Towards an automated approach for detecting *P. ramorum* in Welsh forests with hyperspectral imaging

Matthew Causon

A thesis submitted in partial fulfilment of the requirements for the award of the degree Master of Research (MRes)

Supervised by Prof. Doreen Boyd

Nottingham Geospatial Institute
University of Nottingham
Nottingham
UK
August 2024

Abstract

Forests in the UK are increasingly susceptible to damage from *Phytophthora ramorum*, with larch trees being particularly at risk. In recent years, machine learning has been coupled with remote sensing data for early detection of various forest pests and diseases and has achieved good results. This thesis compares how well three popular machine learning algorithms (random forest, support vector machine and artificial neural network) are able to predict the presence and severity of *Phytophthora ramorum* in larch trees in a forest in Wales, UK. Results indicate that all three models are able to make reasonable predictions, with all models achieving an overall classification accuracy of above 70%. The results of the random forest classifier were most promising, with the model achieving an overall accuracy of 78.4%. A common theme for all models was that healthy larch pixels were more easily identified than larch pixels infected with *Phytophthora ramorum*. In particular, the classifiers struggled with detecting the earlier stages of infection by the pathogen. The pixel-based approach taken for classifying the pixels within the image suffered from noise and it is suggested that future work should address this shortcoming.

Aknowledgements

Firstly, I would like to thank Professor Doreen Boyd whose help and guidance has helped shape these results. Thanks also to Rob Taylor for providing the pre-processed data that has enabled this work. Finally, thanks to the EPSRC for funding this work, and the Geospatial Systems CDT for the teaching and support throughout the academic year.

Contents

1 Introduction

Forest ecosystems are invaluable carbon sinks and key biodiversity hosts, afford leisure opportunities for communities, and provide a valuable source of timber, making them integral to the livelihoods of those working within the timber industry. In recent years, British forests have come under increasing stress from various forest pests and diseases, with one of these diseases being *Phytophthora ramorum (P. ramorum)* which has had a particular impact on larch forests. Traditional methods for understanding the impact and spread of such diseases are time consuming and costly. Fortunately, advances in remote sensing technology and machine learning methods that allow the automation of processing tasks could provide a more cost and time efficient method for understanding the spread of the disease.

The main aim of this project is to understand whether it is possible to move towards an automated approach for detecting the presence and severity of *P. ramorum* in larch trees in the UK. To understand whether this is possible, several machine learning algorithms will be trained and the accuracy of the predictions made with these models compared. As part of this process, relevant machine learning algorithms will be selected and the optimal hyperparameter combinations determined. A key focus of the study will be how well the algorithms can differentiate between different stages of *P. ramorum* infection, and whether the results attained are sufficient to encourage the use of the classifiers in forest health management. The spatial resolution of the training data, and how well it is balanced between classes, will be considered to understand whether higher resolution data leads to improved accuracy and whether the quantity, and balance between classes, of training data impacts the models' ability to accurately predict disease classes. Finally, the merits of using pixel-based classifiers will be examined and a comparison with object-based classifiers made.

2 Background

2.1 Forest pests and diseases

It is widely accepted that the introduction of non-native forest pests and diseases primarily occurs as a result of international trade. Countries with the highest level of imports of plants, raw logs and other wood products are therefore more at risk of pest and disease invasion and a 2014 study by Roy et al. (2014) found that the number of forest pests per country is correlated with its GDP. Additionally, the changing climate across the globe means that pests or diseases that were previously unable to cause damage in certain regions, owing to inhospitable conditions, may now be able to naturalise in these areas and begin to cause damage (Forest Research, 2019*b*).

2.1.1 Phytophthora ramorum

P. ramorum is an oomycete plant pathogen that threatens a large variety of trees and has the potential to cause significant economic and ecological damage (Green et al., 2021). It is thought that *P. ramorum* entered the UK through the commercial plant trade and has since spread into the wider national landscape from there. Over a period of three years, Green et al. (2021) collected 3624 samples from 163 host genera across fifteen plant nurseries and found that approximately 50% of the samples tested positive for

P. ramorum or other closely related oomycetes. In the UK landscape, *P. ramorum* is particularly prevalent in larch trees and in such instances has been called 'larch tree disease' or 'sudden larch death' (Forest Research, 2019*b*). As shown in Figure 1, the disease is known to cause branch dieback, lesions on stems, resinous bleeding and needle necrosis (Brasier and Webber, 2010). Larch is the third most common conifer species in the UK, accounting for 8% of the country's coniferous woodland (Forest Research, 2023). Being the only deciduous conifer in the UK, larch benefits biodiversity and provides interest for recreational purposes, and therefore has a significant presence in the country's landscape (Brasier and Webber, 2010).

Figure 1: (a) Resinous bleeding and (b) needle necrosis in larch caused by *P. ramorum*. Images taken from Forest Research (2019*b*).

Prior to the outbreak of *P. ramorum*, the commercial timber industry was already limited to a small number of viable tree species for plantations in the acidic, upland forestry sites of western Britain (Green et al., 2021). Following the infection of vast swathes of Japanese larch plantations, the UK's forestry industry now finds itself in a vulnerable position, with Sitka spruce the only other commercially viable timber species for these upland forestry sites (Green et al., 2021). There is an appetite within the UK's forestry industry to keep larch as a viable planting option, however, its tall height and significant production of spores provide favourable conditions for the spread of *P. ramorum* over vast distances (Forest Research, 2019*b*). Larch's disease spreading proficiency is particularly concerning considering the unknown damage potential of *P. ramorum* in other plant species. *P. ramorum* is the cause of sudden oak death in the USA - a disease that has reached epidemic proportions in California (Rizzo et al., 2002). Sudden oak death, with symptoms including bleeding cankers on the lower trunk of oaks and rapid deterioration of foliage (Rizzo et al., 2002), has killed tens of thousands of native oak trees in California alone (Rizzo and Garbelotto, 2003) and is the subject of numerous studies looking to predict the distribution of the disease using machine learning (Guo et al., 2005; Václavík et al., 2010). Whilst the disease does not appear to have impacted the UK's native oak population yet, *P. ramorum* has infected over 40 plant species in Europe (Brasier and Webber, 2010), and there are fears that adaptations of the pathogen and changing weather conditions could increase the pathogen's potential for destruction (Forest Research, 2019*b*).

P. ramorum sporulates and spreads most efficiently in areas with high levels of rainfall and humidity (Dun et al., 2024) and the optimum temperature for the growth of *P. ramorum* is $20^{\circ}C$ (O'Hanlon et al., 2017). This means that whilst *P. ramorum* has been found across most regions of the UK, it is generally more often reported in the wetter, western regions of the country as shown in Figure 2 (Forest Research, 2019*b*). Notably, a Welsh Government report setting out its strategy for managing the disease called *P. ramorum* the most serious tree disease to have impacted Welsh forests (Welsh Government, 2019).

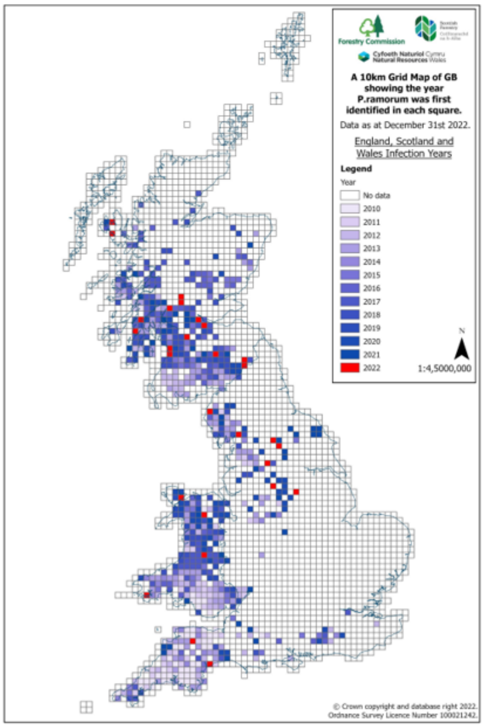

Figure 2: Confirmed locations of *P. ramorum* within the UK. The pathogen clearly favours the wetter, western regions of the country. Image taken from Forest Research (2019*b*).

Whilst there are many studies in literature looking at the remote sensing of diseases in other coniferous species like pine (Kunjie Wu and Lan, 2023; Gensheng Hu and Zeng, 2022; Huan Tao and Jing, 2020; Xia et al., 2021; Lin et al., 2019), there are significantly less studies focussing on larch. A rare study focussing on larch applied k-NN and random forest (RF) classifiers to airborne laser scanning data to detect the presence of *P. ramorum* infection (Barnes et al., 2017). Guo et al. (2005) applied support vector machines (SVMs) to the detection of damage caused by *P. ramorum*, though this study focussed on sudden oak decline - a disease caused by the same pathogen but in oak trees. One factor that differentiates larch from other species of conifer is that it is the only species in the UK that is deciduous (Brasier and Webber, 2010). For this reason, it is not possible to analyse the discolouration or mortality of needles across multiple years and seasonal needle loss must therefore be taken into account when determining the optimal time for data collection (Dun et al., 2024).

2.2 Forest health management

Manual surveillance of forest health requires a significant investment of time and personnel (Forest Research, 2019*a*) and is therefore not a realistic option for landscape-wide monitoring of the ever expanding infected forest stands. For example, the Welsh government have identified limited resources and budget as being key risks to its current strategy (aerial and ground surveys) for monitoring ramorum disease (Welsh Government, 2019).

Additionally, manual surveys are only able to provide a qualitative snapshot of forest health due to uncertainties surrounding the spatiotemporal variability of methodology, varying conditions during observation periods and the biases introduced by manual assessors (Stone and Mohammed, 2017). Remote sensing of forest pests and diseases is not a new concept (Lausch et al., 2017; Stone and Mohammed, 2017; Torres et al., 2021), and can provide forest managers with an alternative option for accurate and cost effective monitoring. In addition to the cost efficiencies realised through the use of remote sensing, big data modelling techniques mean that the accuracy of such methods now exceed both manual field-based and airborne visual mapping (Stone and Mohammed, 2017). It is worth noting that manual forest health assessment will continue to be critical in understanding the spectral traits of plant species that are impacted by pests and diseases (Lausch et al., 2017), as well as proving validation of boundary conditions like temperature, colour of leaves and soil moisture levels (Pause et al., 2016).

Figure 3 demonstrates the relationship between survey capability and technology maturity of the most common remote sensing options (Manfreda et al., 2018). Generally, UAVs (UAS in Figure 3) provide users with higher spatial and temporal resolution but the trade off is that they are less suitable for large-scale surveys currently and they have the greatest data processing complexity. They represent the least developed technology but are considered to have the greatest potential for future advances, which could lead to improvements in survey scale possibilities. The opposite of all of this is true for satellite technology: they present the best large-scale coverage option and have the lowest data processing complexity but the spatial and temporal resolution is generally low. The technology is established and has had significant investment over recent years, meaning future technological advances will likely be limited. The capabilities of occupied aerial systems generally falls between UAVs and satellites offering reasonable spatial and temporal resolution, relatively large scale survey possibilities, and middling data processing complexity.

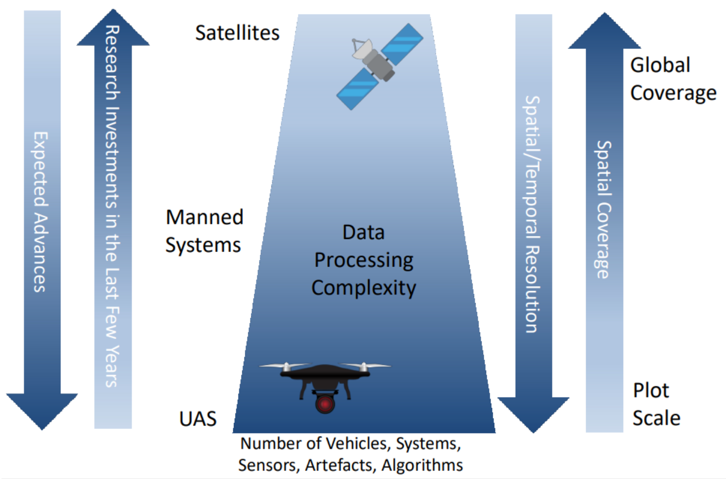

Figure 3: Comparisons between UAV, manned-aircraft and satellite for remote sensing. Image taken from Manfreda et al. (2018).

Torres et al. (2021) conducted a systematic review of literature in the field of forest health remote sensing in 2021 and found that 72.9% of studies utilised satellite data in some capacity. In one study, Townsend et al. (2012) mapped defoliation caused by the gypsy moth by studying changes in vegetation indices produced using Landsat data.

Meddens and Hicke (2014) also used Landsat imagery in their study which aimed to understand the spatial and temporal distribution of tree mortality at stand level caused by mountain pine beetle in North America. Murfitt et al. (2016) utilised satellite-based remote sensing data for segmenting individual tree crowns and predicting whether the trees were infected by emerald ash borer.

Technological advances and the reduction in costs associated with unoccupied aerial vehicles (UAVs) over recent years has paved the way for increased use in applications such as forest health assessment. In particular, the high spatial resolution, low operational costs and quick turnaround associated with the data collection makes UAV-based remote sensing particularly suited to regular time-series monitoring of landscapes (Guimarães et al., 2020). The popularisation of such approaches led Ecke et al. (2022) to carry out a systematic review of UAV-based forest health monitoring in 2022. The cost-benefit of UAVs are most obvious over relatively small study areas ($< 20ha$) (Manfreda et al., 2018), with the main limitation to using UAVs at a wider scale being the payload capacity and battery life of the technology (Guimarães et al., 2020). The high spatial resolution of UAV data is ideal for individual tree level information (Näsi et al., 2018), however, the storage and processing of such spatially abundant data can prove challenging (Manfreda et al., 2018). UAVs generally fly low enough that cloud cover and atmospheric influence do not pose significant issues to data collection, thus reducing the need for complex atmospheric corrections and cloud removal (Jakob et al., 2017). Whilst this makes data collection possible on overcast days, UAVs are vulnerable to other weather conditions like wind and rain, both of which can reduce survey quality or inhibit data collection altogether (Manfreda et al., 2018). The unstable movement of UAVs and the influence of micro-topography at low flight heights does make it necessary for geometric and radiometric pre-processing of hyperspectral data collected on board a UAV (Jakob et al., 2017).

Whilst Näsi et al. (2018) found that UAV-based hyperspectral data were more effective at mapping bark beetle damage in spruce than aircraft-based hyperspectral, their UAV-based data collection was carried out over a much smaller study area (<5ha) than their airborne-based data collection (>100ha). The UAV-based data was of a spatial resolution providing branch level detail, however, the airborne-based data was still sufficiently detailed to provide individual tree crown health assessment. Fassnacht et al. (2014) also utilised airborne hyperspectral data to map bark beetle mortality in the Bavarian Forest National Park. Where the spatial resolution of airborne data is not of sufficient detail, it can be compensated through fusion with LiDAR data which can provide additional structural information (Degerickx et al., 2018). One of the main downsides to using occupied aircraft for remote sensing is that the costs of running flight campaigns can be more prohibitive than for platforms with cheaper initialisation costs like UAV (Näsi et al., 2018). There may however be cost and time efficiencies when using aircraft-based sensors at a wider scale considering that the number of flights required, and therefore the overall time spent collecting data, will be significantly less than if UAV were used. This is particularly true as data collection is scaled to provide a forest wide view as is the case in this study.

2.3 Sensor types

Forest health monitoring uses remote sensing technology to monitor the condition of tree crowns, with changes like defoliation or discolouration of plant matter often indicating damage caused by biotic agents (Stone and Mohammed, 2017). Laser scanning is an often used technology for monitoring changes in the structure of a forest canopy. Where a disease or insect attack causes defoliation, a higher proportion of the laser pulses will pass through the canopy and result in more ground returns (Barnes et al., 2017). For example, the presence of common pine sawfly, which can cause severe growth loss and defoliation in Scots pine and ultimately leads to increased tree mortality, can be detected effectively using laser scanning (Kantola et al., 2010). Detection of forest pests and diseases using laser scanners is limited by the sensors' reliance on structural canopy changes making early detection difficult.

In many cases, it is desirable to catch forest pests and diseases at earlier stages of infection and an alternative indicator of disease presence is therefore necessary. Many remote sensing approaches instead use changes in spectral characteristics to detect plant stress. The most basic spectral sensors (RGB or panchromatic) make use of spectral information in the visible light portion of the spectrum. Del Frate et al. (2007) utilised very high resolution Quickbird satellite images (visible light portion of spectrum) to classify land cover types using ANN, however, it is common to utilise multi- or hyperspectral sensors to provide additional, non-visible spectral information. This additional spectral information is particularly useful when trying to identify forest pests and diseases at an early stage. Multispectral imaging for remote sensing tasks has commonly focussed on classifying land use. Heermann and Khazenie (1992), Paola and Schowengerdt (1997) and C. Huang and Townshend (2002) utilise neural networks for this task, with the latter also assessing SVMs for this application.

Whilst the additional spectral detail provided by multispectral sensors may be useful for certain aspects of remote sensing, like differentiating between different types of land-cover, the sensors' capacity for discriminating between different conditions of the same species (e.g. health classes) is limited (Melgani and Bruzzone, 2004). Carter and Knapp (2001) found that stressed plants commonly exhibit changes in leaf reflectance, transmittance and absorption in the far-red and green-yellow parts of the spectrum. These changes, triggered by stress-induced decreases in leaf chlorophyll levels, can be used to facilitate the study of forest stress at a large scale. Hyperspectral imaging provides a superior option for forestry applications because of the finer spectral resolution it offers researchers (Adão et al., 2017). This increased spectral resolution enables better profiling of plant material and allows the assessment of parameters indicative of forest health that might otherwise go unnoticed using multispectral imaging. The use of hyperspectral remote sensing data for mapping forest pests and diseases is widespread (Näsi et al., 2018; Degerickx et al., 2018; Hu et al., 2015).

Despite the obvious benefits to using hyperspectral data for detecting plant stress, this type of remote sensing data can suffer from the "curse of dimensionality" problem. This issue arises when the dimensions of a dataset are excessively large in comparison to the quantity of data samples, and can lead to poor classification accuracy (Sun and Du, 2019). In remote sensing, where it is often costly and impractical to collect large samples of labelled training data, the high dimensionality of hyperspectral data could limit classification accuracy. One of the key challenges of working with hyperspectral remote sensing data then becomes how to remove spectral redundancy (reduce dimensionality) whilst ensuring that key spectral information is not lost. It is common practice to use some form of feature selection algorithm to remove unnecessary additional information. Fassnacht et al. (2014) utilised the genetic algorithm, which is based on the principal of natural selection, to optimise feature selection. This provides insight into the most useful features for health classification and improves model transferability by reducing the classifier's complexity. Other options include using attention-based convolutional neural networks (CNNs) (Ribalta Lorenzo et al., 2020) or principal component analysis (PCA) (Lin et al., 2019; Chen and Ho, 2008) to select important bands of information. It is also common for researchers to use vegetation indices (VIs) as a method for reducing the dimensions of hyperspectral data by focussing on combining key, complementary wavebands. VIs derived from narrow hyperspectral wavebands offer the greatest sensitivity to the biochemical properties of forest canopies (Huete, 2012).

2.4 Machine learning algorithms used in literature

Supervised machine learning algorithms are commonly applied to the task of classifying remote sensing data for forests. Many of the algorithms utilised in literature have been used to predict the classes of individual pixels within an image. One of the most popular pixel-based classification methods for remote sensing data is the maximum likelihood estimator (MLE) (Sun et al., 2013). This classifier works by building a probability density function for each category and assumes normality of training data (Sun et al., 2013). Remote sensing data often comes from multiple sources and statistical distributions, and for this reason the assumption of normality (or any other assumption of the distribution of the underlying data) generally does not hold (Benediktsson, 1999). For this reason, parametric machine learning algorithms are better avoided when classifying remote sensing data.

Neural networks avoid some of the downfalls of the most widely used classifiers in the field of remote sensing because they avoid making a rigid assumption about the underlying distribution of each data class (C. Huang and Townshend, 2002). This means that neural networks offer a more robust approach for the classification of remote sensing data when the data distributions are not Gaussian (Paola and Schowengerdt, 1997). Bischof et al. (1992) show that neural networks can provide similar results to maximum likelihood in the classification of land cover using multispectral data but without being sensitive to the underlying probability distribution. Neural networks are particularly applicable to remote sensing because, when the network architecture is chosen appropriately, they are less susceptible to the curse of dimensionality and are more accurate than than other statistical pattern recognition techniques when data is less reliable (Chen and Ho, 2008). One of the downsides of using neural networks is that the training process can be very computationally expensive and therefore very slow to converge. This is a particular issue when data has high dimensionality (Benediktsson, 1999). Historically, the most popular neural networks used for classifying remote sensing data are a form of networks called multi-layer perceptrons (C. Huang and Townshend, 2002). Xu et al. (2020) used neural networks to detect the presence of the Masson pine caterpillar, a pest of pine trees, and found positive results. Despite the neural network achieving positive classification results, Xu et al. (2020) found that superior results were obtained using a RF classifier. When detecting and mapping basal stem rot in oil palm plantations, Heri Santoso and Wang (2017) also found RF to be the best classifier (compared with SVM and classification and regression tree (CART) algorithms).

There are also studies that have found SVM to be the most effective remote sensing classifier for forest applications. Melgani and Bruzzone (2004) found that SVMs outperformed other nonparametric classifiers (in their case RBF neural networks and k-nearest neighbours (k-NN)) in terms of classifying hyperspectral remote sensing images. They also found that SVMs were less sensitive to the curse of dimensionality than other classification methods and therefore don't rely on time-consuming feature extraction before use. Fassnacht et al. (2014) found that their SVM model classified later stages of tree mortality caused by bark beetle damage relatively well, but did acknowledge that it failed to reach an operationally sufficient accuracy level (80%) for earlier stages of damage. Guo et al. (2005) utilised SVMs to predict the potential distribution of sudden oak death in California and achieved true positive rates of almost 93%.

One particularly exciting area of the forest health remote sensing field is the recent advances in the capabilities of CNNs (Xia et al., 2021; Deng et al., 2020; Huan Tao and Jing, 2020; Kunjie Wu and Lan, 2023; Gensheng Hu and Zeng, 2022). CNNs are generally used for object classification as opposed to pixel-based classification and these algorithms are therefore capable of classifying whole tree crowns rather than focussing on individual pixels. Chiang et al. (2020) utilised transfer learning, the process of re-using a previously trained model as a starting point for a new problem, and mask region-based CNNs to automate the detection of dead trees in aerial images. Yu et al. (2017) demonstrated that CNNs can be utilised for image classification on small hyperspectral datasets without the need for manual feature selection and whilst avoiding the problem of over-fitting.

As Chen and Ho (2008) note, theory and practice are not always consistent in pattern recognition, so it is difficult to say conclusively that one classification method is better than the others. With adequate training, there could be many classifiers that achieve very similar results, and feature selection or pre-processing of the data may have more of an impact on classification results than the algorithms themselves (Chen and Ho, 2008). For this reason, it is best to trial multiple algorithms to understand which provides the best results for that particular task.

2.5 Challenges associated with automated forest disease detection

Forest health monitoring is currently a distance from being fully automated and it is likely that human-in-the-loop methods will be central to the field in the near future. Much of the research that has been discussed so far requires some form of human input whether that be choosing the features to feed models as input data or using human expertise to label data. This form of human-in-the-loop machine learning is referred to as machine teaching (Mosqueira-Rey et al., 2023). The process of labelling data for use in supervised learning methods is one of the bigger challenges of using machine learning for forest health monitoring. In some instances, data could be labelled using aerial imagery (Fassnacht et al., 2014) although the presence of pests or diseases may not always be evident meaning there is an element of uncertainty to this approach. More likely is that data capture will require an element of field work usually involving a trained specialist inspecting and recording the coordinates of trees (Näsi et al., 2018). Labelling data in this way is both time consuming and expensive.

There is also a requirement for researchers to have a thorough understanding of the visual indicators of forest stress so that they can provide the models with relevant data or features. Features, which might take the form of VIs or prominent bands of spectral information, need to be selected to maximise the discrimination between classes whilst also reducing the dimensionality of the data (Benediktsson, 1999). It is therefore essential for researchers to have a clear understanding of how symptoms of tree crown damage progress over different stages of a plant's reaction to stressors (Stone and Mohammed, 2017). This is so that data in the most discriminating spectral bands can be collected.

Guo et al. (2005) note that, whilst most methods for classifying remote sensing data rely on labelled presence and absence data, the latter is not always reliably available. Often pseudo absence data is generated for use in statistical models (Zaniewski et al., 2002) but there are issues with taking this approach. One class models that only require presence data, such as one class SVMs, have demonstrated positive results previously. When mapping sudden oak death, Guo et al. (2005) found that one-class SVMs had a slightly higher true positive rate than two-class SVMs with pseudo absence data, however, it is likely that the one-class model significantly over predicted the presence of the disease. Lack of true absence data therefore remains a challenge that researchers must overcome when classifying remote sensing data.

A lack of overall labelled data can make it difficult to train models without overfitting. Unless reusing an existing model, it is likely that image recognition tasks may require millions of training examples to work effectively (Geron, 2022). Krizhevsky et al. (2012) propose a method for increasing dataset sizes by synthesising additional data using augmentation. This process involves creating supplementary data by modifying spectral or spatial details of existing labelled data. Spatial modifications to existing data uses translations and reflections of the original labelled training images to produce new data - this will generally be used where whole images are being classified (CNNs). Spectral modifications are produced by modifying the intensity of the training images' RGB channels to mimic variations in the colour and intensity of pixels due to changing illumination. One of the big challenges of needing so much training data, when considered in the field of remote sensing where applications often also require high spatial and spectral detail too, is the data storage and processing resource required (Adão et al., 2017).

Another obvious challenge is that of obtaining data of sufficient quality - if a dataset has outliers, errors or missing data it can make it difficult for algorithms to accurately learn patterns. Data cleansing, in which outliers and errors are removed or edited, is therefore crucial though it often takes a significant amount of time for researchers (Geron, 2022). As mentioned previously, high-dimensionality of data compared with the number of data points available for training can also impact results and so it is often common for researchers to have to remove spectral redundancy in data too before it can be utilised (Lin et al., 2019).

3 Foundations of mathematics

This next section introduces some of the available machine learning classification techniques for processing big remote sensing data. A method for comparing the quality of different algorithms is then presented. Finally, some of the challenges associated with using these automated approaches for data processing are discussed.

3.1 Classification methods

3.1.1 Random forest (RF)

Random forest (RF) classifiers were introduced in their current form in a paper published by Breiman (2001). They fall into a group of machine learning algorithms called ensemble methods which utilise majority voting to predict the most probable classification result using the votes of multiple models. In the case of RF, the underlying models are all decision tree classifiers, with each decision tree casting a vote based on random feature selection and a subset of the training data. One of the benefits of using RF models is that they do not overfit due to the law of large numbers (Breiman, 2001).

To understand RF, it is first necessary to consider decision trees. Decision trees are easy to interpret machine learning models that make use of simple classification rules to fit complex datasets (Geron, 2022). Decision trees are commonly "grown" using the CART algorithm which splits training data on a feature depending on whether a threshold value is exceeded. For example, in land classification of remote sensing data, one split of a decision tree may split instances depending on whether the NDVI score is above or below a value of 0.5. Decision trees built using the CART algorithm are increased in size by adding further splits at each resulting node. This process of splitting subsets further at each new node ceases upon the algorithm failing to find a split that will reduce the impurity within the model further (Geron, 2022). An example decision tree structure can be seen in Figure 4.

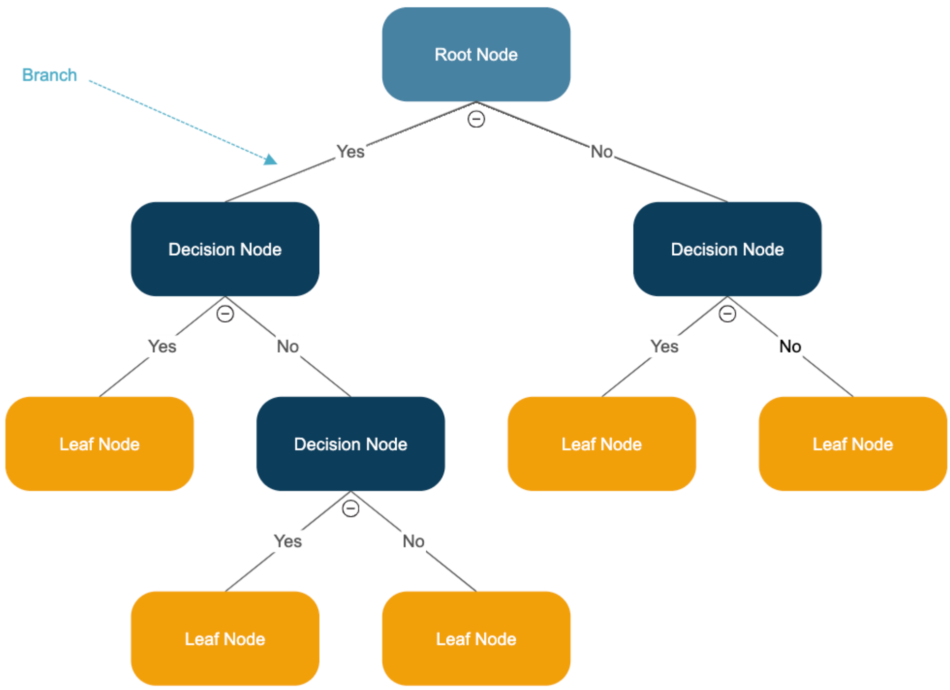

Figure 4: Structure of a decision tree. Image taken from Smartdraw (n.d.).

The exact feature and threshold values that are used to split training instances are determined by finding the purest possible subsets. The purest subsets are those with the lowest number of misclassified instances, weighted by their size at that particular node of the tree. The process of finding how pure a node is uses a measure called *Gini impurity* which is given by the following equation:

$$G_i = 1 - \Sigma_{k=1}^{n} p_{i,k}^2$$

where G_i is the Gini impurity of the i^{th} node and $p_{i,k}$ is the ratio of instances in class k at the i^{th} node in the decision tree. The CART algorithm then decides which splits to make based on minimising the following cost function:

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

where $G_{left/right}$ measures the Gini impurity of the left and right subsets following the proposed split and $m_{left/right}$ is the number of instances in the left and right subsets following the same proposed split (Geron, 2022).

In order to avoid overfitting to the training data, decision trees must have have some restrictions enforced in the form of regularisation parameters. When considering RF models, the regularisation parameters for the underlying decision trees are set at a forest level. Whilst Probst et al. (2019) show that tuning the hyperparameters in a RF can lead to improved model performance, they found that the improvement is not as significant as it is for other machine learning methods like SVM. Probst et al. (2019) do find that a higher number of trees in the forest leads to better results, however, this effect diminishes as forests grow large.

3.1.2 Support vector machine (SVM)

SVMs work by constructing a linear decision surface that best separates training instances into their classes (Cortes, 1995). In practice, there may be many planes capable of separating training instances in adjacent classes successfully, however, it is the planes that are furthest from instances in the training dataset that are chosen as the decision boundaries because this maximises the generalisation capability of the model to new instances. To find the plane that maximises the margin between neighbouring classes, the optimal feature weights and bias term must be determined (Geron, 2022). In the case of multi-dimensional remote sensing classification problems, SVM classifiers have several dividing planes (one between each pair of adjacent classes).

Using hard margins to separate instances means that all data points have to be on the correct side of the margin, however, this leads to issues if the data are not linearly separable or contain outliers. It is better instead to use soft margins to divide data which allow the SVM model flexibility in finding the optimum balance between maintaining a wide margin width and limiting the number of margin violations. This flexibility is controlled by the C hyperparameter, which allows users to define the trade-off they are willing to make between optimising the margin width and limiting the number of margin violations.

In this project we are concerned with a nonlinear classification problem - that is one in which classes are not linearly separable by a flat plane. In some cases, this can be solved by including polynomial features within the model which can transform the dataset into a linearly separable space. In practice, complex datasets can require high polynomial degrees which can result in models becoming overly slow. When working with SVMs, it is possible to apply the *kernel trick* which has the same effect as adding many polynomial features without actually doing so, thus reducing the computational expense and speeding up the model (Geron, 2022). An alternative option for tackling nonlinear problems is to add similarity features which transform our dataset based on how similar each instance is to some landmark defined by the user. The easiest approach to selecting landmarks is to create one at every single instance in the dataset, with this maximising the chance of the transformed data being linearly separable, however adding many similarity features can again result in a very slow model. Like in the polynomial case, this issue can be overcome using the kernel trick but this time using a Gaussian RBF kernel. This produces similar results to models with many similarity features added without having to do so. The Gaussian RBF kernel is commonly used in classification problems and often works very well (Geron, 2022).

The *gamma* hyperparameter controls the range of influence that instances have. Increasing *gamma* reduces the range of influence of each instance, which in turn makes the decision boundary more irregular and therefore more likely to wiggle around a given data point. *Gamma* should be reduced if a model is known to be overfitting, and increased if it is known to be underfitting (Geron, 2022).

3.1.3 Artificial neural network (ANN)

Artificial neural networks (ANNs) are machine learning algorithms inspired by how the human brain works. Interest in ANNs has come in several waves since the 1940s when McCulloch and Pitts (1943) introduced their concept of how neurons in animal brains might combine together for logical computations. Whilst interest in ANNs has fluctuated for various reasons over this time, including due to the successes of other machine learning techniques such as SVMs, it is unlikely that the current interest in these methods will reduce. Some of the reasons for this revitalised interest include the improvements that have been seen in recent years in terms of: the vast quantity and quality of data; the huge advances in computing power, owing in part to the large-scale production of GPUs; and the improvements seen in the algorithms themselves (Geron, 2022).

One of the most basic forms of ANNs is a perceptron, which are themselves based on the threshold logic units (TLU) seen in Figure 5 (a). In linear binary classification problems, TLUs would take a weighted sum of its inputs and if this value exceeds a threshold, determined using the step function, it outputs the positive class (Geron, 2022). Combining multiple TLUs in a single layer and connecting each TLU to every input creates a perceptron - this is often referred to as a fully connected or dense layer (Geron, 2022). An example of a perceptron can be seen in Figure 5 (b).

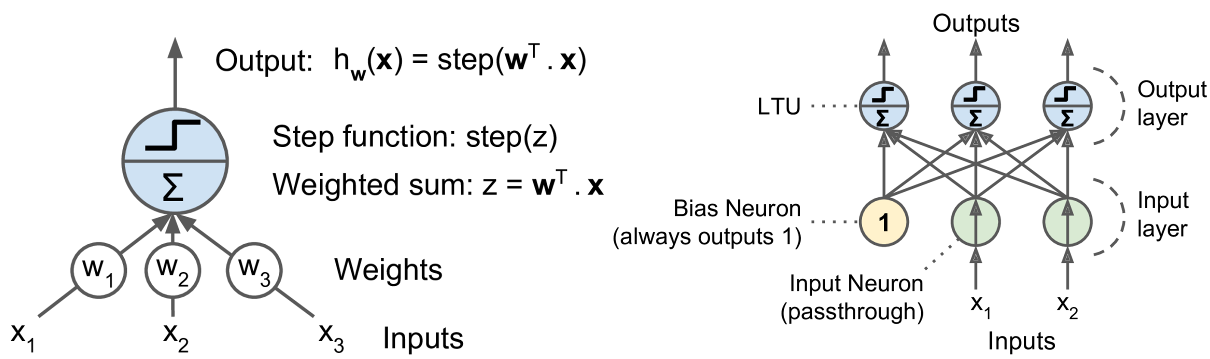

Figure 5: Architecture of **(a)** a threshold logic unit (TLU) and **(b)** a perceptron made of multiple TLUs. Images taken from O'Reilly (2018).

Feedforward neural networks are the product of stacking multiple perceptrons on top of each other, with the output of one perceptron becoming the input for the next perceptron. An example architecture of a basic feedforward network can be seen in Figure 6. This network has one hidden perceptron and one perceptron acting as an output layer. The aim of a feedforward neural network is to map some input $\boldsymbol{x}$ to an output class $\boldsymbol{y}$ by learning the model parameters θ that lead to the best result - this is the objective of the output layer and is specified by the training data (Goodfellow et al., 2016). The other layers in the model are not instructed what to do by the training data and instead the algorithm must determine how best to use those layers to produce the desired output. This process of optimising the connections within the network during training is known as back-propagation (Heermann and Khazenie, 1992). The other layers are referred to as hidden layers because the training data does not provide the desired output for these like it does with the output layer (Goodfellow et al., 2016).

Knowing how many hidden layers, and neurons within these layers, to use is difficult. If too few hidden layers and nodes are chosen within these layers the model will not have sufficient degrees of freedom to fit the training data (Heermann and Khazenie, 1992). Conversely, it is not advisable to use more hidden layers and neurons than is necessary as this can overcomplicate the model, resulting in long training periods and an inability to generalise well to new data (Heermann and Khazenie, 1992). To overcome this issue, a process of training and validating a number of different models should be undertaken and the best performing model then tested on a holdout test set.

In addition to the number of hidden layers and neurons, the learning rate and batch size are other model hyperparameters that impact performance. The learning rate determines the size of steps taken when updating the model parameters during the back-propagation process (Geron, 2022). There is a balance to be found with the learning rate such that the model converges to the optimal with sufficient speed whilst ensuring that the stability of the training process is maintained (Heermann and Khazenie, 1992). Batch size is the number of training examples that the model sees before updating the model weights. Setting the batch size as less than the total number of training instances, and greater than one, results in a process called mini-batch gradient descent. This mini-batch process is less erratic in the parameter space than stochastic gradient descent (batch size equal to one) but significantly faster to run than batch gradient descent (batch size equal to number of training instances) (Geron, 2022).

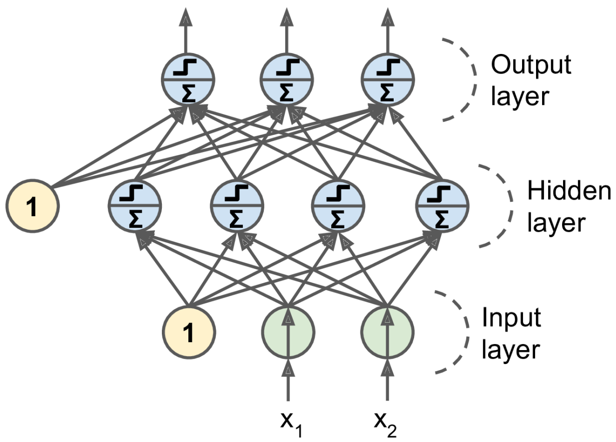

Figure 6: Architecture of a multilayer perceptron made up of one hidden layer of four TLUs and three output TLUs. The model has two inputs. Image taken from O'Reilly (2018).

3.2 Tuning hyperparameters for optimal model performance

The main objective in machine learning is to produce a model that is capable of generalising well to new data (Goodfellow et al., 2016). The training error, a measure of how well the model fits the training data, is first minimised by optimising the model's parameters during the training stage. A model's ability to generalise to new data is then tested by assessing the test error. To provide a fair assessment of the test error, a holdout test set is created by separating a random, representative sample (usually around 20%) from the overall labelled data (Geron, 2022). Models can then be tested on this holdout test set to provide a fair measure of how well it generalises to new data. The challenge of producing good machine learning models is about finding a balance between minimising the training error whilst ensuring the gap between the training and test errors is as small as possible (Goodfellow et al., 2016).

Figure 7 demonstrates three different levels of fit to a training dataset that has been randomly generated from a quadratic distribution. On the left hand side, the linear model does not fit the training data well which indicates that the complexity of the model is not sufficient. This model is said to suffer from underfitting. Increasing the model's complexity by including a quadratic term (middle plot) leads to a superior, and in this case optimal, fit which minimises the training error. Increasing the model's complexity further by adding in a polynomial term of degree nine does not improve the training error any further and the model fails to capture the quadratic nature of the training data. This excessive model complexity has led to what is termed overfitting and means that the model would be poor at generalising to new data (Goodfellow et al., 2016).

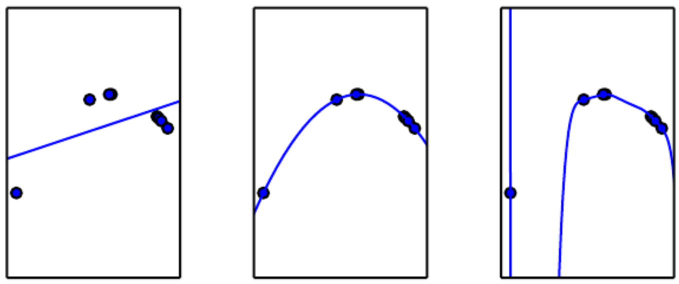

Figure 7: Graphs demonstrating underfitting (left), overfitting (right) and the appropriate level of fit (centre). Image taken from Goodfellow et al. (2016).

In practice, it is difficult to know how complex a model needs to be to avoid overfitting or underfitting. One commonly used method is to train many different models with differing architectures and hyperparameters (set prior to training for regularisation of models) to assess how each performs on unseen data. The holdout test set cannot be used for this purpose because this is reserved for use in testing the optimised models. It is instead common practice to use a validation set, a random selection of the training set (typically another 20% of the overall data) to test the different models (Goodfellow et al., 2016). The hyperparameters resulting in the best performing model on the validation set are then used to train a final model on the full training set (i.e. the reduced training set plus the validation set). The performance of this final model can then be evaluated using the holdout test set to provide the generalisation error estimate (Geron, 2022).

Using a validation set to tune the model hyperparameters results in an underestimation of the generalisation error of the interim models (Goodfellow et al., 2016). To provide a more realistic estimation of the generalisation error, and to therefore enable a fairer assessment of the best model hyperparameters, it is common to use k-fold cross-validation. The training set is divided into k non-overlapping groups (or folds) of approximately equal size and each fold is used as a validation set for a model trained on the remaining folds (Wong, 2015). By averaging the accuracies achieved on each fold, a better estimate of the generalisation error is achieved (Goodfellow et al., 2016). Using k-fold cross-validation for selecting the model hyperparameters should therefore result in selecting a better model. A review of cross-validation methods for enabling efficient hyperparameter selection can be found in (Arlot and Celisse, 2010).

3.3 Assessing effectiveness of ML algorithms

Classification accuracy assesses the proportion of instances for which a model produces the correct classification prediction (Goodfellow et al., 2016). Accuracy is a common measure of the quality of classification algorithms, however, when dealing with skewed datasets it is generally more insightful to use confusion matrices for this purpose (Geron, 2022). The rows in confusion matrices correspond to the true class of an instance, whilst the columns represent the predicted class. Therefore, the main diagonal of a confusion matrix shows how frequently a class is correctly predicted. The remaining cells provide insight into which classes are most commonly confused with each other.

If a more concise metric for assessing model performance is desirable, it is useful to consider the precision, recall and F1 scores. Precision is given by the equation:

$$precision = \frac{TP}{TP + FP}$$

where TP is the number of true positives (the number of instances correctly labelled as being in a class) and FP is the number of false positives (the number of instances incorrectly labelled as being in the same class). Recall is given by the equation:

$$recall = \frac{TP}{TP + FN}$$

where FN is the number of false negatives (the number of instances that have been incorrectly labelled as not belonging in a given class). The F1-score conveniently combines precision and recall into one metric using a harmonic mean. F1-scores will only be high if recall and precision are both high themselves. The metric is calculated using:

$$F_1 = \frac{TP}{TP + \frac{FN+FP}{2}}$$

These three metrics enable an easier comparison between classifiers. The confusion matrices then serve as a complementary assessment to understand whether all models commonly confuse the same classes with each other. This is an important consideration because, depending on the application, there may be a preference to confuse some classes with each other more than others.

4 Methods

Firstly, this section will detail and describe the motivation for using the chosen study area. Next, the data available for use in this study will be discussed. Finally, the methods that have been used to process and interpret the data will be set out.

4.1 Data collection

The focus of this study is a 530-hectare site in the Gwydyr Forest situated in Snowdonia National Park (Wales, UK). This site was chosen due to the known presence of *P. ramorum* within the forest. Data was collected by 2Excel Geo on 20th April 2021 using a manned aircraft flying at an altitude of 1000m. Two Norsk Elektro Optikk (NEO) HySpex hyperspectral sensors were used during the flight campaign (details for which can be found below in Table 1). The raw hyperspectral data had previously been pre-processed and was made available for use in this project as six vegetation index (VI) raster images which are detailed in Table 2. These VI represent the six most informative features from the original hyperspectral dataset (including raw hyperspectral bands and VI) (Taylor, 2022). It has not been possible to source the raw hyperspectral data.

A complementary field campaign was run between April and September 2021 during which field data were collected for 200 reference trees at various stages of *P. ramorum* infection. An experienced observer gathered information relating to the 200 reference trees, rating each tree against the classes presented in Table 3 and recording the location using a Garmin Etrex 32X handheld GPS device (Taylor, 2022). Lateral Flow Devices (LFD) were used to verify the observer's classification of trees - ten trees in each class type were tested using LFD. Part of the study site had previously been damaged by fire - a 'fire damaged' health classification was therefore included as a class type to enable a comparison between foliar damage caused by disease and by fire. The locations and health condition (four classes of *P. ramorum* infection and one fire damaged class) of the 200 reference trees were provided for use in this study as shapefiles. Figure 8 presents the locations and disease classifications for the reference trees within the wider Gwydyr Forest landscape.

Parameter	Hyperspectral sensor	
	HySpex VNIR-1800	HySpex SWIR-384
Spectral range (nm)	400 - 1000	930 - 2500
Spectral bands	182	288
Spectral resolution (nm)	3.26	5.45
Spatial pixels	1800	384
Spatial resolution (m)	0.33	0.71
Bit resolution	16 bit	16 bit

Table 1: Specifications for the two sensors used by 2Excel Geo for the hyperspectral data collection (Taylor, 2022).

4.2 Data cleansing

Data cleansing of the VI rasters included removing anomalous results and ensuring spatial consistency. Four of the VI raster images were produced using information collected by the HySpex VNIR-1800 camera and therefore had a spatial resolution of 0.33m. The two Normalised Larch Disease Indices (NLDI1 and NLDI2), designed specifically for detecting *P. ramorum* in larch trees Taylor (2022), made use of spectral information collected in the 930-2500nm range by the HySpex SWIR-384 camera. The spatial resolution of the raster images for the NLDI1 and NLDI2 was therefore 0.71m. In addition to the irregularities

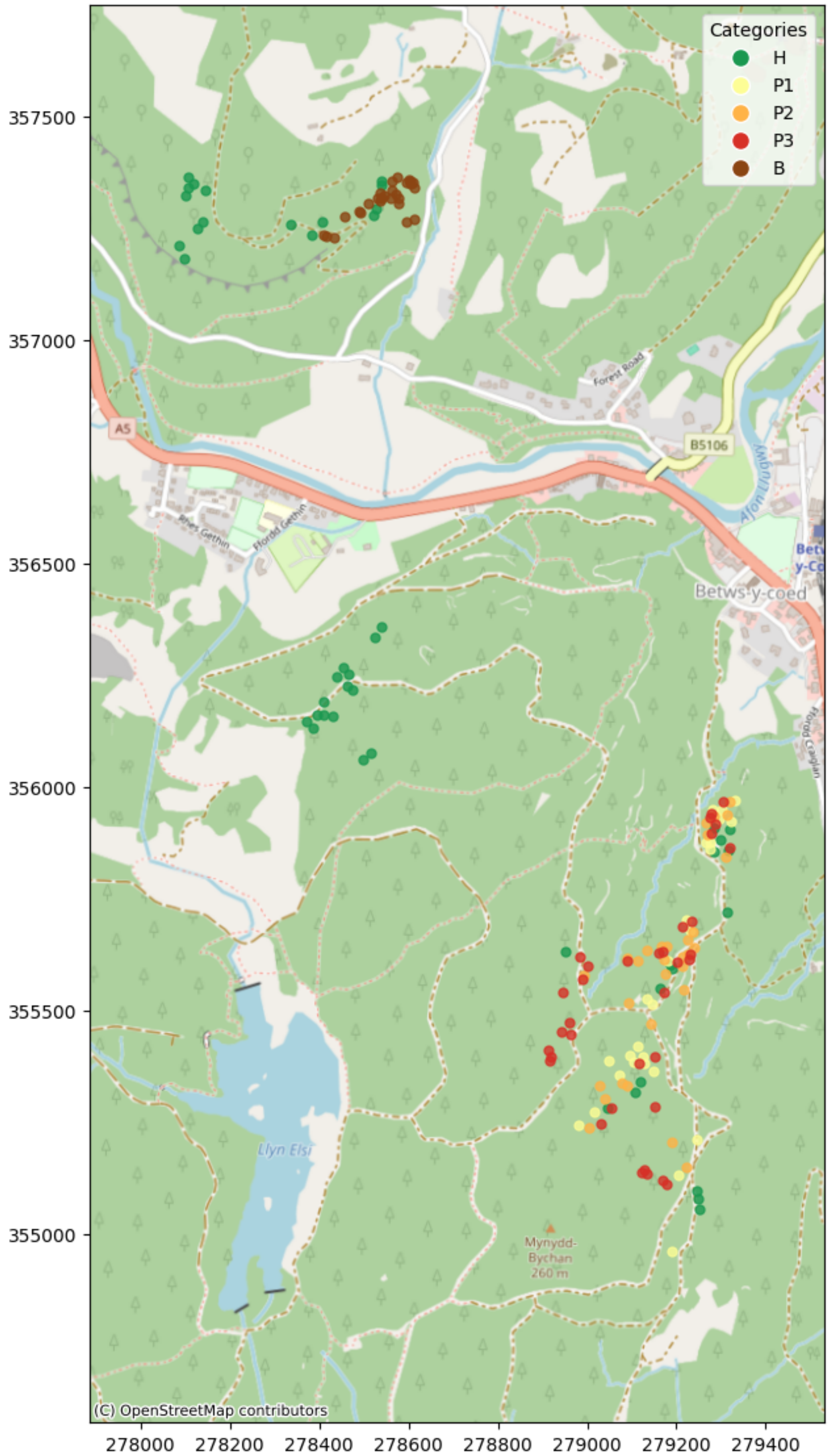

Figure 8: Locations of the 200 reference trees within the Gwydyr Forest, colour coded by health class.

Vegetation index	Spatial resolution (m)	Image dimensions (pixels)	
		Width	Height
Normalised Difference Vegetation Index (NDVI)	0.33	6,478	13,657
Normalised Larch Disease Index 1 (NLDI1)	0.71	2,977	6,418
Normalised Larch Disease Index 2 (NLDI2)	0.71	2,977	6,418
Carotenoid Reflectance Index (CRI)	0.33	6,478	13,657
Water Band Index (WBI)	0.33	5,989	13,518
Transformed Chlorophyll Absorption Reflectance Index (TCAR)	0.33	6,478	13,657

Table 2: Spatial information related to available vegetation indices.

Class ref.	Class type	Live crown proportion	General visual appearance
H	Healthy	80-100%	Lush, dense foliage appearance. Little epicormic sprouting. Little dead wood.
P1	Early symptoms of *P. ramorum* infection	60-80%	Branch tip die back present. Some epicormic sprouting. Minor foliage discolouration present.
P2	Moderate symptoms of *P. ramorum* infection	40-60%	May display dense epicormic sprouting in central crown. Advanced tip dieback and dead branches present. Conspicuous foliage discolouration present.
P3	Advanced symptoms of *P. ramorum* infection	0-40%	Advanced dieback of entire crown. Stem lesions and bleeding may be present.
B	Fire damaged	0-50%	Charred bark on lower stem. Sparse foliage.

Table 3: Larch tree condition visual assessment categories recorded during the field campaign (Taylor, 2022).

in the spatial resolution, the dimensions of the raster images were not always consistent. These differences necessitated resampling and cropping of the raster images.

Initially, the vegetation indices at 0.33m spatial resolution were downsampled to 0.71m resolution and cropped using the raster image with the smallest dimensions (WBI). It was hypothesised that models trained on data resampled to the higher spatial resolution of 0.33m might have superior classification results due to the models having more pure pixels for training. To test this, the vegetation indices at 0.71m were also upsampled to 0.33m resolution and again cropped against the dimensions of the WBI raster image. Statistical summaries and histograms of the pixel values for each of the VI raster images were studied to understand the distribution of the data and to highlight anomalous results. Where necessary, anomalous results were removed by forcing extreme values to take the value of the first percentile for the lowest 1% of values or the value of the 99th-precentile in the case of the highest 1% of values.

4.3 Data transformation

Pixels contained within the 200 reference tree polygons were extracted and flattened into a 2D array with dimensions equal to the number of pixels contained within the polygons (39,484) and the number of layers of information contained within the stacked raster image (six). Table 4 shows the number of pixels falling within each of the health classes. At this point, the data could be used as features (the VI scores for each pixel) and labelled data (pixel health class) for training the classifiers.

Health class	Pixels	
	Number	% of total
H	13,989	35.43%
P1	7,285	18.45%
P2	6,746	17.09%
P3	7,804	19.76%
B	3,660	9.27%
Total	39,484	100%

Table 4: Quantity of pixels in each health class

4.4 Justification of the models and hyperparameters

The data available for this project was limited to six VI raster images and 200 larch tree polygons which had been labelled with a health class during a historic field campaign. It has not been possible to collect any more field data within this project. Additionally, the unavailability of the raw hyperspectral data has meant that adding additional health class labels to the remaining pixels using aerial imagery, like Fassnacht et al. (2014), has not been possible. This lack of absence data has has meant that labelling whole images for use in training CNNs has not been possible and the methods employed in this project have therefore utilised a pixel-based approach to classification. Three machine learning algorithms, that have been widely utilised in the field of remote sensing, have been selected for use in completing this task. An overview of the three chosen algorithms and a description of the process used for tuning the models' hyperparameters will be discussed here.

4.4.1 Random forest (RF)

In this project, *RandomForestClassifier* within Python's *scikit-learn* library was utilised to train our RF model (Pedregosa et al., 2011). *Scikit-learn* uses the CART algorithm to "grow" decision trees, and the hyperparameters used by these underlying decision trees are set at the forest level. It is difficult to know the most suitable hyperparameters for a RF model so an exhaustive grid search of many hyperparameter combinations was utilised to make this decision. With RF having many possible regularisation parameters, which can each take many values, a useful initial step is to find a suitable starting point for the exhaustive grid search. *Scikit-learn* has a function enabling users to complete a randomised search of hyperparameter combinations which was used for this step. A large range of hyperparameters were defined and randomly combined, with each random combination assessed for accuracy using cross-validation. The hyperparameters of the best randomised model, determined using three-fold cross-validation on the training data (a random selection of 80% of the total labelled dataset), was then used to populate the hyperparameter grid for the exhaustive grid search. Three folds were used during the k-fold cross-validation process to limit computational expense. Table 5 shows the hyperparameter grid used for the exhaustive grid search of the RF hyperparameters as well as a summary of those that were found to be optimal (determined using 3-fold cross-validation) for use in the final model. In reality, all hyperparameters apart from "n estimators" defined in Table 5 are concerned with the regularisation of the underlying decision trees. The "n estimators" parameter is the only parameter specific to the RF here and instructs the model on how many decision trees to use in the majority voting process. The optimum hyperparameters were fed into a final model which was then trained on the full training dataset, with this model then tested on a holdout set (the remaining 20% of the total dataset).

Hyperparameter	Values	Optimum values
bootstrap	True	True
max depth	70, 80, 90	70
max features	sqrt	sqrt
min samples leaf	1, 2	1
min samples split	4, 5, 6	5
n estimators	200, 300	200

Table 5: Hyperparameter grid used for the exhaustive grid search and the optimum combination used to train the final RF model.

4.4.2 Support vector machine (SVM)

In this project, the *SVC* class in Python's *libsvm* library has been used to train our SVM model (Chang and Lin, 2011). Similarly to RF, knowing which values to use for the three hyperparameters (kernel type, C value and *gamma* value) to achieve the best results is not intuitive and so an exhaustive grid search was utilised to find the optimum combination. Table 6 shows the hyperparameter grid used for the tuning of the SVM model hyperparameters. With *libsvm* supporting the kernel trick, the Gaussian RBF kernel was chosen to overcome the nonlinear nature of the dataset (Geron, 2022). A random selection of 80% of the total dataset was chosen as the training data. The optimum combination of hyperparameters was determined using a three-fold cross-validation process on the training dataset and are shown in Table 6. Once the optimal hyperparameters had been used to train the final model on the full training dataset, it was tested against the holdout dataset (the remaining 20% of the total dataset).

Hyperparameter	Values	Optimum values
C	1, 10, 100, 500, 1000	1000
gamma	1, 0.1, 0.01, 0.001, 0.0001	0.01
kernel	Gaussian RBF	Gaussian RBF

Table 6: Hyperparameter grid used for the exhaustive grid search and the optimal combination used to train the final SVM model.

4.4.3 Artificial neural network (ANN)

Python's *TF-Keras* API was used to train the neural networks in this project (Chollet et al., 2015). The *Sequential* class was used to stack dense layers of neurons into the ANN models. The input layer shape reflects the six VI raster images that are used as input data, whilst the output layer has five neurons corresponding to the five health classes labelled in the data (see Table 3). Softmax was used as the activation function for the neurons in the ouput later because of the multiclass nature of this application (Geron, 2022). The data has sparse labels, meaning there is a single possible target class for each instance, so sparse categorical crossentropy was chosen as the loss function (Geron, 2022). The Adam optimiser was used for the back-propagation process in all models, owing to its speed and quality of convergence (Kingma and Ba, 2017). The activation function for each hidden layer in all models was chosen as ReLu because it is known to work well with gradient descent and is fast to compute (Geron, 2022). Keras' default layer weight initialiser, the Glorot uniform, was utilised throughout.

Two of the more significant considerations for our model was how many hidden layers to use and how many neurons to include within each of these layers (Heermann and Khazenie, 1992). As with the previous two classifiers, it is difficult to know which network architecture will provide the best results so a process of training and validating many models with different hyperparameters was undertaken. The full dataset was randomly split into training, validation and holdout test sets using a 60:20:20 split. Many different models were trained on this training set and an approximation of the generalisation ability was determined using the validation set. The hyperparameters of the model that generalised best on the validation set was then used to train a final model on the final training dataset (the train and validate sets combined). This final model was then tested on the holdout test set to provide a realistic approximation of the model's classification accuracy. Other hyperparameters that were tuned as part of this iterative process were the learning rate and the batch size. Table 7 shows the different hyperparameter combinations trialed as part of this process, as well as the optimal hyperparameters that were used to train the final model. Models ranged in size from one hidden layer and ten nodes, up to four layers with 400 nodes. Each of these model architectures were tested with each learning rate and batch size options. The optimal model had three hidden layers, each with 400 neurons in, and used a learning rate of 0.001 and a batch-size of 64.

5 Results

5.1 Random forest (RF)

Overall, the optimised RF model achieved a 74.8% accuracy at 0.71m spatial resolution and an accuracy of 78.4% at 0.33m spatial resolution. The confusion matrix in Figure 9 and the classification scores shown in Table 8 better demonstrate the models' ability

Hyperparameter	Values	Optimal values
Hidden layers	1, 2, 3, 4	3
Neurons	10, 20, 40, 200, 400	400
Learning rate	0.1, 0.01, 0.001, 0.0001	0.001
Batch size	32, 64, 128, 256	64

Table 7: Hyperparameter grid used for ANN model hyperparameter tuning, as well as the optimum combination used to train the final model.

to classify pixels across the five health classes. The main diagonals of both confusion matrices show that the models predict the correct class for each pixel more often than they erroneously predict any other class. The recall value for both models show that the RF model predicts healthy (H) larch pixels correctly with the highest consistency (93%). Pixels exhibiting advanced symptoms of *P. ramorum* infection (P3) were the next most accurately predicted class (recall of 0.81% for 0.71m resolution model and 0.84% for the 0.33m resolution model). The two intermediate larch disease classes, P1 and P2, have considerably lower recall values for each model ($< 70\%$ in all cases). Both models failed to identify fire damaged (B) pixels in approximately half of the cases.

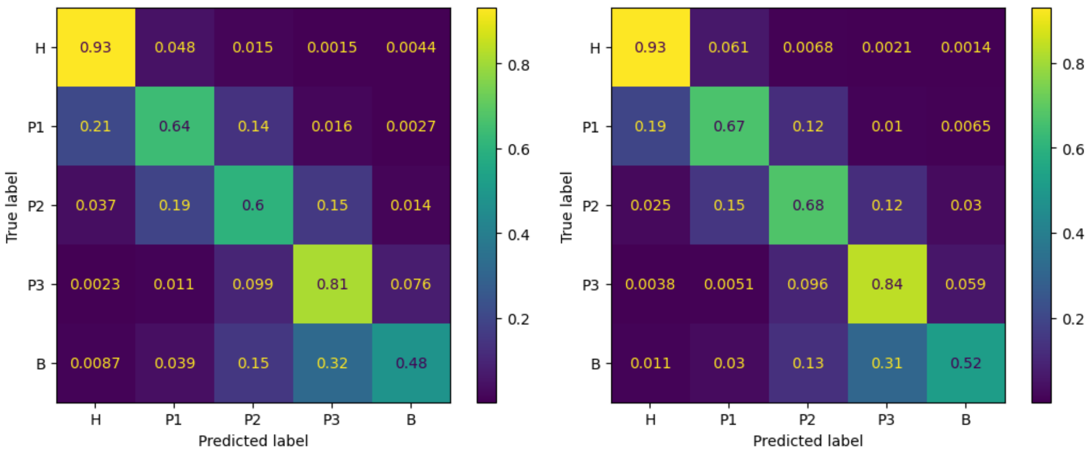

Figure 9: Confusion matrices for RF models at (a) 0.71m and (b) 0.33m spatial resolutions.

Health Class	0.71m spatial resolution			0.33m spatial resolution		
	Precision	Recall	F1-score	Precision	Recall	F1-score
H	0.87	0.93	0.90	0.89	0.93	0.91
P1	0.68	0.64	0.66	0.70	0.67	0.68
P2	0.61	0.60	0.61	0.67	0.68	0.67
P3	0.72	0.81	0.77	0.77	0.84	0.80
B	0.72	0.48	0.58	0.72	0.52	0.60
Accuracy	74.8%			78.4%		

Table 8: Classification scores for RF models.

Another significant measure of a model's effectiveness in classifying pixels is the precision value. This value indicates how many, out of all the pixels it has labelled within each class, have been misclassified (lower values indicate a higher number of incorrectly predicted pixel classes). Table 8 shows that the healthy larch class has the lowest proportion of incorrectly labelled pixels for both models. This class is followed by classes P3, B, P1 and P2 in the case of both models. As an example, 11% of pixels predicted to be healthy by the 0.33m spatial resolution RF model have been misclassified. This proportion is significantly higher at 33% for the P2 health class. One thing that is clear looking at the classification scores and confusion matrices side by side is that the higher spatial resolution model consistently performs better than the lower resolution model. Looking at the F1-score, the harmonic mean of the precision and recall for each model, the 0.33m resolution RF model has higher scores for each health class, with the largest difference between these scores seen in the P2 health class.

The final thing we can understand by examining the confusion matrix is if there are pairs of classes that are commonly confused with each other. The most commonly confused classes for both RF models is the incorrectly classified fire damaged pixels (B) that are predicted to be in the advanced *P. ramorum* symptoms class (P3) (32% of fire damaged pixels for the 0.71m RF model and 31% for the 0.33m model). Fire damaged pixels (B) were also commonly confused with the moderate *P. ramorum* symptoms class (P2) reasonably regularly (15% for 0.71m RF model and 13% for 0.33m RF model). Interestingly, the models do not predict pixels with moderate or advanced symptoms of *P. ramorum* as being fire damaged anywhere near as frequently. Otherwise, neighbouring larch disease severity classes were the most commonly confused classes. The most frequently misclassified of these were pixels in canopies showing early symptoms of *P. ramorum* being predicted to be healthy (21% for 0.71m RF and 19% for 0.33m RF).

5.2 Support vector machine (SVM)

In this section, the results for the SVM model at each of 0.33m and 0.71m spatial resolutions are presented. The model trained on data at the lower spatial resolution achieved an accuracy of 72.4% overall whilst the model trained on data at the higher spatial resolution of 0.33m achieved a marginally improved accuracy of 72.6%. The confusion matrix for each of the models can be seen in Figure 10 and the corresponding classification scores can be seen in Table 9. The main diagonal of the SVM confusion matrices looks similar to those of the RF models - the majority of the larger numbers feature on the main diagonal indicating that the models predict the correct class more than any other class. The one exception for both models where this is not true is for the fire damaged class (B), where both models misclassify fire damaged pixels as being in class P3 more frequently than they correctly classify them as fire damaged (B). The recall value for both models show that the SVM model predicts healthy (H) larch pixels correctly with the highest consistency (92%). Pixels with advanced symptoms of *P. ramorum* infection (P3) were the next most accurately predicted class (recall of 0.86% for 0.71m resolution model and 0.81% for the 0.33m resolution model). The two intermediate larch disease classes, P1 and P2, have lower recall values in the case of both models than the RF model ($\leq$ 60% in all cases).

Table 9 shows that the healthy larch class has the lowest proportion of incorrectly labelled pixels for both SVM models - this is indicated by the highest precision values (86% for both models). This class is followed by classes B, P3, P1 and then P2 in the case of both models. Of particular note is the sizeable portion of pixels that have been labelled as class P2 incorrectly - as many as 42% for the 0.33m resolution model and 40% for the 0.71m model. Unlike with the RF models, where the model trained on higher resolution data performed better (higher overall accuracy and F1-scores), there is no discernible

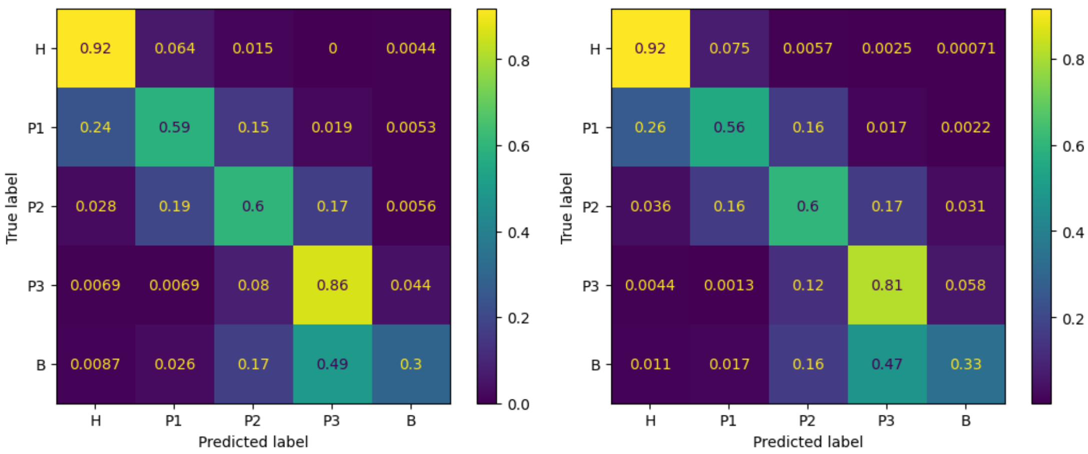

Figure 10: Confusion matrices for SVM models at (a) 0.71m and (b) 0.33m spatial resolutions.

Health Class	0.71m spatial resolution			0.33m spatial resolution		
	Precision	Recall	F1-score	Precision	Recall	F1-score
H	0.86	0.92	0.89	0.86	0.92	0.89
P1	0.64	0.59	0.61	0.64	0.56	0.59
P2	0.60	0.60	0.60	0.58	0.60	0.59
P3	0.68	0.86	0.76	0.69	0.81	0.75
B	0.72	0.30	0.42	0.72	0.30	0.42
Accuracy	72.4%			72.6%		

Table 9: Classification scores for SVM models.

difference between the models at differing resolutions. The overall model accuracy was marginally higher for the 0.33m model (72.6% compared with 72.4% for the 0.71m model), however, for each health class the F1-score for the higher resolution model was less than or equal to the lower resolution model.

Considering the most commonly confused health classes for the SVM models, the most frequent misclassifications were seen for the fire damaged health class (B). The SVM models fail to predict the fire damaged class more frequently than the RF model - in fact, both SVM models confuse fire damaged pixels with advanced symptoms of *P. ramorum* infection considerably more frequently than they correctly identify fire damaged pixels. The models do not confuse the P3 class as being fire damaged anywhere near as frequently. Similarly to the RF models, the SVM models also confuse fire damaged pixels with moderate symptoms of *P. ramorum* infection regularly (17% and 16% for the 0.71m and 0.33m spatial resolution models, respectively). Neighbouring larch disease severity classes seem to be commonly confused - this is consistent with the RF model. The most frequently misclassified of these were pixels in canopies showing early symptoms of *P. ramorum* being predicted to be healthy (24% for 0.71m RF and 26% for 0.33m RF).

5.3 Artificial neural network (ANN)

Finally, the results achieved using ANN models at 0.33m and 0.71m spatial resolutions are presented. The model trained on lower spatial resolution data achieved an overall accuracy of 70.6%, whilst the model trained on data at the higher spatial resolution achieved a minimially improved accuracy of 70.7%. Figure 11 presents the confusion matrix for each of the models and the corresponding classification scores can be seen in Table 10. The ANN confusion matrices looks similar to those of the previous models with the majority of the larger numbers appearing on the main diagonal. This indicates that the models predict the correct class more often than not. The only exception for either model where this is not true is for the fire damaged class (B) of the 0.71m resolution model. In this instance, the model mistakenly labels fire damaged (B) pixels as showing advanced symptoms of *P. ramorum* (P3) as often as it correctly predicts they are fire damaged (B). The recall values for both models show that the ANN model predicts healthy (H) larch pixels correctly with the highest consistency. Pixels with advanced symptoms of *P. ramorum* infection (P3) were the next most accurately predicted class. The classification performance of the ANN models with respect to the two intermediate larch disease classes (P1 and P2) was lower than the RF and SVM models.

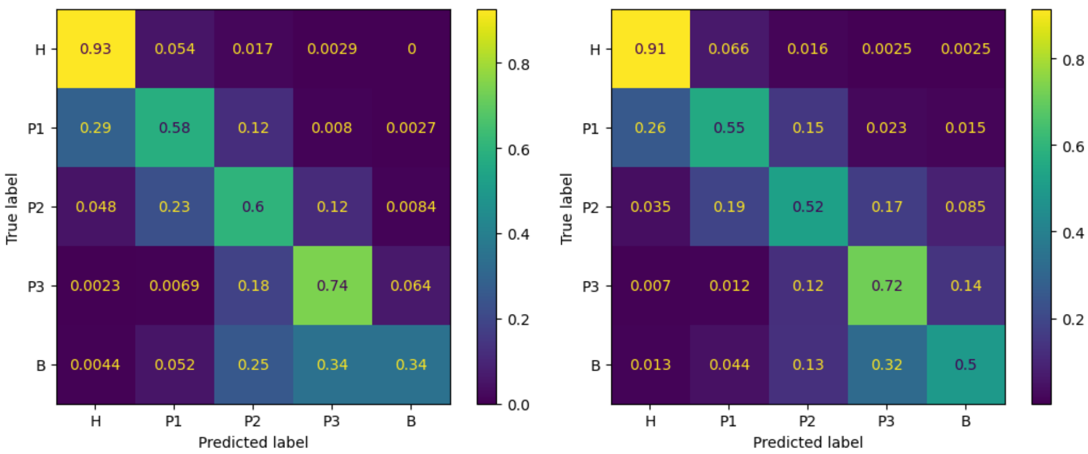

Figure 11: Confusion matrices for ANN models at (a) 0.71m and (b) 0.33m spatial resolutions.

Health Class	0.71m spatial resolution			0.33m spatial resolution		
	Precision	Recall	F1-score	Precision	Recall	F1-score
H	0.83	0.93	0.88	0.86	0.91	0.89
P1	0.62	0.58	0.60	0.62	0.55	0.58
P2	0.52	0.60	0.56	0.55	0.52	0.54
P3	0.72	0.74	0.73	0.70	0.72	0.71
B	0.71	0.34	0.46	0.50	0.50	0.50
Accuracy	70.6%			70.7%		

Table 10: Classification scores for ANN models.

The lowest proportion of incorrectly labelled pixels for both ANN models is within the healthy larch class. This is indicated by the highest precision values (0.83 and 0.86 for the 0.71m and 0.33m resolution models, respectively) in Table 10. The class ranked second in terms of the precision achieved by the models at both resolutions is class P3. In the case of the 0.71m resolution ANN model, the third ranked health classification by this metric is the fire damaged class, for which the model achieved an almost identical precision score to the P3 class. Interestingly, this health class had the lowest precision value when considering the ANN at 0.33m resolution. Overall, when comparing the performance of the two models at different spatial resolutions, it appears that the higher resolution model provides little obvious improvement on the lower resolution model. Whilst the overall accuracy is slightly improved for the higher resolution model, the F1-scores by health class show varied results. This is similar to what is seen in the case of the SVM models.

As has been the case with the previous two algorithms, the most frequently misclassified pixels were in the fire damaged health class (B). The ANN model at 0.71m resolution misclassified fire damaged pixels as belonging to the P3 health class as often as it correctly predicted they were fire damaged. The model trained on data at 0.33m spatial resolution performed much better in comparison, however, it still only correctly identified 50% of fire damaged pixels. Of the fire damaged (B) pixels misclassified by the higher resolution model, the most common confusion was still with the P3 health class (32% of fire damaged pixels erroneously predicted to be P3). Whilst both models do not mislabel pixels in the P3 class as being fire damaged anywhere near as often (which is common for all three algorithms at both resolutions), the ANN model at 0.33m resolution does appear to confuse 14% of the P3 pixels with being fire damaged (approximately double any other model). The ANN model at 0.71m resolution also confuses fire damaged pixels with moderate symptoms of *P. ramorum* infection (P2) regularly - with 25% of fire damaged pixels mislabelled as being P2 class. As seen with the other algorithms, both models mistake neighbouring larch disease severity classes regularly. The most frequently misclassified of these were pixels in canopies showing early symptoms of *P. ramorum* being predicted to be healthy (29% for 0.71m RF and 26% for 0.33m RF).

5.4 Mapping of *P. ramorum*

One of the objectives of this project was to understand if it is possible to map the presence of *P. ramorum* in forests using VI data and pixel-based classification algorithms. Figure 12 shows the predictions of the highest performing model, the 0.33m resolution random forest classifier, across the whole of the Gwydyr Forest as well as enlarged views of the predictions of all three classifiers (at 0.33m resolution) in a four hectare area of the forest. The health classification predictions of each model suffer from noise - this is clear from the majority of tree crowns in each of the enlarged views showing multiple colours. This salt-and-pepper noise is a result of using pixel-based classification methods to classify the tree health. The four hectare area of the forest used in the enlarged views shows that many of the trees in the area are exhibiting symptoms of ramorum disease though the stage of infection is varied. Though there is a lot of noise within the random forest model's predictions, this classifier appears to suffer less from noise than the other classifiers.

6 Discussion

6.1 Best model

Of the three classifiers used in this study, RF appears to provide the best results for predicting the health status of larch trees. At both resolutions, the RF classifier achieved the best overall classification accuracy and the highest F1-score in all classes. The model was particularly effective at classifying healthy larch pixels - correctly identifying 93% of the healthy pixels at both resolutions and having the lowest number of false positives at both resolutions when predicting whether pixels were healthy. The latter point is

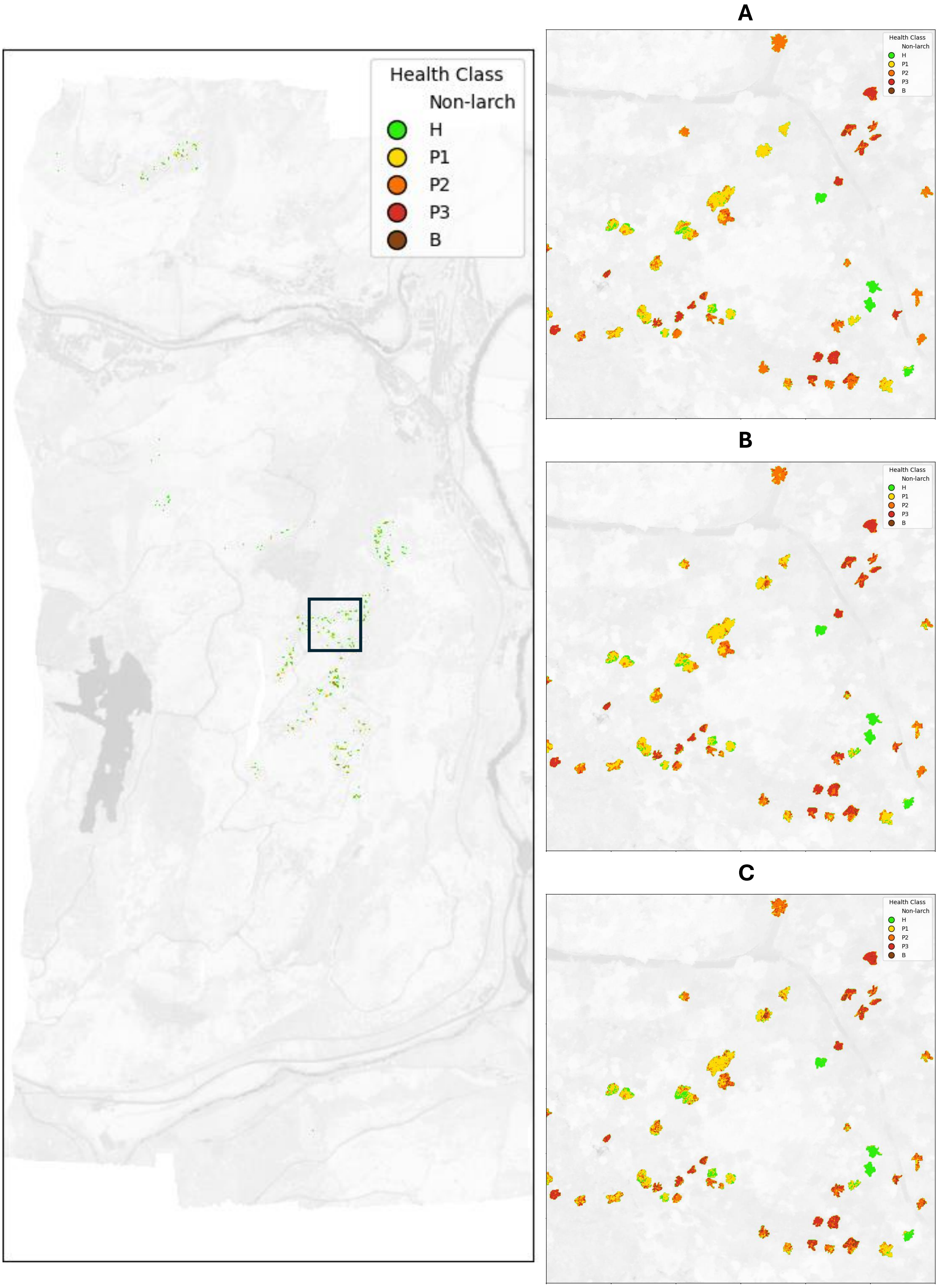

Figure 12: Mapping potential extent of *P. ramorum* in Gwydyr Forest. A whole forest view is provided, making use of the best performing classification model (RF at 0.33m resolution) for predictions, along with zoomed in images detailing the predictions of all three models in a four hectare area of the forest (highlighted by black box on left hand image). **(A)** RF, **(B)** SVM and **(C)** ANN.

of particular importance in this application as it is important that the number of trees that are identified as healthy when they are not is minimised. Whilst the RF classifier generally does a good job of predicting healthy pixels correctly, a slightly contradicting result is seen in the P1 class - approximitely 20% of pixels in this class are false negatives labelled as healthy. For a model to be used with confidence for this application, it would be desirable to minimise this particular category of misclassification so that the disease can be caught early and its impact minimised. Despite the underwhelming performance of the RF classifier in this regard, it still performed better than the other two classifiers. A much lower percentage of pixels in the more advanced stages of *P. ramorum* infection are labelled as healthy, which highlights that it is the early symptoms of the disease that the RF classifier struggles with in particular. Whilst the RF classifier is not as effective at predicting intermediate levels of infection by *P. ramorum* in larch trees (F1-scores < 0.70), it performed better than the alternative classifiers at both resolutions. Confusion between neighbouring larch disease classes is something that Barnes et al. (2017) also experienced when they utilised airborne laser scanning data of tree crown structure to determine *P. ramorum* infection severity in larch trees. They found that isolating instances of moderate and severe infection with this method was effective, however, isolating trees displaying early stages of infection from healthy trees was more difficult. SVMs are particularly useful for classification problems involving small to medium-sized datasets, which is defined by Geron (2022) as being those that contain hundreds to thousands of instances, which could help to explain the SVM classifiers' poor performance on this dataset containing tens of thousands of instances.

6.2 Collective performance across health classes

Considering the visual assessment categories in Table 3, it could be argued that health classes H, P3 and B should be predicted correctly most consistently because the pixels within tree crowns of these classes should be most homogeneous. We do see that health classes H and P3 are the most consistently identified pixel classes, however, health class B has the worst F1-score for all models. Although the classification performance against this health class is disappointing (all F1-scores below 0.60), it could be argued that this is the health class that is least important for the classifiers to be able to identify. This is because tree mortality caused by fire damage is not contagious and mapping of such instances is therefore of less urgency than mapping damage caused by *P. ramorum*. Interestingly, the F1-score rankings by health class for each model are very similar to the rankings of the number of labelled pixels in each health class (Table 4) for all models. This could suggest that there is a relationship between the number of labelled samples in each class and how well a model is able to accurately predict classes.

A large proportion of the pixels belonging to health class B that are misclassified are confused with class P3. This is true for all models and is a significant shortfall of the SVM models in particular. Interestingly, a much lower proportion of pixels belonging to class P3 are confused with class B pixels. Class P3 is the second most common pixel class (19.76%), whilst class B is the least common pixel class (9.27% pixels). This imbalance in the quantity of labelled pixels in the dataset could explain why class B pixels are more commonly confused with class P3. There are plenty of studies that conclude that RF classifiers are sensitive to class imbalances (Dalponte et al., 2013; Chen and Breiman, 2004), though Belgiu and Drăguţ (2016) note that there is no consensus on this effect. Though the preference would be for a classifier that is capable of consistently distinguishing between different classes with high accuracy, it is fortuitous that the confusion between these two classes is in the more desirable direction (fire damaged pixels confused more often with pixels exhibiting advanced symptoms of infection). As noted before, this is because the main objective of these classifiers is to identify cases of *P. ramorum* infection as early as possible to prevent the spread of the disease. False negatives in the fire damaged class are therefore unlikely to have negative consequences in achieving this objective.

The confusion matrices for each model highlight that adjacent classes of *P. ramorum* infection, particularly in the early stages, are commonly mistaken with one another. One possible explanation for this is that the health classes are determined using a whole crown view, so that any crown exhibiting branch tip die back or discolouration of foliage, would have all pixels making up the crown labelled as showing symptoms of *P. ramorum* infection. In reality, perhaps only a fraction of the pixels within a crown might show symptoms of the disease at early stages and this would mean that parts of the crown that appear healthy are labelled as diseased which could cause some confusion when using pixel based classification methods for assessing the health class of whole tree crowns. Another possible explanation is that class P1 and P2 are intermediate classes of *P. ramorum* in larch and could therefore suffer more from inconsistency in the qualitative assessment of tree health during field work. Whilst trees that appear healthy or severely damaged by *P. ramorum* can be identified as fitting in the most healthy (H) and least healthy (P3) classes relatively easily, it could be more difficult for assessors to determine the disease class of trees fitting in the centre of the spectrum. Though these explanations may highlight a weakness in the qualitative approach taken to label tree health, the alternative approach of lab testing tree samples at a large scale to provide conclusive evidence of the disease would be both time consuming and costly. Classification confusion between adjacent health classes is not isolated to this particular host and disease. Fassnacht et al. (2014) used SVMs to identify different stages of tree mortality caused by bark beetle and, similarly to this study, found that late stages of mortality were more consistently identified than earlier stages of damage. The confusion of health classes at earlier stages of damage was attributed to these classes being doubly confused with two neighbouring classes as opposed to one.

With all of this in mind, it may be pragmatic for future work to utilise a more binary approach to mapping the disease like Fassnacht et al. (2014). This would limit the classifiers to only being able to predict the presence, and not the severity, of the disease but if it improved the overall classification accuracy, and maximised precision, recall and F1-scores in the healthy and diseased classes, it may prove beneficial as a management strategy. If keeping multiple classes of *P. ramorum* infection is desirable, utilising contextual information may help improve the classification accuracy. In Figure 12 the enlarged tiles all show signs of "salt and pepper" noise - this describes the presence of isolated, generally misclassified, pixels within larger regions of another class (Bischof et al., 1992). Providing classifiers with contextual information would allow the models to use the statistically dependent relationship between neighbouring pixels within images to improve classification accuracy (Chen and Ho, 2008). Contextual information could be integrated during the classification routine or applied as a post-processing measure (Stuckens et al., 2000). At the classification routine stage, contextual information would be used by the classifiers alongside the spectral information to calculate the class probabilities used to make the predictions. Contextual information provided during post-processing would generally involve using a majority filter to label pixels using the most popular class within their neighbourhood. In the case of neural networks, Paola and Schowengerdt (1997) found that a majority filter smoothing algorithm, which used the most common label within a 3 by 3 window of pixels to classify the central pixel, improved classification accuracy by over 2%. Bischof et al. (1992) found that utilising a neural network based approach to post-classification smoothing also led to an improvement in classification accuracy but without removing isolated pixels that truly belong to a different class to its neighbours. Finally, moving towards a classification approach that focusses on the health of tree crowns as a whole, much like the field-based assessments do, could lead to improved classification performance. This approach would undoubtedly make use of CNNs, which have proven useful in classifying forest health (Kunjie Wu and Lan, 2023; Gensheng Hu and Zeng, 2022; Xia et al., 2021).

6.3 Performance at different resolutions

Owing to the available VI raster images being at different spatial resolutions, it was necessary to upsample the lower resolutions raster images or downsample the higher resolution raster images for consistency. It was hypothesised that using a finer resolution in the models could result in more pure training pixels and therefore a classifier with improved accuracy. The hyperparameters for each classification algorithm at both resolutions were optimised and it was found that the optimal hyperparameters were identical in the case of all three algorithms. Though the 0.33m resolution RF classifier achieved slightly improved results compared with the 0.71m resolution RF classifier, overall there doesn't appear to be a consistent improved performance when the finer resampling resolution is used.

6.4 Limitations

Finally, it is worth noting that the study was somewhat limited by the raw hyperspectral data being inaccessible. Whilst previous work had determined that the six VI features used in this study were the most informative out of a wider selection of 17 features (including further VI rasters and the most prominent wavebands determined using PCA) (Taylor, 2022), it has been found in other studies that including additional features (up to 15 features) improved the overall accuracy of pixel-based classifiers (Fassnacht et al., 2014). Additionally, in order to utilise CNNs to overcome the issue of the noisy classifications seen in the pixel-based approach, we would require additional data labels to create the necessary training images for the models. High quality aerial imagery would therefore be required to enable the adequate labelling of additional pixels. For this reason, it may be useful for future work to source the raw hyperspectral data to enable the inclusion of additional features and data labels.

7 Conclusions

Overall, this study has demonstrated that pixel-based machine learning classifiers are able to detect the presence of *P. ramorum* in larch trees reasonably consistently. The models are less effective at detecting the presence of fire damaged trees and in particular struggle to differentiate the damage caused by fire from the damage caused by advanced *P. ramorum* infection. Random forest classifiers were found to be more effective at detecting *P. ramorum* than SVM or ANN classifiers, though the accuracies achieved by the algorithm in some of the health classes were underwhelming and would make reliance on such methods for managing the disease difficult. Using higher resolution data to train the classifiers does not appear to lead to much improvement in classification accuracy, though the 0.33m and 0.71m resolutions used in this study are similarly fine and so this result may not be so applicable where comparison datasets differ more considerably (for example if using satellite data versus airborne data). All models appear to classify pixels more accurately when they belong to classes containing a larger share of the overall data, though additional work to understand how balanced classes impact results would be necessary to confirm this.

Though there are obvious limitations with the classifiers used here, this work has provided valuable insight in how machine learning methods can be used to aid with management of *P. ramorum* within the UK. The author has a number of suggestions for future work that could provide potential improvements for such methods including:

- Balance the number of training instances in health classes to understand if class imbalance has influenced the level of confusion between classes.
- Utilise contextual information, in the form of post-classification smoothing or during the training routine, to understand if this reduces the salt and pepper noise exhibited within the crowns and if it improves the classification accuracies.
- Consider new training data, for example in a different location where *P. ramorum* is known to exist, to understand if the models generalise well to new cases.
- Understand if adding additional features into the model, for example in the form of other VI or hyperspectral bands, to understand if this improves classification accuracy.
- Consider the effects on classification accuracy of reducing the number of classes to two (one healthy, one infected with *P. ramorum*). If this change reduces the number of instances that are infected with *P. ramorum* but that are predicted to be healthy, this may represent an improvement in terms of usability in forest management.
- Utilise CNNs in future work to understand if image-based classifiers enable a more holistic view on the health of full tree crowns.

References

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R. and Sousa, J. J. (2017), 'Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry', *Remote Sensing* **9**(11).
URL: *https://www.mdpi.com/2072-4292/9/11/1110*

Arlot, S. and Celisse, A. (2010), 'A survey of cross-validation procedures for model selection', *Statistics Surveys* **4**, 40 – 79.
URL: *https://doi.org/10.1214/09-SS054*

Barnes, C., Balzter, H., Barrett, K., Eddy, J., Milner, S. and Suárez, J. C. (2017), 'Airborne laser scanning and tree crown fragmentation metrics for the assessment of phytophthora ramorum infected larch forest stands', *Forest Ecology and Management* **404**, 294–305.
URL: *https://www.sciencedirect.com/science/article/pii/S0378112717309362*

Belgiu, M. and Drăguţ, L. (2016), 'Random forest in remote sensing: A review of applications and future directions', *ISPRS Journal of Photogrammetry and Remote Sensing* **114**, 24–31.
URL: *https://www.sciencedirect.com/science/article/pii/S0924271616000265*

Benediktsson, J. A. (1999), Statistical and neural network pattern recognition methods for remote sensing applications, *in* 'Handbook of Pattern Recognition and Computer Vision', World Scientific, pp. 507–534.

Bischof, H., Schneider, W. and Pinz, A. (1992), 'Multispectral classification of landsat-images using neural networks', *IEEE Transactions on Geoscience and Remote Sensing* **30**(3), 482–490.

Brasier, C. and Webber, J. (2010), 'Sudden larch death', *Nature* **466**, 824–825.
URL: *https://www.nature.com/articles/466824a*

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**, 5 – 32.
URL: *https://doi.org/10.1023/A:1010933404324*

C. Huang, L. S. D. and Townshend, J. R. G. (2002), 'An assessment of support vector machines for land cover classification', *International Journal of Remote Sensing* **23**(4), 725–749.
URL: *https://doi.org/10.1080/01431160110040323*

Carter, G. A. and Knapp, A. K. (2001), 'Leaf optical properties in higher plants: linking spectral characteristics to stress and chlorophyll concentration', *American journal of botany* **88**(4), 677–684.

Chang, C.-C. and Lin, C.-J. (2011), 'Libsvm: A library for support vector machines', *ACM Trans. Intell. Syst. Technol.* **2**(3).
URL: *https://doi.org/10.1145/1961189.1961199*

Chen, C. and Breiman, L. (2004), 'Using random forest to learn imbalanced data', *University of California, Berkeley* .

Chen, C. H. and Ho, P.-G. P. (2008), 'Statistical pattern recognition in remote sensing', *Pattern Recognition* **41**(9), 2731–2741.
URL: *https://www.sciencedirect.com/science/article/pii/S0031320308001647*

Chiang, C.-Y., Barnes, C., Angelov, P. and Jiang, R. (2020), 'Deep learning-based automated forest health diagnosis from aerial images', *IEEE Access* **8**, 144064–144076.

Chollet, F. et al. (2015), 'Keras'.
URL: *https://github.com/fchollet/keras*

Cortes, C. (1995), 'Support-vector networks', *Machine Learning* .

Dalponte, M., Ørka, H. O., Gobakken, T., Gianelle, D. and Næsset, E. (2013), 'Tree species classification in boreal forests with hyperspectral data', *IEEE Transactions on Geoscience and Remote Sensing* **51**(5), 2632–2645.

Degerickx, J., Roberts, D., McFadden, J., Hermy, M. and Somers, B. (2018), 'Urban tree health assessment using airborne hyperspectral and lidar imagery', *International Journal of Applied Earth Observation and Geoinformation* **73**, 26–38.
URL: *https://www.sciencedirect.com/science/article/pii/S0303243418302095*

Del Frate, F., Pacifici, F., Schiavon, G. and Solimini, C. (2007), 'Use of neural networks for automatic classification from high-resolution images', *Geoscience and Remote Sensing, IEEE Transactions on* **45**, 800 – 809.

Deng, X., Tong, Z., Lan, Y. and Huang, Z. (2020), 'Detection and location of dead trees with pine wilt disease based on deep learning and uav remote sensing', *AgriEngineering* **2**(2), 294–307.
URL: *https://www.mdpi.com/2624-7402/2/2/19*

Dun, H. F., MacKay, J. J. and Green, S. (2024), 'Expansion of natural infection of japanese larch by phytophthora ramorum shows trends associated with seasonality and climate', *Plant Pathology* **73**(2), 419–430.
URL: *https://bsppjournals.onlinelibrary.wiley.com/doi/abs/10.1111/ppa.13821*

Ecke, S., Dempewolf, J., Frey, J., Schwaller, A., Endres, E., Klemmt, H.-J., Tiede, D. and Seifert, T. (2022), 'Uav-based forest health monitoring: A systematic review', *Remote Sensing* **14**(13).
URL: *https://www.mdpi.com/2072-4292/14/13/3205*

Fassnacht, F. E., Latifi, H., Ghosh, A., Joshi, P. K. and Koch, B. (2014), 'Assessing the potential of hyperspectral imagery to map bark beetle-induced tree mortality', *Remote Sensing of Environment* **140**, 533–548.

Forest Research (2019*a*), 'About the NFI'.
URL: *https://www.forestresearch.gov.uk/tools-and-resources/national-forest-inventory/about-the-nfi/*

Forest Research (2019*b*), 'Ramorum disease (phytophthora ramorum)'.
URL: *https://www.forestresearch.gov.uk/tools-and-resources/fthr/pest-and-disease-resources/ramorum-disease-phytophthora-ramorum/*

Forest Research (2023), 'Forestry statistics 2023, chapter 1: Woodland area 'i&' planting'.
URL: *https://cdn.forestresearch.gov.uk/2023/09/Ch1_Woodland.pdf*

Gensheng Hu, Tongxiang Wang, M. W. W. B. and Zeng, W. (2022), 'Uav remote sensing monitoring of pine forest diseases based on improved mask r-cnn', *International Journal of Remote Sensing* **43**(4), 1274–1305.
URL: *https://doi.org/10.1080/01431161.2022.2032455*

Geron, A. (2022), *Hands-on Machine Learning With Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, MIT Press.
URL: *http://www.deeplearningbook.org*

Green, S., Cooke, D. E., Dunn, M., Barwell, L., Purse, B., Chapman, D. S., Valatin, G., Schlenzig, A., Barbrook, J., Pettitt, T., Price, C., Pérez-Sierra, A., Frederickson-Matika, D., Pritchard, L., Thorpe, P., Cock, P. J., Randall, E., Keillor, B. and Marzano, M. (2021), 'Phyto-threats: Addressing threats to uk forests and woodlands from phytophthora; identifying risks of spread in trade and methods for mitigation', *Forests* **12**, 1617.
URL: *https://doi.org/10.3390/f12121617*

Guimarães, N., Pádua, L., Marques, P., Silva, N., Peres, E. and Sousa, J. J. (2020), 'Forestry remote sensing from unmanned aerial vehicles: A review focusing on the data, processing and potentialities', *Remote Sensing* **12**(6).
URL: *https://www.mdpi.com/2072-4292/12/6/1046*

Guo, Q., Kelly, M. and Graham, C. H. (2005), 'Support vector machines for predicting distribution of sudden oak death in california', *Ecological Modelling* **182**(1), 75–90.
URL: *https://www.sciencedirect.com/science/article/pii/S0304380004004077*

Heermann, P. and Khazenie, N. (1992), 'Classification of multispectral remote sensing data using a back-propagation neural network', *IEEE Transactions on Geoscience and Remote Sensing* **30**(1), 81–88.

Heri Santoso, H. T. and Wang, X. (2017), 'Random forest classification model of basal stem rot disease caused by ganoderma boninense in oil palm plantations', *International Journal of Remote Sensing* **38**(16), 4683–4699.
URL: *https://doi.org/10.1080/01431161.2017.1331474*

Hu, W., Huang, Y., Wei, L., Zhang, F. and Li, H. (2015), 'Deep convolutional neural networks for hyperspectral image classification', *Journal of Sensors* **2015**.
URL: *https://doi.org/10.1155/2015/258619*

Huan Tao, Cunjun Li, D. Z. S. D. H. H. X. X. and Jing, W. (2020), 'Deep learning-based dead pine tree detection from unmanned aerial vehicle images', *International Journal of Remote Sensing* **41**(21), 8238–8255.
URL: *https://doi.org/10.1080/01431161.2020.1766145*

Huete, A. R. (2012), 'Vegetation indices, remote sensing and forest monitoring', *Geography Compass* **6**(9), 513–532.
URL: *https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-8198.2012.00507.x*

Jakob, S., Zimmermann, R. and Gloaguen, R. (2017), ‘The need for accurate geometric and radiometric corrections of drone-borne hyperspectral data for mineral exploration: Mephysto—a toolbox for pre-processing drone-borne hyperspectral data’, *Remote Sensing* **9**(1).
URL: *https://www.mdpi.com/2072-4292/9/1/88*

Kantola, T., Vastaranta, M., Yu, X., Lyytikainen-Saarenmaa, P., Holopainen, M., Talvitie, M., Kaasalainen, S., Solberg, S. and Hyyppa, J. (2010), ‘Classification of defoliated trees using tree-level airborne laser scanning data combined with aerial images’, *Remote Sensing* **2**(12), 2665–2679.
URL: *https://www.mdpi.com/2072-4292/2/12/2665*

Kingma, D. P. and Ba, J. (2017), ‘Adam: A method for stochastic optimization’.
URL: *https://arxiv.org/abs/1412.6980*

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* F. Pereira, C. Burges, L. Bottou and K. Weinberger, eds, ‘Advances in Neural Information Processing Systems’, Vol. 25, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45*
Paper.pdf

Kunjie Wu, Jiantao Zhang, X. Y. S. W. and Lan, Y. (2023), ‘An improved yolo model for detecting trees suffering from pine wilt disease at different stages of infection’, *Remote Sensing Letters* **14**(2), 114–123.
URL: *https://doi.org/10.1080/2150704X.2022.2161843*

Lausch, A., Erasmi, S., King, D. J., Magdon, P. and Heurich, M. (2017), ‘Understanding forest health with remote sensing-part ii—a review of approaches and data models’, *Remote Sensing* **9**(2).
URL: *https://www.mdpi.com/2072-4292/9/2/129*

Lin, Q., Huang, H., Wang, J., Huang, K. and Liu, Y. (2019), ‘Detection of pine shoot beetle (psb) stress on pine forests at individual tree level using uav-based hyperspectral imagery and lidar’, *Remote Sensing* **11**(21).
URL: *https://www.mdpi.com/2072-4292/11/21/2540*

Manfreda, S., McCabe, M. F., Miller, P. E., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., Ben Dor, E., Helman, D., Estes, L., Ciraolo, G., Müllerová, J., Tauro, F., De Lima, M. I., De Lima, J. L. M. P., Maltese, A., Frances, F., Caylor, K., Kohv, M., Perks, M., Ruiz-Pérez, G., Su, Z., Vico, G. and Toth, B. (2018), ‘On the use of unmanned aerial systems for environmental monitoring’, *Remote Sensing* **10**(4).
URL: *https://www.mdpi.com/2072-4292/10/4/641*

McCulloch, W. S. and Pitts, W. (1943), ‘A logical calculus of the ideas immanent in nervous activity’, *The Bulletin of Mathematical Biophysics* **5**(4), 115–133.

Meddens, A. J. and Hicke, J. A. (2014), ‘Spatial and temporal patterns of landsat-based detection of tree mortality caused by a mountain pine beetle outbreak in colorado, usa’, *Forest Ecology and Management* **322**, 78–88.

Melgani, F. and Bruzzone, L. (2004), 'Classification of hyperspectral remote sensing images with support vector machines', *IEEE Transactions on Geoscience and Remote Sensing* **42**(8), 1778–1790.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. and Fernández-Leal, Á. (2023), 'Human-in-the-loop machine learning: a state of the art', *Artificial Intelligence Review* **56**(4), 3005–3054.

Murfitt, J., He, Y., Yang, J., Mui, A. and De Mille, K. (2016), 'Ash decline assessment in emerald ash borer infested natural forests using high spatial resolution images', *Remote Sensing* **8**(3), 256.

Näsi, R., Honkavaara, E., Blomqvist, M., Lyytikäinen-Saarenmaa, P., Hakala, T., Viljanen, N., Kantola, T. and Holopainen, M. (2018), 'Remote sensing of bark beetle damage in urban forests at individual tree level using a novel hyperspectral camera from uav and aircraft', *Urban Forestry Urban Greening* **30**, 72–83. Special feature: Strategic gardens and gardening: Inviting a widened perspective on the values of private green space.
URL: *https://www.sciencedirect.com/science/article/pii/S1618866716305581*

O'Hanlon, R., Choiseul, J., Grogan, H. and Brennan, J. M. (2017), 'In-vitro characterisation of the four lineages of phytophthora ramorum', *European Journal of Plant Pathology* **147**, 517–525.
URL: *https://doi.org/10.1007/s10658-016-1019-2*

O'Reilly (2018), 'Chapter 1. introduction to artificial neural networks'.
URL: *https://www.oreilly.com/library/view/neural-networks-and/9781492037354/ch01.html*

Paola, J. D. and Schowengerdt, R. A. (1997), 'The effect of neural-network structure on a multispectral land-use/land-cover classification', *Photogrammetric Engineering and Remote Sensing* **63**(5), 535–544.

Pause, M., Schweitzer, C., Rosenthal, M., Keuck, V., Bumberger, J., Dietrich, P., Heurich, M., Jung, A. and Lausch, A. (2016), 'In situ/remote sensing integration to assess forest health—a review', *Remote Sensing* **8**(6).
URL: *https://www.mdpi.com/2072-4292/8/6/471*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Probst, P., Wright, M. N. and Boulesteix, A.-L. (2019), 'Hyperparameters and tuning strategies for random forest', *WIREs Data Mining and Knowledge Discovery* **9**(3), e1301.
URL: *https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1301*

Ribalta Lorenzo, P., Tulczyjew, L., Marcinkiewicz, M. and Nalepa, J. (2020), 'Hyperspectral band selection using attention-based convolutional neural networks', *IEEE Access* **8**, 42384–42403.

Rizzo, D., Garbelotto, M., Davidson, J., Slaughter, G. and Koike, S. (2002), 'Phytophthora ramorum as the cause of extensive mortality of quercus spp. and lithocarpus densiflorus in california', *Plant disease* **86**(3), 205–214.

Rizzo, D. M. and Garbelotto, M. (2003), 'Sudden oak death: endangering california and oregon forest ecosystems', *Frontiers in Ecology and the Environment* **1**(4), 197–204.

Roy, B. A., Alexander, H. M., Davidson, J., Campbell, F. T., Burdon, J. J., Sniezko, R. and Brasier, C. (2014), 'Increasing forest loss worldwide from invasive pests requires new trade regulations', *Frontiers in Ecology and the Environment* **12**(8), 457–465.
URL: *https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/130240*

Smartdraw (n.d.), 'Decision tree'. Accessed: 2024-08-26.
URL: *https://www.smartdraw.com/decision-tree/*

Stone, C. and Mohammed, C. (2017), 'Application of remote sensing technologies for assessing planted forests damaged by insect pests and fungal pathogens: a review', *Current Forestry Report* **3**, 75–92.
URL: *https://doi.org/10.1007/s40725-017-0056-1*

Stuckens, J., Coppin, P. and Bauer, M. (2000), 'Integrating contextual information with per-pixel classification for improved land cover classification', *Remote Sensing of Environment* **71**(3), 282–296.
URL: *https://www.sciencedirect.com/science/article/pii/S0034425799000838*

Sun, J., Yang, J., Zhang, C., Yun, W. and Qu, J. (2013), 'Automatic remotely sensed image classification in a grid environment based on the maximum likelihood method', *Mathematical and Computer Modelling* **58**(3), 573–581. Computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture 2012.
URL: *https://www.sciencedirect.com/science/article/pii/S0895717711006686*

Sun, W. and Du, Q. (2019), 'Hyperspectral band selection: A review', *IEEE Geoscience and Remote Sensing Magazine* **7**(2), 118–139.

Taylor, R. (2022), 'The application of remote sensing to monitor tree pests and diseases in welsh forests'. MRes Thesis.

Torres, P., Rodes-Blanco, M., Viana-Soto, A., Nieto, H. and García, M. (2021), 'The role of remote sensing for the assessment and monitoring of forest health: A systematic evidence synthesis', *Forests* **12**(8).
URL: *https://www.mdpi.com/1999-4907/12/8/1134*

Townsend, P. A., Singh, A., Foster, J. R., Rehberg, N. J., Kingdon, C. C., Eshleman, K. N. and Seagle, S. W. (2012), 'A general landsat model to predict canopy defoliation in broadleaf deciduous forests', *Remote Sensing of Environment* **119**, 255–265.

Václavík, T., Kanaskie, A., Hansen, E. M., Ohmann, J. L. and Meentemeyer, R. K. (2010), 'Predicting potential and actual distribution of sudden oak death in oregon: Prioritizing landscape contexts for early detection and eradication of disease outbreaks', *Forest Ecology and Management* **260**(6), 1026–1035.
URL: *https://www.sciencedirect.com/science/article/pii/S0378112710003580*

Welsh Government (2019), 'Phytophthora ramorum strategy for Wales'.
URL: *https://www.gov.wales/sites/default/files/publications/2021-01/phytophthora-ramorum-management-strategy.pdf*

Wong, T.-T. (2015), 'Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation', *Pattern Recognition* **48**(9), 2839–2846.
URL: *https://www.sciencedirect.com/science/article/pii/S0031320315000989*

Xia, L., Zhang, R., Chen, L., Li, L., Yi, T., Wen, Y., Ding, C. and Xie, C. (2021), 'Evaluation of deep learning segmentation models for detection of pine wilt disease in unmanned aerial vehicle images', *Remote Sensing* **13**(18).
URL: *https://www.mdpi.com/2072-4292/13/18/3594*

Xu, Z., Huang, X., Lin, L., Wang, Q., Liu, J., Yu, K. and Chen, C. (2020), 'BP neural networks and random forest models to detect damage by dendrolimus punctatus walker', *Journal of forestry research* **31**, 107–121.

Yu, S., Jia, S. and Xu, C. (2017), 'Convolutional neural networks for hyperspectral image classification', *Neurocomputing* **219**, 88–98.
URL: *https://www.sciencedirect.com/science/article/pii/S0925231216310104*

Zaniewski, A. E., Lehmann, A. and Overton, J. M. (2002), 'Predicting species spatial distributions using presence-only data: a case study of native new zealand ferns', *Ecological modelling* **157**(2-3), 261–280.