Development of a Fourier Transform Infrared (FT-IR) model for screening susceptible and resistant *Eucalyptus* clones against *Chrysoporthe austroafricana*

By

Valencia Mogashoa

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences Department of Biochemistry, Genetics and Microbiology University of Pretoria
Pretoria

June 2022

Under the supervision of Professor Sanushka Naidoo and co-supervision of Professor Pierluigi Bonello

Declaration

I, Valencia Mogashoa declare that the dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Valencia Mogashoa

June 2022

Table of Contents

Dissertation Summary

Dissertation title: Development of a Fourier Transform Infrared (FT-IR) model for screening susceptible and resistant *Eucalyptus* clones against *Chrysoporthe austroafricana*

Full name: Valencia Mogashoa

Supervisor: Prof. Sanushka Naidoo

Co-supervisors: Prof. Pierluigi (Enrico) Bonello

Collaborator & Mentor: Dr. Anna Conrad

Department: Department of Biochemistry, Genetics and Microbiology, University of Pretoria

Degree: *Magister Scientiae*

Eucalyptus trees are an important source of timber in South Africa. Unfortunately, the trees are susceptible to a number of pests and pathogens. *Chrysoporthe austroafricana* is a significant pathogen of *Eucalyptus* trees in South Africa. The pathogen causes stem cankers which lead to wilting and eventually death. Clonal and hybridization trials have been carried out to improve the genotype of these trees. In order to breed for pest and pathogen resistance, this procedure necessitates the screening of clones and hybrids. Current screening methods rely on inoculation trials and natural infection, both of which are time consuming and, in the case of inoculation trials, destructive. Here I show that resistance phenotyping can be conducted, rapidly and non-destructively, by way of infrared (IR) spectroscopy, which generates chemical fingerprints that can be used to predict and identify resistant and susceptible trees when combined with chemometric analyses. Specifically, I used Fourier transform IR (FT-IR) spectroscopy in combination with machine learning algorithms to distinguish between resistant and susceptible *Eucalyptus* hybrid clones against *C. austroafricana*. The results from the study is a proof of concept on the potential of FT-IR as a tool to screen *Eucalyptus* clones for resistance to *C. austroafricana*.

Acknowledgements

I would like to appreciate and acknowledge the following people, institutions and organizations for their support during my study:

- To my loving mother, thank you for being my pillar of strength. Your constant support and prayers have kept me going. Thank you Mama.

- To Dr K.G. Shadung, thank you for always encouraging me to keep going. Your support and prayers have sustained me during this period and I'm grateful.

- To my supervisor, Prof. Sanushka Naidoo, thank you for believing in me and pushing me to do better as well trusting me with this project. You have been amazing throughout my study. I am grateful for all your efforts, your time, advice and guidance during this time. You are such an inspiration and I am so grateful to have been a part of such an amazing team.

- To my co-supervisor, Prof. Enrico Bonello, thank you for the constant support, advise and of course your warm humour in the meetings. Words cannot express my gratitude towards you. Thank you for all your help and guidance at all times.

- To our collaborator and my mentor, Dr Anna Conrad, thank you for your constant support and help all the time. Thank you for teaching me so much about data analysis, and your willingness to go out of your way to help me. I appreciate your patience with me and your kindness always.

- To my amazing friends: Tsholanang Moselane, Laura Malivhoho, Morongwa Mathipa and Charmaine Malebe. I am blessed to have you in my life, thank you for your constant support and love.

- To Demissew Teshome, Lorraine Mhoswa, Erik Visser, Shae Swanepoel, thank you for always lending a hand and being available whenever I need help.

- To all past and present members of the *Eucalyptus* Pest Pathogen Interactions (EPPI), and Forestry Molecular Genetics (FMG) groups. Thank you for always willing to lend a hand.

- I would like to appreciate my funding bodies, the University of Pretoria, Sappi, Mondi, the Forest Sector Innovation Fund, the Department of Science and Innovation and the Technology Innovation Agency.

- To the department of Biochemistry, Genetics and Microbiology and the Forestry and Agricultural Biotechnology Institute. Thank you for providing a wonderful research environment.

List of abbreviations

ATR	attenuated total reflectance
AUC	area under the curve
BER	balanced error rate
CLO	coast live oak
CV	cross validation
ET	ethylene
FDA	functional data analysis
FN	false negative
FP	false positive
FT-IR	Fourier transform infrared spectroscopy
IR	infrared radiation
JA	jasmonic acid
LOOCV	leave one out cross validation
MEA	malt extract agarose
NIR	near infrared spectroscopy
RFE	recursive feature elimination
ROC	receiver operating characteristic
SA	salicylic acid
SAR	systemic acquired resistance

SD	standard deviation
SIMCA	soft independent modelling of class analogy
sPLS-DA	sparse partial least square discriminant analysis
SVM	support vector machine
TN	true negative
TP	true positive
UVE	uninformative variable elimination
VSURF	variable selection using random forests

Chapter 1

Literature review

Fourier transform infrared (FT-IR) spectroscopy as a phenotyping tool for discriminating between disease resistant and susceptible phenotypes in forestry

1.1. Introduction

Forests are important for ecological, economic, and social functions, and they continuously contribute to climate change mitigation (Sturrock *et al.*, 2011). However, forest trees have suffered significant losses throughout the years as a result of recurrent disease outbreaks caused by climate change and globalization. This has created difficulties for forest trees, endangering natural habitats, tree yield, and economic viability (Bonan, 2008; Boyd *et al.*, 2013; Hassan *et al.*, 2005; Mbow *et al.*, 2014).

When a disease is introduced into a new habitat where the native trees have little or no resistance to the pathogen, host shift events may occur (Brassier, 2008). Due to the lack of opportunity for hosts to co-evolve with introduced pathogens and acquire resistance, exotic infections gain an advantage, resulting in significant damage and even extinction of some tree species in a plantation or landscape (Hansen, 2008). This is typically a result of globalization, in which invasive and non-native pathogens are accidentally introduced into new environments as a result of the import and export of goods between countries (Hurley *et al*., 2016).

Climate change is also a significant factor in forest tree epidemics, as it has the potential to alter tree physiology and defense mechanisms, as well as shift pathogen distributions (Kirilenko & Sedjo, 2007). In both circumstances, trees that exhibit pathogen tolerance and resistance are desired for disease management strategies. Trees with these characteristics can be used in breeding strategies to select for pathogen resistance (Wingfield *et al.*, 2015). Current disease resistance screening approaches continue to rely on artificial inoculation and natural infection, which are not only time consuming and labor costly, but also lack throughput (Neale & Kremer, 2011; Conrad & Bonello, 2016).

This review focuses on Fourier transform infrared (FT-IR) spectroscopy in forestry and its potential application in screening *Eucalyptus grandis* hybrid clones for resistance to the stem canker pathogen *Chrysoporthe austroafricana.*

1.2. *Eucalyptus* plantation forestry

The genus *Eucalyptus* was first described in 1788 by botanist L' Heritier with further classification contributions of species within the genus by Maiden (1903-1933) and Blakely (1934; FAO, 1979; Turnbull, 1999). The trees are native to Australia and some pacific islands, and are widespread in South America, North America, southern Europe, the Middle East, Africa, China, and the Indian subcontinent (Eldridge *et al.*, 1993; FAO, 1979; Turnbull, 1999). More than 700 *Eucalyptus* species have been classified with the most common being *E. saligna*, *E. camaldulensis*, *E. globulus*, *E. grandis* and *E. urophylla* (Boland *et al.,* 2006; Friis, 1995). *Eucalyptus* trees are known for their ability to grow fast as well as adapt to different environments. Over the last century, many *Eucalyptus* species have been introduced to different parts of the world and quickly became important in the forest industry (Davidson, 1993).

Additionally, they have become important and valuable in forest plantations where they are used for a wide range of purposes such as honey making, pulp and paper production, poles for electricity transmission, construction, mining, as well as charcoal and biofuel. Essential oils from these trees can be extracted and used for both pharmaceutical and perfumery purposes (FAO, 2006; 1985 & 1979). The paper industry depends heavily on *Eucalyptus* trees because they are an excellent source of pulp (Magaton *et al.*, 2009). In South Africa, the trees were first introduced via transportation in containers from Mauritius in 1807 (FAO, 1979).

1.2.1. *Eucalyptus* plantations in South Africa

Commercial plantations in South Africa are made up of mostly pine and eucalypts with a smaller amount of wattle trees. According to a 2019 Forestry South Africa (FSA) report, *Eucalyptus* plantations are predominant in the northern regions of the country (Zululand, Kwa-Zulu Natal province), accounting for over 296 000 hectares (57%) compared to pine, wattle and other commercial tree species in that province. The report also shows that most of these trees are used for pulpwood production, which has generated over R14 billion since 2018. Pathogens have become more prevalent in South African plantations as a result of increased imports and exports of commodities, as well as a changing climate. Both these factors have had a detrimental

effect on *Eucalyptus* productivity which is already vulnerable to a variety of pests and diseases (Poore & Fries, 1985; Wingfield *et al.*, 2015). Some of the pests found in South African eucalypt plantations include *Leptocybe invasa*, *Procantha semipuncata*, and *Coryphodemia tristis*. Examples of important fungal pathogens include *Austropuccinia psidii*, *Ceratocystis* spp., *Teratosphaeria destructans*, *Botryosphaeria* spp., and *C. austroafricana* (Wingfield *et al.*, 2008; Crous *et al*., 2009; Paine *et al.*, 2011; Greyling *et al.*, 2016).

1.2.2. Chrysoporthe austroafricana as a pathogen on *Eucalyptus*

Previously known as *Chrysoporthe cubensis*, *C. austroafricana* is a fungal pathogen that causes cankers on the stems of *Eucalyptus* and *Tibouchinia* species (Wingfield *et al.*, 1989; Myburg *et al.*, 2002). It was detected for the first time in South Africa during disease inspection surveys on young *E. grandis* plants in the late 1980s. At the time, *C. austroafricana* was discovered causing cankers on trees in forest plantations north of Kwa Zulu Natal. DNA sequence analysis of *C. austroafricana* revealed that its β-tubulin, ITS regions and histone H3 genes, as well as morphological differences such as longer asci, are distinct from those of *C. cubensis* (Gryzenhout *et al.*, 2004).

Early infection symptoms in *Eucalyptus* spp. include sunken outer bark at the base, whereas severe disease progression symptoms include colonization of the cambium and girdling, which results in tree mortality (Wingfield *et al.*, 1989). In young trees, disease symptoms typically include stem girdling at the root collar and rapid death with leaves intact on the tree (Conradie *et al.*, 1990). According to Wingfield (2003), susceptible young trees are most likely to be sensitive to the disease during their first year of growth, dying in significant numbers. Surviving trees typically have large basal cankers, and they can fall over in a windstorm or die gradually over a five-to ten-year rotation. Due to the lesion on the stems, it was previously assumed that the pathogen was necrotrophic; however, microscopic examination of its interaction with *E. grandis* plant cells revealed that it may be hemi-biotrophic (Mangwanda *et al.*, 2016; Zwart *et al.*, 2017).

The interaction between *C. austroafricana* and *E. grandis* is an example of a well-established pathosystem that has been employed to study defense mechanisms in *Eucalyptus* against fungal pathogens (Naidoo *et al.*, 2013).

1.2.3. *Eucalyptus grandis*-*C. austroafricana* pathosystem

The *E. grandis*-*C. austroafricana* pathosystem has a well-developed controlled inoculation protocol that has assisted in providing reliable screening of resistant and susceptible *E. grandis* clones and hybrids (Wingfield *et al.*, 1989; Roux *et al.*, 2003; Van Heerden *et al*., 2005). Van Heerden *et al* (2005) carried out a screening study for disease resistance of clonal *E. grandis* trees and hybrids against *C. austroafricana*. Two *E. grandis* clones, namely ZG14 and TAG5 showed varied levels of susceptibility to the pathogen, with ZG14 showing high susceptibility, and TAG5 displaying moderate resistance to the pathogen. Other *E. grandis* hybrids showed varied levels of resistance to *C. austroafricana* (Van Heerden *et al*., 2005).

Due to lack of fully resistant *E. grandis* genotypes, hybrid clones have been widely used in South African plantations to effectively control *C. austroafricana* (Wingfield *et al.*, 2001). *Eucalyptus grandis* x *E. urophylla* hybrids have been specifically used to control this stem canker pathogen (Denison & Kietzka, 1993). The hybrids combine *E. grandis*' rapid growth and wood quality with *E. urophylla*'s stronger resistance to *C. cubensis* (Gominho *et al.*, 2001). Backcrossing, a type of recurrent hybridization, can be used to create lines that are almost identical to the recurrent parent with the addition of the gene of interest through breeding (Vogel, 2009). The Urograndis backcross population is an example of recurrent hybridization, whereby *E. grandis* is crossed with *E. urophylla*, the progeny is then crossed back with *E. grandis* (GUxG) (Mondi Tree Breeding Program). In this case, *E. urophylla* is the donor parent with the desirable characteristics, while *E. grandis* is the recurrent parent. The availability of the *E. grandis* genome has also allowed for the modelling of the *C. austroafricana*-*E. grandis* pathosystem to study defense mechanisms in *Eucalyptus grandis* (Myburg *et al.*, 2014).

1.2.4. *Eucalyptus* specific defenses against *C. austroafricana*

Defense responses in stem and leaf tissues of *E. grandis* to *C. austroafricana* have been extensively studied. Naidoo *et al.* (2013) employed the *C. austroafricana*-*E. grandis* pathosystem for the first time to investigate the role of plant phytohormones salicylic acid (SA) and jasmonic acid (JA) in host resistance to the fungal pathogen. The antagonistic link between SA and JA was shown by marker genes expressed in these phytohormonal signaling pathways. Significant reductions in lesion length on ZG14 plants supported the hypothesis that SA is involved in mediating disease resistance. Mangwanda *et al.* (2015) investigated the role of phytohormones in defense responses, revealing that phytohormone responses are triggered in TAG5 and ZG14 in response to *C. austroafricana* infection. SA basal levels were significantly higher in the moderately resistant TAG5 than in ZG14, suggesting that TAG5 may induce earlier systemic acquired resistance (SAR).

SA has been implicated in TAG5 defense responses against *C. austroafricana*. This finding contradicted the assumption that *C. austroafricana* was a necrotrophic pathogen. According to literature, JA/ET-mediated responses are related to defense responses against necrotrophic pathogens, whereas SA-mediated responses are associated with responses against biotrophic pathogens (Glazebrook, 2005). Microscopy studies were conducted on *E. grandis* clones infected with *C. austroafricana* to validate the fungal pathogen's lifecycle. Microscopy studies of *E. grandis* stems infected with *C. austroafricana* confirmed the pathogen's suspected lifestyle. Hyphae discovered in living cells of *E. grandis* showed the possibility of a biotrophic phase in the fungus' life cycle, implying that *C. austroafricana* is a hemi-biotroph instead of a necrotroph (Mangwanda *et al.*, 2016; Zwart *et al.*, 2017).

Genome sequencing of *E. grandis* by Myburg *et al*. (2014) revealed that *E. grandis* has the largest number of terpene synthase genes among all the sequenced plant genomes (n= 113). Visser *et al*. (2015) studied foliar levels of terpenes and the expression of biosynthetic genes in ZG14 and TAG5 infected with *C. austroafricana*. A variation in constitutive terpenoid levels between ZG14 and TAG5 was demonstrated by the presence of p-cymene and pinocavone in TAG5. Elevated p-cymene levels in inoculated TAG5 plants vs mock-inoculated plants revealed that *C.*

austroafricana elicits systemic responses in *E. grandis*. These findings showed that constitutive and induced chemical defenses are involved in defense against this stem canker pathogen.

1.3. Chemical defenses in plants

There is a knowledge gap in the chemical defenses involved in the interactions between *Eucalyptus* spp. and *C. austroafricana*. Plants have pre-formed or constitutive defense systems that protect them from herbivory, infections, and prevent full-blown invasions. Physical and chemical barriers are examples of these defences (Fahn, 1988; Bonello *et al.*, 2006). Chemical defenses, both pre-formed and induced, include specialized metabolites such as resin acids, terpenes, and phenolic compounds (Honkanen *et al.*, 1999; Campos *et al.*, 2008). Induced defense responses can be either direct, directly affecting the invader, or indirect, drawing the invader's natural opponent to the tree. Induced defense responses are only activated in the presence of the invader and are less costly to the plants because they are not always activated (Frost *et al.*, 2008).

Specialized metabolites are a wide class of compounds that are critical in protecting plants from biotic and abiotic stressors (Inderjit, 1999; Kosuge, 1969; Rice, 1984). These compounds are divided into three general groups: terpenes, phenolics, and compounds containing nitrogen or sulfur, such as alkaloids and glucosinolates (Koornneef & Pieterse, 2008; Mazid *et al.*, 2011).

Phenolics is an umbrella term that refers to a diverse group of about 8000 known compounds that have an aromatic ring with one or more hydroxyl substituents and functional moieties such as esters, methyl esters, and glycosides (Ho, 1992). The first step in the biosynthesis of phenolics is the deamination of phenylalanine or tyrosine. The amino acids are transformed into cinnamic acids, which are then introduced into the phenylpropanoid pathway, where they are changed into a variety of phenols via the addition of hydroxyl and other groups to the phenyl ring(s) (Pereira *et al.,* 2009). These compounds are then classified into various other groups, with flavonoids being the most prevalent and well known. The other groups also include phenolic acids, stilbenes, lignans, coumarins, tannins, and anthocyanins (Pereira *et al.*, 2009).

Phenolics have been demonstrated to play a role in plant defense against herbivores, insects, and fungal and bacterial diseases in many plant species including *Eucalyptus* (Batish *et al*., 2008; Lattanzio *et al.*, 2008).

1.3.1. Specialized metabolites in eucalypts

An array of diverse specialized metabolites are known to exist in *Eucalyptus* spp. Monoterpenes, diterpenes, triterpenes, sesquiterpenes, triketones, and steroidal chemicals are abundant in essential oils from *Eucalyptus* tree species, which are mostly extracted from leaves but are also present to some extent in seeds, flowers, and bark (Brezáni & Šmejkal, 2006). These chemicals have been proven to have anti-inflammatory, antibacterial, and antifungal activities (Brezáni & Šmejkal, 2006). Santos *et al.* (2012) discovered a range of phenolic chemicals in *E. grandis*, *E. urograndis*, and *E. maindenii* bark extracts. In comparison to *E. maidenii*, gallic acid, ellagic acid, ellagic acid-rhamnoside, and epicatechin were detected in greater concentrations in *E. grandis* and *E. urophylla* bark extracts. Epicatechin was the most abundant phenolic component in both *E. grandis* and *E. urograndis*. A year later, Santos *et al*. (2013) discovered that gallic acid, catechin, ellagic acid, and ellagic acid pentoside were present in amounts greater than 2 $mg.g^{-1}$ in all three *Eucalyptus* species. This provides insights on the phenolic chemical profiles that are present in the three studied *Eucalyptus* species, especially those of *E. grandis* and *E. urograndis* which are of interest to this review.

There is limited information available about the role of specialized constitutive and induced metabolic defenses against fungal pathogens in *Eucalyptus*. While the majority of studies focus on chemical defenses against herbivory in these plants, these studies can still provide insights into their roles against other biotic stressors. Foley and Moore (2005) showed that phenolic compounds such as condensed tannins can play a role in resistance against marsupial and cottonwood beaver herbivory in *E. globulus*. This was also confirmed by results from a study by O'Reilly-Wapstra *et al*. (2005), where they demonstrated that elevation of tannin levels in four *E. globulus* hybrids protects against browsing damage by *Trichosumus vulpecula.* This suggests that chemicals associated with defense in *Eucalyptus* hybrid clones may be useful for

selection of resistant phenotypes against fungal pathogens e.g., through the use of chemical phenotyping techniques (Conrad & Bonello, 2016).

1.3.2. Chemical phenotyping for disease resistance screening

Fungal pathogens are rapidly emerging in *Eucalyptus* plantations due to climate change and increased global trade. Therefore, there is a critical need for rapid screening tools for early (pre-symptomatic) disease diagnosis and targeted disease management, as well as rapid, non-destructive (i.e., without inoculation) identification of resistant phenotypes for breeding programs (Conrad & Bonello, 2016; Conrad *et al.*, 2020). Natural infection and artificial inoculation are still primarily employed to screen trees for disease resistance and select resistant genotypes (Neale & Kremer, 2011). While these methods have proven to be successful, they are time consuming, labor intensive, expensive, and dependent on the age and size of the trees (Martin *et al.,* 2005). Although genetic and genomic techniques such as the selection of quantitative features for disease resistance are promising, they are very expensive and limiting in circumstances where the genetic structure of host resistance is unknown for the pathosystem being studied (Taoutaou *et al.*, 2012; Muranty *et al.*, 2014).

Compared to artificial inoculations and molecular marker techniques, constitutive metabolic profiling of hosts can quickly reveal chemical features that are associated with resistance, allowing us to classify trees as resistant or susceptible (Conrad & Bonello, 2016). The technique of constitutive metabolic profiling is simply a classification tool, i.e., it does not provide functional information attributable to the chemical constituents being measured. The metabolic profile, also known as a chemical fingerprint, is a representation of the various metabolites or chemicals in a sample (Sumner *et al.*, 2003), which makes it possible to phenotype plants (Roesnner & Bowne, 2018).

1.4. Chemical fingerprinting techniques

Infrared (IR) spectroscopy, a type of vibrational spectroscopy, is a promising method for obtaining plant chemical fingerprints. The approach determines the wavelength

and intensity of a material's infrared light absorption (Putzig *et al.*, 1994). This is achieved through the absorption of infrared light by the chemical compounds within the sample, whose energy, in turn causes chemical bonds within the sample to vibrate in a bond-specific manner. This generates an interference pattern that can be amplified to produce an IR spectrum. According to Berthomieu and Hienerwadel (2009), the IR absorption bands are related to specific functional groups and their associated vibrations. Infrared is classified into three ranges: near IR (13000-4000 cm^{-1}), mid IR (4000-400 cm^{-1}), and far IR (400-10 cm^{-1}; Li-Chan, 2010). A spectrum is a unique chemical fingerprint that is typical of specific functional groups and molecules within that sample, hence no two samples may have the same spectrum (Dutta, 2017). The spectrum can be used to identify known and unknown materials by comparing spectral repositories and interpreting distinctive absorbances (Tyner & Francis, 2017).

Near-infrared (NIR) and mid-IR spectroscopy using Fourier transform infrared (FT-IR) are methods for chemical fingerprinting (Smith, 2011). The NIR range (13000- 4000 cm^{-1}) is known to feature overtones and a combination of CH, NH, and OH vibrations (Bokobza, 1998; Weyer, 1985). Raman spectroscopy, on the other hand, uses a laser source of monochromatic radiation, but is similar to FT-IR in that it employs light scattering to detect the vibrational frequencies of molecules and the lattice vibration of crystalline materials. However, FT-IR is reliant on light absorption by molecules (Murphy *et al.*, 1998). Depending on the nature of samples analyzed, all of these procedures can be non-destructive, quick, and can be used to obtain chemical fingerprints (Carden & Morris, 2000). This review will solely look at FT-IR as a tool for chemical fingerprinting.

1.4.1. Fourier transform infrared spectroscopy (FT-IR)

Fourier transform infrared (FT-IR) spectroscopy is a useful method for identifying functional groups and molecular bonds within an analyte or mixed sample (Christy *et al.*, 2001). Because of the scarcity of fast scanning instruments, the tool was created to measure IR frequencies simultaneously (Thermofisher Technical Report, 2013). The equipment is mostly used to investigate the mid-IR region (4000-700 cm^{-1}) and can quantify and identify organic and inorganic materials (Diem, 2015). The technique has numerous advantages which contribute to its common use.

1.4.1.1. Advantages of FT-IR spectroscopy

There are several reasons why FT-IR spectroscopy is widely used and popular in a variety of industries and fields. One reason is that its developments have significantly enhanced the potential of IR. When frequencies are measured, the task is reduced to a single scan, which accelerates the process. This significantly reduces the time required to capture several scans (Thermofisher Technical Report, 2013). The sensitivity of the FT-IR instrument improves the signal-to-noise ratio. The majority of interferometers employ the fewest available mirrors to limit reflection losses. This allows for increased energy to reach the sample, boosting the signal and improving the technique's signal-to-noise ratio and sensitivity (Perkins, 1987).

FT-IR spectrometers have the advantage of combining and averaging scans via co-addition of interferograms or spectral data, which is desirable because just one interferogram is calculated (Thermofisher Technical Report, 2013). Lasers are employed as an internal reference to ensure the instrument's precision and accuracy. This maintains wavelength precision, hence eliminating wavelength drift issues. All of this, combined with the technique' potential for non-invasive analysis and minimum sample preparation requirements, makes FT-IR a desirable tool for spectral data processing (Ismail *et al.*, 1997; Diem, 2015; Levine *et al.*,1989).

1.4.1.2. The principle of FT-IR

The use of interferometry based on the optical principle of the Michelson interferometer gives FT-IR its desirable properties compared to other dispersive instruments (Kaufmann & Galizzi, 2002). When IR radiation is passed onto a sample, some of the radiation is absorbed, causing molecules in the sample to vibrate at certain frequencies (Hollas, 2013). Molecular bonds have characteristic vibrational frequencies that are determined by the stiffness of the bonds as well as the molecular weight of the atoms at the edge of each bond (Ismail *et al.*, 1997).

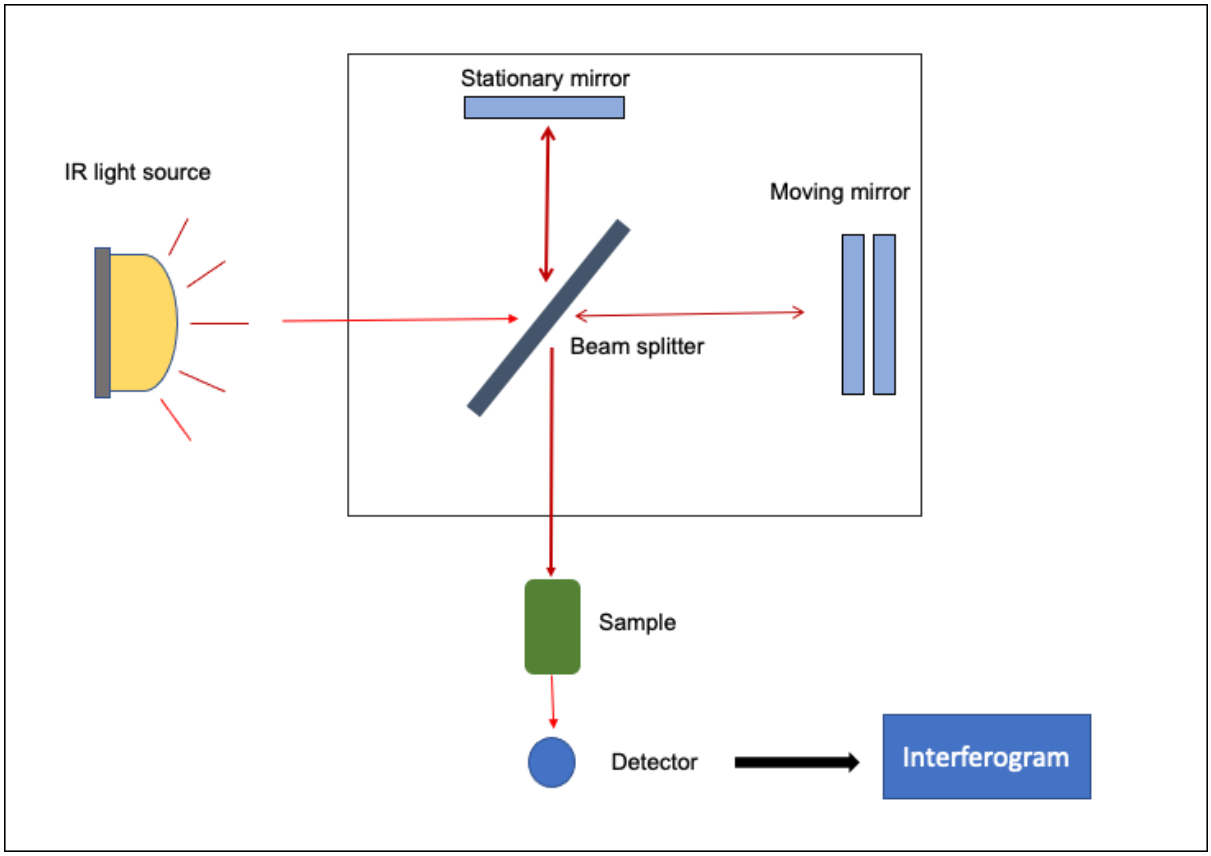

Figure 1.1: A schematic diagram of an FT-IR interferometer (adapted from Lee *et al.,* 2014).

1.4.1.3. Instrumentation

In a Michelson interferometer, a light beam is split into two pathways and then recombined with the introduction of a path difference (**Figure 1.1**; Griffith & de Haseth, 2007). The Michelson interferometer, according to Markovich *et al*. (1991), consists of two mirrors at right angles to each other, one movable and the other stationary. The interferometer's third component, the beam splitter, is located between the two mirrors at a 45-degree angle. An interferometer receives a parallel light beam form a light source. The beam is split into two parts: transmitted and reflected lights (Diem, 2015). An optical interference wave is created when the divided beam's light components reunite. The interference pattern differs due to the two components' different paths passing through the sample and may change during this interaction depending on the sample's attributes (Faix, 1992).

The energy that reaches the detector has a different intensity depending on the sample. When the interference pattern reaches the detectors, it is digitalized in real time by a computer, resulting in an interferogram that encodes spectral information, i.e., the intensity of light in relation to the optical path difference (Ismail *et al.*, 1997; Mohamed *et al.*, 2017; Schmitt & Fleming, 1998). The computer is critical for spectral data collecting and storage. Additionally, it performs post-scanning operations such as calibration, resolution enhancement, spectral visualization, and correlation equation calculation (Faix, 1992). Prior to analysing samples, a background measurement under the same conditions as the sample is important, as there may be interfering solvent and dissolved gas traces that contribute information unrelated to the sample (Moraes *et al.*, 2008). FT-IR can be coupled with an accessory tool to improve sample processing and minimize sampling time (Kazarian and Chan, 2006).

1.4.1.4. Attenuated total reflectance FT-IR

An FT-IR output image can be collected in three ways: transmission, reflection, and attenuated total reflectance (ATR). Depending on the type of sample being analysed, these sample measurement modes are utilized in various fields. The first two have been utilized in biological tissue analysis, but ATR has been demonstrated to be more adaptable (Kazarian & Chan, 2006). ATR enables samples to be analysed directly with little or no preparation; the accessory is useful for analysing strong absorbing or thick material samples, which are frequently difficult to process (Thermofisher Technical Report, 2013). ATR crystal materials can include diamond, KRS-5, ZnSe, thallium bromide, and optical solid solutions. The substance used to make the crystal can affect the differences and placements of the peak intensity (Shimadzu Corporation, 2015). ATR attachments are frequently used in conjunction with FTIR for fast and accurate analysis of solid and liquid samples, including pastes, as well as difficult-to-handle samples (Kazarian & Chan, 2006). Following the acquisition of FT-IR spectra (**Figure 1.2A-B**), wavenumbers relevant to disease resistance are analysed using multivariate analysis and machine learning algorithms (**Figure 1.2 A-K**). This procedure consists of several steps, which are detailed in the next section.

1.4.2. FT-IR data analysis with machine learning tools

Machine learning is defined as the process of applying statistical methods to large and complex datasets to discover patterns (Sperschneider, 2020). There are two types of learning methods: supervised and unsupervised. Supervised learning makes use of labelled data and is typically used for classification problems, whereas unsupervised learning (not reviewed here) makes use of unlabelled data and may include some exploratory analysis (Witten *et al.*, 2016). For classification purposes, critical components include a training data set with both positive and negative examples, feature selection for model training, and a validation test set that is independent of the training data set (Sperschneider, 2020). Numerous machine learning algorithms are available for extracting and interpreting useful information from FT-IR spectra (Deisentroth *et al.*, 2020).

1.4.2.1 Spectral data pre-processing

To build strong and reliable predictive models, pre-processing spectral data is required (Leger & Ryder, 2006; **Figure 1.2C)**. Pre-processing of spectral data follows Beer's rule, which requires that absorbance and concentration are linearly related, according to Rinnan *et al*. (2009). The procedure has three stages: noise reduction, data transformation, and peak handling (Xu *et al.*, 2020). Data analysis requires removing outliers and this can be accomplished in numerous ways. Functional data analysis (FDA) is a method for screening and removing outliers from spectral data (Febrero-Bande & de la Fuente, 2012; Heim *et al.*, 2018). Smoothing and noise reduction are common in spectral data, and Savitzky-Golay filtering is a popular approach used for this purpose. It uses the "Savgol" filter to remove noise and scattering as well as first and second derivative transformations to reduce errors and adjust baselines. These are reviewed by Press & Teukolsky (1990); Brown *et al.* (2000); Luo *et al*. (2005).

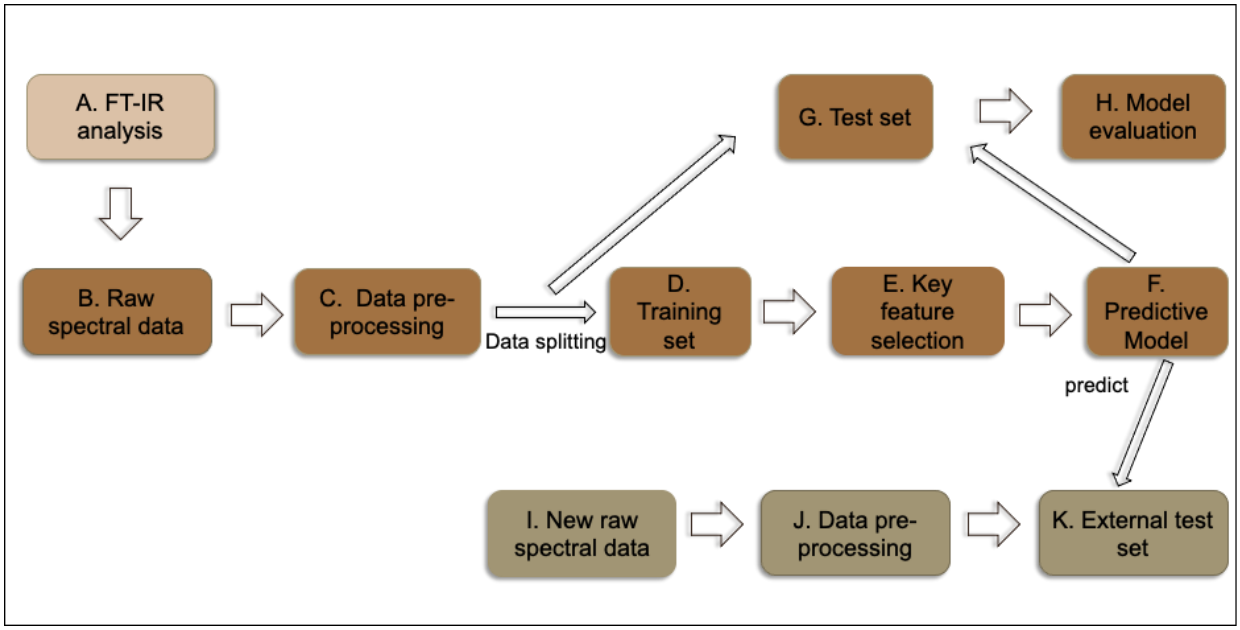

Figure 1.2: Machine learning workflow for predictive modelling of FT-IR spectral data.

1.4.2.2. Data splitting

Following pre-processing, which serves as the foundation for predictive modelling and classification, the splitting of spectral data into training and testing datasets is a necessary step (**Figure 1.2D & 1.2G**). The training data is used to construct the model by incorporating key variables, while the test dataset is used to assess the model's performance and overfitting (Conrad & Bonello, 2016). When developing a predictive model, it is not always necessary to utilize the entire FT-IR spectrum. Certain wavenumbers or regions of the spectrum may not be useful for prediction. Thus, critical and optimal variables containing important information or that are most correlated to the response (e.g. resistance) should be chosen and used for modelling (Moros *et al.,* 2010).

1.4.2.3. Variable selection

Anderson and Bro (2010) assert that variable selection can occasionally aid in improving model performance while reducing model complexity (**Figure 1.2E**). This procedure can be advantageous when it comes to removing similar redundant variables. However, it is critical to consider the model's nature. Numerous techniques, including recursive feature elimination (RFE), uninformative variable elimination

(UVE), data binning, and variable selection using random forests (VSURF), can be used to identify critical features for predictive modelling. (Centner & Massart, 1996; Guyon *et al.*, 2002; Genuer *et al.*, 2010; 2015).

RFE is a popular machine learning feature selection wrapper technique that is both simple and effective at identifying important features. As a wrapper strategy, it utilizes an external search method to select distinct subsets of the complete predictor set for model evaluation, thus separating the feature selection and model fitting processes. The method is well-suited for small sample sizes and was initially applied to a study of cancer classification using microarrays with fewer than 100 training samples but over a thousand characteristics (Chen & Jeong, 2007). The approach is used in conjunction with random forest to ensure that irrelevant variables are removed (Kuhn & Johnson, 2019; Machine learning Mastery, 2020).

VSURF is another variable selection approach that makes use of random forests. The selection of variables is accomplished in three stages using random forests: thresholding, interpretation, and prediction (Genuer *et al.*, 2019). The first phase of the strategy is to sort the variables in descending order of their importance using a variable importance measure. This is followed by the interpretation stage, which involves the selection of variables, and finally the prediction step, which involves the refinement of the variables from the previous step by removing redundancies in preparation for predictive modelling (Grenuer *et al.* 2015). A random forest predicts an outcome by averaging the outputs of hundreds or thousands of trees (Franklin, 2005). Additionally, the approach takes into consideration of the grouping variable (Y) which is either a class label or numerical response (Genuer *et al*., 2015).

1.4.2.4. Predictive modelling with support vector machine (SVM)

Subsequent to splitting data and selecting key variables as seen on **Figure 1.2**, machine learning tools and multivariate analysis can be used to analyse, classify, visualize data, and build predictive models (**Figure 1.2F**; Conrad & Bonello, 2016). A predictive model should be constantly validated and revised to incorporate changes in the underlying data, meaning that it is not fixed (Ali, 2021).

Machine learning techniques can also be used to develop classification predicting models. Support vector machine (SVM) is one such algorithm. SVM is intended for categorization between two classes or more (James, 2013). A hyperplane is used to categorize data, and the data points closest to the hyperplane are referred to as support vectors (Boser *et al.*, 1992). Cross-validation can be used determine the SVM model's performance. SVM can be useful due to the algorithm's high dimensionality and versatility as a result of a variety of kernel properties (Karatzoglou *et al,* 2006). SVM has recently been employed in the early detection of the rice sheath blight fungus *Rhizoctonia solani* in rice plants (Conrad *et al.* 2020). The tool was also used to develop classification models which could distinguish between healthy and diseased sugar beet plants (Rumpf *et al.,* 2010).

Certain tuning parameters must be considered when developing an SVM model. Imbault and Lebart (2004) suggest that tuning an SVM model is critical for classification. The tuning procedure uses a generalized error estimate to get the ideal parameters that maximize model accuracy.

1.4.2.5. Tuning parameters for predictive modelling

Tuning parameters include kernel, cost, and gamma. A kernel fits a support vector classifier to a higher dimensional space that includes polynomials of degree, assisting in the solution of non-linear problems using linear classifiers (James, 2013). There are different types of kernels: linear, polynomial, Gaussian radial basis, sigmoid and Bassel. The linear kernel is mostly used when there are more features, the polynomial type is less efficient and is less preferred. The radial Gaussian kernel is mostly used in SVM and preferred when there is no prior knowledge about the data. Sigmoid kernel is preferred for neural networks while Gaussian kernel is usually used when there is no prior knowledge about a particular data set (Awasthi, 2020).

The tuning parameter cost specifies the cost of a violation of the margin. When the cost is small, the hyperplane margins become wide with many support vectors on the margin or a violation of the margin; the opposite is true for large cost values, where the margins become narrow with few support vectors on the margin (Ben-Hur & Weston, 2010). The gamma parameter is only used with the radial Gaussian kernel.

A higher value gamma value introduces flexibility to the decision boundary and can lead to overfitting, while small values keep the decision boundary linear (James *et al.*, 2013). All these parameters play a crucial role in how accurately the SVM model classifies the two classes. Important metrics are required to ensure that the SVM model trains the data well.

1.4.3. Predictive model evaluation and validation

Over or underfitting is a major issue in predictive modelling (Domingos, 2012). Both can affect model performance. An overfit model memorizes the entire dataset, even the noise, and fits the model too well, whereas an underfit model cannot represent the training data, and both models can produce low accuracies for the testing set (Ying, 2019). Model evaluation verifies the trained model's accuracy and efficiency for application purposes (**Figure 1.2G-H**). A model that has not been properly validated may underperform and would be unable to adapt to new stress circumstances; it may also overfit, unable to receive and utilise new inputs. This is done for the training model using the test data set in order to validate the model and prevent it from being overfit (Lupoi *et al.,* 2014). Additionally, testing the model against a new, independent dataset can aid in the selection of the ideal model (**Figure 1.2I-K**; Witten *et al.,* 2016)

1.4.3.2. Cross-validation of predictive models

Cross validation (CV) is also used to assess the model's accuracy and errors or overfitting (Foley *et al.,* 1998). It consists of three steps: first, a subset of the original dataset is removed; then the remaining dataset is used to train the model. Finally, the model is validated using the removed dataset (Stone, 1974). There are several cross-validation approaches for FT-IR models, the most frequently used being K-fold CV and leave one out CV (LOOCV; Wong, 2015). Hawkins et al. (2003) define LOOCV as a strategy for fitting a model to n-1 samples of the dataset and then evaluating the model on the basis of the solitary omitted data point. This is time costly for large data sets but does not need to be repeated because the splits are identical. Another shortcoming of the method is that it cannot accurately determine the proportion of examples in each class that should be included in the test set (Hawkins *et al.*, 2003).

The k-fold CV divides a dataset into k folds uniformly, where k is the number of k folds chosen. Ten-fold CV has been frequently utilized by researchers (Bengio & Grandvalet, 2004). When the correct k-fold number is chosen, the technique is less expensive than LOOCV and can yield the lowest predicted cross-validation error rate (Mahmood & Khan, 2009). Cross-validation and a testing data set are typically employed to calibrate the model. However, in some circumstances with a small sample size, the complete data set may be used to train the model, with CV used for validation, without a true testing set (James *et al*, 2013). To achieve an accurate estimation of the error rate, the k-fold CV should be repeated ten times, e.g. the tenfold CV should be repeated ten times, followed by the results being averaged (Witten *et al.*, 2016).

1.4.3.3. Confusion matrices for error rates estimation

A metric accounting for true and false positives is important in predictive modelling (Indata Labs, 2021). Confusion matrices allow for calculation of model accuracy which informs on model performance. Accuracy is equivalent to correct model predictions divided by total model predictions (Machine learning crash course, 2020). The confusion matrix is a classification output metric that produces four classification outputs, the first of which is true positive (TP), which indicates that the predicted positive is true. A false positive (FP) indicates how many times the model forecasted negative values as positive, whereas a true negative (TN) indicates when the actual negative value is equivalent to the predicted negative values. Finally, a false negative (FN) value indicates how frequently the model identified negative values as positives. In this situation, the accuracy of correctly classified observations can be computed by adding TP and TN, then dividing by TP+ TN+ FP+ FN (Witten *et al.*, 2016).

In some circumstances, higher accuracies may be found when the distribution of classes within the dataset is uneven, as a result of the classifier predicting the dataset's predominant negative class (Oded & Lior, 2010). A classifier's discrimination ability can be evaluated using performance measurement methodologies such as receiver operating characteristic (ROC) curves with a range of trade-offs between true positive and false positive error rates. The area under the curve (AUC) is a frequently used metric in ROC curves (Bradley, 1997). The AUC can be used to assist classifiers in establishing a dominance relationship. Sensitivity and precision are performance

indicators in ROC curves that indicate the true positive rate for a binary classification problem, which is the chance of correctly predicting a positive/true case. Specificity is a term that refers to a true negative rate that indicates the chance of making an accurate prediction in the presence of a false/negative case (Raschka, 2014). All these different approaches are important in ensuring the development of good predictive models.

1.4.4. What makes a good predictive model suitable for field use?

A good predictive model should be measurable, repeatable, and comparable (Ge, 2020). The problem that the model is intended to answer should be well defined in order to choose the most appropriate prediction measures and metrics (Indata Labs, 2021). Such a model should enable accurate data validation using the aforementioned model assessment techniques. It is important that the model can tolerate changes in datasets and evaluation on new unseen diverse data sets (Witten *et al.*, 2016). Another critical component that contributes to the accuracy of prediction models is a sufficient amount of data. Utilizing sufficient and diverse data enables the model to learn more effectively and hence improves the model's accuracy (Domingos, 2012). The quality of the data used in predictive modelling is essential to guarantee that predictive models can be deployed in the field, as good data results in better models (Miah, 2017).

All processes and steps described in this section (**Figure 1.2**) are essential in the development of a reproducible, robust and transferable classification model that can be used to differentiate between resistant and susceptible groups using FT-IR. A number of studies discussed in the next section have demonstrated how FT-IR coupled with statistical methods can be used in disease resistance management of forest trees when challenged with various pathogens. Inoculation trials of tree species under study were important in the establishment of symptomatic and non-symptomatic phenotypes before processing plant samples for FT-IR analysis**.** This is essential when supervised learning methods are used for spectral data analysis in order to create labelled datasets.

1.4.5. Tree disease resistance screening with FT-IR

Due to the scarcity of tools for disease management in forest trees, the adoption of methods for selecting resistant phenotypes has gained considerable attention. FT-IR spectroscopy can be utilized to differentiate resistant and susceptible tree genotypes as well as to quantify chemical markers that might be associated with resistant phenotypes (Conrad *et al*., 2014; Villari *et al*., 2018 ; Mukrimin *et al*., 2019). Martin *et al.* (2005) used this technique to assess elm resistance to *Ophiostoma novo-ulmi*, the fungus that causes Dutch elm disease, which manifests as a wilt. The spectra of non-inoculated trees revealed significant differences in the phenotypes of resistant and vulnerable trees. Additionally, the syringyl to guaiacyl (S/G) ratio was found to be greater in resistant elms than in susceptible elms, implying biosynthesis of syringyl monomers following inoculation.

A more recent and better developed study involved coast live oak (CLO), an important evergreen forest tree species that is native to the state of California, in the U.S.A., and Mexico. The trees are susceptible to sudden oak death caused by *Phytophthora ramorum*. Conrad *et al*. (2014) employed two FT-IR instruments and chemometrics to identify resistant CLO in their native stands. The study correctly classified 100% of resistant trees and 97% of susceptible trees in trimmed data sets. Additionally, two spectral areas with critical characteristics for differentiating resistant and susceptible trees were found (1250-1350 cm^{-1}; 1700-1800 cm^{-1}). The technique was then also used to quantify two putative CLO resistance biomarkers against *P. ramorum* from these two spectral areas, namely ellagic acid and an uncharacterized flavonoid FLV2. The predicted concentration was shown to be significantly correlated with the measured one, demonstrating that FT-IR can be utilized to assess phytochemicals associated with resistant phenotypes. The models developed in this study could be used to predict the resistance of non-infected CLO to *P. ramorum*.

In another study, European ash trees were screened for resistance against *Hymenoscyphus fraxineus*- a fungal pathogen causing ash dieback, a disease that has led to devastating losses in Europe. Multivariate analysis coupled with FT-IR was used to develop a predictive model using chemical markers to distinguish between resistant and susceptible phenotypes. Samples of leaves and bark were processed

and analysed using chemometric approaches to build predictive models. Overall model accuracy was an impressive 87%, showing that the tool can classify resistant and susceptible groups, respectively (Villari *et al.*, 2018).

Mukrimin *et al*. (2019) also demonstrated that stem tissues and leaves can be used in predictive models to discriminate between resistant and susceptible Norway spruce against *Heterobasidion annosum*. In this study, FT-IR spectroscopy was used to discriminate between phloem, xylem and needle extracts of source trees. The technique was able to correctly classify 97% of these samples. The classification of samples into non-symptomatic and symptomatic was also performed. It was discovered that SIMCA could not correctly classify phloem samples. Xylem samples were classified into susceptible and resistant with an accuracy of 97%, while needles samples correctly classified 91% of the samples. The technique was also used to predict the concentration of tannins, which are associated with resistant responses in spruce trees between the 1680-1279 cm^{-1} regions. Findings from the study showed that FT-IR spectroscopy was able to differentiate between symptomatic and asymptomatic trees naturally infected with *Heterobasidion* spp. in the field. All these studies highlight the progress of FT-IR spectroscopy in the field of tree resistance phenotyping.

From these studies, it is clear that the foundation of predictive modelling for disease resistance screening using FT-IR spectroscopy has been laid. Coupled with chemometrics, it offers the potential of being used in disease resistance breeding programs and forest trees disease management. The technique shows so much potential in this field that we decided to test whether it can be used to differentiate between resistant and susceptible *Eucalyptus* against *C. austroafricana.*

1.5. Conclusion

Studies covered in this review have outlined the potential of FT-IR spectroscopy as a phenotyping tool. It has also been shown that chemical signatures from forest trees can be used to associate chemical fingerprints with resistant phenotypes. The elevated emergence of pathogens in South African forest plantations has affected productivity in forest trees, including *Eucalyptus* which is a primary source of timber in

the country. Additionally, as far as interactions between *Eucalyptus* and fungal pathogens are involved, there is a lack of defined diagnostic tools for screening resistant and susceptible genotypes.

This calls for efficient and rapid disease resistance screening tools for successful disease management. Artificial inoculation and natural infection have been effective as phenotyping tools, however, they are not efficient as far as time, labour and high throughput are involved. Predictive classification models can be set up using FT-IR that may allow for rapid identification of resistant *Eucalyptus* trees, as well as aid in our understanding of chemical groups that are associated with disease resistance against fungal pathogens including *C. austroafricana*. Beyond this, the technique holds the potential to transform field phenotyping protocols through the use of handheld devices, thus allowing for high throughput and mass disease resistance screening in field.

References

Ahanger, R.A., Bhat, H.A., Bhat, T.A., Ganie, S.A., Lone, A.A., Wani, I.A., Ganai, S.A., Haq, S., Khan, O.A., Junaid, J.M., & Bhat, T.A. (2013). Impact of climate change on plant diseases. *International Journal of Modern Plant & Animal Sciences*, 1(3):105-115.

Ali, R. (2020). Predictive modelling: Types, benefits and algorithms. Netsuite. https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml *~accessed on 20 November 2021*.

Andersen, C.M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, 24(11-12):728-737.

Awasthi, S. (2020). Seven most popular SVM kernels. https://dataaspirant.com/svm-kernels/ *~accessed on 10 November 2021.*

Batish, D.R., Singh, H.P., Kohli, R.K., & Kaur, S. (2008). *Eucalyptus* essential oil as a natural pesticide. *Forest Ecology & Management*, 256(12):2166-2174.

Beenken, L. (2017). *Austropuccinia*: a new genus name for the myrtle rust *Puccinia psidii* placed within the redefined family Sphaerophragmiaceae (Pucciniales). *Phytotaxa*, 297(1):53-61.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. In: Carugo, O. and Eisenhaber, F., Eds., Data mining techniques for the life science. *Methods in Molecular Biology (Methods and Protocols),* 609:223-239.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research,* 5:1089-1105.

Berthomieu, C., & Hienerwadel, R. (2009). Fourier transform infrared (FT-IR) spectroscopy. *Photosynthesis Research,* 101**:**157-170. https://doi.org/10.1007/s11120-009-9439.

Bokobza, L. (1998). Near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 6(1):3-17.

Boland, D.J., Brooker, M.I.H., Chippendale, G.M., Hall, N., Hyland, B.P.M., Johnston, R.D., Kleinig, D.A., McDonald, M.W., & Turner, J.D. (2006). Forest trees of Australia, 5th edition. Collingwood, Melbourne, Australia: CSIRO Publishing.

Bonan, G.B. (2008). Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science*, 320(5882):1444-1449.

Bonello, P., Gordon, T.R., Herms, D.A., Wood, D.L., & Erbilgin, N. (2006). Nature and ecological implications of pathogen-induced systemic resistance in conifers: A novel hypothesis. *Physiological and Molecular Plant Pathology*, 68(4-6):95-104.

Boots, M., & Best, A. (2018). The evolution of constitutive and induced defenses to infectious disease. *Proceedings of Royal Society B*, 285:6-58.

Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifier. In Proceedings of the 5th ACM Workshop on Computational Learning Theory, pages 144-152, Pittsburgh, PA.

Boyd, I.L., Free-Smith, P.H., Gilligan, C.A., & Godfray, H.C.J. (2013). The consequence of tree pests and disease for ecosystem services. *Science*, 342(6160):123577.

Bradley, A.P. (1997). The use of the area under the ROC Curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(6):1145-1159.

Brassier, C.M. (2008). The biosecurity threat to the UK and global environment from international trade in plants. *Plant Pathology* 57(5):792-808.

Brezáni, V., & Šmejkal, K. (2006). Secondary metabolites isolated from the genus *Eucalyptus*. *Chemistry (Weinheim an der Bergstrasse, Germany)*, *1*.

Brooker, M.I.H. (2000). A new classification of the genus of *Eucalyptus* L'Hér. (Myrtaceae). *Australian Systemic Botany*, 13(1):79-148.

Brown, C.D., Vega-Montoto, L., & Wentzell, P.D. (2000). Derivative pre-processing and optimal corrections for baseline drift in multivariate calibration. *Applied Spectroscopy*, 54(7):1055-1068.

Campos, W.G., Faria, A.P., Oliveira, M.G., & Santos, H.L. (2008). Induced responses against herbivory by chemical information transfer between plants. *Brazilian Journal of Plant Physiology*, 20(4):257-266.

Carden, A., & Morris, M.D. (2000). Application of vibrational spectroscopy to the study of mineralized tissues (review). *Journal of Biomedical Optics*, 5(3):259-269.

Centner, V., & Massart, D.L. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry*, 68:3851-3858.

Charkraborty, S., & Datta, S. (2003). How will pathogens adapt to host plant resistance at elevated CO^2 under a changing climate? *New Phytology*, 159:733-742.

Chen, X. & Jeong, J.C. (2007). "Enhanced recursive feature elimination, "*Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pp. 429-435, doi: 10.1109/ICMLA.2007.35

Christy, A.A., Ozaki, Y., & Gregoriou, V.G. (2001). *Modern Fourier transform infrared spectroscopy* (Vol. 2001). Amsterdam: Elsevier.

Conrad, A.O., Li, W., Lee, D.Y., Wang, G.L., Rodriguez-Saona, L., & Bonello, P. (2020). Machine learning-based pre-symptomatic Detection of rice sheath blight using spectral profiles. *Plant Phenomics*, 2020:10.

Conrad, A.O., & Bonello, P. (2016). Application of infrared and Raman spectroscopy for the identification of disease resistant trees. *Frontiers in Plant Science*, 6:1152.

Conrad, A.O., Rodriquez-Saona, E.L., McPherson, B.A., Wood, D.L., & Bonello, P. (2014). Identification of *Quercus agrifolia* (coast live oak) resistant to the invasive pathogen *Phytophthora ramorum* in native stands using Fourier-transform infrared (FT-IR) spectroscopy. *Frontiers in Plant Science*, 5:521.

Conradie, E., Swart, W.J., & Wingfield, M.J. (1990). *Cryphonectria* canker of *Eucalyptus*, an important disease in plantation forestry in South Africa. *South African Forestry Journal*, 152(1):43-49.

Cozzolino, D. (2014). Use of Infrared Spectroscopy for in-field measurements and phenotyping of plant properties: Instrumentation, data analysis and examples. *Applied Spectroscopy Reviews*, 49:564-584.

Crous, P.W., Groenewald, J.Z., Summerell, B.A., Wingfield, B.D., & Wingfield, M.J. (2009). Co-occurring species of *Teratosphaeri*a on *Eucalyptus. Persoonia*, 22:38-48.

Davidson, J. (1993). "Ecological aspects of *Eucalyptus* plantations". *Proceedings of Regional Expert Consultation on Eucalyptus*, 1:35-72. RAPA/FAO, Bangkok, Thailand.

De Hoffman, E., & Stroobant, V. (2007). Mass spectrometry: Principles and applications. 3rd Edition. John Wiley & Sons, West Sussex, England.

Deisentroth, M.P., Faisal, A.A., & Ong, C.S. (2020). Mathematics for machine learning. Cambridge University Press, 1-417.

Denison, N.P., & Kietzka, J. E. (1993). The use and importance of hybrid intensive forestry in South Africa. *South African Forestry Journal*, 165(1):55-60.

Diem, M. (2015). Modern vibrational spectroscopy and micro-spectroscopy. John Wiley & Sons, West Sussex, United Kingdom.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55:78-87.

Duchesne, L. C., Hubbes, M., & Jeng, R. S. (1985). Accumulation of phytoalexins in *Ulmus americana* in response to infection by a non-aggressive and an aggressive strain of *Ophiostoma ulmi*. *Canadian Journal of Botany*, 63:678-680.

Dutta, A. (2017). Fourier transform infrared spectroscopy. *Spectroscopic methods for nanomaterials characterization*, 73-93.

Eldridge, K., Davidson, J., Harwood, C., & Wyk, G.V. (1994). *Eucalypt domestication and breeding*. Clarendon Press.

Ellis, D., & Goodacre, R. (2006). Metabolic fingerprinting in disease diagnosis: Biomedical applications of infrared and Raman spectroscopy. *The Analyst*, 13(8):875-885.

Eyles, A., Bonello, P., Ganley, R., & Mohammed, C. (2009). Induced disease resistance to pests and pathogens in trees. *New Phytologist Tansley Review*, 185(4):893-908.

Fahn, A. (1988). Secretory tissues in vascular plants. *New phytologist*, 108(3):229-257.

Faix, O. (1992). Fourier transform infrared spectroscopy. In: Methods in lignin chemistry. Springer, Berlin, Heidelberg, pp. 83-109.

Food and Agriculture Organization (FAO). (1985). The ecological effects of *Eucalyptus*. FAO Forestry Paper No.59. FAO, Rome, Italy.

Food and Agriculture Organization (FAO). (2006). Global Forest Resources Assessment 2005. Progress towards sustainable forest management. FAO Forestry Paper:147. Rome, Italy.

Febrero-Bande, M., & de la Fuente, M.O. (2012). Statistical computing in functional data analysis: The R package fda. usc. *Journal of statistical Software*, 51(1):1-28.

Fiehn, O. (2002). Metabolomics-the link between genotypes and phenotypes. *Plant Molecular Biology*, 48:155-171.

Foley, W.J., & Moore, B.D. (2005). Plant secondary metabolites and vertebrate herbivores–from physiological regulation to ecosystem function. *Current Opinion in Plant Biology*, 8(4):430-435.

Foley, W.J., McIlwee, A., Lawler, I., Aragones, L., Woolnough, A.P., & Berding, N. (1998). Ecological applications of near infrared reflectance spectroscopy–a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. *Oecologia*, 116(3):293-305.

Food and Agriculture Organization (FAO). (2011). State of the world's forests. Food and Agriculture Organization. Rome, Italy, pp 179.

Food and Agriculture Organization (FAO). (2010). Global forest resources assessment-main report. Food and Agriculture Organization. Rome, Italy, pp 378.

Food and Agriculture Organization (FAO). (1999). State of the world's forests. Food and Agriculture Organization. Rome, Italy.

Food and Agriculture Organization (FAO). (1979). Establishment techniques for forest plantations. FAO Forestry Paper 8. Rome, Italy.

Forestry in South Africa (FSA). (2019). Forestry in South Africa: Introducing commercial forestry. https://www.forestrysouthafrica.co.za/info-graphics/homepage/introducing-commercial-forestry/ *~accessed on 15 November 2021*.

Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83-85.

Friis, I. (1995). Flora of Ethiopia and Eritrea Vol. 2. *Addis Ababa and Uppsala University,* Uppsala.

Frost, C.J., Mescher, M.C., Carlson, J.E., & de Moraes, C.M. (2008). Plant defense priming against herbivores: getting ready for a different battle. *Plant Physiology*, 146(3):818-824.

Galvin-King, P., Haughey, S.A., & Elliott, C.T. (2020). The detection of substitution adulteration of paprika with spent paprika by the application of molecular spectroscopy tools. *Foods*, 9(7):944.

Ge, S. 2020. What makes a good predictive model? TMCNET Feature. https://www.tmcnet.com/topics/articles/2020/10/09/446805-what-makes-good-predictive-model.html ~*accessed on 11 December 2021.*

Genuer, R., Poggi, J.M., Tuleau-Malot, C., & Genuer, M.R. (2019). Package 'vsurf'. *Pattern Recognition Letters*, 31(14):2225-2236.

Genuer, R., Poggi, J.M., & Tuleau-Malot, C. (2015). VSURF: An R package for variable selection using random forests. *The R Journal,* 7(2):19-33.

Genuer, R., Poggi, J.M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters,* 31(14):2225-2236.

Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annual Review of Phytopathology*, 43:205-227.

Glen, M., Alfenas, A.C., Zauza, E.A.V., Wingfield, M.J., & Mohammed, C. (2007). *Puccinia psidii*: A threat to the Australian environment and economy. A review. *Australasian Plant Pathology*, 36:1-16.

Gominho, J., Figueira, J., Rodrigues, J.C., & Pereira, H. (2001). Within-tree variation of heartwood, extractives and wood density in the eucalypt hybrid urograndis (*Eucalyptus grandis* x *E. urophylla*). *Wood and Fiber Science*, 33(1):3-8.

Greyling, I., Wingfield, M.J., Coetzee, M.P., Marincowitz, S. & Roux, J. (2016). The *Eucalyptus* shoot and leaf pathogen *Teratosphaeria destructans* recorded in South Africa. *Southern Forests: A Journal of Forest Science*, 78(2):123-129.

Griffiths, P.R., & de Haseth., J.R. (2007). Fourier Transform infrared spectrometry. Second edition. John Wiley & Sons, Hoboken, New Jersey.

Gryzenhout, M., Myburg, H., Van der Merwe, N.A., Wingfield, B.D. & Wingfield, M.J. (2004). *Chrysoporthe*, a new genus to accommodate *Cryphonectria cubensis*. *Studies in Mycology*, 50:119-142.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46**:**389-422. https://doi.org/10.1023/A:1012487302797

Hansen, E.M. (2008). Alien forest pathogens: *Phytopthora* species are changing world forests. *Boreal Environment Research*, 13:33-41.

Hassan, R., Scholes, R., & Ash, N. (2005). Ecosystems and human well-being: Current state and trends. Nature 443:743-750.

Hawkins, D.M., Basak, S.C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, 43(2):579-586.

Heim, R.H.J., Wright, I.J., Chang, H.C., Carnegie, A.J., Pegg, G.S., Lancaster, E.K., Falster, D.S., & Oldeland, J. (2018). Detecting myrtle rust (*Austropuccinia psidii*) on lemon myrtle trees using spectral signatures and machine learning. *Plant Pathology*, 67(5):1114-1121.

Ho, C.T. (1992). Phenolic compounds in food. In Huang, M.T., Ho, C.T. & Lee, C.Y. (Eds), *Phenolic compounds in food and their effects on health*. ACS Symposium Series*,* 507:2-7.

Hollas, J. M. (2013). *High resolution spectroscopy*. Butterworth-Heinemann.

Honkanen, T., Haukioja, E., & Kitunen, V. (1999). Responses of *Pinus sylvestris* branches to stimulated herbivory are modified by tree sink/source dynamics and by external resources. *Functional Ecology*, 13:126-140. doi:10.1046/j.1365-2435.1999.00296.x.

Hurley, B.P., Garnas, J., Wingfield, M.J., Branco, M., Richardson, D.M., & Slippers, B. (2016). Increasing numbers and intercontinental spread of invasive insects on eucalypts. *Biological Invasions*, 18(4):1-17.

Imbault, F., & Lebart, K. (2004). A stochastic optimization approach for parameter tuning of support vector machines. *Proceeding of the 17th International Conference on Pattern Recognition*, 4:596-600.

Indata Labs. (2021). Predictive performance and their performance evaluation. https://indatalabs.com/blog/predictive-models-performance-evaluation-important~ *accessed on 19 November 2021.*

Inderjit, D.K. (1999). Bioassays for allelopathy: Interactions of soil organic and inorganic constituents. *Principles and practices in plant ecology: Allelochemical interactions. CRC Press LLC, Boca Raton, FL*, 35-44.

Ismail, A.A., van de Voort, F.R., & Sedman, J. (1997). Fourier transform infrared spectroscopy: Principles and applications. In *Techniques and instrumentation in analytical chemistry*, 18:93-139. Elsevier.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer, New York, pp. 18.

James, S.A., & Bell, D.T. (2000). Influence of light availability on leaf structure and growth of two *Eucalyptus globulus* ssp. Globulus provenances. *Tree Physiology*, 20(15):1007-1018.

Jones, J.D., & Dangl, J.L. (2006). The plant immune system. *Nature*, 444:323-329.

Ju, W., Lu, C., Liu, C., Jiang, W., Zhang, Y., & Hong, F. (2020). Rapid identification of atmospheric gaseous pollutants using Fourier-Transform infrared spectroscopy combined with independent component analysis. *Journal of Spectroscopy*, 1-14.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software*, 15(1):1-28.

Kaufmann, G.H., & Galizzi, G.E. (2002). Phase measurement in temporal speckle pattern interferometry: comparison between the phase-shifting and the Fourier transform methods. *Applied Optics*, 41(34):7254-7263.

Kazarian, S.G., & Chan, K.L.A. (2006). Applications of ATR-FTIR spectroscopic imaging to biomedical samples. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 175(7):858-867.

Kirilenko, A.P., & Sedjo, R.A. (2007). Climate change impacts on forestry. *Proceedings of the National Academy of Sciences USA*, 104(50):19697-19702.

Koornneef, A., & Pieterse, C.M. (2008). Cross talk in defense signalling. *Plant Physiology*, 146(3):839-844.

Kosuge, T. (1969). The role of phenolics in host response to infection. *Annual Review of Phytopathology*, 7(1):195-222.

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. https://bookdown.org/max/FES/recursive-feature-elimination.html ~ *accessed on 10 November 2021*.

Kunkel, B.N. & Brooks, D.M. 2002. Cross signalling pathways in pathogen defense. *Current Opinion in Plant Biology*, 5(4):325-331.

Lattanzio, V., Kroon, P.A., Quideau, S., & Treutter, D. (2008). Plant phenolics—secondary metabolites with diverse functions. *Recent Advances in Polyphenol Research*, 1:1-35.

Lee, S., Park, S., & Lee, L. (2014). Optical methods in studies of olfactory system. In: Park T. (eds). Bioelectronic Nose. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-8613-3_11

Lee, S.S. (2000). Noisy replication in skewed binary classification. *Computational Statistics and Data Analysis*, pp.34.

Leger, M.N., & Ryder, A.G. (2006). Comparison of derivative pre-processing and automated polynomial baseline correction method for classification and quantification of narcotics in solid mixtures. *Applied Spectroscopy,* 60(2):182-193. doi:10.1366/000370206776023304

Levine, S.P. Levine, Li-Shi, Y., Strang, C.R. & Hong-Kui, X. (1989). Advantages and disadvantages in the use of Fourier transform infrared (FT-IR) and filter infrared (FIR) spectrometers for monitoring airborne gases and vapors of industrial hygiene concern. *Applied Industrial Hygiene*, 4(7):180-187.

Chalmers, J.M., & Griffiths, P.R. (2010). Vibrational spectroscopy: sampling techniques and fiber-optics probes. *Applications of vibrational spectroscopy in food science. Wiley, Chichester*, 47-88.https://doi.org/10.1002/0470027320.s8934

Luo, J., Ying, K., He, P., & Bai, J. (2005). Properties of Savitzky–Golay digital differentiators. *Digital Signal Processing*, 15(2):122-136.

Lupoi, J.S., Singh, S., Davis, M., Lee, D.J., Shepherd, M., Simmons, B.A., & Henry, R.J. (2014). High-throughput prediction of eucalypt lignin syringyl/guaiacyl content using multivariate analysis: a comparison between mid-infrared, near-infrared, and Raman spectroscopies for model development. *Biotechnology for Biofuels*, 7(1):1-14.

Machine Learning Crash Course. (2020). Google Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy *~ accessed on 15 March 2022.*

Machine Learning Mastery. (2020). https://machinelearningmastery.com/rfe-feature-selection-in-python/ ~ *accessed on 10 November 2021.*

Magaton, A.D.S., Colodette, J.L., Gouvêa, A.D.F.G., Gomide, J.L., Muguet, M.C.S., & Pedrazzi, C.R.I.S.T.I.A.N.E. (2009). *Eucalyptus* wood quality and its impact on kraft pulp production and use. *Tappi Journal*, 8(8):32-39.

Magwanda, R., Zwart, L., Van der Merwe, N.A., Moleleki, L.N., Berger, D.K., Myburg, A.A., & Naidoo, S. (2015). Localization and transcription responses of Chrysoporthe austroafricana in *Eucalyptus grandis* identify putative pathogenicity factors. *Frontiers in Microbiology*, 7:1953.

Mahmood, Z. & Khan, S. (2009). On the use of K-Fold cross-validation to choose cut-off values and assess the performance of predictive models in stepwise regression. *The International Journal of Biostatistics*, 5(1). DOI: 10.2202/1557-4679.1105

Mangwanda, R., Myburg, A.A., & Naidoo, S. (2015). Transcriptome and hormone profiling reveals *Eucalyptus grandis* defense responses against *Chrysoporthe austroafricana*. *BMC Genomics*, 16(1):1-13.

Mangwanda, R., Zwart, L., van der Merwe, N.A., Moleleki, L., Berger, D., Myburg, A., & Naidoo, S. (2016). Localization and transcriptional responses of *Chrysoporthe austroafricana* in *Eucalyptus grandis* identify putative pathogenicity factors. *Frontiers in Microbiology*, 7:1953.

Mark, H. (1992). Data analysis: Multilinear regression and principal component analysis. In Burns, D.A. & E.W. Ciurczak (Eds). *Handbook of Near Infrared analysis* (pp. 129-184). Taylor and Francis, 2001.

Markovich, G., Giniger, R., Levin, M., & Cheshnovsky, O. (1991). Photoelectron spectroscopy of negative ions solvated in clusters. *Zeitschrift für Physik D Atoms, Molecules and Clusters*, 20(1):69-72.

Martin, J.A., Solla, A., Coimbra, M.A., & Gil, L. (2005). Metabolic distinction of *Ulmus minor* xylem tissues after inoculation with *Ophiostoma novo-ulmi*. *Phytochemistry*, 66:2458-2467.

Matkovic, S.R., Valle, G.M., & Briand, L.E. (2005). Quantitative analysis of Ibuprofen in pharmaceutical formulations through FT-IR spectroscopy. *Latin American Applied Research*, 35:189-195.

Mazid, M., Khan, T.A., & Mohammad, F. (2011). Role of secondary metabolites in defense mechanisms of plants. *Biology and Medicine*, 3(2):232-249.

Mbow, C., Van Noordwijk, M., Luedeling, E., Neufeldt, H., Minang, P. A., & Kowero, G. (2014). Agroforestry solutions to address food security and climate change challenges in Africa. *Current Opinion in Environmental Sustainability*, 6:61-67.

Mendel, Z., Protasov, A., Fisher, N., & La Salle, J. (2004). Taxonomy and biology of *Leptocybe invasa* gen. & sp.n. (Hymenoptera: Eulophidae) an invasive gall inducer on *Eucalyptus*). *Australian Journal of Entomology*, 43:101-113.

Miah, E. (2017). Key factors in the successful use of machine learning. https://www.datasciencecentral.com/profiles/blogs/key-factors-in-the-successful-use-of-machine-learning *~accessed on 04 December 2021.*

Mohamed, M.A., Jaafar, J., Ismail, A.F., Othman, M.H.D., & Rahman, M.A. (2017). Fourier transform infrared (FT-IR) spectroscopy. In *Membrane Characterization* (pp. 3-29). Elsevier.

Moraes, L.G.P., Rocha, R.S.F., Menegazzo, L.M., Araújo, E.B.D., Yukimito, K., & Moraes, J.C.S. (2008). Infrared spectroscopy: A tool for determination of the degree of conversion in dental composites. *Journal of Applied Oral Science*, 16:145-149.

Moros, J., Garrigues, S., & de la Guardia, M. (2010). Vibrational spectroscopy provides a green tool for multi- component analysis. *Trends in Analytical Chemistry*, 29(7):578-591.

Movasaghi, Z., Rehman, S., & uh Rehman, I. (2008). Fourier transform infrared spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 43(2):134-179.

Mukrimin, M., Conrad, A.O., Kovalchuk, A., Julkunen-Tiitto, R., Bonello, P., & Asiegbu, F.O. (2019). Fourier-transform infrared (FT-IR) spectroscopy analysis discriminates asymptomatic and symptomatic Norway spruce trees. *Plant Science*, 289:110-247.

Muranty, H., Jorge, V., Bastien, C., Lepoittevin, C., Bouffier, L., & Sanchez, L. (2014). Potential for marker-assisted selection for forest tree breeding: Lessons from 20 years of MAS in crops. *Tree Genetics & Genomes* 10:1491-1510.

Murphy, P.J., Stevens, G. & Lagrange, M.S. (1998). Geological applications of Raman spectroscopy and the use of Raman spectroscopy in the study of Gold speciation in Fluids. *Economic Geology Research Unit Information Circular no,* 321:1-45.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., & Goodstein, D.M. (2014). The genome of *Eucalyptus grandis*. *Nature*, 510(7505):356-362.

Myburg, H., Gryzenhout, M., Wingfield, B.D., & Wingfield, M.J. (2002). □- tubulin and histone H3 gene sequences distinguish *Chryphonectria cubensis* from South Africa, Asia and South America. *Canadian Journal of Botany*, 80:590-596.

Naidoo, R., Ferreira, L., Berger, D.K., Myburg, A.A., & Naidoo, S. (2013). The identification and differential expression of *Eucalyptus grandis* pathogenesis-related genes in response to salicylic acid and methyl jasmonate. *Frontiers in Plant Science*, 4:43.

Neale, D.B., & Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, 2(2):111-122.

O'Reilly-Wapstra, J.M., Potts, B.M., McArthur, C., & Davies, N.W. (2005). Effects of nutrient variability on the genetic-based resistance of *Eucalyptus globulus* to a mammalian herbivore and on plant defensive chemistry. *Oecologia*, 142(4):597-605.

Oded, M., & Lior, R. (2010). Data mining and knowledge discovery handbook. *Chapter 45: Data Mining for Imbalanced Datasets: An Overview*, 875.

Organization For Economic Co-operation and Development. (2014). Consensus Document on the Biology of *Eucalyptus* spp. *Series on Harmonisation of Regulatory Oversight in Biotechnology*, 58:1-81.

Paine, T.D., Steinbauer, M.J., & Lawson, S.A. (2011). Native and exotic pests of *Eucalyptus*: A worldwide perspective. *Annual Review of Entomology*, 56,181-201.

Particle Sciences. (2012). Vibrational Spectroscopy in pharmaceutical development. Technical Brief, 7.

Pereira, D.M., Valentão, P., Pereira, J.A., & Andrade, P.B. (2009). Phenolics: From chemistry to biology. *Molecules*, 14(6):2202-2211.

Perkins, W.D. (1987). Fourier transform infrared spectroscopy. Part II. Advantages of FT-IR. *Journal of Chemical Education*, 64(110):269-271.

Poore, M.E.D., & Fries, C. (1985). The ecological effects of *Eucalyptus*. FAO Forestry paper 59, Rome.

Press, W.H., & Teukolsky, S.A. (1990). Savitzky-Golay smoothing filters. *Computers in Physics*, 4(6):669-672.

Putzig, C.L., Leugers, M.A., McKelvy, M.L., Mitchell, G.E., Nyquist, R.A., Papenfuss, R.R., & Yurga, L. (1994). Infrared spectroscopy. *Analytical Chemistry*, 66:1226-1266.

Raschka, S. (2014). Predictive modelling, supervised machine learning and pattern classification- the big picture. https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html *~accessed on 20 November 2021*.

Rice, E.L. (1984). Allelopathy, 2nd edition. Academic Press, London.

Rinnan, A.A., Nørgaard, L., van der Berg, F., Thygesen, J., Bro, R. & Engelsen, S.B. (2009). Data pre-processing: Infrared Spectroscopy for food quality analysis and control. Elsevier.

Roesnner, U. & Bowne, J. (2018). What is metabolomics all about? *Biotechniques*, 46(5).

Roux, J., Myburg, H., Wingfield, B.D., & Wingfield, M.J. (2003). Biological and phylogenetic analyses suggest that two *Cryphonectria spp*. cause cankers of *Eucalyptus* in Africa. *Plant Diseases*, 87:1329-1332.

Rozefelds, A.C. (1996). *Eucalyptus* phylogeny and history: A brief summary. *TasForests*, 8:15-26.

Rumpf, T., Mahlein, A.K., Steiner, U., Oerke, E.C., Dehne, H.W., & Plümer, L. (2010). Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 74(1):91-99.

Salari, A. & Young, R.E. (1998). Application of attenuated total reflectance FT-IR spectroscopy to the analysis of mixtures of pharmaceutical polymorphs. *International Journal of Pharmaceutics*, 163(1-2):157-166.

Sankaran, S., Mishra, A., Ehsani, R., and Davis, C. (2010). A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture,* 72:1-13.

Santos, S.A., Vilela, C., Freire, C.S., Neto, C.P., & Silvestre, A.J. (2013). Ultra-high performance liquid chromatography coupled to mass spectrometry applied to the identification of valuable phenolic compounds from *Eucalyptus* wood. *Journal of Chromatography B*, 938:65-74.

Santos, S.A., Villaverde, J.J., Freire, C.S., Domingues, M.R.M., Neto, C.P., & Silvestre, A.J. (2012). Phenolic composition and antioxidant activity of *Eucalyptus*

grandis, *E. urograndis* (*E. grandis*× *E. urophylla*) and *E. maidenii* bark extracts. *Industrial Crops and Products*, 39:120-127.

Schmitt, J., & Flemming, H.C. (1998). FT-IR-spectroscopy in microbial and material analysis. *International Biodeterioration & Biodegradation*, 41(1):1-11.

Shimadzu Corporation. (2015). Introduction to Quest single reflection ATR accessory. https://www.shimadzu.com/an/sites/shimadzu.com.an/files/pim/pim_document_file/applications/application_note/10485/jpa215005.pdf~ a*ccessed on 26 September 2021.*

Smith, B.C. (2011). Fundamentals of Fourier transform infrared spectroscopy, 2nd Edition. CRC Press, USA, 7-11.

Sperschneider, J. (2020). Machine learning in plant-pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytologist*, 228(1):35-41.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion). *Journal of the Royal Statistical Society*, *Series B,* 36:111-147.

Sturrock, R.N., Frankel, S.J., Brown, A.V., Hennon, P.E., Kilejunas, J.T., Lewis, K.J., Worall, J.J., & Woods, A.J. (2011). Climate change and forest diseases. *Plant Pathology*, 60(1):133-149.

Sumner, L.W., Mendes, P., & Dixon, R.A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62:817-836.

Taoutaou, A., Socaciu, C., Pamfil, D., Fetea, F., Balazs, E., & Botez, C. (2012). New markers for potato late blight resistance and susceptibility using FT-IR spectroscopy.

Thermofisher Technical Report. (2013). Introduction to Fourier transform infrared spectroscopy.https://tools.thermofisher.com/content/sfs/brochures/BR50555_E_0513M_H_1.pdf ~accessed *on 17 September 2021*.

Turnbull, J.W. (1999). Eucalypt plantations. *New Forests*, 17: 37-52.

Tyner, T., & Francis, J. (2017). ACS Reagent Chemicals (Specifications and Procedures for Reagents and Standard-Grade Reference Materials). *Infrared Spectroscopy,* 1-3.

Van de Voort, F.R. (1992). Fourier transform infrared spectroscopy applied to food analysis. *Food Research International,* 25(5):397-403.

Van Heerden, S.W., Amerson, H.V., Preisig, O. Wingfield, B.D., & Wingfield, M.J. (2005). Relative pathogenicity of *Chryphonectria cubensis* on *Eucalyptus clones* differing in resistance to *C. cubensis*. *Journal of Plant Disease*, 89:659-662.

Villari, C., Dowkiw, A., Enderle, R., Ghasemkhani, M., Kirisits, T., Kjær, E.D., ... & Cleary, M. (2018). Advanced spectroscopy-based phenotyping offers a potential solution to the ash dieback epidemic. *Scientific Reports*, 8(1):1-9.

Vishin, A.P., & Sachin, A.N. (2014). A review on *Eucalyptus globulus*: A divine medicinal herb. *World Journal of Pharmacy and Pharmaceutical Sciences*, 3(6):559-567.

Visser, E.A., Magwanda, R., Becker, J.V.W., Külheim, C., Foley, W.J., Myburg, A.A., & Naidoo, S. (2015). Foliar terpenoid levels and corresponding gene expression are systemically and differentially induced in *Eucalyptus grandis* clonal genotypes in response to *Chrysoporthe austroafricana*. *Plant Pathology*, 64:1320-1325.

Vogel, K.E. (2009). Backcross breeding: In Transgenic Maize. Humana Press, Totowa, NJ, pp 161-169.

Weyer, L.G. (1985). Near-infrared spectroscopy of organic substances. *Applied Spectroscopy Reviews*, 21(1-2):1-43.

Wingfield, M.J., Brockerhoff, E.G., Wingfield, B.D., & Slippers, B. (2015). Planted forest health: the need for a global strategy. *Science*, 349(6250):832-836.

Wingfield, M.J., Roux, J., Govender, P., & Wingfield, B.D. (2001) Plantation disease and pest management in the next century. *Southern African Forestry Journal*, 190:67-71.

Wingfield, M.J. (2003). Daniel McAlpine Memorial Lecture. Increasing threat of diseases to exotic plantation forests in the Southern Hemisphere: lessons from *Chryphonectria* canker. *Australasian Plant Pathology*, 23:133-139.

Wingfield, M.J. Swart, W.J. & Abear, B.J. (1989). First record of *Chryphonectria* canker of *Eucalyptus* in South Africa. *Phytophylactica*, 21:311-313.

Wingfield, M.J., Slippers, B., Hurley, B.P., Coutinho, T.A., Wingfield, B.D., & Roux, J. (2008). Eucalypt pests and disease: Growing threats to plantation productivity. *Southern Forests*, 70(2):139-144.

Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2016). Data mining, fourth edition: practical machine learning tools and techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109-130.

Xu, X., Chen, S., Xu, Z., Yu, Y., Zhang, S. & Dai, R. (2020). Exploring Appropriate Pre-processing Techniques for Hyperspectral Soil Organic Matter Content Estimation in Black Soil Area. *Remote Sensing*, 12(22): 3765. https://doi.org/10.3390/rs12223765

Ying, X. (2019). An overview of overfitting and its solutions. *OP Conf. Series: Journal of Physics: Conference Series* ,1168.

Zhang, A., Sun, H., Wang, P., Han, Y., & Wang, X. (2012). Modern analytical techniques in metabolomics. *The Analyst*, 137(2):293-300.

Zhang, C., Feng, X., Wang, J., Liu, F., He, Y., & Zhou, W. (2017). Mid-infrared spectroscopy combined with chemometrics to detect Sclerotinia stem rot on oilseed rape (*Brassica napus* L.) leaves. *Plant Methods*, 13(1):1-9.

Zhang, S., & Klessig, D.F. (2001). MAPK cascades. *Trends in Plant Science*, 6(11):520-527.

Zipfel, C., & Felix, G. (2005). Plants and animals: A different taste for microbes? *Current Opinion in Plant Biology*, 4(8):353-360.

Zwart, L., Berger, D.K., Moleleki, L.N., Van der Merwe, N.A., Myburg, A.A., & Naidoo, S. (2017). Evidence for salicylic acid signalling and histological changes in the defense responses of *Eucalyptus grandis* to *Chrysoporthe austroafricana*. *Scientific Reports*, 7:45402.

Chapter 2

Predicting resistance of *Eucalyptus* hybrid clones to *Chrysoporthe austroafricana* using FT-IR

2.1. Abstract

The elevated occurrence of pests and pathogens in South African forest tree plantations due to global trade and climate change has been alarming over the past couple of years. This has resulted in serious losses and limitations in the productivity of these plantations. With this, the need for rapid disease resistance screening tools has become pertinent for disease management. Fourier transform infrared (FT-IR) spectroscopy is an approach that can be used to rapidly screen plants for disease resistance when combined with chemometric analysis. This can be done through the generation of chemical fingerprints using FT-IR and thus allowing for the identification of tree phenotypes that are resistant to specific pests and pathogens. The objective of this study was to develop a predictive FT-IR model to discriminate among *C. austroafricana* resistant and susceptible *Eucalyptus* hybrid clones. Stem tissue harvested from *Eucalyptus* hybrid clones was analyzed with FT-IR. The chemical fingerprints from FT-IR were further analyzed with machine learning tools to compare spectra from resistant and susceptible clones before infection as well as after infection. Results showed that FT-IR is able to correctly classify between resistant and susceptible *Eucalyptus* hybrid clones prior to and after inoculation *with C. austroafricana*. The constitutive SVM model correctly classified 74% (repeated CV mean accuracy) of the ramets with the training data while both the induced SVM models' repeated CV mean accuracies were 80% for clone ID grouping and 81% for lesion length grouping. The study shows the potential of FT-IR as a tool for disease resistance screening of *Eucalyptus* hybrid clones. Classification models developed from this study provided good proof of concept on the potential of FT-IR coupled with machine learning for discriminating between resistant and susceptible clones. Further studies will focus on improving the models' accuracies and reproducibility in field.

2.2. Introduction

Eucalyptus trees are an important source of timber and pulpwood in South Africa with *Eucalyptus grandis* being the most extensively planted eucalypt species (Albaugh *et al*., 2013; FSA, 2019). The trees boast good rooting abilities, rapid growth and the ability to adapt to a wide range of environments (FAO, 1979). However, despite these qualities they are susceptible to a wide range of pests and pathogens (Poore & Fries, 1985; Wingfield *et al*., 2015). One particular fungal pathogen known to attack eucalypt trees is *Chrysoporthe austroafricana*. The pathogen causes a stem canker disease which is characterized by stem girdling at the root collar and wilting (Wingfield *et al*.,1989; Conradie *et al*., 1990; Gryzenhout *et al*., 2004).

Hybrids of *Eucalyptus grandis* x *Eucalyptus urophylla* have been used specifically to control the stem canker pathogen (Denison & Kietzka, 1993). The interaction between *Eucalyptus grandis* and *C. austroafricana* has been used as a model pathosystem with a controlled inoculation protocol to explore defense responses in *Eucalyptus* species against fungal pathogens over the years. The emergence of pests and pathogens in forest plantations has increased over the years (Wingfield *et al.,* 2015).

Trees that exhibit disease resistance are important for disease management strategies and traditional breeding (Conrad & Bonello, 2016; Villari *et al*., 2018). Current disease resistance screening methods are still reliant on time-consuming natural infection and artificial inoculation protocols (Neale & Kremer, 2011; Conrad & Bonello, 2016). Molecular markers can be used; however, they demand time for field validation and are both expensive and limiting, particularly in cases where the genetic basis of host resistance in the pathosystem under study is unclear (Taoutaou *et al*., 2012; Muranty *et al*., 2014). Spectroscopic tools such as Fourier transform infrared (FT-IR) have shown a lot of potential in screening trees for disease resistance (Conrad *et al*., 2014).

Fourier transform infrared (FT-IR) spectroscopy, is a form of vibrational spectroscopy commonly employed to investigate the mid infrared region (4000-700cm^{-1}), that provides a method for identification of resistant trees (Christy *et al.*, 2001). It can generate chemical fingerprints, also referred to as metabolic profiles, which represent

an array of different metabolites or chemicals contained within a sample (Sumner *et al*., 2003). The procedure is quick and cost effective, with high sensitivity (Martin *et al*., 2005). Due to the role of specialized metabolites in pest and pathogen defense, chemical fingerprinting may be useful for identifying chemical signatures associated with disease resistance and can therefore be used to generate predictive models for resistance, and also save time from artificial inoculations (Conrad & Bonello, 2016). Coupled with machine learning, spectral data from FT-IR analysis can be used to develop predictive models.

Machine learning is a branch of artificial intelligence that enables the creation of classification and predictive models from large and complex datasets (Singh *et al*., 2016). The supervised machine learning tool, support vector machine (SVM) uses a hyperplane to separate two classes for categorization purposes (Boser *et al*., 1992). A support vector classifier can be fit to a higher dimensional space using a kernel to distinguish between healthy and diseased plants during the early stages of infection, for example (Rumpf *et al*., 2010). SVM has recently been used in olive trees to classify between asymptomatic and symptomatic trees infected with the dangerous plant bacteria, *Xylella fastidiosa* (Zarco-Tejada *et al*., 2018). SVM can be coupled with spectroscopic tools such as FT-IR to develop classification models (Li *et al*., 2013; Morais *et al*., 2017).

Previous research has demonstrated that spectroscopic techniques like near infrared (NIR) and FT-IR spectroscopy can be coupled with chemometrics and machine learning tools to build classification models that distinguish between tree phenotypes against specific fungal pathogens as well as early disease diagnosis. Conrad *et al.* (2020) used chemical fingerprints from NIR coupled with SVM for early detection of rice sheath blight disease in rice plants. Ash dieback, a terrible disease that has resulted in significant ash tree losses in Europe, is caused by the fungus *Hymenoscyphus fraxineus*. Villari *et al*. (2018) tested European ash trees for resistance against the pathogen. Without conducting an inoculation trial, a prediction model using chemical fingerprints to discriminate between resistant and susceptible ash tree phenotypes was developed using multivariate analysis and FT-IR (Villari *et al*., 2018). There's a lack of cost effective, accurate and rapid diagnostic tools for screening resistant and susceptible phenotypes within *Eucalyptus* plantations and

other trees species. Therefore, we hypothesized that FT-IR spectral data from *E. grandis* hybrid clones can be used to generate predictive models that differentiate between resistant and susceptible clones against *C. austroafricana*.

To test the hypothesis, FT-IR spectra was analyzed with machine learning tools to generate predictive models that can distinguish between resistant and susceptible *Eucalyptus* hybrid clones against *C. austroafricana*. This was achieved by setting up an inoculation trial of *E. grandis* x *E. urophylla* hybrid clones with varying resistance to *C. austroafricana*. Following that, FT-IR spectral data from inoculation trial stem tissue, were used to develop predictive models capable of distinguishing between resistant and susceptible hybrid clone phenotypes using machine learning tools.

2.3. Materials and methods

2.3.1. Study site, plant material & inoculation

The study was conducted at the University of Pretoria Innovation Africa campus (Koedoespoort 456-Jr, Pretoria, 25°44'49.9"S 28°15'31.6"E). The inoculation trial included ramets of 39 *Eucalyptus* hybrid clones, *E. urophylla* and *E. grandis*, together with their parental clones (GU210, CG096) and *E. grandis* clones ZG14 and TAG5 (Mondi, South Africa**, Figure 2.1**). Three treatments were applied: inoculated (induced), mock-inoculated, and control/non-inoculated (constitutive). Each treatment included 43 clones, with eight ramets per clone for inoculation, four for mock-inoculation, and three for control/non-inoculation. Each ramet was considered as an individual sample during the experiment and therefore a biological replicate. During FT-IR analysis each ramet also had two technical replicates.

The CMW2113 *C. austroafricana* isolate (Forestry & Agricultural Biotechnology Institute, South Africa; Roux *et al*., 2003) was cultured for seven days at 28°C on 2% MEA agar and then used to inoculate trees with a stem diameter of at least 1 cm. To inoculate the trees, the bark of the stem was removed 20 cm above the root collar with a 5 mm cork borer. The wound was filled with a 5 mm agar plug containing mycelia, followed by the replacement of the removed bark onto the wound with agar plug, and the wound site was sealed with parafilm. For mock-inoculation, identical steps as described above were employed with the use of a sterile agar plug without mycelia

(Naidoo *et al*., 2013). Control (non-inoculated) trees were kept in the greenhouse and left as is, with some distance from inoculated trees to avoid cross-contamination.

2.3.2. Lesion measurements and harvesting of stem material

Disease progression was examined 28 days post inoculation, where the presence or absence of lesions on the stems was determined (**Figure 2.2**). This was accomplished by removing the parafilm surrounding the wound site and scraping the bark with a scalpel to expose the lesions. Additionally, the lesion lengths were measured using a tape measure. Finally, stem tissue (4 cm in length) from the inoculation site of the trees was harvested, frozen with liquid nitrogen, and stored at -80°C for future processing.

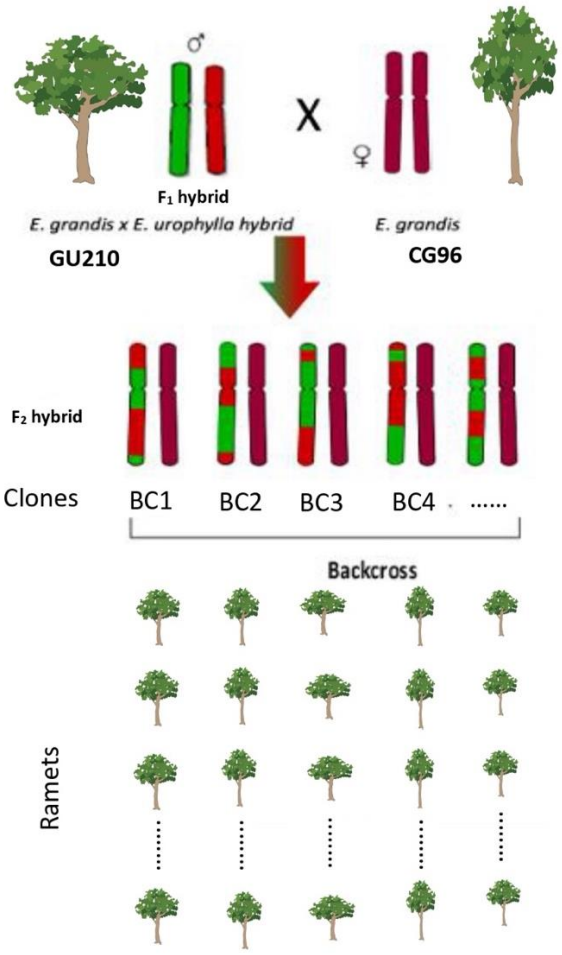

Figure 2.1: ***E. grandis* x *E. urophylla* (GU) backcross population**. Second generation (F_2) backcross clones and their ramets from F_1 hybrid and *E. grandis* recurrent hybridization. *E. grandis* x *E. urophylla* (GU210) progeny crossed back with *E. grandis* (CG96). The F_2 individuals, termed clones, are labelled as BC1, BC2, BC3

etc. Ramets are individual members of a clone produced through vegetative propagation and thus carry the same genotype. A set of ramets were inoculated in this study and labelled as inoculated (I) I1-I8 while some were not inoculated and labelled as S1-S3.

2.3.3. Determining clone phenotypes

The 43 clones were classified into three categories based on statistical analysis of lesion measurements from ramets: resistant, intermediate, and susceptible (**Figure 2.3**). Eight inoculated and three non-inoculated ramets from ten resistant (N= 110) and ten susceptible (N= 110) clones were selected. Additionally, the parental clones (GU210 and CG96), the moderately resistant TAG5 and the highly susceptible ZG14 were selected (N = 44). This yielded a total of 264 samples for FT-IR analysis (**Table 2.1**).

2.3.4. Stem tissue preparation

During tissue preparation, stem samples previously stored at -80°C were retained in a liquid nitrogen cold box before being placed in an IKA analytical electrical grinder (IKAWerke, Staufen, Germany) pre-cooled with liquid nitrogen for primary grinding of stem tissue. The roughly ground stem tissue samples were then transferred to a mortar immersed in liquid nitrogen and ground into a fine powder in a mortar and a pestle. A weight of approximately 200 ± 1 mg of finely ground tissue was transferred to individual 2 mL microcentrifuge tubes and kept at -80°C until extraction.

2.3.5. Phenolic compounds extraction and purification

The extraction of phenolic compounds was carried out in accordance with the methodology described by Villari *et al*. (2018). A solution of acetone containing 700 µL of 70% HPLC grade acetone (Merck) and 30% Milli-Q water was added to each microcentrifuge tube containing finely ground powder, followed by vortexing and sonication with sweep setting for 30 minutes at room temperature. Thereafter, the tubes were centrifuged (Sigma, Lasec) at 16 000 rcf for 8 minutes at room temperature. The supernatants were transferred to new 2 mL chloroform-resistant microcentrifuge tubes and centrifuged for 2 minutes at 10°C. Working under a fume hood, 800 µL of chloroform was added to the tubes with the supernatant, which were

then inverted and vortexed for 15 seconds to mix the chloroform with the extract. Following that, the tubes were placed on ice, and the top aqueous layer containing phenolic compounds was collected and transferred to new 1.5 mL tubes and stored at -20°C until purification.

To begin the purification procedure, samples were thawed for approximately 30 minutes on ice. At room temperature, remnants of chloroform were evaporated from the samples using a SpeedVac (Genevac-miVac, United Scientific) on organic mode. A single syringe volume (1 mL) of HPLC grade methanol was used to activate a C18 column (Waters Plus Short 360 mg/0.7 mL). After equilibrating the column with Milli-Q water, the sample was gradually pushed through the column. Phenolic extracts cohered to the column and formed a yellow/brown coloured band during this procedure.

The extract was then washed twice with water and dried with a twofold volume of air. Methanol was added to the sample in a single volume (1 mL). Pressure was slowly applied to the syringe plunger to move the coloured band down. The first drops were discarded, and only yellow or brown drops were collected in a new 1.5 mL tube. A volume of 200 μL of the sample was dried down using a SpeedVac to evaporate the methanol and then shipped on dry ice to the Department of Pathology at The Ohio State University, Columbus, OH, USA.

2.3.6. FT-IR spectral collection from phenolic extracts

The extracts were concentrated tenfold by adding 20 μL of methanol. After vigorously shaking the tubes for 10 seconds, they were left to stand at room temperature for 15-20 minutes. Samples were then spun and sonicated for 5 minutes, followed by a final round of shaking and spinning before being stored at -20°C. To obtain FT-IR spectra of phenolic extracts, a volume of 2 μL was analysed using a single-bounce zinc selenide attenuated total reflectance (ATR) accessory on a portable Cary 630 FT-IR spectrometer (Agilent Technology; Conrad *et al*, 2014). FT-IR spectral data was acquired in the range 4000-700 cm^{-1}, with a minimum of two technical replicates (reads) for each sample extract. The spectral resolution of the equipment was set at 4 cm^{-1} with 64 scans co-added for each sample.

Table 2.1: Sample material for FT-IR analysis. Twenty *Eucalyptus* backcross clones selected based on lesion length from resistant and susceptible extreme groups. Each clone's three non-inoculated ramets are labelled S1, S2, S3, and all eight inoculated clones' ramets are labelled as I1-I8.

	Non- inoculated	**Inoculated**
Clone ID	**Ramet ID**	
BC010 (resistant) BC090 (susceptible) BC118 (resistant) BC016 (resistant) BC024 (resistant) BC030 (susceptible) BC036 (resistant) BC044 (resistant) BC058 (susceptible) BC105 (resistant) BC113 (susceptible) BC127 (susceptible) BC134 (susceptible) BC148 (susceptible) BC158 (susceptible) BC181 (resistant) BC230 (susceptible) BC234 (resistant) BC250 (susceptible) BC252 (susceptible) CG096 (Parental) GU210 (Parental) TAG 5 (Control-moderately resistant) ZG14 (Control-highly susceptible)	S1, S2, S3	I1, I2, I3, I4, I5, I6, I7,I8

2.3.7. FT-IR spectral collection from phenolic extracts

The extracts were concentrated tenfold by adding 20 μL of methanol. After vigorously shaking the tubes for 10 seconds, they were left to stand at room temperature for 15-20 minutes. Samples were then spun and sonicated for 5 minutes, followed by a final round of shaking and spinning before being stored at -20°C. To obtain FT-IR spectra of phenolic extracts, a volume of 2 μL was analysed using a single-bounce zinc selenide attenuated total reflectance (ATR) accessory on a portable Cary 630 FT-IR spectrometer (Agilent Technology; Conrad *et al*, 2014). FT-IR spectral data was acquired in the range 4000-700 cm^{-1}, with a minimum of two technical replicates (reads) for each sample extract. The spectral resolution of the equipment was set at 4 cm^{-1} with 64 scans co-added for each sample.

2.3.8. Spectral data pre-processing and analysis

The raw FT-IR spectral data was pre-processed and analysed according to the approach described by Conrad *et al*. (2020) which is detailed on the workflow shown in **Figure S1**. The data was first exported to R statistical software version 4.1.0 (R Core Team, 2021). Samples with poor quality or methanol peaks were removed based on visual inspection, and outliers were detected and removed using the approach by Heim *et al*. (2018). Functional data analysis and utilities ("fda", "fda.usc") packages with depth measures (dfunct = depth.FM, nb= 10, smo = 0.1, trim= 0.06) were used to detect and trim outliers from the dataset (Febrero-Bande & de la Fuente, 2012; Ramsay *et al.,* 2021). Following outlier removal, the dataset contained 495 observations and 1782 variables (wavenumbers). The data was then subset into two categories: constitutive (non-inoculated) and induced (inoculated). After averaging technical replicates for each biological replicate (clone ID), the phenotypic data was added to the spectral dataset using the "dplyr" package (Wickham *et al*., 2021). The constitutive dataset consisted of 69 observations and 183 for the induced dataset. Both datasets were processed and analysed independently.

2.3.9. Data splitting and transformation

For each of the two datasets, the “caret” package (Kuhn, 2021) was used to randomly split the dataset into training and testing datasets. This results in balanced data splits within each class through random sampling, while retaining the data's overall class

distribution (Kuhn, 2019). Due to the sample size of the constitutive data, the complete dataset was used for the training set. The induced dataset had two grouping variables, clone ID and lesion length, both these were split into training (70%) and testing set (30%; **Table 2.2**). To construct a new dataset with lesion length as a grouping variable, lesion lengths less than 3.0 cm were used to identify resistant trees and marked as quartile 1(q1) whereas lesion lengths equal to or greater than 7.0 cm were used to identify susceptible trees and denoted as "q4".

Table 2.2: Sample sizes for each model class, split into training and testing datasets. The induced datasets, grouped by clone ID and lesion length, were split into a training (70%) and a testing (30%) dataset. For constitutive data, the entire dataset was used to train the model, and data was grouped by clone ID.

Dataset	Number of samples			
	Training		Testing	
	Resistant	Susceptible	Resistant	Susceptible
Constitutive	30	39	—	—
Induced- Clone ID	58	71	24	30
Induced- Lesion length	30	35	12	15

The new dataset merged data from the described “q1” and “q4”. The clone ID dataset consisted of clone IDs of the ramets analysed with FT-IR. The datasets were transformed using second derivative transformation with the “mdatools” package (Kucheryavskiy, 2020; **Figure S2 & S3**) (constitutive width filter= 19; induced width filter for clone ID= 17; induced width filter for lesion length= 5; porder= 2; dorder= 2).

2.3.10. Selection of key features

Two variable selection methods were used to identify key variables that included useful information for predictive modelling. The "caret" package was used to select predictors using recursive feature elimination (RFE), a wrapper technique that makes use of the random forest selection function (Kuhn, 2021). Variable selection utilizing random forests ("VSURF" package) was also employed as a secondary method to select key wavenumbers important for classification (Genuer *et al.,* 2019). VSURF

returns two sets of variables based on the interpretation and prediction steps. Variables from the prediction step are more refined and free of redundancy than those from the interpretation step, and are thus more ideal for predictive modelling (Geneur *et al*., 2015).

Support vector machine (SVM), a supervised machine learning approach, was used to create classification models using the "e1071" package (Meyer *et al*., 2021) and the datasets with VSURF and RFE selected variables, respectively. The optimal tuning settings were determined using a radial kernel and a 10-fold cross validation (CV; **Table S1**). The "MLmetrics" package was used to evaluate the model's performance on the training and testing sets, with the exception of the constitutive model, which was trained with 100% of the data (Yan, 2016). To determine the model's accuracy, training sets were cross-validated tenfold. For an accurate estimation of the error rate, 100 runs of a 10-fold CV were performed, followed by a calculation of the repetitions' mean accuracy and standard deviation. Additionally, model performance measurements were made using receiver operating characteristic (ROC) curves and the area under the curve (AUC) metric to evaluate the binary classifier's performance and the trade-offs between true positive and false positive error rates (package: "ROCR"; Singh *et al*., 2016; **Figure S4**).

2.3.11. Sparse partial least square discriminant analysis (sPLS-DA)

The three datasets were analysed for important spectral features and classification using sPLS-DA with the "mixOmics" package (Rohart *et al.*, 2017). Firstly, the optimal number of variables for each component were selected by performing a 5-fold CV repeated 50 times. The optimal number of components and variables were chosen for the final sPLS-DA model. The model's performance was validated using a 5-fold CV that was repeated 50 times with the error rate calculated. Finally, with the exclusion of the constitutive dataset, the testing sets were used to further assess the model's accuracy using the balanced error rate (BER) of prediction on the testing sets.

2.4. Results

2.4.1. Disease progression in *Eucalyptus* hybrid clones

At 28 days post-inoculation, ramets from the backcross clones showed varying levels of resistance to *C. austroafricana* (**Figure 2.2**). Susceptible ramets displayed wilting and high death rates in comparison to resistant ramets which showed minimal to no lesions around the inoculation site. Mock-inoculated ramets showed no evidence of infection and were included in the inoculation study to assess the response to wounding but were excluded from the FT-IR analysis. During the experiment, plants that were not inoculated remained healthy. Lesion measurements from these ramets were divided into three groups: resistant (<4.2 cm), intermediate (4.25-5.5 cm), and susceptible (> 5.5 cm) as seen in **Figure 2.3**.

Figure 2.2: Canker lesions on stems of *Eucalyptus* ramets inoculated with *C. austroafricana.*

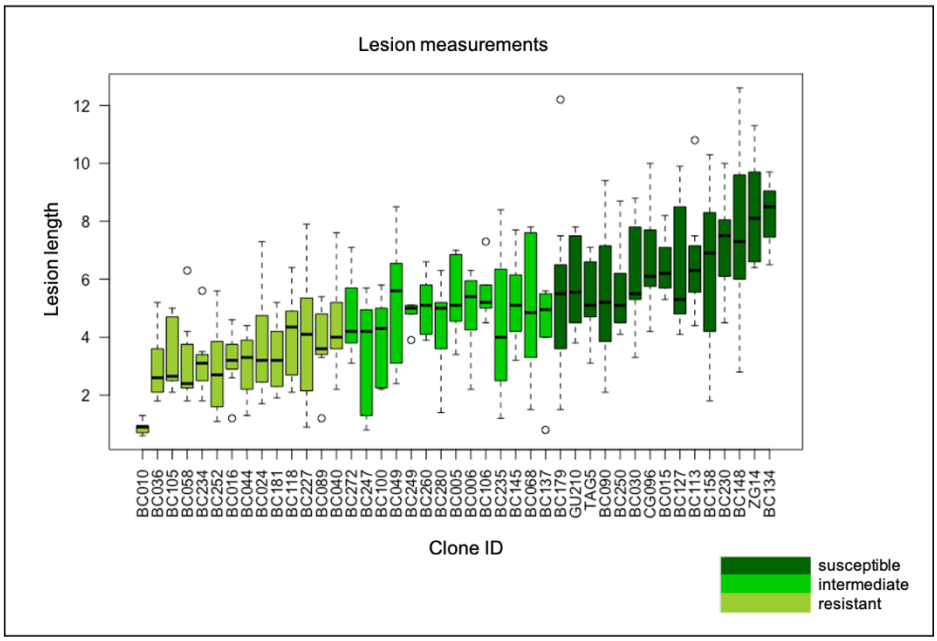

Figure 2.3: Classification of *Eucalyptus* hybrid clones, parental lines and *E. grandis* control clones into phenotypic groups using median lesion length measurements. Solid bars show the interquartile range of lesion length measurements, broken lines show the maximum and minimum values.

2.4.2. FT-IR spectral analysis

The average raw FT-IR spectra for *Eucalyptus* hybrid clones measured in the 4000-700 cm^{-1} spectral region are shown in **Figure 2.4.** There were differences in spectra and absorbance between non-inoculated and inoculated trees. Spectra from non-inoculated trees revealed differences in absorbances between resistant and susceptible trees in the spectral region between 900 and 1591 cm^{-1}. Further spectral variations between the two treatments were seen in the 3520-4000 cm^{-1} spectral region (**Figure 2.4A & 2.4B**). A PCA plot was also done to visualize the variation with the groups (**Figure S4**). The spectral range 1774-889 cm^{-1}, which is a subset of the mid-IR spectral region, was used as the fingerprinting region as it had the most informative chemical signatures and was thus used for all spectral analyses.

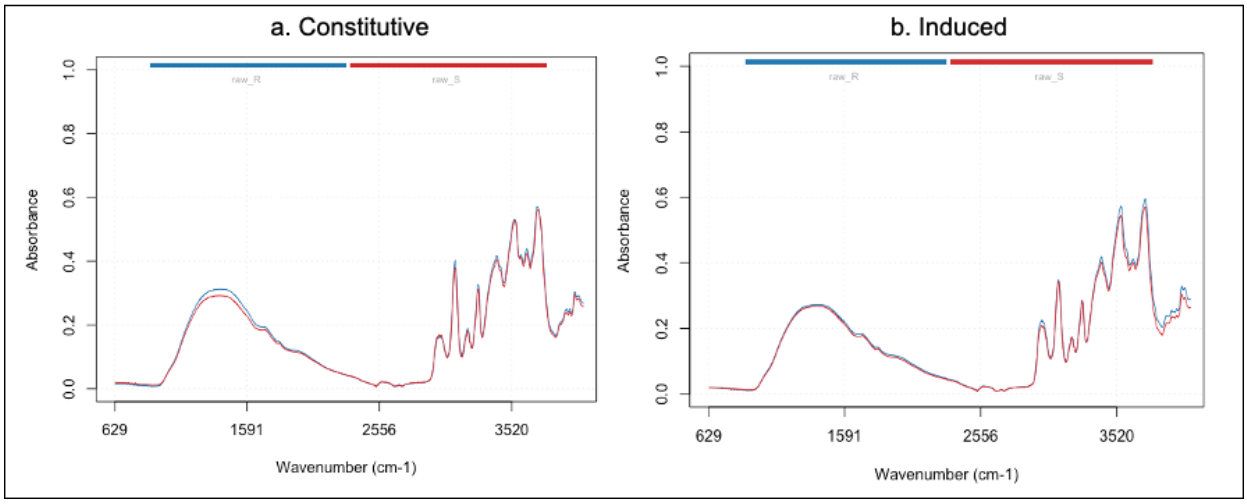

Figure 2.4: Average raw FT-IR spectra of *Eucalyptus* hybrid clones from 4000- 700 cm^{-1}. (a) Constitutive (non-inoculated) FT-IR spectra of resistant (blue) and susceptible (red) hybrid clones. (b) Induced (inoculated) FT-IR spectra of resistant (blue) and susceptible (red) hybrid clones.

2.4.3. Important feature selection and classification models

RFE and VSURF were used to determine essential classification variables for the three data groups. VSURF produces two sets of variables: interpretation and prediction. The prediction variable sets were used to generate the models, which included from three to seven selected variables, with none of them being shared across the three models (**Table 2.3**). The same was observed for RFE, with the top five selected variables shown in **Table 2.4**, with none of the selected wavenumbers common throughout the models. Three wavenumbers were chosen for the constitutive SVM model by the VSURF prediction step: 978, 1385, and 1346 cm^{-1}. Among the three datasets, the induced SVM model based on clone ID grouping had the most wavenumbers selected (1016, 1370, 1368, 911, 926, 1465, 1524 cm^{-1}). The four wavenumbers used for lesion length-based grouping were 1349, 947, 1754, and 964 cm^{-1} (**Table 2.3**). In RFE, the top five wavenumbers selected for the constitutive SVM model ranged from 978 to 1014 cm^{-1}. The top five variables chosen for induced samples, grouped by clone ID, varied from 1370-1016 cm^{-1}. Lastly, the wavenumbers chosen for the induced lesion length-based grouping ranged from 919 to 1347 cm^{-1} (**Table 2.4**).

Table 2.3: VSURF-selected wavenumbers. Key variables for modelling selected using variable selection using random forests (VSURF). Key variables selected at interpretation and prediction steps, variables from prediction step used for classification modelling with support vector machine (SVM).

	Selected wavenumbers (cm^{-1})	
Model	Interpretation	Prediction
Constitutive	978, 1385, 1346	978, 1385, 1346
Induced- clone ID	1016, 1370, 1366, 1139, 1368, 911, 913, 1107, 1014, 1325, 1364, 926, 1465, 1524	1016, 1370, 1368, 911, 926, 1465, 1524
Induced-lesion length	1349, 1347,1360,1754,964	1349, 947,1754,964

Table 2.4: Recursive feature elimination (RFE) top five (5) selected variables for predictive modeling with SVM.

Model	**Selected wavenumbers (cm^{-1})**
Constitutive	978, 980, 977, 986, 1014
Induced- clone ID	1370, 1364, 1368, 1016, 1229
Induced-lesion length	1347, 1349, 1366, 964, 919

2.4.4. Constitutive classification model

In comparison to RFE selected variables, the predictive constitutive model based on VSURF–selected variables was more accurate (**Table 2.5**). The constitutive SVM model based on VSURF–selected variables correctly classified 74% of the training set data (repeated CV mean accuracy) in comparison to the induced clone-ID and lesion length grouped models whose accuracies were 89 and 90% respectively (**Table 2.5; Table 2.6**). The model's ability to correctly classify phenotypic categories was also tested. According to VSURF, the confusion matrix derived from the model's training non-CV accuracy demonstrated that the model could accurately classify 100% of resistant samples and 95% of susceptible samples (**Table 2.7**).

Table 2.5: SVM classification performance based on constitutive FT-IR spectral analysis.

Variable selection method	Training accuracy (%)	Mean CV accuracy (%)	Repeated mean CV □□SD accuracy (%)	Training precision (%)
VSURF	97.1	76.8	74.4 □ 2.6	100.0
RFE	89.9	68.1	70.3 □ 3.0	89.7

2.4.5. Induced classification models

The SVM model based on VSURF selected variables (repeated mean CV accuracy) correctly classified 81% of samples in the training set when using clone ID as a grouping variable and 80% for lesion length-based groupings (**Table 2.6**). The testing accuracies for clone ID grouping based on VSURF were 61% and 59% for lesion length grouping respectively (**Table 2.6**). Additionally, the SVM models of the two groupings demonstrated the ability to classify samples correctly into their corresponding phenotypic groups (**Table 2.7**). The training dataset for the SVM model for clone ID-based grouping correctly classified 89% of resistant and 87% of susceptible samples. In the same model, the testing data correctly classified 56% of samples as resistant and 66% as susceptible. The SVM model for lesion length-based grouping correctly classified 90% of the samples as resistant and 94% as susceptible in the training dataset. In the testing dataset, 56% of the samples were correctly classified as resistant, whereas 65% were correctly classified as susceptible (**Table 2.7**). ROC curves were used as a metric to analyse the performance of the models using training and testing sets (**Figure S5).**

2.4.6. Sparse partial least square discriminant analysis models (sPLS-DA)

SVM classification models using VSURF selected variables were confirmed by sPLS-DA (**Table 2.8**). The results from this analysis showed that 57% samples were correctly classified in the constitutive training data. Across the three datasets, the clone ID-based grouping achieved high classification accuracies and low BER values, with 71% correctly classified in the training set and 61% in the testing set. The training

dataset for lesion length-based groupings correctly classified 60% of samples in the training set and 53% of samples in the testing set.

Table 2.6: SVM classification performance based on analysis of induced FT-IR spectra.

Grouping variable	Variable selection method	Training accuracy (%)	Mean CV accuracy (%)	Repeated mean CV ± SD accuracy (%)	Testing accuracy (%)	Training precision (%)	Testing precision (%)
Clone ID	VSURF	87.6	82.2	81.2± 1.5	61.1	88.9	56.0
	RFE	76.0	65.9	65.0 ±1.7	50.0	84.6	42.5
Lesion length	VSURF	92.3	80.0	82.9 ±1.9	59.3	90.3	55.5
	RFE	96.9	81.5	80.1 ±2.2	63.0	96.7	58.3

Table 2.7: Proportions of sample phenotypes (resistant and susceptible) correctly classified by SVM for three treatment sets based on VSURF selected variables.

Model	**Dataset**	**Correctly classified as resistant (%)**	**Correctly classified as susceptible (%)**
Constitutive	Training	100	95.1
	Testing	—	—
Clone ID	Training	89.0	87.0
	Testing	56.0	66.0
Lesion length	Training	90.3	94.1
	Testing	56.0	65.0

Table 2.8: sPLS-DA model performance for constitutive and induced training and testing datasets.

Model	Number of components	Training BER	Testing BER	Proportion correctly classified (training)	Proportion correctly classified (testing)
Constitutive	4	0.434	—	0.566	—
Clone ID	4	0.295	0.395	0.705	0.605
Lesion length	3	0.401	0.467	0.599	0.533

2.5. Discussion

Chemical fingerprinting of *Eucalyptus* hybrid clones was accomplished under greenhouse conditions, through the inoculation of hybrid clones with *C. austroafricana* using a standard controlled inoculation protocol (Roux *et al*., 2013). In contrast to previous studies by Mukrimin *et al*. (2019) and Conrad *et al*. (2014), the inoculation trial was conducted using ramets of the clones. Post inoculation, ramets displayed variability in lesion lengths. This random variation is not uncommon, and has been observed in other studies. This could be due to environmental factors in the greenhouse (Pillbeam *et al*., 2011). Ramets can also be used for field application. This might result in uniform growth of the ramets as well as improved rooting abilities which are mostly limited in greenhouse conditions (Muchovej *et al*., 2008).

Phenolics were extracted from samples and analysed using FT-IR to generate chemical fingerprints, since phenolics are known to play a key role in plant defense (Pichersky & Lewinsohn, 2011). FT-IR has also been effectively employed for phenotyping in different pathosystems, indicating its use as a tool for screening resistant phenotypes that is both faster and more convenient than inoculation trials and natural infection (Martin *et al*., 2005; Conrad *et al*., 2014; Villari *et al*., 2018; Mukrimin *et al*., 2019; Conrad *et al*., 2020). Coupled with machine learning, chemical fingerprints analysis enabled the classification of resistant and susceptible stem samples with varying levels of accuracy depending on the model.

Prior to developing the models, spectral data was cleaned to exclude samples with low absorbance values and methanol peaks which can interfere with model development (Conrad *et al*., 2014). The data was then smoothed to resolve spectral peaks and minimize noise (Leger & Ryder, 2006). Classification models were developed using the SVM algorithm and the performance was assessed using 10-fold CV repeated 100 times and testing sets. Due to the limited sample size (N=69) of the constitutive samples, the model was trained using 100% of the samples. The model was then validated using a 10-fold CV that was repeated 10 times, as recommended by the literature in circumstances where the sample size is small (James *et al*., 2013; Lupoi *et al*., 2014; Witten *et al*., 2016). This is vital as it verifies the model's accuracy and its efficiency for application in large scale settings (Lupoi *et al*., 2014).

An FT-IR spectrum is divided into four regions: single bond (4000-2500 cm^{-1}) region, triple bond (2500-2000 cm^{-1}) region, double bond (2000-1500 cm^{-1}) region, and fingerprint region (1500-600 cm^{-1}; Nandiyanto *et al*., 2019). The fingerprint region is unique to each compound and cannot be imitated. This can help with the identification of unique molecules that make up the compounds present in resistant trees that may not absorb in these areas in susceptible trees. Phenolics are known for their role in defence responses in many plants including *Eucalyptus* species (Batish *et al*., 2008; Lattanzio *et al.*, 2008). Gallic acid, ellagic acid-rhamnoside, epicatechin, and gallic acid were discovered at higher concentrations in *E. grandis* and *E. urophylla* bark extracts (Santos *et al.,* 2012; Santos *et al*., 2013). These are known for their antifungal, anti-inflammatory properties as well as their roles in *Eucalyptus* defence responses and may be useful for chemotyping (Brezáni & Šmejkal, 2006; Conrad & Bonello, 2016).

Spectral differences were observed between inoculated and non-inoculated samples in the 900-1591 cm^{-1} and 3520-4000 cm^{-1} spectral region. Differences were also visible between resistant and susceptible phenotypes in the spectral regions mentioned above. The visible differences in constitutive and induced spectra show that FT-IR is effective in detecting chemical composition differences in inoculated and non-inoculated trees (Conrad *et al*., 2014). Several wavenumbers were selected from VSURF and RFE for the constitutive dataset based on non-inoculated samples and the two induced grouping variables from the inoculation trial.

There were some wavenumbers shared between VSURF and RFE in the three models (constitutive, clone ID and lesion length). This demonstrated that the two techniques are effective in selecting key variables for modelling. And that they may have similarities as they both use random forests to select key variables (Franklin, 2005; Kuhn & Johnson, 2019). RFE selected wavenumbers (978-1014 cm^{-1}) could also be associated with carbohydrate stretching (C□O) along with the VSURF selected wavenumber selected from the same model (978 cm^{-1}) and CH_2 stretching (1346, 1385 cm^{-1}) wherein proteins, fatty acids, DNA, RNA and phosphorus molecules are found (Taoutaou *et al*., 2012). Additionally, carbohydrate detection is unlikely because the protocol used for sample extraction is specific for phenolics (Villari *et al*., 2018).

Wavenumbers selected by VSURF for the clone ID grouping were associated with CH wagging vibrations (911, 926,1016 cm^{-1}), carbonyl (C□O; 1370,1368 cm^{-1}) and benzene stretching (C□C; 1465, 1524 cm^{-1}; Larkin, 2017). For lesion length- based groupings, VSURF based wavenumbers were associated with carbohydrate (C□O; 947 & 964 cm^{-1}), carbonyl (C□O; 1349 cm^{-1}) and ester groups (RCOOR', 1754 cm^{-1}) stretching (Martin *et al*., 2005; Taoutaou *et al*., 2012; Conrad *et al*., 2014).

In *Eucalyptus* wood, spectral peaks 1370-1365 cm^{-1} are associated with CH_2 bending in cellulose and hemicellulose while the spectral region 1470-1460 cm^{-1} is associated with C-H deformation and lignin groups (Popescu *et al*., 2007; Gonultas and Candan, 2018). Lignin is a polyphenol that acts as a barrier against pathogens (Boudet, 2000). Lignin was associated with defense responses in *E. nitens* leaf shoots during *Mycosphaerella* infection (Smith *et al*., 2007). According to Coates (2000), compounds with phenol groups tend to have absorption bands in the following regions: phenol with an OH stretch (3620-3540 cm^{-1}), phenol or alcohol with an OH bend (1410-1310 cm^{-1}) and phenol with C-O stretch (~1200 cm^{-1}). Although this provides information about the functional groups that are present in resistant and susceptible clones, further studies are necessary to identify the unique and specific chemical markers responsible for resistance in the hybrid clones against the stem canker disease. This is a long term objective for this study.

The constitutive SVM model based on VSURF selected variables from the prediction step was able to accurately classify non-inoculated samples with a repeated mean CV

accuracy of 74%. Furthermore, the model correctly classified 100% of resistant samples and 95% of susceptible samples in the training set. A total of three ramets were sampled from each non-inoculated clone. This was enough to train and cross-validate the model without a testing dataset, but more ramets are recommended in future studies for not only training the model but also validating its efficiency with a testing dataset. This is because the more data that is used to train the model, the better the model performs (James *et al*., 2013). Although, there is no "golden rule" as to exactly what percentage of data is required for training a good model, having more data to work with is ideal more so to allow for further validation of the model on a testing dataset which was not possible in this study due to the sample size (Appen Blog, 2020). A recent study by Mukrimin *et al*. (2019) used 75 and 33 samples to train and test their FT-IR their model. The model correctly classified 97% of the training data and 83% of the testing data, thus validating the efficiency of their model.

The constitutive SVM model in this study, shows how FT-IR coupled with machine learning approaches can be used to distinguish between resistant and susceptible samples before infection. This approach could allow for rapid disease screening of *Eucalyptus* clones as well as possible field applications against *C. austroafricana* (Conrad & Bonello, 2016). In a recent study, Villari *et al*. (2018) developed a classification model without inoculation trials with samples collected in the field. This was advantageous as it is rapid, robust and less destructive.

Contrary to the above mentioned study, greenhouse inoculations were performed in this study as the resistance and/or susceptibility of the ramets to *C. austroafricana* was not known. In future studies, we can adopt the sampling methods used in the previously mentioned studies to generate predictive models in the absence of inoculation trials. Additionally, the results provide a good "proof of concept" on the potential of FT-IR in identifying resistant and susceptible *Eucalyptus* hybrid trees against the stem canker fungus without an inoculation trial.

The SVM results for induced FT-IR spectra grouping based on clone ID and lesion length from the training dataset accurately classified 80 and 83% of infected samples, respectively. In the testing data, the accuracy was 61% for clone ID grouping and 59% for lesion length grouping models, respectively. SVM models for the two grouping

variables were able to correctly classify resistant and susceptible samples, although classification accuracies in the testing datasets were higher for susceptible samples (66 & 65%) than resistant samples (56 & 56%). These results show that the models may be more effective in identifying hybrid clones that are susceptible to *C. austroafricana*.

An area under the ROC curve (AUC) plot was produced by plotting the true positive rate (TPR) and false positive rate (FPR) of the models. The AUC-ROC curve measures the accuracy of the model's predictions regardless of the classification threshold used. This value ranges from 0 to 1, with values closer to 1 indicating that the model can accurately classify observations into their classes (Marrocco *et al*., 2008). There are a number of studies that have employed AUC-ROC curves to verify model classification accuracies. One study, by Galvin-King *et al* (2020), created predictive models to detect adulterated paprika using data from both FT-IR and NIR. In their study, AUC-ROC curves were used to validate the performance of the NIR and FT-IR models. The AUC values for both NIR and FT-IR dataset were 0.951 and 0.907 respectively, thereby showing how effective both models were in detecting adulterated samples. In this study, the constitutive model had a ROC value of 1.00, indicating that the model can distinguish between resistant and susceptible ramets. The values for the induced and lesion length grouping models were lower than that of the constitutive, showing that the models still need to be improved upon.

The next step would be to identify novel chemical markers that are associated with resistance or susceptibility, using the spectral data from the inoculated trees. This can be done through the use of ultra-high performance liquid chromatography (UHPLC) which allows for the identification of phenolic compounds (Santos *et al*., 2013). Additionally, we can employ partial least square discriminant analysis (PLS-DA) to predict the concentration of the phenolic markers associated with resistance (Conrad *et al.,* 2014).

Near infrared (NIR) spectroscopy has also shown a lot of potential as a phenotyping tool and has the advantage of being non-destructive (Bokobza, 1998). Advances in miniaturized NIR spectrometers have recently increased, making field application easier (Yan & Siesler, 2018) In a recent study, the technique was coupled with

machine learning algorithms for early detection of *Rhizoctonia solani* which causes rice sheath blight in rice (Conrad *et al*., 2020). We can therefore, test NIR in future for its use in discriminating resistant and susceptible *Eucalyptus* clones against *C. austroafricana.*

2.6. Conclusion and future work

The results discussed above suggest that FT-IR has the potential to screen resistant and susceptible *E. grandis* hybrid clones against *C. austroafricana*. When combined with machine learning, the technique could classify non-inoculated *Eucalyptus* hybrid clones. The approach has the potential to be rapid, sensitive, and less destructive compared to artificial inoculation making it suitable for field phenotyping, and could save costs and time associated with inoculation experiments.

To our knowledge, this is the first study in South Africa which demonstrates the value of FT-IR spectroscopy in screening *Eucalyptus* hybrid clones for the identification of resistant phenotypes against *C. austroafricana*. This could potentially be integrated into resistance breeding programs in *Eucalyptus* plantations under changing climate conditions, with the potential to be transferred to other pathosystems, to screen trees for disease resistance against other fungal pathogens.

Further studies are required to ensure the models reproducibility in field. To improve model performance, trials with larger sample sizes for both training and testing sets are necessary. New independent datasets can also be used to test the robustness of existing models. Subsequently, the identification of chemical markers associated with disease resistance is critical. Key variables chosen for predictive modelling can be utilized as a guide to study these markers.

2.7. References

Albaugh, J.M., Dye, P.J., & King, J.S. (2013). *Eucalyptus* and water use in South Africa. *International Journal of Forestry Research*, 2013.

Appen Blog. (2020). What is training data?. https://appen.com/blog/training-data/ *~accessed on 23 May 2022*.

Bonello, P., Gordon, T.R., Herms, D.A., Wood, D.L., & Erbilgin, N. (2006). Nature and ecological implications of pathogen-induced systemic resistance in conifers: a novel hypothesis. *Physiological and Molecular Plant Pathology*, 68(4-6):95-104.

Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152).

Boudet, A.M. (2000). Lignins and lignification: selected issues. *Plant Physiology and Piochemistry*, 38(1-2):81-96.

Campos, W.G., Faria, A.P., Oliveira, M.G., & Santos, H.L. (2008). Induced responses against herbivory by chemical information transfer between plants. *Brazilian Journal of Plant Physiology*, 20(4):257-266.

Christy, A.A. , Gregoriou, V., & Ozaki, Y. (2001). Modern Fourier transform infrared spectroscopy. Elsevier, Amsterdam.

Coates, J. (2000). Interpretation of infrared spectra, a practical approach. In Encyclopedia of Analytical Chemistry R.A. Meyers (Ed.). John Wiley & Sons Ltd, Chichester, pp 10815-10837.

Conrad, A. O., & Bonello, P. (2016). Application of infrared and Raman spectroscopy for the identification of disease resistant trees. *Frontiers in Plant Science*, 6:1152.

Conrad, A. O., Li, W., Lee, D.Y., Wang, G.L., Rodriguez-Saona, L., & Bonello, P. (2020). Machine learning-based pre-symptomatic detection of rice sheath blight using spectral profiles. *Plant Phenomics*, 2020:10.

Conrad, A.O., Rodriguez-Saona, E.L., McPherson, B.A., Wood, D.L., & Bonello, P. (2014). Identification of *Quercus agrifolia* (coast live oak) resistant to the invasive pathogen *Phytophthora ramorum* in native stands using Fourier- transform infrared (FT-IR) spectroscopy. *Frontiers in Plant Science*, 5:521.

Conradie, E., Swart, W.J., & Wingfield, M.J. (1990). *Cryphonectria* canker of *Eucalyptus*, an important disease in plantation forestry in South Africa. *South African Forestry Journal*, 152(1):43-49.

Denison, N.P., & Kietzka, J.E. (1993). The use and importance of hybrid intensive forestry in South Africa. *South African Forestry Journal,* 165(1):55-60.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78-87.

Fahn, A. (1988). Secretory tissues in vascular plants. *New Phytologist*, 108(3):229-257.

Food and Agriculture Organization (FAO). (1979). Establishment techniques for forest plantations. FAO Forestry Paper 8; Rome, Italy.

Febrero-Bande, M., & de la Fuente, M.O. (2012). Statistical computing in functional data analysis: The R package fda. usc. *Journal of statistical Software*, 51(1):1-28.

Forestry in South Africa (FSA). (2018). Forestry in South Africa: Introducing commercial forestry. https://www.forestrysouthafrica.co.za/info-graphics/homepage/introducing-commercial-forestry/ *accessed on 15 November 2021*.

Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83-85.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., & Genuer, M.R. (2019). Package 'vsurf'. *Pattern Recognition Letters*, 31(14):2225-2236.

Genuer, R., Poggi, J.M., & Tuleau-Malot, C. (2015). VSURF: An R Package for variable selection using random forests. *The R Journal*, 7(2):19-33.

Gonultas, O., & Candan, Z. (2018). Chemical characterization and FT-IR spectroscopy of thermally compressed eucalyptus wood panels. *Maderas. Ciencia y Tecnología*, 20(3):431-442.

Gryzenhout, M., Myburg, H., Van der Merwe, N.A., Wingfield, B.D., & Wingfield, M.J. (2004). *Chrysoporthe*, a new genus to accommodate *Cryphonectria cubensis*. *Studies in Mycology*, 50:119-142.

Harfouche, A., Meilan, R., Kirst, M., Morgante, M., Boerjan, W., Sabatti, M., & Mugnozza, G.S. (2012). Accelerating the domestication of forest trees in a changing world. *Trends in Plant Science*, 17(2):64-72.

Heim, R.H.J., Wright, I.J., Chang, H.C., Carnegie, A.J., Pegg, G.S., Lancaster, E.K., Falster, D.S. & Oldeland, J. (2018). Detecting myrtle rust (*Austropuccinia psidii*) on lemon myrtle trees using spectral signatures and machine learning. *Plant Pathology*, 67(5):1114-1121.

Honkanen, T., Haukioja, E., & Kitunen, V. (1999). Responses of *Pinus sylvestris* branches to stimulated herbivory are modified by tree sink/source dynamics and by external resources. *Functional Ecology*, 13:126-140. doi:10.1046/j.1365-2435.1999.00296.x.

Indatalabs. (2021). Predictive performance and their performance evaluation. https://indatalabs.com/blog/predictive-models-performance-evaluation-important. *~accessed on 19 November 2021*.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An introduction to statistical learning. Springer New York, Dordrecht, London.

Kucheryavskiy, S. (2020). Mdatools–R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 198:103937.

Kuhn, M. (2015). Caret: classification and regression training. *Astrophysics Source Code Library*, ascl-1505.

Kuhn, M. (2021). caret: Classification and Regression Training. R package version 6.0-89. https://CRAN.R-project.org/package=caret.

Kuhn, M. 2019. The caret package. https://topepo.github.io/caret/data-splitting.html. *~accessed on 10 December 2021*.

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. https://bookdown.org/max/FES/recursive-feature-elimination.html *~accessed on 10 November 2021.*

Larkin, P. (2017). Infrared and Raman spectroscopy: principles and spectral interpretation. Elsevier.

Leger, M.N., & Ryder, A.G. (2006). Comparison of derivative pre-processing and automated polynomial baseline correction method for classification and quantification of narcotics in solid mixtures. *Applied Spectroscopy,* 60(2):182-193. doi:10.1366/000370206776023304.

Li, Q., Wang, W., Ling, X., & Wu, J.G. (2013). Detection of gastric cancer with Fourier transform infrared spectroscopy and support vector machine classification. *BioMed Research International*, 2013:942427.

Lupoi, J.S., Singh, S., Davis, M., Lee, D.J., Shepherd, M., Simmons, B. A., & Henry, R.J. (2014). High-throughput prediction of eucalypt lignin syringyl/guaiacyl content

using multivariate analysis: A comparison between mid-infrared, near-infrared, and Raman spectroscopies for model development. *Biotechnology for Biofuels*, 7(1):1-14.

Marrocco, C., Duin, R.P., & Tortorella, F. (2008). Maximizing the area under the ROC curve by pairwise feature combination. *Pattern Recognition*, 41(6):1961-1974.

Martin, J.A., Solla, A., Woodward, S., & Gil, L. (2005). Fourier transform-infrared spectroscopy as a new method for evaluating host resistance in the Dutch elm disease complex. *Tree Physiology*, 25(10):1331-1338.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. C. (2014). e1071: Misc functions of the department of statistics (e1071), TU Wien. *R Package Version*, 1(3).

Morais, C. L., Costa, F. S., & Lima, K. M. (2017). Variable selection with a support vector machine for discriminating Cryptococcus fungal species based on ATR-FTIR spectroscopy. *Analytical Methods*, 9(20):2964-2970.

Muchovej, J.J., Onokpise, O.U., & Bambo, S. (2008). Comparison of ramet and “wild type” establishment for cogongrass, *Imperata cylindrica*. In *Proceedings of the Florida State Horticultural Society*, 121:383-384.

Mukrimin, M., Conrad, A.O., Kovalchuk, A., Julkunen-Tiitto, R., Bonello, P., & Asiegbu, F.O. (2019). Fourier-transform infrared (FT-IR) spectroscopy analysis discriminates asymptomatic and symptomatic Norway spruce trees. *Plant Science*, 289:110247.

Muranty, H., Jorge, V., Bastien, C., Lepoittevin, C., Bouffier, L., & Sanchez, L. (2014). Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. *Tree Genetics & Genomes,* 10:1491-1510.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., & Goodstein, D.M. (2014). The genome of *Eucalyptus grandis*. *Nature*, 510(7505):356-362.

Naidoo, R., Ferreira, L., Berger, D.K., Myburg, A.A., & Naidoo, S. (2013). The identification and differential expression of *Eucalyptus grandis* pathogenesis-related genes in response to salicylic acid and methyl jasmonate. *Frontiers in Plant Science*, 4:43.

Nandiyanto, A. B. D., Oktiani, R., & Ragadhita, R. (2019). How to read and interpret FTIR spectroscope of organic material. *Indonesian Journal of Science and Technology*, *4*(1):97-118.

Neale, D.B., & Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, *12*(2):111-122.

Pichersky, E., & Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology*, 62:549-566.

Pilbeam, R.A., Howard, K., Shearer, B.L., & Hardy, G.E.S.J. (2011). Phosphite stimulated histological responses of *Eucalyptus marginata* to infection by *Phytophthora cinnamomi. Trees*, 25(6):1121-1131.

Poore, M.E.D., & Fries, C. (1985). The ecological effects of *Eucalyptus*. FAO Forestry paper 59, Rome.

Popescu, C.M., Popescu, M.C., Singurel, G., Vasile, C., Argyropoulos, D.S., & Willfor, S. (2007). Spectral characterization of *Eucalyptus* wood. *Applied Spectroscopy*, 61(11):1168-1177.

R Core Team. (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.r-project.org/ *~accessed on 10 November 2021.*

Ramsay, J.O., Graves, S. & Hooker, G. (2021). fda: Functional data analysis. *R package version 5.4.0*.https://CRAN.R-project.org/package=fda.

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752.

Roesnner, U. & Bowne, J. (2018). What is metabolomics all about? *Biotechniques*, 46(5).

Roux, J., Myburg, H., Wingfield, B.D., & Wingfield, M.J. (2003). Biological and phylogenetic analyses suggest that two *Cryphonectria spp*. cause cankers of *Eucalyptus* in Africa. *Plant Disease*, 87(11):1329-1332.

Rumpf, T., Mahlein, A.K., Steiner, U., Oerke, E.C., Dehne, H.W., & Plümer, L. (2010). Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 74(1):91-99.

Santos, S.A., Vilela, C., Freire, C.S., Neto, C.P., & Silvestre, A.J. (2013). Ultra-high performance liquid chromatography coupled to mass spectrometry applied to the identification of valuable phenolic compounds from *Eucalyptus* wood. *Journal of Chromatography B*, 938:65-74.

Santos, S.A., Villaverde, J.J., Freire, C.S., Domingues, M.R.M., Neto, C.P., & Silvestre, A.J. (2012). Phenolic composition and antioxidant activity of *Eucalyptus grandis*, *E. urograndis* (*E. grandis*× *E. urophylla*) and *E. maidenii* bark extracts. *Industrial Crops and Products*, 39:120-127.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940-3941.

Singh, A., Ganapathysubramanian, B., Singh, A.K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, 21(2):110-124.

Smith, A.H., Gill, W.M., Pinkard, E.A., & Mohammed, C.L. (2007). Anatomical and histochemical defence responses induced in juvenile leaves of *Eucalyptus globulus* and *Eucalyptus nitens* by *Mycosphaerella* infection. *Forest Pathology*, 37(6):361-373.

Sumner, S.C., Williams, C.C., Snyder, R.W., Krol, W.L., Asgharian, B., & Fennell, T.R. (2003). Acrylamide: A comparison of metabolism and hemoglobin adducts in rodents following dermal, intraperitoneal, oral, or inhalation exposure. *Toxicological Sciences*, 75(2):260-270.

Taoutaou, A., Socaciu, C., Pamfil, D., Fetea, F., Balazs, E., & Botez, C. (2012). New markers for potato late blight resistance and susceptibility using FTIR spectroscopy. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, 40(1):150-154.

Villari, C., Dowkiw, A., Enderle, R., Ghasemkhani, M., Kirisits, T., Kjær, E.D., Marčiulynienė, D., McKinney, L.V., Metzler, B., Muñoz, F., & Nielsen, L.R. (2018). Advanced spectroscopy-based phenotyping offers a potential solution to the ash dieback epidemic. *Scientific Reports*, 8(1):1-9.

Visser, E.A., Magwanda, R., Becker, J.V.W., Külheim, C., Foley, W.J., Myburg, A.A., & Naidoo, S. (2015). Foliar terpenoid levels and corresponding gene expression are systemically and differentially induced in *Eucalyptus grandis* clonal genotypes in response to *Chrysoporthe austroafricana*. *Plant Pathology*, 64:1320-1325.

Wickham, H., François, R., Henry, L. & Müller, K. (2021). dplyr: A grammar of data manipulation. *R package version1.0.7*. https://CRAN.R-project.org/package=dplyr *~accessed 10 November 2021.*

Wingfield, M. J. (2003). The 2003 Daniel McAlpine Memorial Lecture. Increasing threat of diseases to exotic plantation forests in the Southern Hemisphere: lessons from *Cryphonectria* canker. *Australasian Plant Pathology*, 32(2):133-139.

Wingfield, M.J., Brockerhoff, E.G., Wingfield, B.D., & Slippers, B. (2015). Planted forest health: The need for a global strategy. *Science*, 349(6250):832-836.

Wingfield, M.J., Roux, J., Govender, P. & Wingfield, B.D. (2001) Plantation disease and pest management in the next century. *Southern African Forestry Journal*, 190:67-71.

Wingfield, M.J. Swart, W.J. & Abear, B.J. (1989). First record of *Chryphonectria* canker of *Eucalyptus* in South Africa. *Phytophylactica*, 21:311-313.

Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2016). Data mining, fourth edition: practical machine learning tools and techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yan, H., & Siesler, H.W. (2018). Hand-held near-infrared spectrometers: State-of-the-art instrumentation and practical applications. *NIR News*, 29(7):8-12.

Yan, Y. (2016). MLmetrics: Machine learning evaluation metrics. *R package version*, *1*(1).

Zarco-Tejada, P.J., Camino, C., Beck, P.S.A., Calderon, R., Hornero, A., Hernández-Clemente, R., Kattenborn, T., Montes-Borrego, M., Susca, L., Morelli, M. & Gonzalez-Dugo, V. (2018). Previsual symptoms of *Xylella fastidiosa* infection revealed in spectral plant-trait alterations. *Nature Plants*, 4(7):432-439.

2.8. Supplementary material

Table S1: Optimal SVM parameters selected for two variable selection methods.

Dataset	**Variable selection method**	**Kernel**	**Cost**	**Gamma**
Constitutive	VSURF	radial	100	0.5
	RFE		10	0.5
Induced by Clone ID	VSURF	radial	10	0.05
	RFE		100	0.05
Induced by lesion length	VSURF	radial	1	0.5
	RFE		1	0.5

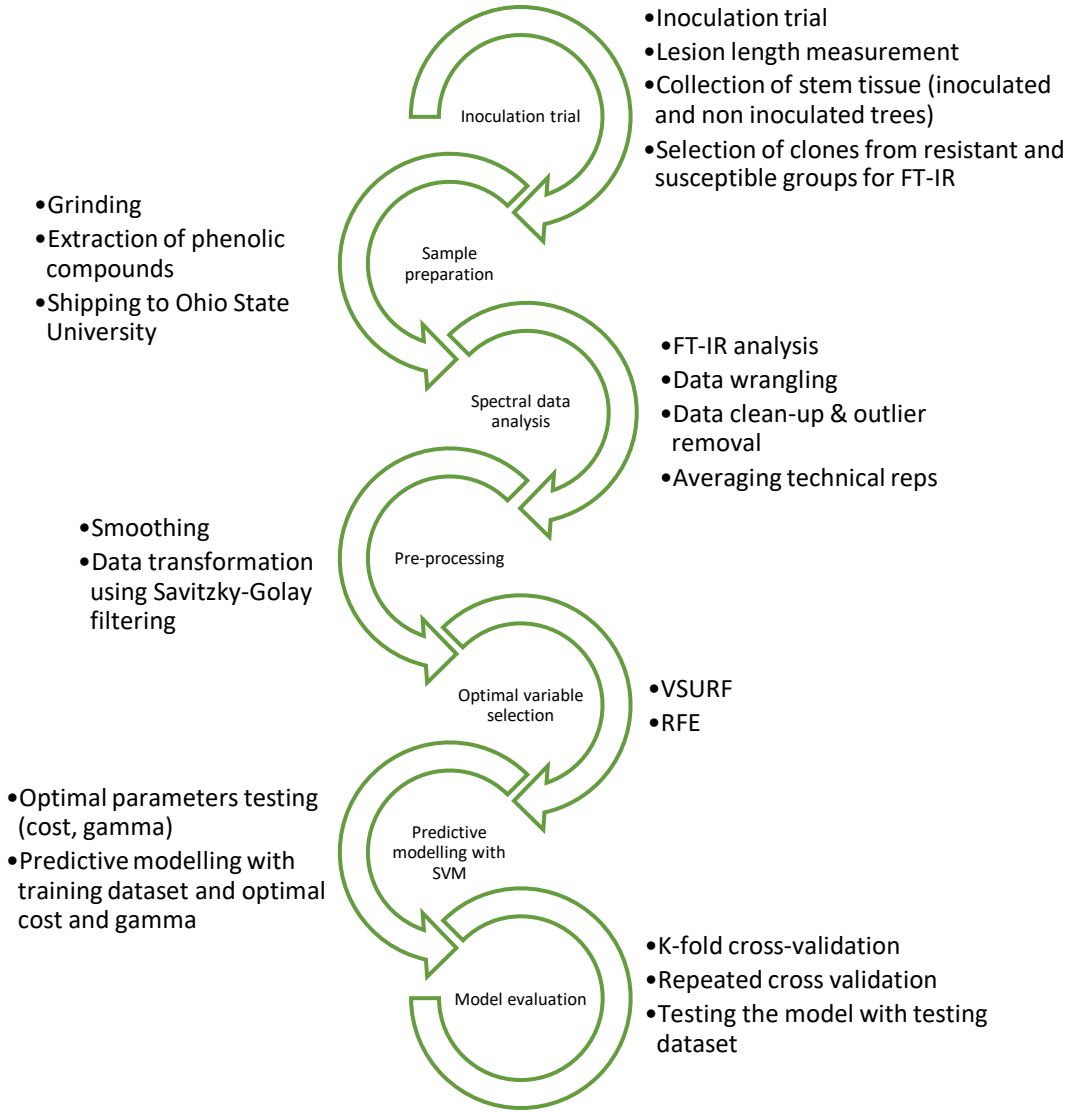

Figure S1: Workflow for FT-IR classification modelling with machine learning.

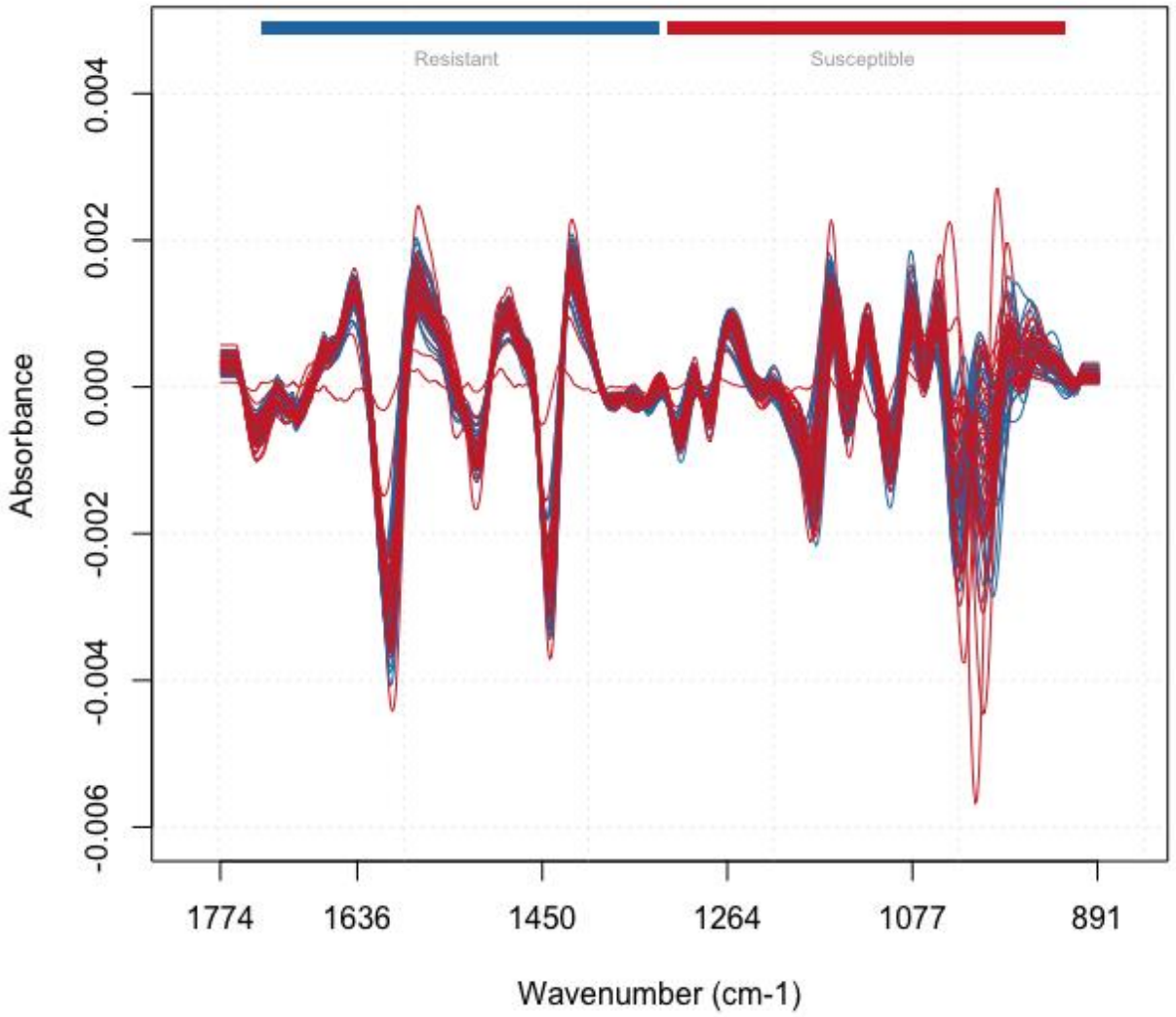

Figure S2: Second derivative transformed constitutive FT-IR spectra with resistant (blue) and susceptible (red) samples.

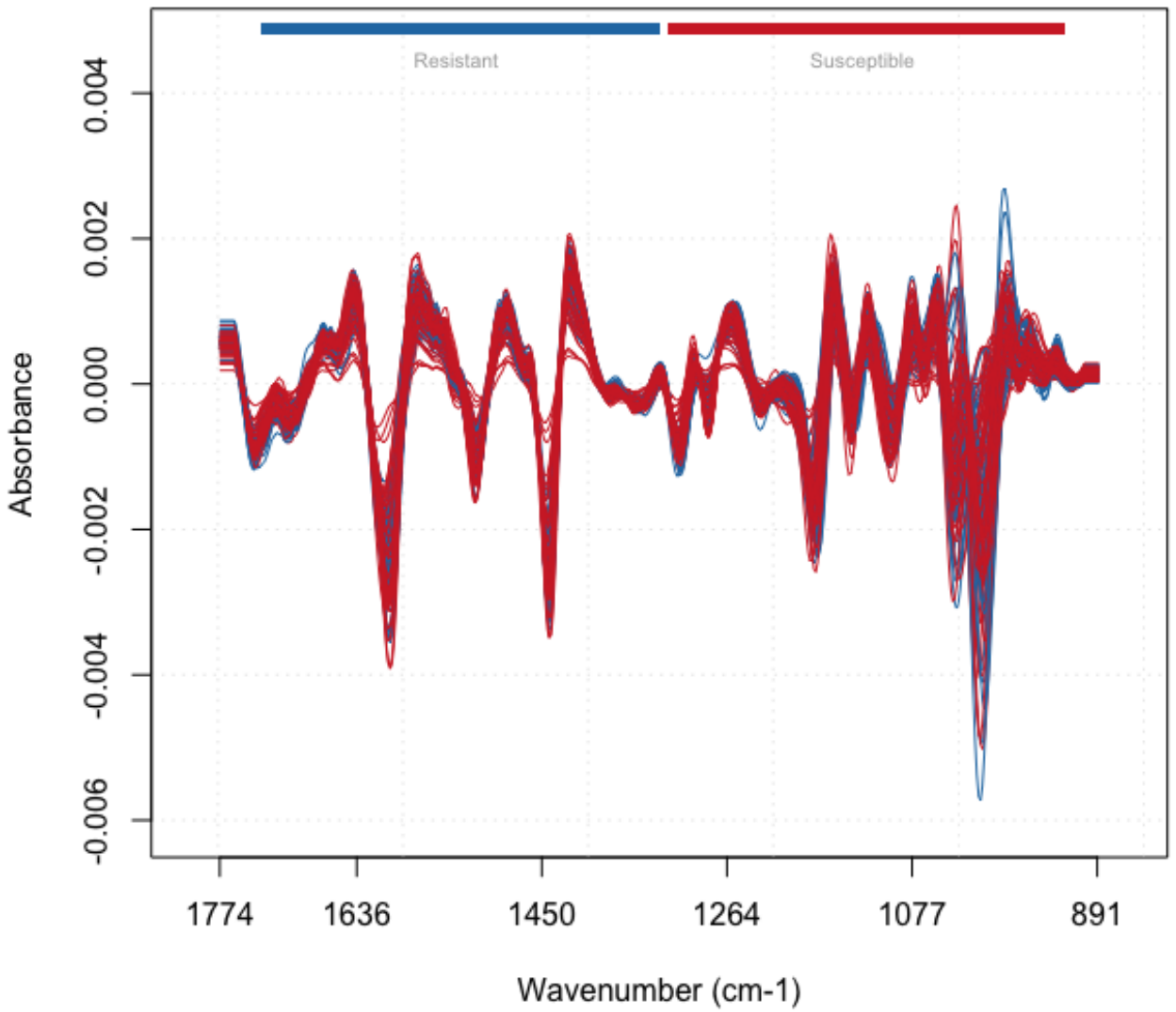

Figure S3: Second derivative transformed induced FT-IR spectra with resistant (blue) and susceptible (red) samples.

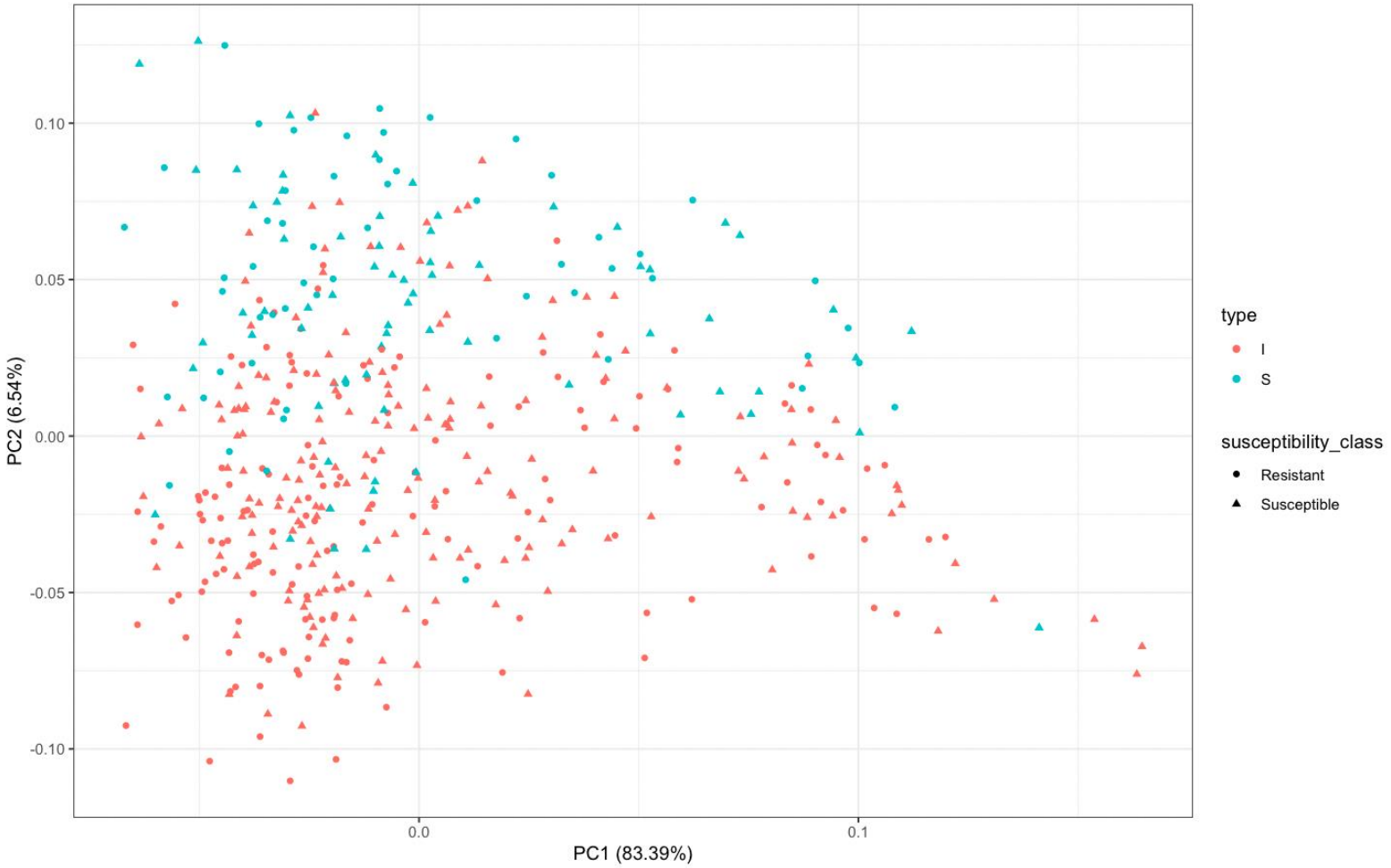

Figure S4: PCA plot of all non-inoculated (S- blue color) and inoculated ramets (I-red). Ramets are classed based on their phenotype: resistant (closed circle) and susceptible (closed triangle).

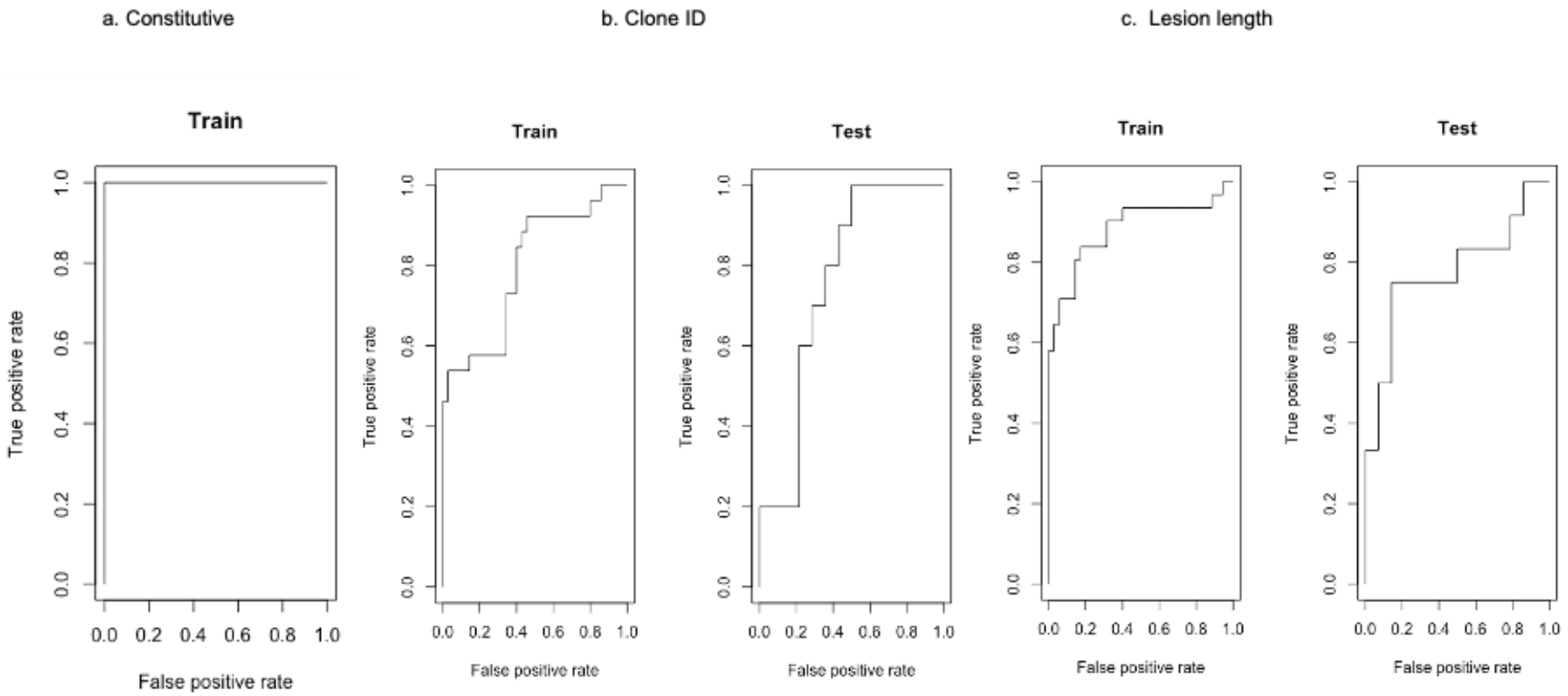

Figure S5:Area under curve-Receiver operating characteristic (AUC-ROC) curves for SVM classification constitutive **(a)**, induced grouped by clone ID **(b)** and lesion length **(c)** models**.**