066409.
- Stewart B, Cieszewski CJ & Smith EL 2006. Preliminary results of spatial modelling of selected forest health variables in Georgia. Proceedings of 8<sup>th</sup> Annual Forest Inventory and Analysis Symposium, 16-19 October, Monterey, California, USA.
- Stiglic G & Kokol P 2007. Effectiveness of Rotation Forest in Meta-learning Based Gene Expression Classification. Proceedings of 20<sup>th</sup> IEEE International Symposium on Computer-Based Medical Systems (CBMS), 20–22 June, Maribor, Slovenia. doi:10.1109/CBMS.2007.43.
- Stone C & Mohammed C 2017. Applications of remote sensing technologies for assessing planted forests damaged by insect pests and fungal pathogens: a review. *Current Forestry Reports* 3, 75-92.
- Stone C, Chisholm LA & McDonald S 2003. Spectral reflectance characteristics of *Pinus radiata* needles affected by dothistroma needle blight. *Canadian Journal of Botany* 81, 560-569.
- Stone C & Coops NC 2004. Assessment and monitoring of damage from insects in Australian eucalypt forests and commercial plantations. *Australian Journal of Entomology* 43, 283-292.



- Stone C & Haywood A 2006. Assessing canopy health of native eucalypt forests. *Ecological Management and Restoration* 7, S24-S30.
- Storer AJ, Gordon TR & Clark SL 1998. Association of the pitch canker fungus, *Fusarium circinatum* f.sp. pini, with Monterey pine seeds and seedlings in California. *Plant Pathology* 47, 649-656.
- Storer AJ, Wood DL & Gordon TR 2002. The epidemiology of pitch canker of Monterey pine in California. *Forest Science* 48, 694-700.
- Strobl C, Boulesteix A-L & Augustin T 2007a. Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis* 52, 483-501.
- Strobl C, Boulesteix A, Zeileis A & Hothorn T 2007b. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8. doi:10.1186/1471-2105-8-25.
- Strobl C, Malley J & Tutz G 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323-348.
- Sumsion GR, Bradshaw MS, Hill KT, Pinto LDG & Piccolo SR 2019. Remote sensing tree classification with a multilayer perceptron. *PeerJ* 7. doi:10.7717/peerj.6101.
- Sun J, Zhang Y, Mao H, Cong S & Wu X 2018. Research of moldy tea identification based on RF-RFE-Softmax model and hyperspectra. *Optik* 153, 156-163.
- Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP & Feuston BP 2003. Random forest: A classification and regression tool for compound classification and QSAR modelling. *Journal of Chemical Information and Computer Sciences* 43, 1947-1958.
- Tang, Jiliang & Alelyani, Salem & Liu, Huan. (2014). Feature selection for classification: A review. In Aggarwal CC (Ed), *Data Classification: Algorithms and Applications*, 37-64. Chapman & Hall/CRC: Boca Raton, Florida, USA.
- The Mathworks Inc. 2019. MATLAB. <https://www.mathworks.com/products/matlab.html> [last accessed 22 October 2019].
- Thenkabail PS, Mariotto I, Gumma MK, Middleton EM, Landis DR & Huemmrich KF 2013. Selection of hyperspectral narrowbands (HNBS) and composition of hyperspectral twoband vegetation indices (HVIs) for biophysical characterization and discrimination of crop types using field reflectance and Hyperion / EO-1 data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6, 427-439.
- Thomas V, Treitz P, McCaughey JH, Noland T & Rich L 2008. Canopy chlorophyll concentration estimation using hyperspectral and LiDAR data for a boreal mixedwood forest in northern Ontario, Canada. *International Journal of Remote Sensing* 29, 1029-1052.

- Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M & Van Hijum SAFT 2012. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics* 14, 315-326.
- Trimble Geospatial 2019. eCognition Developer. <http://www.ecognition.com/suite#ecognition-developer> [last accessed 22 October 2019].
- Trojak M & Skowron E 2017. Role of anthocyanins in high-light stress response. *World Scientific News* 81, 150-168.
- Tsai C-F & Hsiao Y-C 2010. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems* 50, 258-269.
- Tucker CJ 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* 8, 127-150.
- Tuv E, Borisov A, Runger G & Torkkola K 2009. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research* 10, 1341-1366.
- Urrea V & Calle ML 2013. Variable selection with random forests and the area under the curve. In: Package 'AUCRF'. The Comprehensive R Archive Network: Vienna, Austria. Available online: <http://cran.r-project.org/web/packages/AUCRF/index.html> [last accessed 1 November 2019].
- Ustin SI & Gamon JA 2010. Remote sensing of plant functional types. *New Phytologist* 186, 795-816.
- Vainio EJ, Bezos D, Bragança H, Cleary M, Fourie G, Georgieva M, Ghelardini L, Hannunen S, Ioos R, Martín-García J, Martínez-Álvarez P, Mullet M, Oszako T, Papazova-Anakieva I, Piškur B, Romeralo C, Sanz-Ros AV, Steenkamp ET, Tubby K, Wingfield MJ & Diez JJ 2019. Sampling and detection strategies for the pine pitch canker (PPC) disease pathogen *Fusarium circinatum* in Europe. *Forests* 10, 723. doi:10.3390/f10090723.
- Vigier BJ, Pattey E & Strachan IB 2004. Narrowband vegetation indexes and detection of disease damage in soybeans. *IEEE Geoscience and Remote Sensing Letters* 1, 255-259.
- Viljoen A, Wingfield MJ & Marasas WFO 1994. First report of *Fusarium subglutinans* f. sp. pini on pine seedlings in South Africa. *Plant disease* 78, 309-312.
- Viljoen A, Wingfield MJ, Marasas WFO & Coutinho TA 1997. Pitch canker of pines – A contemporary review. *South African Journal of Science* 93, 411-413.
- Vincenzi S, Zucchetta M, Franzo P, Pellizzato M, Pranovi F, De Leo GA & Torricelli P 2011. Application of a random forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon, Italy. *Ecological Modelling* 222, 1471-1478.

- Vincini M, Frazzi E & D'Alessio P. Angular dependence of maize and sugar beet VIs from directional CHRIS / Proba data. Proceedings of 4th ESA CHRIS PROBA Workshop, 23-27 April, Frascati, ESRIN, Italy.
- Vogelmann JE, Rock BN & Moss DM 1993. Red edge spectral measurements from sugar maple leaves. *International Journal of Remote Sensing* 14, 1563-1575.
- Walker P 2009. Guidelines for post processing ASD FieldSpec Pro and FieldSpec 3 spectral data files using the FSF MS Excel template, V03.1. Natural Environment Research Council Field Spectroscopy Facility, Edinburgh, UK.
- Wang H, Tang Y, Jia Z & Ye F 2019. Dense adaptive cascade forest: a self-adaptive deep ensemble for classification problems. *Soft Computing*. doi:10.1007/s00500-019-04073-5.
- Wang J, Sammis TW, Gutschick VP, Gebremichael M, Dennis AO & Harrison RE 2010. Review of satellite remote sensing use in forest health studies. *The Open Geography Journal* 3, 28-42.
- Wang J, Shi T, Liu H & Wu G 2016. Successive projections algorithm-based three-band vegetation index for foliar phosphorus estimation. *Ecological Indicators* 67, 12-20.
- Wang L, Gong P & Biging S 2004. Individual Tree-Crown Delineation and Treetop Detection in High-Spatial-Resolution Aerial Imagery. *Photogrammetric Engineering & Remote Sensing* 70, 351-357.
- Wang Z, Skidmore AK, Wang T, Darvishzadeh R, Heiden U, Heurich M, Latifi H & Hearne J 2017. Canopy foliar nitrogen retrieved from airborne hyperspectral imagery by correcting for canopy structure effects. *International Journal of Applied Earth Observation and Geoinformation* 54, 84-94.
- Waske B, Fauvel M, Benediktsson JA & Chanussot J 2009. Machine learning techniques in remote sensing data analysis. In Camps-Valls G & Bruzzone L (Eds), *Kernel methods for remote sensing data analysis*. Chichester: John Wiley & Sons, Ltd.
- White JC, Wulder MA, Brooks D, Reich R & Wheate RD 2005. Detection of red attack stage mountain pine beetle infestation with high spatial resolution satellite imagery. *Remote Sensing of Environment* 96, 340-351.
- Whiteside T & Boggs G 2009. Object oriented image analysis. Proceedings of Surveying & Spatial Sciences Institute Biennial International Conference, 28 September-2 October, Adelaide, Australia.
- West JS, Bravo C, Obertu R, Lemaire D, Moshou D & McCartney HA 2003. The potential of optical canopy measurement for targeted control of field crop diseases. *Annual Review of Phytopathology* 41, 593-614.

- Wingfield MJ, Coutinho TA, Roux J & Wingfield BD 2002. The future of exotic plantation forestry in the tropics and Southern Hemisphere: lessons from pitch canker. *South African Forestry Journal* 195, 79-82.
- Wingfield MJ, Hammerbacher A, Ganley RJ, Steenkamp ET, Gordon TR, Wingfield BD & Coutinho TA 2008. Pitch canker caused by *Fusarium circinatum* – a growing threat to pine plantations and forests worldwide. *Australasian Plant Pathology* 37, 319-334.
- Wingfield MJ, Wingfield BD, Coutinho TA, Viljoen A, Britz H & Steenkamp ET 1999. Pitch canker: a South African perspective. Proceedings of IMPACT Monterey Workshop, 30 November-3 December 1998, Monterey, California, USA. CSIRO Forestry and Forest Products Technical Report No. 112. pp 62-69.
- Witten IH, Frank E & Hall MA 2011. Data mining: Practical machine learning tools and techniques. 3<sup>rd</sup> Ed. Morgan Kaufmann Publishers: Burlington, MA, USA.
- Woodall CW, Morin RS, Steinman JR & Perry CH 2010. Comparing evaluations of forest health based on aerial surveys and field inventories: oak forests in the Northern United States. *Ecological Indicators* 10, 713-718.
- Woolley JT 1971. Reflectance and transmittance of light by leaves. *Plant Physiology* 47, 656-662.
- Wright MN, Wager S & Probst P 2019. A fast implementation of random forests. In Package ‘ranger’. The Comprehensive R Archive Network: Vienna, Austria. Available online: <https://cran.r-project.org/web/packages/ranger/index.html> [last accessed 10 September 2019].
- Wright MN & Ziegler A 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77. doi:10.18637/jss.v077.i01.
- Wu Y, Boos DD & Stefanski LA 2007. Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* 102, 235–243.
- Wu X & Kumar V (Eds) 2009. *The top ten algorithms in data mining*. Chapman & Hall/CRC: Boca Raton, Florida, USA. pp 201.
- Wu C, Niu Z, Tang Q & Huang W 2008. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology* 148, 1230-1241.
- Wulder MA, Dymond CC, White JC, Leckie DG & Carroll AL 2006. Surveying mountain pine beetle damage of forests: A review of remote sensing opportunities. *Forest Ecology and Management* 221, 27-41.
- Wulder MA, White JC, Bentz BJ, Alvarez MF & Coops NC 2006a. Estimating the probability of mountain pine beetle red-attack damage. *Remote Sensing of Environment* 101, 150-166.

- Wulder MA, White JC, Bentz BJ & Ebata T 2006b. Augmenting the existing survey hierarchy for mountain pine beetle red-attack damage with satellite remotely sensed data. *The Forestry Chronicle* 82, 187-202.
- Wulder MA, White JC, Coops NC & Butson CR 2008. Multi-temporal analysis of high spatial resolution imagery for disturbance monitoring. *Remote Sensing of Environment* 112, 2729-2740.
- Xia J, Chanussot J, Du P & He X 2015. Spectral-spatial classification for hyperspectral data using rotation forests with local feature extraction and Markov Random Fields. *IEEE Transactions on Geoscience and Remote Sensing* 53, 2532-2546.
- Xia J, Du P, He X & Chanussot J 2014. Hyperspectral remote sensing image classification based on rotation forest. *IEEE Geoscience and Remote Sensing Letters* 11, 239-243.
- Xia J, Ghamisi P, Yokoya N & Iwasaki A 2018. Random forest ensembles and extended multiextension profiles for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 56, 202-216.
- Xu B, Huang JZ, Williams G, Wang Q & Ye Y 2012. Classifying very high dimensional data with random forests built from small subspaces. *International Journal of Data Warehousing and Mining* 8, 44-63.
- Xulu S, Peerbhay K, Gebreslasie M & Ismail R 2019. Unsupervised clustering of forest response to drought stress in Zululand region, South Africa. *Forests* 10, 531. doi:10.3390/f10070531.
- Yang X, Fan W & Yu Y 2010. Leaf and canopy chlorophyll content retrieval from hyperspectral remote sensing imagery. Proceedings of IEEE Sensors Applications Symposium, 23-25 February, Limerick, Ireland. doi:10.1016/j.rse.2008.04.005.
- Yarbrough LD, Easson G & Kuzsmaul JS 2005. QuickBird 2 tasseled cap coefficients: a comparison of derivation methods. Proceedings of American Society for Photogrammetry and Remote Sensing (ASPRS/Pecora), 23-27 October, Sioux Falls, South Dakota, USA.
- Yu G, Yuan J & Liu Z 2013. Action search by example using randomized visual vocabularies. *IEEE Transactions on Image Processing* 22, 377-390.
- Yuan H, Yang G, Li C, Wang Y, Liu J, Yu H, Feng H, Xu B, Zhao X & Yang X 2017. Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: analysis of RF, ANN, and SVM regression models. *Remote Sensing* 9. doi:10.3390/rs9040309.
- Zarco-Tejada PJ, Berjón A, López-Lozano R, Miller JR, Marin P, Cachorro V, González MR & De Frutos A 2005. Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sensing of Environment* 99, 271-287.

- Zarco-Tejada PJ, Berni JAJ, Suárez L, Sepulcre-Cantó G, Morales F & Miller JR 2009. Imaging chlorophyll fluorescence with an airborne narrowband multispectral camera for vegetation stress detection. *Remote Sensing of Environment* 113, 1262-1275.
- Zarco-Tejada PJ, González-Dugo V & Berni JAJ 2012. Fluorescence, temperature and narrowband indices acquired from a UAV platform for water stress detection using a micro-hyperspectral imager and a thermal camera. *Remote Sensing of Environment* 117, 322-337.
- Zarco-Tejada PJ, Miller JR, Mohammed GH & Noland TL 2000. Chlorophyll fluorescence effects on vegetation apparent reflectance: I. Leaf-level measurements and model simulation. *Remote Sensing of Environment* 74, 582-595.
- Zarco-Tejada PJ, Miller JR, Noland TL, Mohammed GH & Sampson PH 2001. Scaling-up and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* 39, 1491-1507.
- Zarnoch SJ, Bechtold WA & Stolte KW 2004. Using crown condition variables as indicators of forest health. *Canadian Journal of Forest Research* 34 1057-1070.
- Zhang Y, Cao G, Li X & Wang B 2018. Cascaded random forest for hyperspectral image classification. *IEEE Journal of Selected Topics in Earth Observations and Remote Sensing* 11, 1082-1094.
- Zhang C, Liu Y, Kovacs JM, Flores-Verdugo F, de Santiago FF & Chen K 2012. Spectral response to varying levels of leaf pigments collected from a degraded mangrove forest. *Journal of Applied Remote Sensing* 6. doi:10.1117/1.JRS.6.063501.
- Zhang R, Ma J, Chen X & Tong Q 2009b. Feature selection for hyperspectral data based on modified recursive support vector machines. Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 12-17 July, Cape Town, South Africa. doi:10.1109/IGARSS.2009.5418228.
- Zhang H, Yu C-Y & Singer B 2003. Cell and tumor classification using gene expression data: construction of forests. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 100, 4168-4172.
- Zhang H, Wang M & Chen X 2009a. Willows: a memory efficient tree and forest construction package. *BMC Bioinformatics* 10. doi:10.1186/1471-2105-10-130.
- Zhang C-X & Zhang J-S 2008. RotBoost: a technique for combining rotation forest and AdaBoost. *Pattern Recognition Letters* 29, 1524-1536.
- Zhou 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC: Boca Raton, Florida, USA. pp 218.

- Zhou 2015. Ensemble learning. In Li SZ & Jain AK (Eds), *Encyclopaedia of Biometrics*. Springer: Boston, MA. pp 1452. doi:10.1007/978-1-4899-7488-4.
- Zhou L, Wang Q, Yin P, Xing W, Wu Z, Chen S, Lu X, Zhang Y, Lin X & Xu G 2012. Serum metabolomics reveals the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver diseases. *Analytical and Bioanalytical Chemistry* 403, 203-213.
- Zhou Z-H & Feng J 2017. Deep forest: towards an alternative to deep neural networks. Proceedings of Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), 19-25 August, Melbourne, Australia. doi:10.24963/ijcai.2017/497.
- Zhao D, Redy KR, Kakani VG, Read JJ & Koti S 2005. Selection of optimum reflectance ratios for estimating leaf nitrogen and chlorophyll concentrations of field-grown cotton. *Agronomy Journal* 97, 89-98.