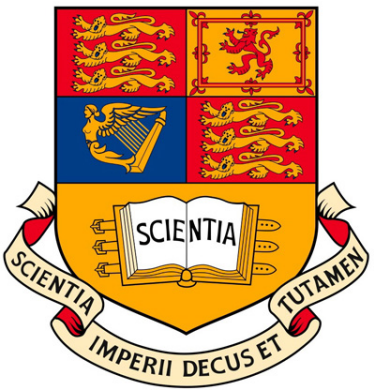
Classificaiton of complex stresses in plants using machine learning

by

Abigail Baines (CID: 00742373)
Supervisor : Oliver Windram, Imperial College London,
o.windram@imperial.ac.uk

Department of Life Sciences
Imperial College London
London SL5 7PY
United Kingdom



Thesis submitted as part of the requirements for the award of the MRes in Computational Methods in Ecology and Evolution, Imperial College London, 2017-2018

Formatted in the style of frontiers in Plant Science

Word count: 5455

Declarations:

Data provided by Dr Oliver Windram, Chris Adams and Abigail Baines

Data processing, cleaning and augmenting by Abigail Baines

All mathematical models developed by Abigail Baines
All interpretations of results and analyses were developed by Abigail Baines

Abstract

Determining the incidence and types of plant stresses has been of interest for many years, particularly due to their negative impact on agricultural production. Despite this, current methods are very qualitative, with many relying on manual detection which can be highly variable. With this in mind, our aim for this project was to see whether we could use machine learning techniques, in particular Random Forests (RFs) and Convoltional Neural Networks (CNNs), to classify complex plant stresses from RGB and colour-pass filtered leaf images. Two plant leaf datasets were collected; an infection duration study conducted to assess the impact of *Botrytis cinerea* on *Arabadopsis thaliana* over a 72 hour period, and a complex multi-stress study, involving drought, nitrogen deficiency and *Botrytis cinerea*, conducted on tomato leaves. We found strong support that machine learning can correctly identify both infection duration and complex/multi stresses in an image; both RFs and CNNs produced training accuracies of >85%. Finally, it was noted that the use of colour-pass filters improve model performance. Overall the results obtained from this study will be greatly beneficial to both the botanical and horticultural communities, and to the future of crop production.

Keywords: Digital Pathology, Random Forests, Python, Disease detection, Multispectral features, Machine Learning.

1 Introduction

Recently, the field of digital pathology has exploded; searching digital pathology on Web of Science returns over 3,000 hits, and 2013 onwards returns over 250 papers per annum. Despite this, plant pathology lags behind in using this technology; searching for "plant" within digital pathology only yields 18 results, in comparison to 455 for searching "human". This discrepancy is in part due to simpler disease detection methods, such as manual identification, being so heavily relied upon within the field of plant pathology up until fairly recently. (Lowe, Harrison and French 2017).

Plant Pathology

The study of plant pathology has always received much attention, due to it's link with agricultural yield and crop productivity (Waller et al. 2005; Donatelli et al. 2017). Plant pathology is similar to other streams of pathology; susceptibility is the antithesis of resistance, with resistance being that a plant can overcome or exclude pathogenic effects/organisms. The cycle of resistance and susceptibility is powered by natural selection; plants become able to attack invading pathogens via their immune response, then pathogens evolve so that they can suppress this triggered immunity. This perpetual cycling results in a constant struggle to maintain resistance and virulence in host and pathogen respectively (Zheng et al. 2012; Lapin and Van den Ackerveken 2013; Pel and Pieterse

2013). Artificial selection in plants has routinely been used in agriculture, in order to introduce resistance to specific diseases into crops (Dennis et al. 2008).

Despite the existence of numerous control measures for many crop diseases (Wood 1996) , those that still manage to impact upon crop production can be devastating (Woodham-Smith 1962; McCook 2006), with a prominent example being that of the Great Famine in Ireland, caused by potato blight. The total death toll during the 4 year period (1845-1849) reached over a million, with many more emigrating(Woodham-Smith 1962).

It has been well documented that abiotic stress conditions also cause extensive losses to crop production (Ron Mittler 2006); nitrogen deficiencies, drought, and excess UV exposure are amongst several other factors influencing both plant immune system responses and plant growth (Karimi et al. 2006; Santos et al. 2018). Although these abiotic conditions receive much research attention, the research conducted is often on individual stresses (Zhao et al. 2005; Santos et al. 2018). In the field, crops are routinely being subjected to various combinations of stresses, both biotic and abiotic (Ron Mittler 2006), meaning complex stress experiments are becoming essential research priorities.

Due to the complex nature of stresses affecting plant growth, it can often be difficult to identify the cause of the stress; this is particularly apparent when multiple factors are influencing the phenotype expression. This is further complicated by the fact that different pathogen races and variable environmental conditions can radically affect the degree of disease/ stress severity (Dangl and Jones 2006; Suzuki and Ron Mittler 2006). Recent studies have revealed that responses of plants exposed to multiple stresses are unique, and cannot be extrapolated from that of responses of plants exposed to the individual stresses (see: Rizhsky, Liang and R. Mittler 2002; Ron Mittler 2006). These findings make stress/ infection identification particularly complicated in the field; even when stress combination experiments are conducted, these primarily occur in the lab, with conditions that do not compare well to the field.

Past methods in detecting plant disease and other stress-inducing factors rely on experts simply observing the plant (V. Singh and Misra 2017); despite the rise in the use of machine learning for classification, many still rely on manual detection in this field. Manual detection, despite being easily applied in the field, is highly qualitative, and therefore is subject to interpretation. It is also noted that this method cannot quantify the level of disease severity, nor infection duration, accurately (Lowe, Harrison and French 2017). Manual identification also cannot correctly identify complex stresses or masking/ hidden factors, such as underlying root rot. Finally, manual detection limits the ability to compare data and results, due to the nature of the identification method.

Machine Learning

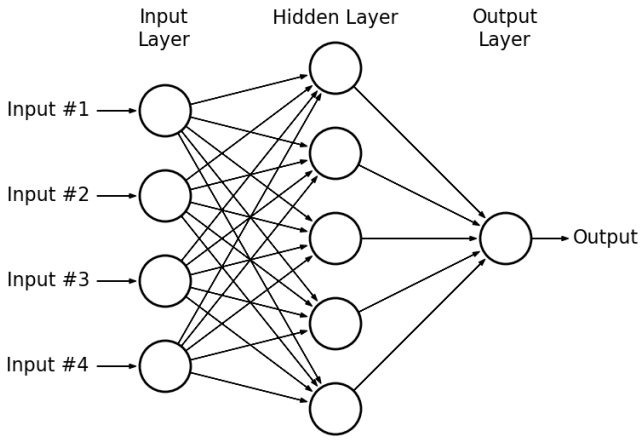
Due to the limitations of manual detection, more recent studies have been using Machine Learning algorithms (Awate et al. 2015; Baranowski et al. 2015; Mohanty, Hughes and

Salathe 2016; Lowe, Harrison and French 2017). Mohanty *et al* (2016) used a CNN to detect the presence/absence of 26 diseases, including leaf blight and tomato mosaic virus, in 14 crop species. The dataset used was open source (obtained from PlantVillage), and contained 54,306 images in total; the highest accuracy achieved was 99.35%, whilst a model trained on images under one condition and tested under another yielded results of only 31.4%. This discrepancy was primarily due to the factors affecting the image being taken, such as daylight, light exposure and type of camera being used; using more images and normalising images is supposed to minimise this discrepancy (Ciresan, Meier and Schmidhuber 2012; Krizhevsky, Sutskever and E. Hinton 2012). Another study, conducted by Zhu *et al* 2016), looked into early detection and classification of Tobacco Mosaic Virus; a range of machine learning algorithms were used, including Random Forests, Support Vector Machines and Back Propagation Neural Networks, with most models returning accuracy ratings of 85% or higher.

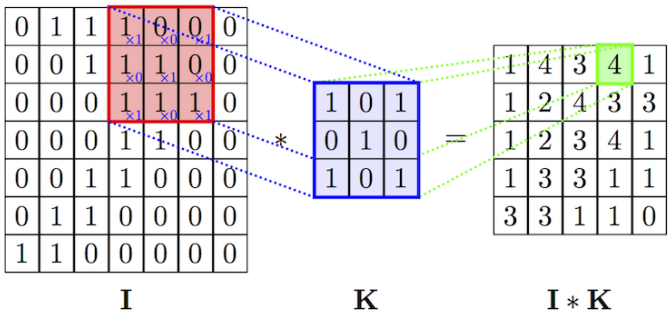
Neural Networks (NNs) have become particularly important in the field of Plant Pathology, and Pathology in general, in more recent years (Awate et al. 2015; Sladojevic et al. 2016). A NNs basic structure consists of 3 types of layers, an input layer, one or many hidden layers and, an output layer, with each layer made up of 'neurons' (Figure 1a). Convolutional Neural Networks (CNNs), a subclass of NNs, have received much attention in the field of image recognition; in recent years, CNN's have out-ranked many other classification algorithms and achieved best performances on many major image recognition benchmarks (Simard, Steinkraus and Platt 2003; Krizhevsky, Sutskever and E. Hinton 2012). One such benchmark is the ImageNet classification challenge, where models are built to classify 10,000,000 labelled images containing 10,000+ classes; top error rates are primarily achieved by CNNs (Krizhevsky, Sutskever and E. Hinton 2012; Russakovsky et al. 2015).

CNNs also consist of 'neurons'; these neurons primarily receive inputs in the form of 2/3D matrices (e.g. single channel images, RGB images, image data), apply specific functions and biases, and return an output. CNNs contain layers called convolutional layers (these are within the aforementioned hidden layer(s)); within this layer a matrix dot function is applied using a matrix of size h (height) x w (width) (known as a convolution kernel/matrix, K in Figure 1b) on an image or image data, I (Figure 1b).

By overlaying the kernel on top of the image in all possible ways, the convolutional layer is able to extract valuable information from the image, including edges and shape (Zeiler and Fergus 2014). This is possible in part due to the convolutional matrix, K; the matrix values are calculated by model to extract useful information for processing and classifying (Figure 1b). This is process is aided by the non-linear activation functions applied to CNN layers, such as the ReLu (Rectified Linear unit) function. These functions allow for true learning to take place, by allowing different input transformations to occur, which are learnt from the dataset itself (Punjani and Abbeel 2015).



(a) Basic example of a Neural Network, with a simple input, hidden and output layer. The Input Layer contain the split image(s), which then progress onto the Hidden layers. The output layer is a single neuron containing a prediction made by the hidden layers. This example is most likely of a simple regression or binary classification model; multi-class classifiers require one output neuron for each class.



(b) Basic example of how a convolutional layer is created. A matrix dot function is applied to the input layer (I) using a convolutional matrix (K); the resulting output layer is the matrix I*K.

Figure 1: Figures showing the basic layout of a Neural Network, (a), and a convolutional layer (b), found within the hidden layer of a network. All neurons within each layer connect to previous layers, as well as to future layers; as the layers 'learn', interaction strengths vary depending on how well each neuron classifies the data input into the model.

Despite their excellent performance, CNNs require more resources in terms of computing power, as well as more time to train in comparison to other machine learning algorithms (e.g., K-nearest neighbours, Random Forests). This problem is further exacerbated because CNNs are difficult to design correctly, with many networks being created in the pursuit of the final model. Often, a balance must be struck between model factors (such as number of layers, layer type and parameters used) and the risk of over-fitting.

Another commonly used algorithm in Plant Pathology is a Random Forest model (Baranowski et al. 2015). Despite returning lower accuracies than CNNs (Baranowski et al. 2015), they are still used and relied upon as a classifier; Ilastik (Sommer et al. 2011), an 'interactive learning and segmentation toolkit', is developed in Python and utilises Random Forest Classifiers. It is also a primary tool in classification in both Human and Plant Pathology (see: Kleesiek et al. 2014; Haubold et al. 2016; Visschers, Dam and Peters 2018).

Image data

Despite the excellent track records of Machine Learning algorithms as image classifiers, the images used are still essential to creating a correctly classifying model. To ensure that the classifier does not overfit the data, images are often augmented prior to training (Simard, Steinkraus and Platt 2003; Ciresan, Meier, Masci et al. 2011; Ciresan, Meier and Schmidhuber 2012; Sladojevic et al. 2016); the images are often augmented in such a way as to retain the structure of the image (i.e. transformations, rotations) whilst also providing additional test data. Images can also be normalised in an attempt to further reduce over-fitting and to allow comparison between datasets.

The quality of the image is also essential when it comes to building a classifier; Mou *et al* (2017) note that hyperspectral imagery has started collecting considerable attention in the past few decades, in part due to it's high spectral resolution. Both multispectral

(a) Dataset taken under RGB sensor

(b) Dataset taken under RGB sensor and CalColor 90 pink colour-pass filter

(c) Dataset taken under RGB sensor and CalColor 90 pink colour-pass filter

Figure 2: Samples of multi-stress datasets taken under different filters. Using colour-pass filters (b and c) highlight chlorosis occurring in the leaves more evidently than in a.

and hyperspectral image data have been used in CNNs, and have achieved great results (Harsanyi and Chang 1994; Delalieux et al. 2007; Baranowski et al. 2015; H. Zhu et al. 2016; Lowe, Harrison and French 2017; Mou, Ghamisi and X. X. Zhu 2017). Hyperspectral and multispectral imaging work by acquiring the light intensity for a large number of spectral bands, as opposed to just three (Red, Green and Blue) for a normal camera; this therefore allows for more data to be collected per image. Although these methods have attracted attention over the recent decade, these imaging techniques are still costly. Another method for extracting valuable information from data are band-pass (Interference filters) and colour-pass filters (high pass filters; Figure 2); these work by filtering image wavelengths, and allowing only specific wavelengths to pass whilst others are blocked. Band-pass filters are non-coloured and can allow very narrow wavelengths through; colour-pass filters are coloured glass and work according to the principle of subtractive colour mixing (Sischka 2018). These filters allow for greater contrast within the image (Figure 2), and are gaining in popularity in image analysis (Piron et al. 2008; Knoth et al. 2013).

Study - Organisms and Relevance/Outline

Several methods are currently being used to study disease classification; more advanced models (H. Zhu et al. 2016) are producing higher classification accuracies, and better quality images allow for additional information to be extracted (Baranowski et al. 2015; Lowe, Harrison and French 2017). This paper will utilise both advanced models (CNNs and RFs), and high quality images produced using colour-pass filters.

Previous studies have used machine learning techniques to classify simple stresses; this study will be classifying the impact of multi-stresses on tomato (*Solanum lycopersicum*) leaves. *Botrytis cinerea*, a common fungal infection, and both drought and nitrogen deficiency were used as the stresses in this image dataset. A further study will look at the effect of infection duration on *Arabadopsis thaliana*.

Arabadopsis thaliana and Tomatoes were used as they are common study hosts in both host-pathogen and host-stress studies (Tank and Saraf 2010; A. Singh et al. 2016; Choi et al. 2017); both have also been used in studies involving *Botrytis cinerea* infection and resistance (Diaz, Have and Kan 2002; Choi et al. 2017).

Botrytis cinerea is a commonly studied disease with distinguishable characteristics (Awate et al. 2015); by using this fungal infection, we should obtain classifiable results, whilst also adding to the growing body of research in plant-host susceptibility and resistance. Finally, the abiotic stresses were chosen due to their notable effect on plant growth (drought:Santos et al. 2018, nitrogen: Zhao et al. 2005), and due to their ease of application.

Aims of Project

This study was undertaken to see whether machine learning can correctly classify instances of complex stress exposure in tomato leaves, as well as duration of infection of *Botrytis cinerea* on *Arabidopsis thaliana*.

Our study aims to address five questions: Can machine learning classify instances of complex stress (question 1) and infection duration (question 2), will including colour-pass filtered images affect the model outcomes (question 3), will models trained on normalised images produce different results (question 4), and, will CNNs differ in model outcomes in comparison to RFs (question 5)?

Based on previous studies looking into machine learning algorithms as plant disease classifiers (Awate et al. 2015; Baranowski et al. 2015; Sladojevic et al. 2016; V. Singh and Misra 2017; Ubbens and Stavness 2017), and on other image classification benchmarks that have been achieved using these algorithms (see:Krizhevsky, Sutskever and E. Hinton 2012), we believe both CNNs and RFs will be able to correctly classify (>85% training accuracy) both complex stresses (hypothesis 1) and infection duration (hypothesis 2) images. Including colour-pass filtered images should improve the classifiers accuracy (hypothesis 3); band-pass filters have already been used successfully in plant pathology (see: Piron et al. 2008) with colour-pass filters applying similar, but subtler, techniques. Models trained on the normalised dataset should have an improved accuracy rating in comparison to the original model accuracies (hypothesis 4); normalising the data should reduce the potential for over-fitting, and should also allow the model to be used on different datasets without requiring additional information. Due to the staggering performances of the top CNN models in many image classification challenges (for example, see: Krizhevsky, Sutskever and E. Hinton 2012), we also hypothesise that CNNs will produce better model outcomes in comparison to RFs (hypothesis 5).

2 Materials & Methods

Data and composition

The data used was collected during 2009 and 2017; the 2009 data was collected by Dr Oliver Windram, and the 2017 data was collected by myself and Chris Adams.

The 2009 data was collected to look at the effect of *Botrytis cinerea* on *Arabodopsis thaliana* (Figure 4); the data was collected over a period of 72 hours, allowing us to look at the effect of duration of infection. The time periods for this dataset were 10hpi (hours post infection), 18hpi, 26hpi, 34hpi, 42hpi, 48hpi, 72hpi and non-infected/ control. Images in this dataset were taken under a RGB sensor only.

The 2017 data was collected to look at the effect of multiple stressors on tomato leaves (Figure 5); a combination of drought stress, Nitrogen deficiency, and *Botrytis cinerea* were applied (see Table 1 for each treatment type) and left for 72 hours, after which several images were taken, including a RGB image and a range of images under colour-pass filters (appendix table 1).

Both datasets were then cropped to contain individual leaf images; once this was done each individual image was processed, as laid out in Figure 3, ready for image classification.

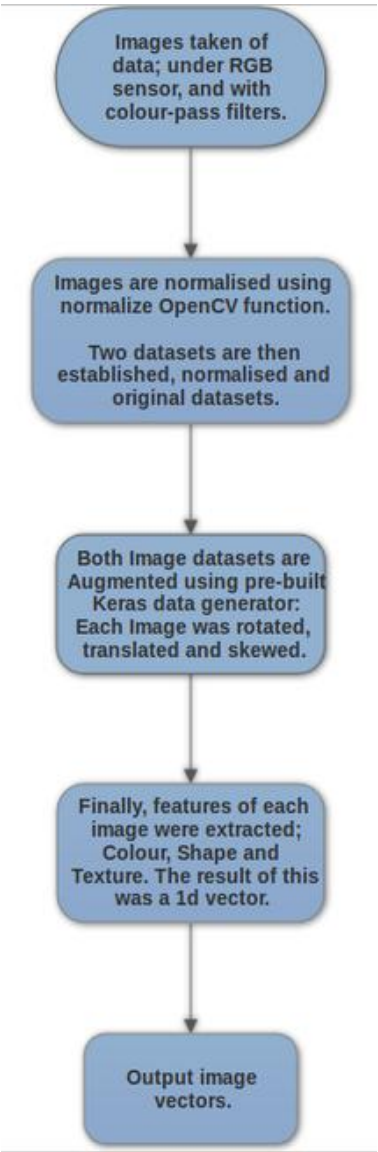


Figure 3: Outline of the workflow from raw images to final datasets.

Colour-pass filters

Part of this papers aim was to ascertain whether colour-pass filters bring out more visual data than just RGB images alone; to test this, three multi-stress datasets were created, with increasing numbers of colour-pass filters used in each. The colour-pass filters were held over a normal RGB sensor camera (Figure 3).

The three datasets consisted of a RGB image dataset (dataset 1), a RGB, purple and magenta image dataset (dataset 2), and a dataset consisting of RGB images and all color-pass filters (dataset 3). The colour-pass filter colours used are noted in Table 1 in the Appendices.

The duration infection dataset did not contain additional filters; only one dataset (dataset 4) was created, containing RGB images.

Table 1: Treatment types for tomato leaves in multi-stress study. Stresses used were Drought, Nitrogen deficiency and *Botrytis cinerea* infection.

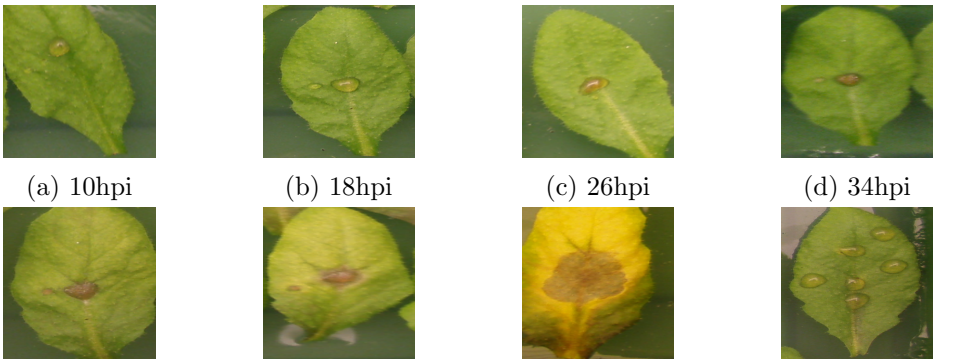
		Drought			
		Present	Absent	Present	Absent
Nitrogen	Present	+++	++ -	+ - +	+ - -
	Absent	- + +	- + -	- - +	- - -
		Present		Absent	
		Infection			

Image Normalisation

In order to allow for easier comparison between datasets, Dataset 3 was normalised using the normalize function in OpenCV Python (Figure 3). Each image was normalised between absolute black (0) and absolute white (255); the original images were also kept to allow for results comparison. This dataset is referred to as Dataset 3-NORM.

Data Augmentation - Reducing over-fitting

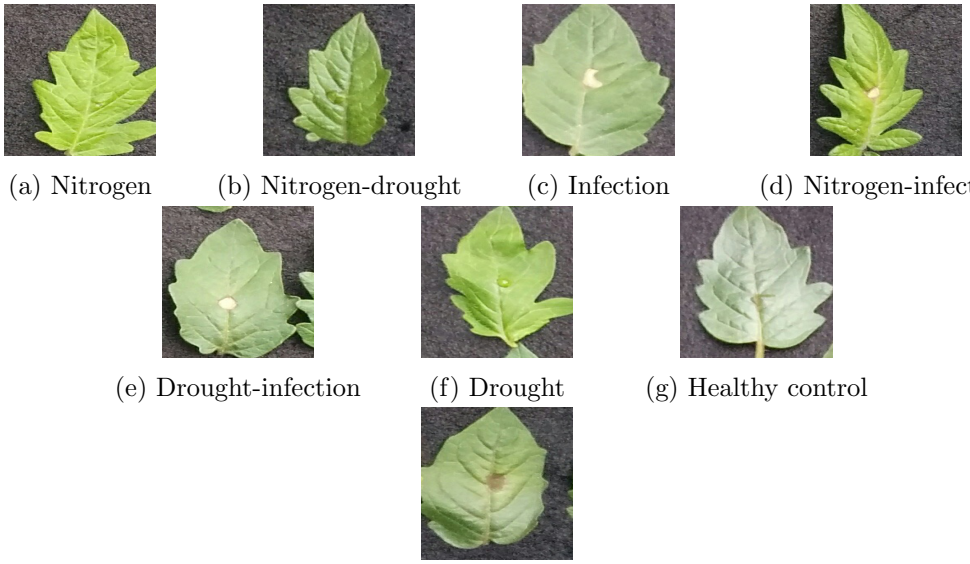
One of the easiest and most common methods of minimising the risks of over-fitting on image data, is to artificially enlarge the dataset(s) (Figure 3: Simard, Steinkraus and Platt 2003; Krizhevsky, Sutskever and E. Hinton 2012; Sladojevic et al. 2016; Ubbens and Stavness 2017). Another benefit of this approach is that data generation/ augmentation has been noted to improve classifier performance (see:Yaeger, Lyon and Webb 1996, or Krizhevsky, Sutskever and E. Hinton 2012), by increasing the chance for the classifier to learn appropriate features of each treatment (Sladojevic et al. 2016). We augmented the



(a) 10hpi (b) 18hpi (c) 26hpi (d) 34hpi

(e) 42hpi (f) 48hpi (g) 72hpi (h) Healthy control

Figure 4: Sample images of *Arabodopsis thaliana* infected with *Botrytis cinerea*. Individual image captions show duration of infection.



(a) Nitrogen (b) Nitrogen-drought (c) Infection (d) Nitrogen-infection

(e) Drought-infection (f) Drought (g) Healthy control

(h) Nitrogen-drought-infection

Figure 5: Sample images of tomato leaves infected with *Botrytis cinerea*, and stressed with drought and nitrogen deficiency. Individual image captions show combination of stresses.

dataset using a pre-built data generator in the Keras library (Chollet 2015); each image was randomly augmented 20 times using simple distortions such as translations, rotations, and skewing.

Feature extraction

Prior to training, features of each image dataset (Figure 3: shape, colour, texture) were extracted; this was done as preliminary training was returning poor classification results when trained on the original image dataset. This method of feature extraction is also done to reduce the computational workload, and to further minimise over-fitting. After the features were extracted, each image's data consisted of a 532 x 1 array.

Test images for each dataset were kept separate from the training datasets; these images consisted of data from the original experiment and from other experiments with the same treatment types.

The final datasets consisted of 3,600 duration infection image vectors; 450 for each stage of infection, and 7,952 multi-stress image vectors; 994 for each treatment. These datasets were split by 80:20 for training/validation of the models. There were also 42 multi-stress test images, and 209 duration infection test images not input into any models; these were used to test the final models for a final accuracy/prediction score.

Data modelling

Image classification problems have utilised various machine learning algorithms in the past, ranging from Linear Regression to Deep CNNs (Naseem, Togneri and Bennamoun

2010; Krizhevsky, Sutskever and E. Hinton 2012; Tripathi and Maktedar 2016). The two main classification algorithms investigated in this project are Random Forest classifiers, and Convolutional Neural Networks.

Both algorithms have been successfully applied for image classification tasks (Krizhevsky, Sutskever and E. Hinton 2012; Malof et al. 2016), and are currently being utilised in the field of plant pathology (Awate et al. 2015; A. Singh et al. 2016).

Machine Learning

To get the best classifier comparison, several classifiers were trained on the image data, including Random Forests, K-Nearest-Neighbours and Support Vector Machines. An ensemble approach was also conducted, consisting of Decision trees, Random Forests and KNNs. Random Forests produced the best results, and so further testing was using RF alone.

Each RF classifier had the same random state, and consisted of 100 trees unless otherwise stated. All RF classifiers underwent a ten fold cross validation to ensure over-fitting was minimised; accuracy of each classification was used as the scoring method.

The CNN used in this study (Appendix Figure 3) consisted of many layers including several convolutional layers, a dropout layer and, a dense layer. A Relu activation function was used in all convolutional and dense layers; this activation function was used to model nonlinearity in the data, as it can be computed faster than more traditional functions, such as sigmoid or hyperbolic tangent functions. The dropout layer had a 40% dropout function, and the learning rate was 15% unless otherwise stated. The Tensorflow framework (Abadi et al. 2016) within Python was used to create this CNN. The batch size (the number of training examples in one forward pass through the model) was 100.

(a) Original RGB Image (b) Masked Image

Figure 6: Original RGB Image before and after applying the background mask. By masking the background, Individual leaf pixel intensities can be gathered.

All CNN models were stopped between 3000 and 5000 training iterations (one pass through the model, with each pass using batch size number of examples); accuracies (number of correct predictions in the batch, over batch size) had exceeded 90% by this time, and model loss (the number of inaccurate predictions) was <20% (Figure 5). Allowing the model to run for excess iterations did not improve the training accuracy nor did it drastically improve the loss; stopping the models allowed for over-fitting to be minimised whilst also allowing model losses to remain low.

Models were adjusted to assess impacts on model accuracies; the number of trees in the Random Forest model were altered, as were the CNN models learning rates (Table 2).

Pixel Intensities

In order to further validate the predictions the classifiers made, and to ensure that there are valid differences between each treatment, pixel intensities were gathered.

To do this, the background was first masked, so that only the leaves were visible (Figure 4). After this each leaf was labelled in relation to the treatment, and pixel intensities of each treatment leaf were gathered using Python 3.7 (Rossum and Drake 2001). These Pixel Intensities were then compared using an ANOVA.

Model comparison

To compare between classifiers, separate test datasets, consisting of 209 duration images, and 74 treatment images were used to obtain predictions. These predictions consisted of a predicted class and a percentage of certainty, which were also used in comparison analysis.

Statistical analysis

All statistical analyses were carried out in R version 3.2.3 (Team 2013). ANOVAS (Analysis of Variance) and subsequent Tukey post-hoc tests were used to explore variance between treatments within both datasets.

3 Results

It was noted prior to training that a dataset containing 8 classes can only achieve an average accuracy of 12.5% if randomly guessing. Human expert accuracies could not be reliably obtained as all individuals were involved in the data collection, but it is estimated that a 75% accuracy rating could be concluded. Table 2 shows all accuracy outputs obtained from all models; even the lowest training accuracy (CNN trained on Dataset 3 - RGB images and all colour-pass filtered images, learning rate 0.15 - 88.3%) supersedes both of the aforementioned accuracies, showing strong promise for the use of deep learning in future classification problems.

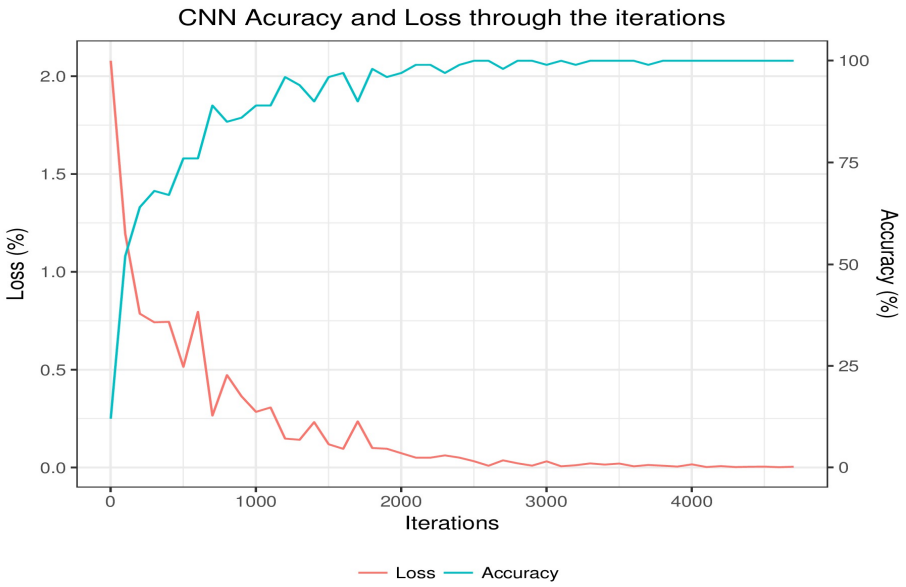


Figure 7: CNN accuracy and loss plotted over 4800 training iterations. The CNN plotted was classifying infection duration, which consisted of 8 treatment groups. The graph shows that after 2500 iterations, both accuracy and loss begin to level off, and by 3500 iterations, they both reach a plateau.

Testing accuracies were obtained by using the final models to predict the classes of unseen data (42 multi-stress images, 209 duration infection images). Testing accuracies were lower than training accuracies, but still exceeded that of random chance and human accuracy.

Increasing the number of trees in the RF model proved computationally more expensive, with final accuracies only slightly differing from the original outcomes (Table 2). Lowering the learning rate of the CNN proved worthwhile; testing/ validation accuracy increased 5%, with the risk of over-fitting also being minimised.

Table 2 shows that both RF and CNN models produced very similar results; when True Positive accuracy ratings were compared between RF and CNN models, CNN come out on top (Figure 6).

The pixel intensities gathered from the RGB images in the multi-stress and infection duration dataset show that the models were classifying correctly; all treatments were significantly different, thus allowing for further support for deep learning methods in future studies.

Multi-stress treatments

We wanted to test the ability of both CNNs and RFs to classify multi-stress images; all models were built and trained on the same set of multi-stress images, and tested on a separate set of images. The four versions of the multi-stress image dataset (1 - RGB, 2 -

RGB and two colour filters, 3 - RGB and all filters, 3-NORM - normalised dataset 3) show slight variation in performance (Table 2). All training accuracies were >90%, superseding the theoretical random classifier (12.5%) and human expert (75%).

Both RFs and CNNs returned high (>88%) training accuracies for all datasets; training accuracies decreased as datasets contained more colour-pass filter images (Table 2). Despite this, testing/ validation accuracies increased as more colour-pass images were included.

Models trained on the normalised dataset (Dataset 3-NORM) returned lower accuracies for both testing and training in comparison to Dataset 3 (Table 2).

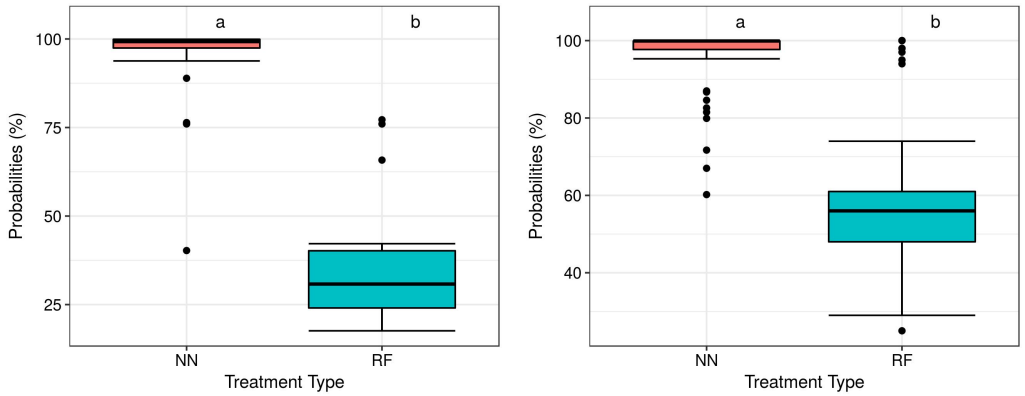
As both CNN and RF models were returning similar accuracy ratings, the percentage certainties of correct predictions for were compared (Figure 6a). The outcome of this comparison was that the CNN models predictions were significantly more certain than that of the RFs (ANOVA: $F_{1,50} = 218.1$, p <0.0001).

Duration Infection

To test the ability of both CNNs and RFs in their ability to classify infection duration images, several models were built and trained on the same set of infection duration images, and tested on a separate set of images. The infection duration dataset (dataset 4) was correctly (>85%) classified by both models; both testing and training accuracies were >90%, far greater than random chance and human accuracies.

Table 2: Outputs of all models trained on all datasets (1-4); training and testing accuracies were compared between models to assess model fit.
Dataset 1 - Multi-stress images, RGB only, Dataset 2 - Multi-stress RGB and 3 colour-pass filters, Dataset 3 - Multi-stress RGB and all colour-pass filters and Dataset 4 - Infection duration dataset, RGB only.

Model Type	***Training Images***	**Training Accuracy (%)**	**Test Accuracy (%)**
CNN *Learning rate 0.2*	Dataset 1	99.1	31
	Dataset 2	99	48
	Dataset 3-NORM	92.2	55
	Dataset 3	93.9	57
	Dataset 4	98.6	94
CNN *Learning rate 0.15*	Dataset 3	88.3	62
Random Forest *100 trees*	Dataset 1	99.8	42
	Dataset 2	99.4	45
	Dataset 3-NORM	95.9	57
	Dataset 3	96	62
	Dataset 4	97.9	94
Random Forest *500 trees*	Dataset 2	99.4	48
	Dataset 3	96.3	60



(a) Accuracy results for treatment models trained on dataset 3. CNNs produced significantly more certain (ANOVA: $F_{1,50} = 218.1$, p <0.0001) accuracy results in comparison to RFs.

(b) Accuracy results for duration models trained on dataset 4. CNNs produced significantly more certain (ANOVA: $F_{1,120} = 216.3$, p <0.0001) accuracy results in comparison to RFs.

Figure 8: True Positive test accuracy results for RF and CNN models trained on multi-stress dataset 3 and duration dataset 4. True positive accuracies were obtained from test results of the models; correctly classified image accuracies were compiled for each individual model. The letters plotted on the graphs show the Tukey post-hoc comparisons.

Both RF and CNNs managed to return a >97% training accuracy and a 94% testing accuracy (Table 1); when true positive prediction certainties were compared (Figure 6b), CNNs were found to be significantly more certain in comparison to the RF model (ANOVA: $F_{1,120} = 216.3$, p <0.0001).

Pixel Intensities

Pixel intensities were extracted from dataset 1 (multi-stress, RGB only) and dataset 4 (infection-duration, RGB) images. These were extracted by masking the images, as shown in Figure 6, and extracting pixel intensity values for the RGB channels. Pixel intensities for both multi-stress and infection duration treatments show that all treatments are statistically different from one another, allowing for further support of the models classification results.

When pixel intensities were compared between treatments, both multi-stress and duration infection treatments returned significant results (ANOVA multi-stress: $F_{7,4646161} = 38676$, p <0.0001; ANOVA duration infection: $F_{2,1628925} = 32894$, p <0.0001); Figures 7 and 8 show that all treatment types were significantly different from one another. Further comparison shows that channels significantly differ both within and between treatments (ANOVA duration infection: Appendix Figure 1, $F_{8,1628919} = 3437540$, p <0.0001; ANOVA multi-stress: Appendix Figure 2, $F_{23,4646145} = 17195$, p <0.0001).

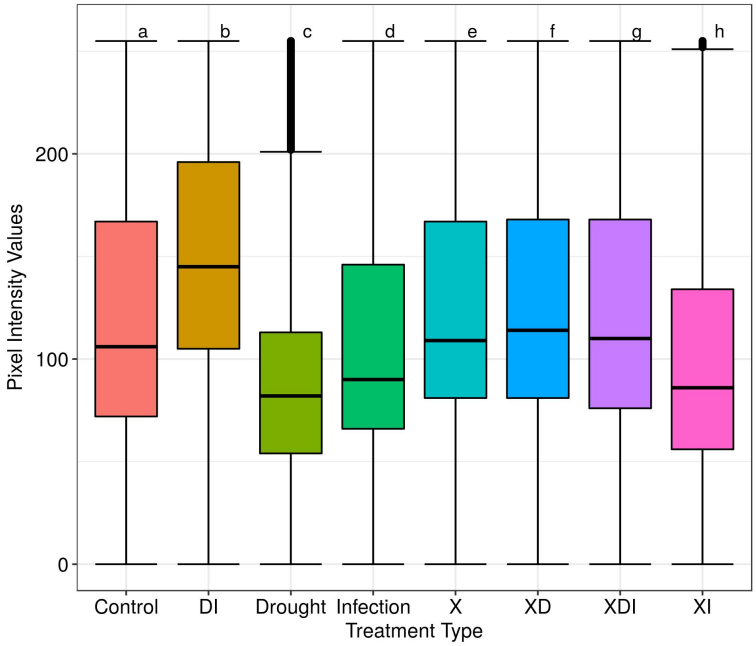


Figure 9: Pixel Intensities of the multi-stress dataset (RGB images only); each boxplot contains pixel intensities from all channels (RGB) of the treatment image. All treatments were proven to be statistically different from one another(ANOVA: $F_{7,4646161} = 38676$, p <0.0001). The letters on the graph show the Tukey post-hoc comparisons.

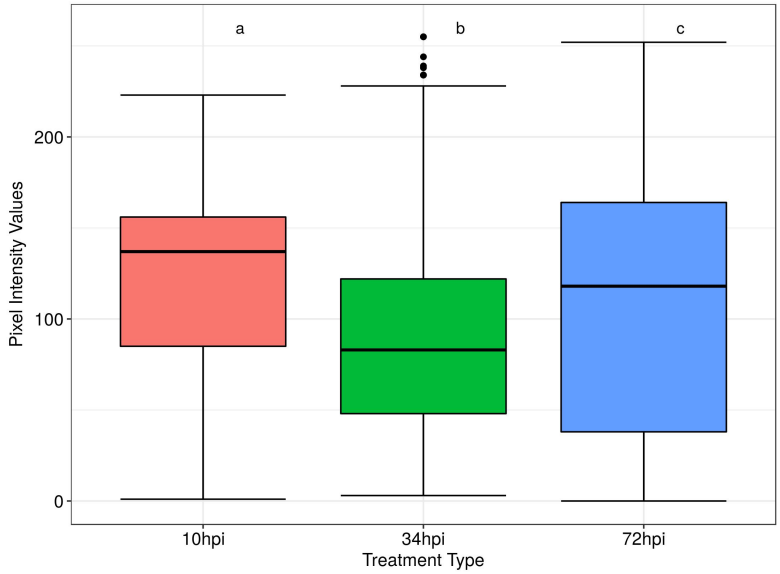


Figure 10: Pixel Intensities for part of the infection duration dataset (dataset 4); each boxplot contains pixel intensities from all channels (RGB) of the treatment image. All treatments were proven to be statistically different from one another (ANOVA: $F_{2,1628925}$ = 32894, p <0.0001). The letters on the graph show the Tukey post-hoc comparisons.

4 Discussion

All classification models achieved >88% training accuracy, with testing accuracy ranging from 31% to 94%. These accuracies are higher than both random chance and predicted human accuracy. Four out of the proposed hypotheses were supported by the data collected; hypothesis 4 (Normalised images will improve model accuracies) was not supported by the data obtained. Furthermore, a website containing some of these models was also built to allow others to test their own images using the models and to allow people to use these techniques in future research (Appendix Figures 4-6).

Multi-stress treatments

The results obtained from the multi-stress classifiers support hypothesis 1; both machine learning algorithms will be able to correctly classify the images with a training accuracy of >85%. Several studies have looked into classifying simple treatment images (see: Baranowski et al. 2015; Mohanty, Hughes and Salathe 2016; Sladojevic et al. 2016; V. Singh and Misra 2017) with classification accuracies reaching up to 99.35%; Mohanty *et al* (2016) note that their study was made harder by using a 38 class dataset, consisting of both crop species and disease, but that ultimately their CNNs still achieved accuracies of >85%.

Despite numerous studies looking into simple stresses, and achieving high classification accuracies, there are no studies clearly documenting multi-stress classification experiments. This is a vital subject area, as in the real world plants are routinely subjected to various combinations of both biotic and abiotic stresses; it has also been noted that combination stresses cannot be identified by extrapolating from individual stresses (Rizhsky, Liang and R. Mittler 2002). Due to this, we believe that our results will be extremely significant both in the field and in future experiments, by allowing individuals to identify multiple stessors that may be impacting upon their crop's health. Knowing the impact that both biotic and abiotic factors have on plant growth, particularly when combined, will become extremely useful in disease identification as well as future treatment measures.

Validation accuracies for models trained on datasets 1-3 were obtained from a mixed test set; several images were collected from various datasets. This may partly explain the lower testing accuracy (Table 2); despite this, the results obtained are indicative of a good overall model, with the potential to be used in future treatment studies. A Website was created containing the models created, allowing individuals to classify their own images. We aim to develop this into a fully functioning application to be used in future studies, whilst also allowing other users to add improvements and further train the models on new datasets.

Despite the obvious benefits of CNNs in comparison to RFs, classification results were almost identical (Table 2); the prediction certainties of true positive test results (Figure 5a) show that CNNs produce a significantly higher percentage certainty in comparison to

RFs. This then allows support for hypothesis 5; CNNs are better classifiers in comparison to RFs. This has already been noted in several studies, with Baranowski *et al* (2015) noting a classification rate of 55% for a RF model, and 80% for a Back-Propagation NN. With many ways to interpret a models output, it can often be difficult to settle on one statistic; it is believed that by producing several model comparison statistics, a more general view of the model can be obtained. By calculating percentage certainties, it can be better ascertained how an individual will respond to a classification accuracy; if two models correctly classify an unknown image (both to the models and the individual classifying the image), but one returns the prediction with a 25% certainty, and the other with a 98% certainty, the individual is far more likely to trust the later certainty. Our results favour CNNs, despite their increased requirements in comparison to RFs; future studies could build on previous CNNs, reducing build time as well as training time.

The use of colour-pass filters increased the models accuracy (Table 2); hypothesis 3 was thereby supported by this study. Colour-pass filters block out their colour specific wavelengths, and have been used before in an ecological classification survey (Knoth et al. 2013). By using them in classification studies, the object in question can come into focus whilst also allowing certain aspects of the object to become more apparent, as shown in Figure 2. This is likely due to metamerism, whereby different colours appear the same under certain conditions, with the colours in question referred to as metamers. By taking photographs of the data under various colour-pass filters, we can capture a range of colour information, with a higher likelihood of capturing the true colour of the dataset.

Due to the price and accessibility of colour-pass filters in comparison to Hyperspectral imaging cameras, as well as the results obtained from this study, we recommend that colour-pass filters be included in more classification studies to allow for greater data to be brought out of a dataset. Due to the ease of applying a colour-pass filter to a dataset, this method will also allow for the wider botanical and agricultural community to apply these techniques in future research.

Training and testing on a normalised dataset resulted in models of a lower accuracy than identical models trained on non-normalised images; normalised images appeared to have a negative effect on classification rates, thus resulting in the rejection of hypothesis 4. It is possible that the process of normalising an image distorts the feature space of the image, thus making it harder for a model to classify. It is also possible that the method used to normalise the images was not sufficient; future studies should look into a wider array of normalising approaches, so as to ensure sufficient evidence has been obtained before making any compelling statements.

Duration Infection

The results of both models trained on the duration infection dataset support hypothesis 2; RFs and CNNs will be able to correctly classify the images with a training accuracy of

>85%. Another study, conducted by Zhu *et al* 2016, also looked into classifying infection duration images; tobacco leaves were infected with Tobacco mosaic virus (TMV), and images were taken at 2 days post infection (DPI), 4DPI and 6DPI, along with healthy tobacco leaf images. Classification accuracies of up to 95% were established, using hyperspectral images and a back-propagation NN. Their study noted that is it hyperspectral imaging that has the potential to aid presymptomatic disease detection; our results indicate that even simple RGB images have the ability to advance detection. We also note that by 3 days (72 hpi), leaves were severely infected, and so believe that classifying many hours before this stage is essential if the crop is to be treated. Being able to classify between healthy and up to 72hpi leaves will greatly improve presymptomatic treatment; applying this technique in the field will allow for more precise pesticide and insecticide application, which is beneficial to both the community and the environment.

As in multi-treatment classification, the results for both models were almost identical (Table 2), whilst the prediction certainties (Figure 5b) show that CNNs produce a significantly higher percentage certainty in comparison to RFs. This then allows for further support for hypothesis 5.

Due to the image data for this study being collected in the laboratory, future studies should focus on conducting studies using field data; this would allow for more realistic results to be obtained, and subsequently, a better suited algorithm for field classification.

Pixel Intensities

Pixel intensities for both datasets show that each treatment was significantly different to all other treatments within that dataset (Figures 7 and 8); this allows for further support of the classifiers built and the hypotheses supported, by acknowledging that they are classifying on relatively discrete and distinct treatments.

Appendix figures 1 and 2 show a more detailed view of the pixel intensities for each treatment. Two clear patterns are observed in Appendix Figure 1; the blue channel in the infection duration study significantly decreases as time increase, whilst the green and red channel significantly increase at 72hpi. This is likely an indication of ongoing infection, with the overall leaf colour changing from a green in 10hpi to a yellow/brown in 72hpi. Between 34 and 72hpi it is likely that chlorosis began, thus resulting in an increase in intensity in the red and green channels in the 72hpi treatment.

5 Conclusions

There has been a noteworthy increase in recent years in the use of digital pathology techniques. Plant disease detection and management are important activities for both agriculture and horticulture; by utilising digital techniques to identify diseases, it allows for quicker interventions on the infected individual(s).

This study shows that is it possible for relatively simple machine learning methods to identify both complex stress treatments, and duration of infection. These techniques will allow farmers, growers and horticultural enthusiasts to begin earlier treatment, thus minimising crop loss and reduced crop quality.

This study also sheds light on the impact that both colour-pass filters and image normalising techniques have on digital classification. These findings should benefit future research and agricultural practises alike.

References

Abadi, Martin et al. (2016). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: *arXiv:1603.04467 [cs]*. arXiv: 1603.04467.

Awate, A. et al. (2015). "Fruit disease detection using color, texture analysis and ANN". In: *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*. 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 970–975.

Baranowski, Piotr et al. (2015). "Hyperspectral and Thermal Imaging of Oilseed Rape (Brassica napus) Response to Fungal Species of the Genus Alternaria". In: *PLOS ONE* 10.3, e0122913.

Choi, Bosung et al. (2017). "Positive regulatory role of sound vibration treatment in Arabidopsis thaliana against Botrytis cinerea infection". In: *Scientific Reports* 7.1, p. 2527.

Chollet, Francois et al. (2015). *Keras*. https://keras.io.

Ciresan, Dan C., Ueli Meier, Jonathan Masci et al. (2011). "High-Performance Neural Networks for Visual Object Classification". In: *Computing Research Repository - CORR*.

Ciresan, Dan C., Ueli Meier and Juergen Schmidhuber (2012). "Multi-column Deep Neural Networks for Image Classification". In: *ArXiv e-prints*. arXiv: 1202.2745.

Dangl, Jeffery L. and Jonathan D. G. Jones (2006). "The plant immune system". In: *Nature* 444.7117, p. 323.

Delalieux, Stephanie et al. (2007). "Detection of biotic stress (Venturia inaequalis) in apple trees using hyperspectral data: Non-parametric statistical approaches and physiological implications". In: *European Journal of Agronomy* 27.1, pp. 130–143.

Dennis, Elizabeth S et al. (2008). "Genetic contributions to agricultural sustainability". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1491, pp. 591–609.

Diaz, Jose, Arjen ten Have and Jan A. L. van Kan (2002). "The Role of Ethylene and Wound Signaling in Resistance of Tomato to Botrytis cinerea". In: *Plant Physiology* 129.3, pp. 1341–1351.

Donatelli, M. et al. (2017). "Modelling the impacts of pests and diseases on agricultural systems". In: *Agricultural Systems* 155 (Supplement C), pp. 213–224.

Harsanyi, J. C. and C. I. Chang (1994). "Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach". In: *IEEE Transactions on Geoscience and Remote Sensing* 32.4, pp. 779–785.

Haubold, Carsten et al. (2016). "Segmenting and Tracking Multiple Dividing Targets Using *ilastik*". In: *Focus on Bio-Image Informatics*. Advances in Anatomy, Embryology and Cell Biology. Springer, Cham, pp. 199–229.

Karimi, Y. et al. (2006). "Application of support vector machine technology for weed and nitrogen stress detection in corn". In: *Computers and Electronics in Agriculture* 51.1, pp. 99–109.

Kleesiek, Jens et al. (2014). "ilastik for Multi-modal Brain Tumor Segmentation". In: *MICCAI-BraTS*, pp. 12–17.

Knoth, Christian et al. (2013). "Unmanned aerial vehicles as innovative remote sensing platforms for high-resolution infrared imagery to support restoration monitoring in cut-over bogs". In: *Applied Vegetation Science* 16.3. Ed. by Sarah Goslee, pp. 509–517.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Neural Information Processing Systems* 25.

Lapin, Dmitry and Guido Van den Ackerveken (2013). "Susceptibility to plant disease: more than a failure of host immunity". In: *Trends in Plant Science* 18.10, pp. 546–554.

Lowe, Amy, Nicola Harrison and Andrew P. French (2017). "Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress". In: *Plant Methods* 13, p. 80.

Malof, J. M. et al. (2016). "A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery". In: *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*. 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), pp. 650–654.

McCook, Stuart (2006). "Global rust belt: *Hemileia vastatrix* and the ecological integration of world coffee production since 1850". In: *Journal of Global History* 1.2, pp. 177–195.

Mittler, Ron (2006). "Abiotic stress, the field environment and stress combination". In: *Trends in Plant Science* 11.1, pp. 15–19.

Mohanty, Sharada P., David P. Hughes and Marcel Salathe (2016). "Using Deep Learning for Image-Based Plant Disease Detection". In: *Frontiers in Plant Science* 7.

Mou, L., P. Ghamisi and X. X. Zhu (2017). "Deep Recurrent Neural Networks for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.7, pp. 3639–3655.

Naseem, I., R. Togneri and M. Bennamoun (2010). "Linear Regression for Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.11, pp. 2106–2112.

Pel, Michiel J. C. and Corne M. J. Pieterse (2013). "Microbial recognition and evasion of host immunity". In: *Journal of Experimental Botany* 64.5, pp. 1237–1248.

Piron, A. et al. (2008). "Selection of the most efficient wavelength bands for discriminating weeds from crop". In: *Computers and Electronics in Agriculture* 62.2, pp. 141–148.

Punjani, Ali and Pieter Abbeel (2015). "Deep learning helicopter dynamics models". In: IEEE, pp. 3223–3230.

Rizhsky, L., H. Liang and R. Mittler (2002). "The combined effect of drought stress and heat shock on gene expression in tobacco". In: *Plant Physiology* 130.3, pp. 1143–1151.

Rossum, G van and F. L Drake (2001).

Russakovsky, Olga et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3, pp. 211–252.

Santos, Emerson Alves dos et al. (2018). "Path analysis of phenotypic traits in young cacao plants under drought conditions". In: *PLOS ONE* 13.2, e0191847.

Simard, Patrice, David Steinkraus and John Platt (2003). "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis." In: pp. 958–962.

Singh, Arti et al. (2016). "Machine Learning for High-Throughput Stress Phenotyping in Plants". In: *Trends in Plant Science* 21.2, pp. 110–124.

Singh, Vijai and A. K. Misra (2017). "Detection of plant leaf diseases using image segmentation and soft computing techniques". In: *Information Processing in Agriculture* 4.1, pp. 41–49.

Sischka, Nicholas (2018). *Using optical filters to enhance image contrast.* URL: `https://www.vision-systems.com/articles/print/volume-19/issue-4/features/using-optical-filters-to-enhance-image-contrast.html` (visited on 20/07/2018).

Sladojevic, Srdjan et al. (2016). *Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification.* Computational Intelligence and Neuroscience. URL: `https://www.hindawi.com/journals/cin/2016/3289801/` (visited on 29/06/2018).

Sommer, C. et al. (2011). "Ilastik: Interactive learning and segmentation toolkit". In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro.* 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 230–233.

Suzuki, Nobuhiro and Ron Mittler (2006). "Reactive oxygen species and temperature stresses: A delicate balance between signaling and destruction". In: *Physiologia Plantarum* 126.1, pp. 45–51.

Tank, Neelam and Meenu Saraf (2010). "Salinity-resistant plant growth promoting rhizobacteria ameliorates sodium chloride stress on tomato plants". In: *Journal of Plant Interactions* 5.1, pp. 51–58.

Team, R Core (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Tripathi, M. K. and D. D. Maktedar (2016). "Recent machine learning based approaches for disease detection and classification of agricultural products". In: *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*. 2016 International Conference on Computing Communication Control and automation (ICCUBEA), pp. 1–6.

Ubbens, Jordan R. and Ian Stavness (2017). "Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks". In: *Frontiers in Plant Science* 8.

Visschers, Isabella G. S., Nicole M. van Dam and Janny L. Peters (2018). "An objective high-throughput screening method for thrips damage quantitation using Ilastik and ImageJ". In: *Entomologia Experimentalis et Applicata*.

Waller, Frank et al. (2005). "The endophytic fungus Piriformospora indica reprograms barley to salt-stress tolerance, disease resistance, and higher yield". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.38, pp. 13386–13391.

Wood, R. K. S. (1996). "Sustainable agriculture: the role of plant pathology". In: *Canadian Journal of Plant Pathology* 18.2, pp. 141–144.

Woodham-Smith, C. (1962). "The Great Hunger: Ireland 1845-9." In: *The Great Hunger: Ireland 1845-9*.

Yaeger, Larry, Richard Lyon and Brandyn Webb (1996). "Effective Training of a Neural Network Character Classifier for Word Recognition". In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. NIPS'96. Cambridge, MA, USA: MIT Press, pp. 807–813.

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *Computer Vision - ECCV 2014*. European Conference on Computer Vision. Lecture Notes in Computer Science. Springer, Cham, pp. 818–833.

Zhao, D. et al. (2005). "Nitrogen deficiency effects on plant growth, leaf photosynthesis, and hyperspectral reflectance properties of sorghum". In: *European Journal of Agronomy* 22.4, pp. 391–403.

Zheng, Xiao-yu et al. (2012). "Coronatine Promotes Pseudomonas syringae Virulence in Plants by Activating a Signaling Cascade that Inhibits Salicylic Acid Accumulation". In: *Cell Host & Microbe* 11.6, pp. 587–596.

Zhu, Hongyan et al. (2016). "Early Detection and Classification of Tobacco Leaves Inoculated with Tobacco Mosaic Virus Based on Hyperspectral Imaging Technique". In: American Society of Agricultural and Biological Engineers.

6 Appendix

Table 1: Colour-pass filters used in the multi-stress study on tomato leaves.

Roscolux Filter Number	*Colour Name*
#3407	Roscosun
#4590	CalColor 90 Yellow
#12	Straw
#318	Mayan Sun
#4690	CalColor 90 Red
#4890	CalColor 90 Pink
#4760	CalColor 60 Magenta
#4790	CalColor 90 Magenta
#39	Skeleton Exotic Sangria
#41	Salmon
#47	Light Rose Purple
#4930	CalColor 30 Lavender
#59	Indigo
#3220	Double Blue
#65	Daylight Blue
#2005	Storaro Cyan
#77	Green Blue
#378	Alice Blue
#80	Primary Blue
#4490	CalColor 90 Green
#93	Blue Green

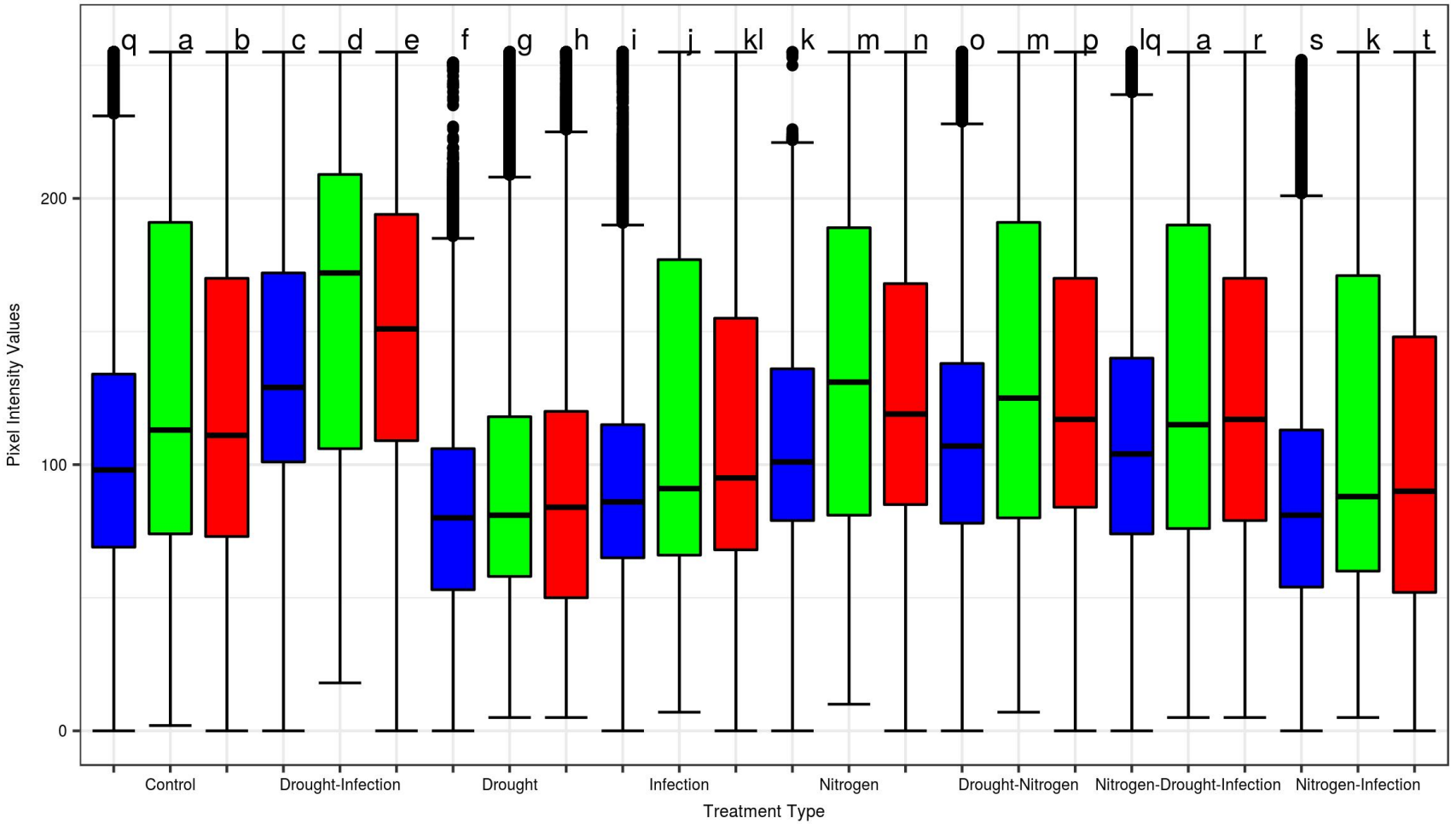


Figure 1: Pixel intensity values for all treatments in multi-stress dataset (RGB only); the graph shows individual channel (RGB) intensities. Letters above each boxplot correspond to an ANOVA result; most channels were significantly different from one another ($p < 0.0001$).

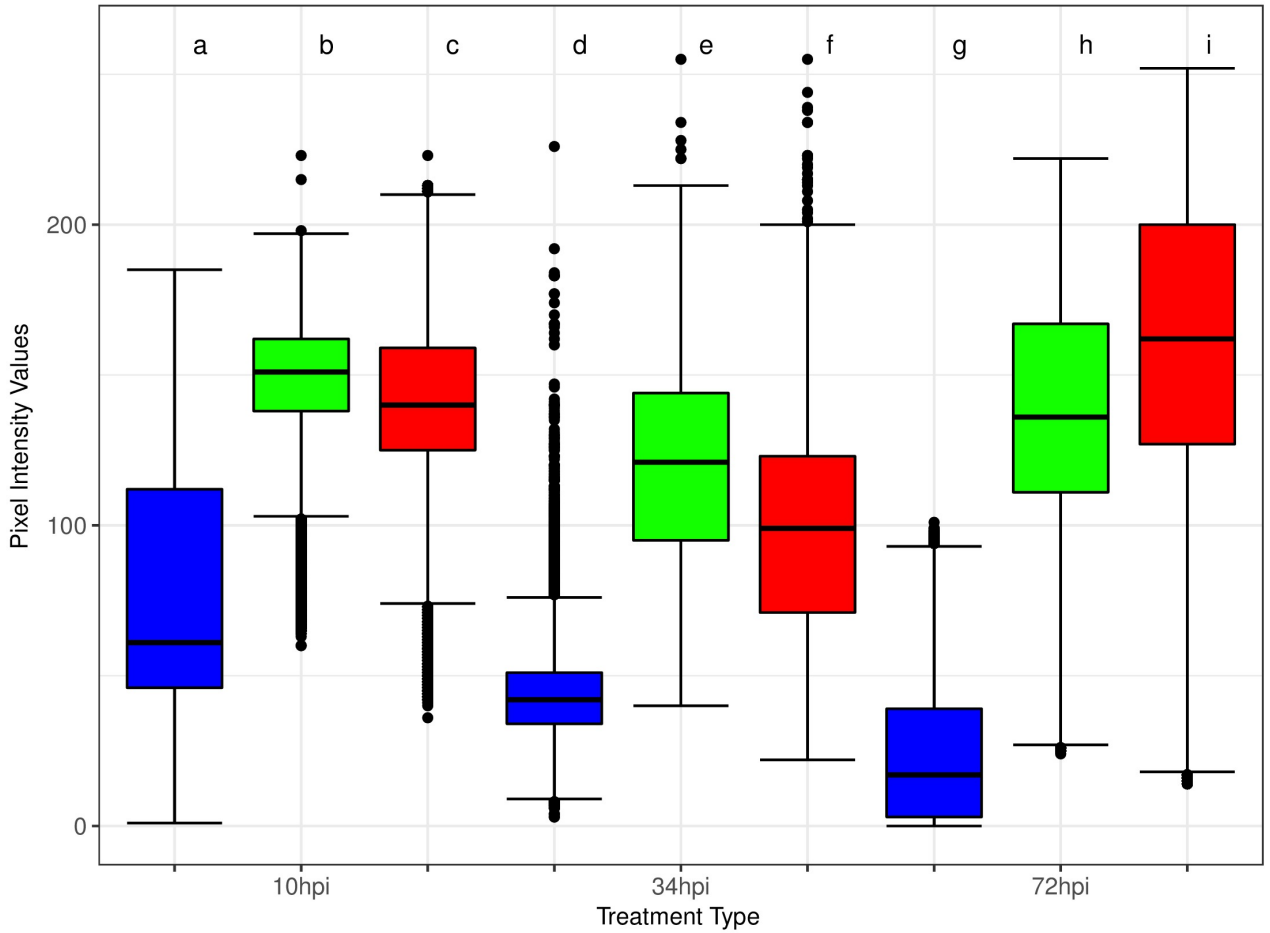


Figure 2: Pixel Intensity values for 10,34 and 72hpi. These intensities were obtained from the infection duration dataset, and show individual channel (RGB) intensities.Letters above each boxplot correspond to an ANOVA result; most channels were significantly different from one another ($p < 0.0001$).

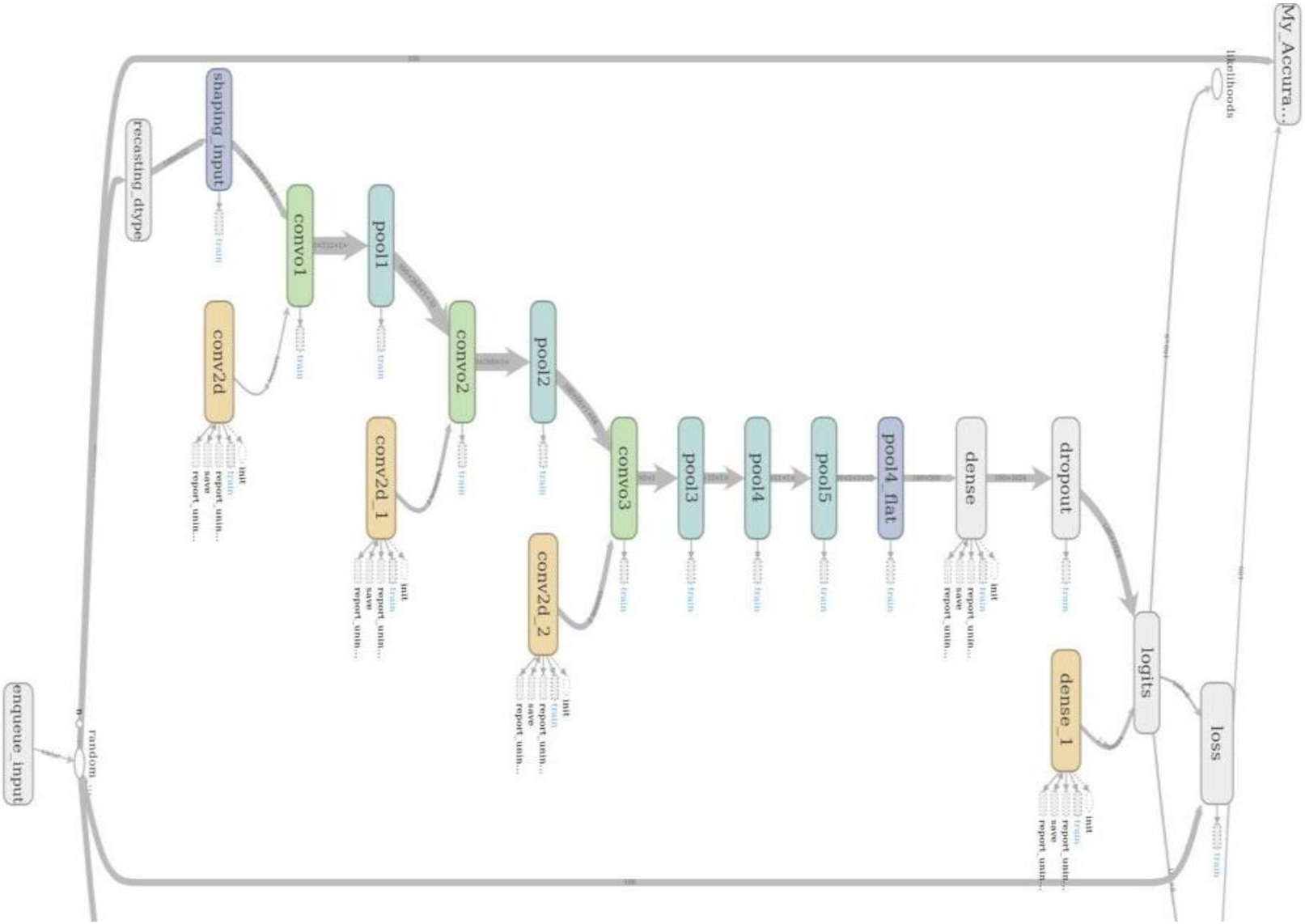


Figure 3: Layout for the Convolutional Neural Network model used. The model begins which a reshaping and recasting layer, followed by several convolutional and pooling layers. These layers are conducted to extract classifying information and features from the training images. After all pooling layers, several flattening and a dense layer are run; these are run to convert the convolutional layer outputs into probabilities for each classifying class (8 classes for both the multi-stress and infection duration studies). The dropout layer is included to reduce the risk of over-fitting. Finally, accuracy and loss calculations were conducted after each iteration of the model, which are plotted in Figure 5.

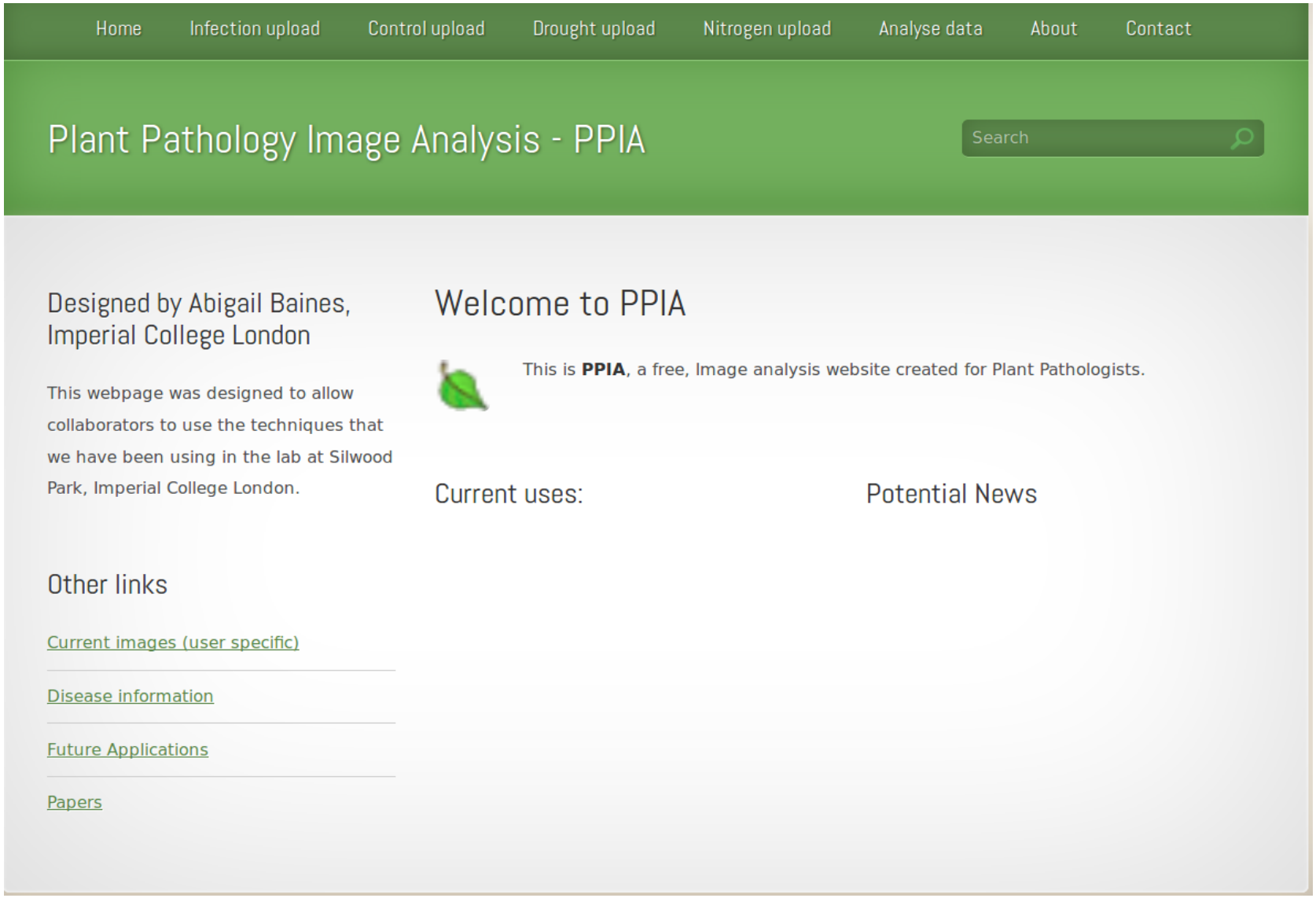


Figure 4: Main page of the website created. On the top bar there are multiple links to upload images, with the side bar for previewing already uploaded images, disease information and potential published papers referencing the website.

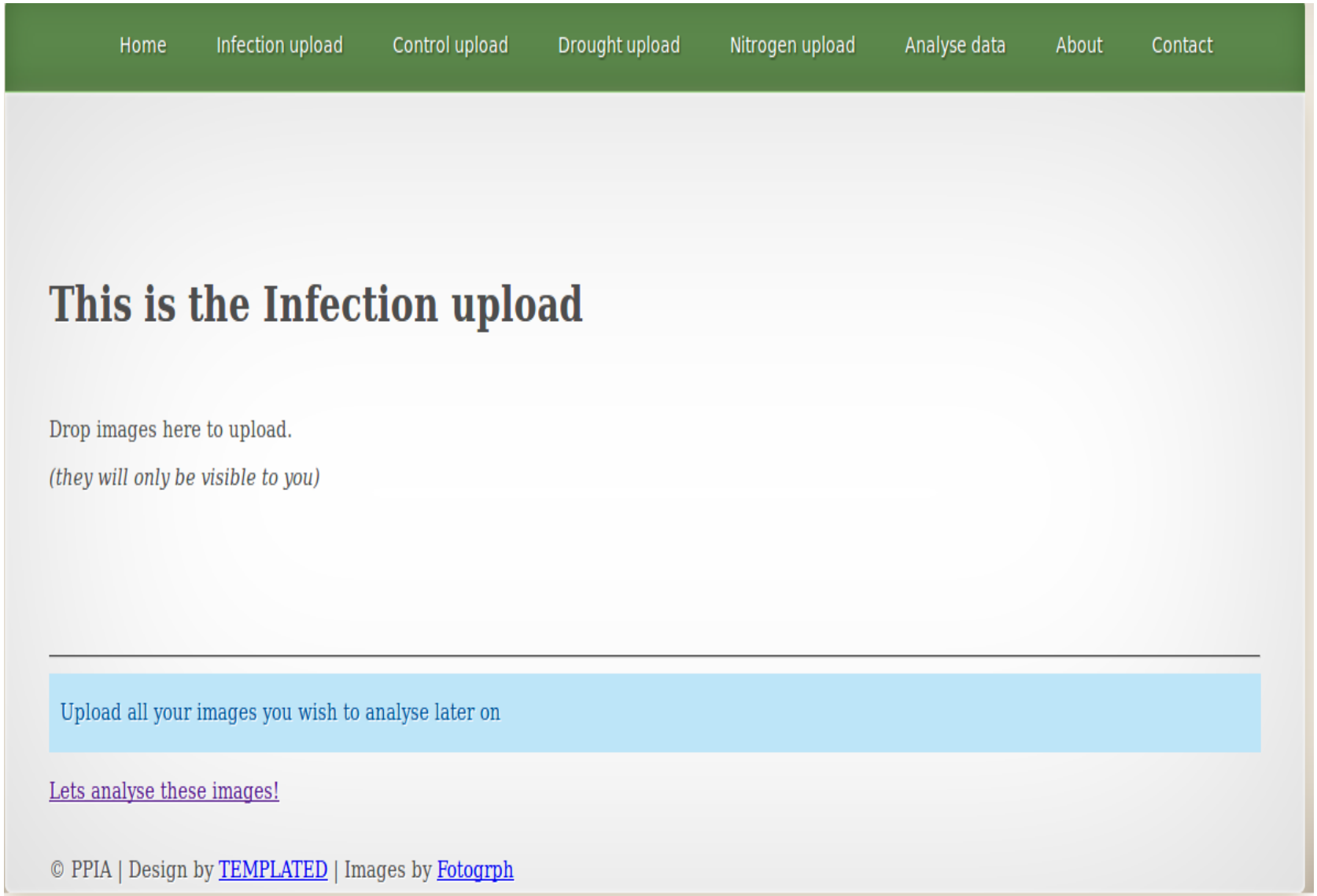


Figure 5: Upload page for the website. A drag and drop feature was enabled so that multiple images and folders can be uploaded easily.

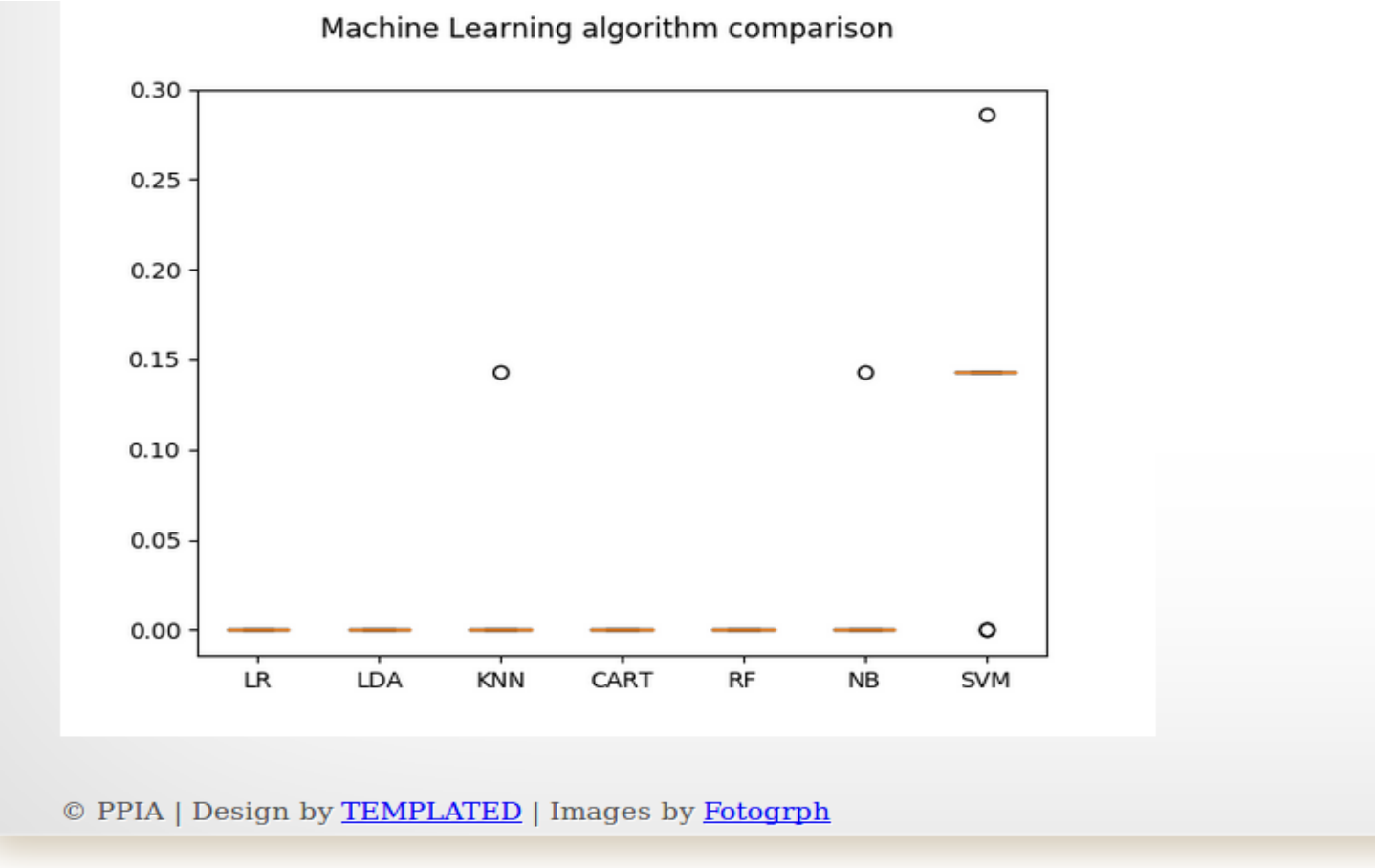


Figure 6: Sample output graphic from the website created from several test images. The current algorithm allows for users to test their images against multiple ML algorithms, such as KNN and RF. We aim to enable testing against the CNNs in the next website update.