ADVANCES IN AGRICULTURAL AND FOOD RESEARCH JOURNAL

Review Article

Detection and Classification of Basal Stem Rot Disease in Oil Palm Using Machine Learning Techniques: A Mini Review

Nur Azuan Husin[1,2*], Mohd Hamim Abd Aziz[3], Siti Khairunniza-Bejo[1,2,4]

[1]Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia, nurazuan@upm.edu.my

[2]Smart Farming Technology Research Centre, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia, skbejo@upm.edu.my

[3]Department of Science and Technology, Faculty of Humanities, Management and Science, Universiti Putra Malaysia, Bintulu Campus, Bintulu 97008, Malaysia, m_hamim@upm.edu.my

[4]Laboratory of Plantation System Technology and Mechanization (PSTM), Institute of Plantation Studies, Universiti Putra Malaysia, Serdang 43400, Malaysia

*Corresponding author: Nur Azuan Husin, Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia; nurazuan@upm.edu.my

Abstract: The oil palm grown around the world to meet the demand for food and bio-fuels, is threatened by a fatal disease known as basal stem rot (BSR). Application of machine learning (ML) in agriculture keeps increasing with the advancement of technology, especially in disease detection. This manuscript presents a mini-review of the different methods relevant to BSR disease classification and detection using ML. The steps were discussed, including pre-processing and approaches used. Various algorithms, feature extractions and classification methods were discussed in the review. The review results revealed that the adoption of disease detection and classification methods for BSR disease in oil palm using ML approaches is still in its early stages of research. Hence, new tools are needed to fully automate the detection and classification processes for practical, operational, fast and accurate systems to be used in vast oil palm plantations.

Keywords: Disease detection; Oil palm; Basal stem rot; Machine learning; Classification

Received: 14th March 2022

Received in revised form: 24th November 2022

Available Online: 5th January 2023

Published: 19th September 2023

Citation: Husin, N. A., Abd Azis, M. H., Khairunniza-Bejo, S. Detection and classification of basal stem rot disease in oil palm using machine learning techniques: A mini review. *Adv Agri Food Res J* 2023; 4(2): a0000365. https://doi.org/10.36877/aafrj.a0000365

1. Introduction

The oil palm is a member of the monocotyledonous palm family (Arecaceae) and palm oil is the most consumed edible oil in the world, representing about 40% of the world's vegetable oil supply (USDA, 2019). It is a perennial tree crop, which better resembles a forest tree than other agricultural crops. As an industrial crop, oil palms are planted in a monoculture fashion and a majority of planted oil palms are tissue culture clones (McMorrow, 2001; Shafri *et al.*, 2011). Plants are strongly associated with plenty of nutritional benefits and lower the threats of diseases, where oil palm contains a high amount of antioxidants, vitamin A, vitamin E and beta carotene (Ebong *et al.*, 1999) that may reduce blood cholesterol, protects against heart disease, blocks the growth of cancer cells and enhance the anti-cancer medications (Fife, 2017). Demand for palm oil to feed the world and for non-fossil fuels continues to increase every year and the production of oil palm has kept increasing precede of other major oil crops since 2003 (Mielke, 2018). The world population is expected to increase to more than 9 billion in 2050 (Clay, 2011) and a larger population highlights the need for greater demand for food and biofuels. One of the factors that can help to cater to the increasing demand for palm oil over the coming decades is to overcome the disease problem.

Oil palm is the most efficient oil-bearing crop in the world, where it is 8 to 10 times more productive than other major oilseed crops (Abdullah *et al.*, 2010). The two types of oil produced from oil palm are palm oil from the mesocarp (a layer of the fruit wall) and palm kernel oil from the seed or kernel. Both of these oils are widely used in food applications and non-food applications (e.g., in oleochemical industries, animal feed and biofuel). Harvested year-round and having an economic lifespan of around 25 to 30 years, oil palm trees produce on average 10 tonnes of fruit per hectare. For every 10 tonnes of fresh fruit bunches (FFB), approximately 6 tonnes of crude palm oil (CPO) is obtained and about 1 ton of palm kernel oil (PKO) is produced per hectare per year (MPOC, 2017). However, oil palm diseases related to algal, bacterial and fungal diseases, macro and micronutrient deficiency problems and genetic disorders are seriously affected in the production of palm oil (Pornsuriya *et al.*, 2013).

Among the major diseases is Ganoderma or the common name basal stem rot (BSR) BSR disease is caused by Ganoderma fungal, where *Ganoderma boninense* (GB) species is known as the most devastator species to cause a great economical effect in the palm oil industry especially in Southeast Asia (Naher *et al.*, 2015; Chong *et al.*, 2017). The disease spreads through root-to-root contact, via airborne spores and from independent secondary inoculum in the soil (Abdullah *et al.*, 2000). The specific diagnostic feature of the disease is the presence of basidiocarps fruiting bodies on the stems (Rees *et al.*, 2007) (Figure 1). As the infection progresses the young unfolded leaves become chlorotic and may reduce in length with necrotic tips. It is followed by the flatting down of leaves fronds and the appearance of several fully extended but unopened spears in the centre of the crown. (Turner, 1965; Flood *et al.*, 2010; Azuan *et al.* (2019); Husin *et al.* (2020); Husin *et al.* (2021)). The next stage is the rotting of the basal stem along with the collapsing of the plant's crown giving the appearance of a narrow waist at the canopy whereas the spear leaves remain unopened. At the final stage, the oil palm easily collapses, leaving stumps and diseased bole tissue in the ground (Paterson, 2007).

Figure 1. Symptoms of BSR disease

Plant diseases are responsible for major economic and agricultural production losses. Identification of plant disease is vital to avoid fatalities in the quality and quantity of agricultural products. Research in image processing for plant disease detection has grown rapidly over the past decades (Golhani *et al.*, 2018). Machine learning (ML) has been applied in various fields including bioinformatics, aquaculture, food and precision farming, presently also termed as digital farming (Liakos *et al.*, 2018). ML approach has emerged to facilitate monitoring and early information on plant health for strategic management strategies. In the agriculture industry, plant diseases are primarily responsible for the reduction in production which causes national economic losses.

In this paper, we present a mini-review of the ML applications in the detection and classification of BSR disease in oil palm fields. A limited number of relevant papers are presented, despite the importance of this subject. The structure of the present work is as follows: Section 2 discusses the ML terminology, definition and common learning models. Section 3 describes the application of ML for BSR disease detection and classification. Finally, Section 4 concludes the paper with further research directions. Due to the high number of acronyms used in related scientific studies, Table 1 lists all the acronyms that are used in this work.

Table 1. Acronyms use in ML

Acronyms	Description
NN	Neural network
DT	Decision tree
SVM	Support vector machine
kNB	Kernel Naïve Bayes
CNN	Convolution neural network
LDA	Linear discriminant analysis
LS	Least squares
MLR	Multiple linear regression
PCA	Principal component analysis
PLS	partial least squares
RF	Random forest
DBSCAN	Density-based spatial clustering of applications with noise
SVD	Singular value decomposition
GMM	Gaussian mixture model
kNN	k-nearest neighbour
HCA	Hierarchical cluster analysis
GA	Genetic algorithm
CART	Classification and regression tree

2. An Overview of Machine Learning (ML)

ML is a form of artificial intelligence (AI) that simulates the way humans learn, which uses software applications and historical input to predict outcomes and gradually improve its accuracy without being explicitly instructed. Depending on the nature of learning, ML tasks are often categorised into multiple broad groups such as learning types – supervised or unsupervised, and learning models – dimensional reduction, classification or regression. A general diagram of the ML approach is shown in Figure 2.

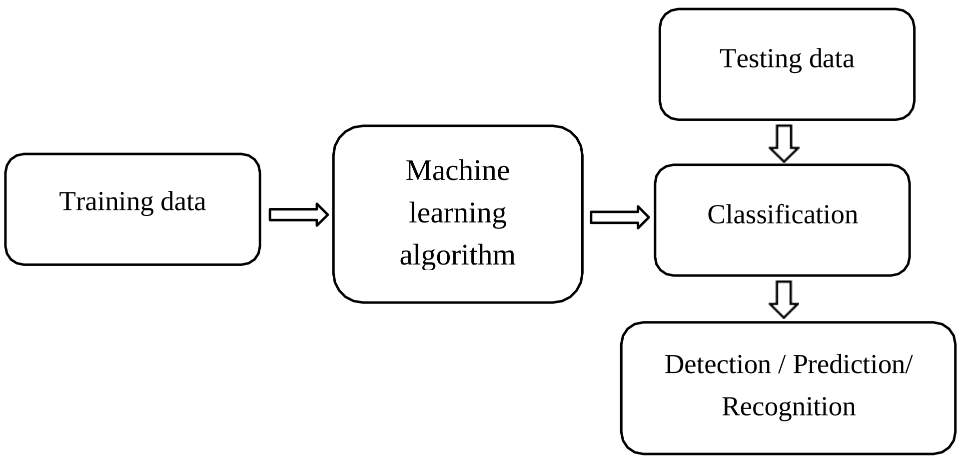

Figure 2. A general schematic flow diagram of ML approach

Basically, there are two ML paradigms — supervised and unsupervised learning (Figure 2). Supervised learning is created to use labelled training datasets that were designed to “supervise” algorithms in classifying or predicting outcomes. Supervised learning models can be further grouped for two main tasks — classification and regression. Classification is the process of grouping the output into different categories or classes based on one or more input variables and mapping it to a discrete value. Linear classifiers, SVMs, DTs and RFs are common types of classification algorithms. Regression is another type of supervised learning method that uses mathematical methods to understand the relationship between a single dependent/output variable and one or more other independent/input variables. It is used when the value of the output variable is continuous or real, and some popular regression algorithms are linear regression, logistic regression and polynomial regression.

Table 2. Examples of ML algorithms’ classification

ML Algorithms	**Continuous**		**Categorical**	
Supervised	Classification: DT, RF	Regression: LDA, Linear, Polynomial	Classification: kNN, Trees, Naïve-Bayes, SVM	Regression: Logistic, PLS
Unsupervised	Clustering: k-Means, DBSCAN, GMMs, Hierarchical Clustering	Dimensional Reduction: PCA, SVD	Association Analysis: Apriori, FP-Growth	Hidden Markov Model

Meanwhile, unsupervised learning works with unlabelled datasets, therefore the algorithms discover hidden patterns without supervision. Unsupervised learning models can be further grouped into three main tasks — clustering, association and dimensionality reduction. Clustering is a data mining technique used when we want to find the inherent groups from the data. It is a way to group unlabelled data and find hidden patterns based on their similarities or differences. Association is another type of unsupervised learning method that uses different rules to find relationships among variables in a given dataset within a large dataset. This learning algorithm's primary goal is to identify the dependencies between data items and map the variables accordingly. Dimensionality reduction is used to reduce the number of data inputs, features or dimensions to a manageable size while also preserving the data integrity. Usually, this technique is used in the pre-processing data stage or to remove noise from data and to improve the data.

2.1 Commonly Used ML Methods

Mostly, ML contains the pre-processing of the plant structure consisting of segmentation, colour extraction and specific data extraction like spectral data and filtration of images. Numerous well-known ML models have been created and used by researchers for categorisation and prediction. The sections below provide explanations for some of them which are SVM, NN and PCA.

2.1.1 Support Vector Machine (SVM)

SVM is currently considered one of the most widely used and best-performing classification techniques (O'Grady, 2013). SVM classifies data by finding the best hyperplane in a number of dimensional spaces that maximises the separation of the data (Li *et al.*, 2013). The largest margin between the classes is considered the best hyperplane. For example, Medium Gaussian means the margin of the separation between the classes was set to medium distinctions and the Gaussian works as the kernel. The value of the kernel depends on the distance from the origin or some setting point. SVM uses a subset of training points in the decision function called support vectors. Using different kernels, it can work through the separation between classes (Kaestner, 2013). Two concepts used in SVM for the study are soft margin and kernel trick. The soft margin attempts to compromise the separations that minimises the misclassification and maximises the margin of the classes. The kernel trick is to separate the non-linear problem in higher dimensional space i.e., it converts not separable problem to separable problem. Figure 3 shows examples of kernels such as linear, polynomial and Gaussian (radial basis function) (Ben-Hur, 2008). It shows that the Gaussian kernel provides more flexibility in separating the classes compared to linear and polynomial kernels. Then, based on the Gaussian kernel Equation 1, different values of parameter sigma (σ) were used to find the best fit in reducing the generalisation error and over-fitting problem. Sometimes, there were some points that were misclassified, since it was difficult to find a classifier that is perfect and ideal for every data.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x - \mu)^2}{2\sigma^2}) \tag{1}$$

where p is density with respect to the Lebesgue
μ and σ^2 are the mean and variance of *x*.

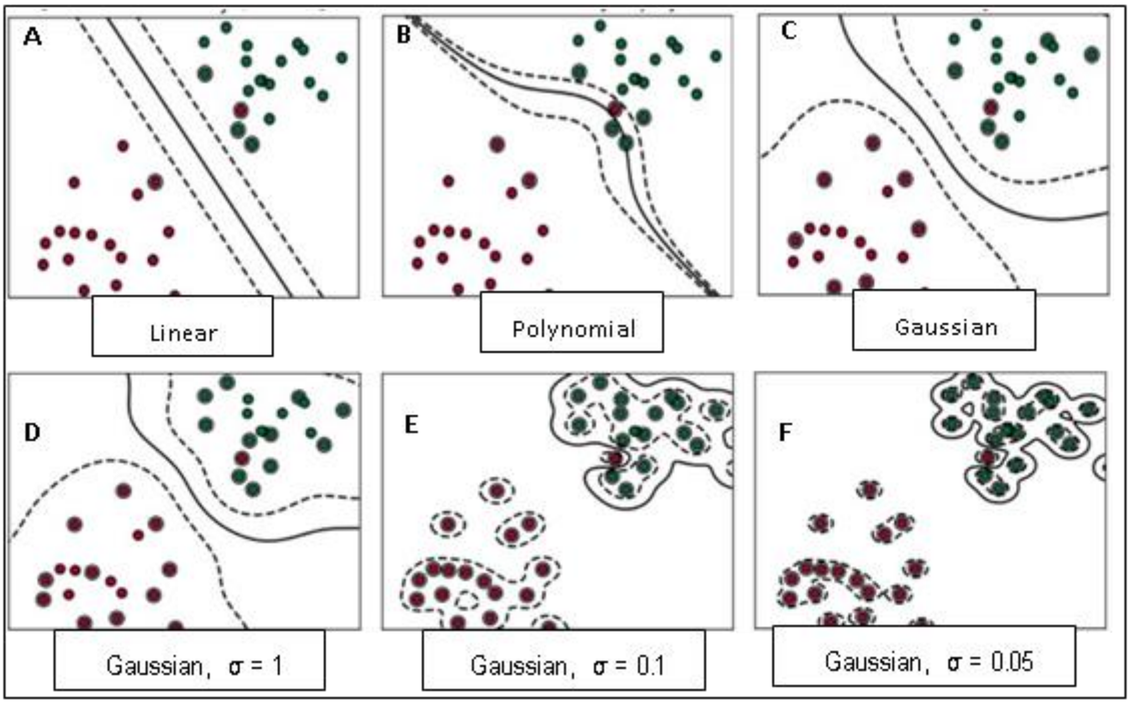

Figure 3. Examples of soft margin and kernel tricks in SVM (Ben-Hur, 2008)

2.1.2 Neural Networks (NNs)

NN is based on the inception of the human brain with neuron nodes interconnected like a web, which is complex and nonlinear for processing information (Shanmuganathan & Samarasinghe, 2016). A NN is characterised by its pattern of connections between the neurons (architecture) and its method of determining the weights on the connections (algorithm) (Fausett. 1994). A large number of new and modified NNs were developed to solve problems and used for data modelling and function approximation (Basheer & Hajmeer, 2000). In order to solve the nonlinearly separable problems, the input layer and the output layer are needed, which construct the multilayer perceptron (MLP). The hidden layer is used to process the information from the input layer and pass it to the output layer (Figure 4). The hidden layer does not interact with external factors, while the addition of the layers further its ability to solve more complex problems (Basheer & Hajmeer, 2000). One example of the artificial neural network (ANN) approach is backpropagation (BP). The term BP refers to the way the error computed at the output side is propagated backward to the hidden layer, and finally to the input layer. Error term will be estimated when compared with single training with the actual output values. The error will be back propagated to the network and used to adjust the error term until the minimum error is achieved, which is an iterative procedure. Detailed explanations of the working principle of ANN can be found in Haykin (1994).

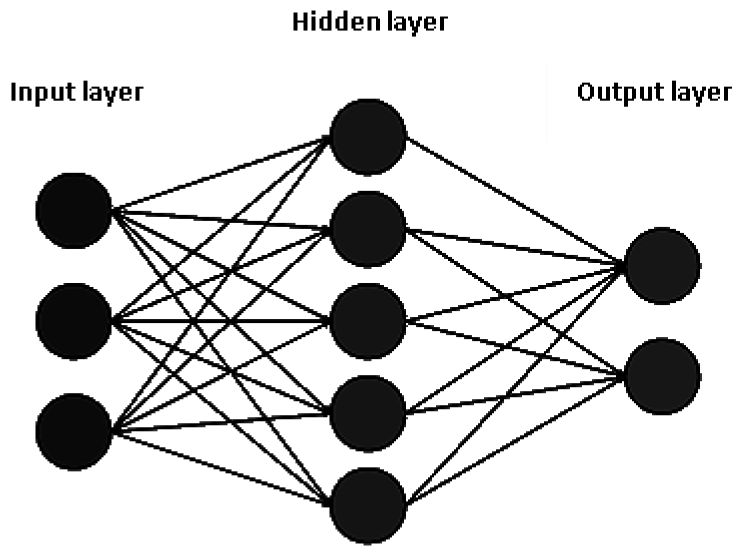

Figure 4. Structure of ANN

2.1.3 Principle component analysis (PCA)

PCA is based on the transformation of the original data into a set of linearly uncorrelated variables, called Principal Components (PCs) (Panigrahi, 2014). The aim is to reduce the amount of data and to create predictive models. In PCA, the eigenvectors of the PCs determine the direction of the new feature space, and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes. Meanwhile, the loading matrix is a correlation between the original variables (parameters) and the PCs. The closer the value is to one, the greater the effect of the PC on the parameter. To infer a correlation, there should be a clustering on a two-dimensional loading plot, and the squared cosine should be greater than one-half (David & Jacobs, 2014). The greater the squared cosine, the greater the relationship with the corresponding PCs. Several conditions have been proposed to determine how many PCs should be retained and excluded. One common condition is to ignore PCs at the point at which the next PC offers little increase in the total variance explained. A second condition is to include all those PCs up to a predetermined total percentage variance explained from the original data, such as 90% (Bonate, 2011). A third condition is to ignore the last PCs whose variance explained is all roughly equal (Holland, 2008). A fourth condition is to use a scree plot, which is a line plot of the eigenvalues of factors or PCs in an analysis (Figure 5). Following the criteria set out by Cattell and Jaspers (1967), which suggests using all the PCs up to and including the first one before the break, whereby three PCs were retained to which eigenvalues greater than unity were attained.

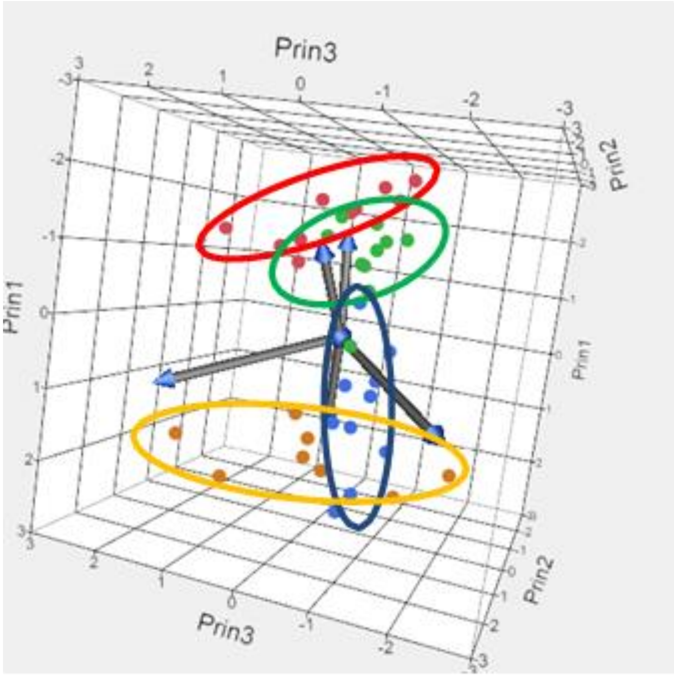

Figure 5. Example of PCA

3. Disease Detection and Classification

The reviewed articles that implemented ML for BSR disease detection have been classified into five approaches; namely electric nose, electrical properties and remote sensing that encompasses multispectral, hyperspectral, radio detecting and ranging (RADAR) as well as light detection and ranging (LiDAR). Details of the approaches are discussed in the following section.

3.1 Electronic Nose (e-nose) System

In the e-nose system, the odour parameter was used to differentiate between healthy and unhealthy trees. A commercial e-nose sensor, Cyranode 320 (C-320 by Cyrano Sciences Inc. (USA)) used by Markom *et al.* (2009) is a handheld e-nose instrument that has 32 sensing elements including components for sampling, sensory and signal processing (Boilot *et al.*, 2002). The collected data sets were normalised and separated into training and testing. ML methods that were used are PCA and ANN. PCA was used to reduce the dimensionality of the data and used as input for the pattern recognition tool, ANN. Three parts from each of the oil palm trees in every category were used for laboratory analysis i.e., bored trunk, surrounding trunk and soil point. There were differences in the laboratory results with the on-site because of physiological changes in the test samples during transit. Thus, only on-site data were utilised. The odour profiles recorded were able to give specific odour fingerprints of healthy and unhealthy trees and are consistent for each tree. Meanwhile, Abdullah *et al.* (2012) invented a handheld e-nose sensor and the basidiocarp samples were taken to the lab for testing without field tests. Similar ML method, PCA was used to reduce the dimensionality of the data. Ten input sensors were chosen, which have high values of principal loading and total variance higher than 90%. ML methods comprised of HCA and LDA were utilised to find the separation between the samples. The results from ML methods used such as LDA and HCA were able to separate the odours of the basidiocarp samples and the ambient air as the reference. The challenges are the detection has to be done with the existence of GB basidiocarp, where some of the infected oil palm trees did not exhibit the basidiocarp. The invented handheld e-nose may give different results when tested on a plantation due to different environments compared to the laboratory. The study can be improved by developing biosensors using chemical components based on conducting polymers. Both outcomes from the odour input showed that the e-nose device was possible to segregate infected and healthy trees but regarding different levels of infection, further research is needed.

3.2 Electrical Properties

Khaled *et al.* (2018) used electrical properties such as impedance, capacitance, dielectric constant and dissipation factor to detect BSR disease in oil palm trees at an early stage. ML methods used to choose the most significant frequencies input were SVM-FS, RF and GA, while the BSR severity was classified using SVM and ANN methods. The results showed that SVM-FS was the best model with the highest accuracies compared with RF and GA models. SVM-FS is the tool in combination with the ranks search method, which represents the results in feature ranking format in line with the predictive model score. The score is based on SVM recursive feature elimination (SVMRFE) to sort represented by the weight vector (Guyon *et al.*, 2002). In addition, RF is a decision tree bagging method that is effective when considering high dimensional data due to its ability to prevent missing and unbalanced data and reduce noise (Samsudin *et al.*, 2015). Meanwhile, GA is a search algorithm to reduce the computational issues associated with a large number of features. Only 56 samples from mature oil palm trees were selected with 14 oil palm trees for every four levels of infections. Leaflet from frond number 17 was randomly collected, which gathered 224 samples in total. Spectral data of foliar samples were scanned using a portable spectroradiometer at a 1.45 nm interval and range of 273 to 1,100 nm with a resolution of 5 nm. Data were separated between training and testing with a ratio of 70:30. The results show that SVM performed slightly better than ANN but no significant difference was found using analysis of variance (ANOVA) at a 5% significant level.

Additionally, Duncan test found a significant difference between the healthy and mild infection levels but no significant difference was found between moderate and severe infection levels. Impedance values produced the highest accuracies, while dissipation values produced the lowest accuracies. It was because of the electrical properties of leaf changes as a result of stress on plant water content and structure in infection level. The challenges are to differentiate the moderate and severe classes of infection and to make the approach practical and operational in the oil palm plantation because the experiments were conducted in the laboratory.

3.3 Remote Sensing

Remote sensing data integrated with ML methods were used by several researchers to detect and classify BSR disease in oil palm plantations. A remote sensing system is defined as a method for the acquisition of information on their properties through analysis of the data on the objects by sensors without physical contact with the objects (Okamoto, 2001). The approaches comprise multispectral, hyperspectral, RADAR and LiDAR imaging systems using airborne, ground and handheld platforms at leaf and canopy scales.

3.3.1 Multispectral image

Thermal imaging technique was used to detect BSR disease in 53 healthy and 53 infected palms (Bejo *et al.*, 2018). ML method used for this study was PCA which showed two distinguishable trend lines for healthy and BSR-infected trees in score plots PC1 and PC3. Then, two ML methods - SVM and kNN were used, where SVM results gave a higher accuracy compared to kNN with 89.2% accuracy during training and 84.4% during testing. Three thermal images were captured randomly for each tree, in three different angles focused on the canopy section using a handheld thermal camera. Four values were extracted from the images, i.e., the maximum, minimum, mean and standard deviation of pixel intensity. For each sample of the tree, the data were averaged into a single value. Then, statistical analysis was used to determine significant information for the classification of healthy and unhealthy trees. To improve the classification, three different indices values were formed - the average value of the mean intensity of healthy trees, the average value of mean intensity value of unhealthy trees and the average value of the mean intensity of all trees. Data were separated between training and testing with a ratio of 70:30. It was proposed that another type of image such as visible (VIS) image can be integrated for better classification and development of new vegetative index. Meanwhile, Santoso *et al.* (2017) employed Quickbird satellite images with five available bands visible red, green, blue, near-infrared and panchromatic for BSR disease classification in Sumatera, Indonesia. A similar ML approach was used, PCA for pre-processing data and then other ML methods were used, RF, SVM and CART to process the data. The results showed that the RF classifier was the best classifier compared to SVM and CART models with the highest producer accuracy of 91%, user accuracy of 83% and overall accuracy of 91%. The sites consisted of 144 oil palm trees of different ages ranging from 10 to 21 years old and were divided into only two classes - 99 healthy trees (99) and unhealthy trees (45). Segmentation of the image was processed in ENVI and ArcGIS (ESRI) to identify the spatial pattern of BSR disease and pixel extraction was performed in Rstudio (RStudio 2015) software. Data were separated in the ratio of 75% for training (109 trees) and 25% for testing (35 trees). The training models used 10 repeated cross-validations (10 folds). Subsequently, images from the WorldView-3 satellite that has a panchromatic resolution of 1.31 m, 8 bands with 1.24 m resolution and a revisit interval of less than 1 day were used for BSR severity classification (Santoso *et al.*, 2019). Similar ML algorithms were applied as Santoso *et al.* (2017), using RF and SVM, only the CART method was replaced by DT. As a result, the SVM approach was the best classifier to differentiate all four classes with a moderate overall accuracy of 54%. Tuning parameters were used with the training data to process the prediction models and to classify the severity classes in the testing data set. Jeffries-Matusita (JM) distance was used to calculate the classes' separability and stepwise variable selection was used to obtain significant variables to separate the classes. The reflectance values of each satellite band showed significant differences among the four classes of BSR infection, but the JM distance showed low values, which indicated low separability among all class pairs. Image pixels were converted from digital numbers to top-of-atmosphere spectral radiance at a minimum. 1923 oil palm trees were selected based on four severity labels as healthy (H), initial unhealthy (UH1), moderate unhealthy (UH2) and severely unhealthy (UH3), comprising 36% H (695), 23% UH1 (432), 18% UH2 (348), and 23% UH3 trees (448). The mean of the nine-pixel values was extracted from the adjustable square polygon used to cover different oil palm crown sizes. Data were separated between training and testing with a ratio of 75:25 and normalisation, resampling and training and prediction were performed using R software. The images were pre-processed for radiometric correction and geometric correction using the rational polynomial coefficient (RPC) technique. In the future, instead of using satellite data, the method could be tested using unmanned aerial vehicles (UAVs) attached with multispectral cameras.

3.3.2 Hyperspectral image

NN analysis method was applied by Ahmadi *et al.* (2017) for separating and classifying spectral data of healthy and infected oil palm trees. The NN method used in this study was BP and multilayer owed to the ability to determine nonlinear combinations of raw, first, and second derivative spectral datasets. A total of 1016 foliar samples, which is 416 from the first trial and 600 from the second trial were obtained from frond numbers 9 and 17. Spectral data of foliar samples were scanned using a portable spectroradiometer at a 1.45 nm interval and range of 273 to 1,100 nm with a resolution of 5 nm. Data were separated between training and testing with a ratio of 70:30. Derivative data were calculated by division of the differences between spectral reflectance. Best results for differentiating T1 and T2 occurred in the visible range of green wavelength with an accuracy of 83.3%, and 100.0% for 540 nm to 550 nm, respectively. Other stresses i.e., lack of nutrition and weather could cause chlorophyll deficiency, which may affect the spectroscopy reflectance. Detection of BSR in oil palm had also been explored in North Sumatra, Indonesia using PP-SYSTEMS in the range of 310 to 1130 nm, with 256 bands and 10 nm resolutions (Lelong *et al.*, 2010). ML classification method using PLS-DA was applied and the results have shown that the proposed method could discriminate the healthy and infected trees with 98% accuracy and for four levels of classification with 94% accuracy. The systems were mounted on a 2 m shaft on top of a scaffold to measure the canopy of 95 oil palm trees consisting of 36 healthy trees, 18 trees of level 1, 38 trees of level 2 and 3 trees of level 3. The reflectance values were averaged from six to ten repetitions resulting in 202 spectral bands in the range of 450 to 1100 nm. Spectra pre-processing is essential to reduce the noise due to the difference in background reflectance and illuminations. In this research, Savitzky- Golay (SG) filtering was used as an unweighted linear least square fit method for data smoothing. Then, a large set of combinations were tested consisting of nine different window sizes, derivatives and polynomial fits, where the best result was a smoothing window of 26 nm with the second-order derivative of a third-degree polynomial. The challenge is to differentiate between levels 1 (mild) and level 2 (moderate), where the boundaries for these two levels are difficult to determine and only depend on field observation. The method could be advanced through satellite image application, while a larger sample size with an on-site tissue culture technique could be used to validate the findings for field application. In addition, a handheld portable hyperspectral spectroradiometer was used to collect leaf reflectance data of frond no. 17 from 47 healthy, 55 slightly damaged trees, 48 moderately damaged, and 40 heavily damaged oil palm trees (Liaghat *et al.*, 2014). Among the ML models used i.e., LDA and QDA, kNN and NB, PCA and kNN classification models resulted in 97% average overall accuracy with the second derivative dataset. Field data collection was performed during optimum illumination between 11:00 a.m. and 3:00 p.m. Three parts of a frond with 10 replications for each part were averaged. Meanwhile, the spectral range was 325 nm to 1040 nm resulting in 716 field reflectance values. The absorbance spectra also were normalised and baseline before running the SG filter for smoothing the data (Liaghat *et al.*, 2011). The filter window size was set to seven and the spectra were derived using second order to generate the first and second derivation datasets. Data were separated between training and testing with a ratio of 75:25. The results of the study showed that vigour discrimination techniques along with the ML approach are needed to support a consistent classification of the disease. Future work will involve a wider variation of oil palm trees in terms of ages of trees, soil types and varieties. Furthermore, a hyperspectral camera was attached to an aerial vehicle to scan 130 hectares of oil palm plantations in Colombia (Pinto *et al.*, 2019). The ML methods employed in this study are SU and CNN based on the spectral and spatial properties of the hyperspectral images of the oil palms. SU is the process of decomposing the spectral signature of a mixed pixel into a collection of endmembers and their corresponding fractions, or abundances. Simulation results showed that the best overall accuracy was up to 89% using only 20% of the training samples. The main characteristic of a CNN is the weight-sharing mechanism that can drastically decrease the number of parameters, and thus prevent the emergence of overfitting while reducing the complexity of the NN model (Gu *et al.*, 2018). The results are images with 299 x 294 pixels spatial resolutions, 160 hyperspectral bands and a range spectral of 400 nanometres to 1000 nanometres. The images were acquired through 430 meters of flight altitude and nine flight lines, while the ground truth was manually created with 3 colours i.e., healthy palms in green, diseased palms in red, and background in blue. Three ratios of data separation were used for training and validation – 10:90, 15:85 and 20:80.

3.3.3 RADAR

ALOS PALSAR 2 is a Synthetic Aperture Radar (SAR) sensor, which emits L-band microwave radio waves and was employed by Hashim *et al.* (2018a). Two polarisations from the satellite images, HH (horizontal-horizontal) and HV (horizontal-vertical) for reception and emission of radar acquisition were used. ML classification using DT classifier showed that the overall accuracies for HV and HH backscatter classification were 56.52% and 45.65%, respectively. The SAR data acquisition is independent of solar illumination is unaffected by the presence of cloud cover. Data from 55 healthy and 37 unhealthy oil palm trees were pre-processed to filter out the noise using the Sentinel Application Platform (SNAP). SNAP is open-source software for processing SAR data, including from ALOS PALSAR 2. The images were imposed radiometric calibration to remove the noise and patches. Then, the backscatter coefficients were converted from linear to dB and lastly were filtered using Enhanced Lee filter with 3x3 windows to remove the high-frequency components. Further analysis was conducted by Hashim *et al.* (2018b) to classify the oil palm trees into four different classes consisting of 55 healthy (T0), 11 mild infected (T1), 15 moderate infected (T2) and 11 severe infected (T3). Similar ML techniques were used by Hashim *et al.* (2018a), and only different classifier models were used – MLP and K-star. MLP is a class of NNs comprised of one or more layers of neurons. Meanwhile, K-star or K* is an instance-based classifier, where the classification is done by comparing it to a database of pre-classified examples with similar instances will have similar classifications. (Mahmood, 2013). The results revealed that the MLP classifier for HV polarisation accomplished better results than the Kstar classifier with 77% accuracy. In future studies, full polarisation, as well as multi-temporal data, can be investigated to develop an early detection technique of 419 *G. boninense* disease.

3.3.4 LiDAR

Terrestrial LiDAR was used to scan 40 oil palm trees in four levels of severity (Husin *et al.*, 2020). PCA was used to reduce the dimensionality of the original data and cross-validation was used to validate the dataset. The results showed that the classification learners of kNB using the first and second PCs (PC1 and PC2) attained the best results, where classification of healthy (T0) and mild infected trees (T1) was done with 100% accuracy, which is good for early detection of the disease. Five parameters were extracted from the scans: C200 (canopy stratum 200 cm from the top); C850 (canopy stratum 850 cm from the top); crown area (number of pixels inside the crown); frond angle and frond number. The registration process was done to merge and synch laser hits resulting from the four scan positions for each tree. C200 and C850 were generated from the stratification method of the 3D point cloud, while crown area, frond angle and frond number were generated from the top view image. The data were divided into a ratio of 80:20 for testing and training. The classification models with accuracy levels higher than 70% were considered for evaluation using another 40-prediction data. Furthermore, the classification model achieved an average model accuracy of 74.5% during training, an average accuracy of 85% when classifying four levels of infections and an average accuracy of 90% when classifying two levels of infections, healthy–unhealthy trees. In future research, the database will be broadened to improve classification accuracy. All the reviewed ML approaches used for BSR detection have been summarised in Table 3.

Table 3. Summary of ML approaches used for BSR detection

Oil palm part	Platform	Input data	ML techniques	Levels	Sampling details	Highest accuracy	References
Canopy	Satellite	Spectral (multispectral)	RF	2	144 palms - 99 healthy trees and 45 unhealthy. Ages from 10 to 21 years old	91%	(Santoso *et al.*, 2017)
	Satellite	Spectral (multispectral)	SVM	4	1923 palms - 695 healthy, 432 level 1, 348 level 2, and 448 level 3 Ages from 2 to 14 years old	54%	(Santoso *et al.*, 2019)
	Satellite	Spectral (RADAR)	MLP	4	92 palms - 55 healthy, 11 level 1, 15 level 2 and 11 level 3. Age 9 years old.	77.17%	(Hashim *et al.*, 2018b)
	Ground	Spectral (thermal)	SVM	2	106 palms - 53 healthy trees and 53 unhealthy.	89.2%	(Bejo *et al.*, 2018)
	Ground	Spectral (hyperspectral)	PLS-DA	4	95 oil palm trees – 36 healthy, 18 level 1, 38 level 2 and 3 level 3	94%	(Lelong *et al.*, 2010)
	Ground	Spectral (hyperspectral)	kNN	4	190 palms - 47 healthy, 55 level 1, 48 level 2, and 40 level 2 from frond no. 17	97%	(Liaghat *et al.*, 2014)
	Ground	Spectral (LiDAR)	KNB	4	40 palms - 10 trees in each level Age 9 years old	74.5%	(Husin *et al.*, 2020)

Oil palm part	Platform	Input data	ML techniques	Levels	Sampling details	Highest accuracy	References
Leaf	Aerial	Spectral (hyperspectral)	CNN	2	Spatial dimensions of 299 x 294 pixels and 160 hyperspectral bands	88.79%	(Pinto *et al.*, 2019)
	Ground	Electrical impedance	SVM, RF, GA, ANN	4	32 palms - 8 trees in each level 180 samples from frond no. 17	80.79%	(Khaled *et al.*, 2018)
	Ground	Spectral (hyperspectral)	ANN	4	374 palms 1,016 samples from frond numbers 9 and 17. Age 12 years old	100%	(Ahmadi *et al.*, 2017)
Trunk	Ground	Odour	LDA	2	n/a	100%	(Abdullah *et al.*, 2012)

4. Conclusions

In this paper, a summary of the selected useful and effective ML applications in the detection and classification of BSR disease in oil palm fields was presented. Throughout the survey, it was concluded that pre-processing techniques i.e., PCA is a common method used to reduce the dimensionality of data. SVM is perhaps a more popular method in ML because of its kernel ability in non-linear separation problems and its effectiveness in high dimensional spaces. Future predictions for ML applications are significantly greater adoption of ML models, opening the door to integrated and useful tools. The fusion of automated data collection, data analysis, ML deployment, and decision-making or assistance will offer useful benefits consistent with so-called knowledge-based agriculture for the solutions of oil palm disease problems. In addition, all the developed models were reasonably accurate in classification using both training and testing datasets. Different features were utilised by various ML algorithms for BSR disease detection; however, no model generally outperforms the others because the choice of approach will depend on constraints i.e., size of data and nature of the classification problems. BSR remains a serious problem in oil palm fields, thus, there are need to implement an effective, fast, accurate and automatic system, which can be fully utilised for disease detection in oil palm fields. The integration of the ML approach in the techniques will continue to grow and the application of such techniques to solve problems will be on the rise.

Acknowledgements: The authors would like to acknowledge the Ministry of Higher Education, Malaysia (MOHE) and Universiti Putra Malaysia (UPM) for sponsoring this research under grant number GP-IPM/2021/9697200

Conflicts of Interest: The authors declare no conflict of interest.

References

Abdullah, A. H., Adom, A. H., Shakaff, A. Y. M., *et al.* (2012). Hand-held electronic nose sensor selection system for basal stamp rot (BSR) disease detection. In 2012 Third International Conference on Intelligent Systems Modelling and Simulation (pp. 737–742). Kota Kinabalu, Malaysia, IEEE.

Abdullah, F., Flood, J., Bridge, P. D., *et al*. (2000). Spatial and sequential mapping of the incidence of basal stem rot of oil palms (Elaeis guineensis) on a former coconut (Cocos nucifera) plantation. *Ganoderma Diseases of Perennial Crops*, 183–194.

Abdullah, R., & Wahid, M. B. (2010). *World palm oil supply, demand, price and prospects: focus on Malaysian and Indonesian palm oil industry*. Malaysian Palm Oil Board Press. Malaysia.

Ahmadi, P., Muharam, F. M., Ahmad, K., *et al*. (2017). Early detection of Ganoderma basal stem rot of oil palms using artificial neural network spectral analysis. *Plant Disease*, *101*(6), 1009–1016. https://doi.org/10.1094/PDIS-12-16-1699-RE.

Azuan, N. H., Khairunniza-Bejo, S., Abdullah, A. F., *et al*. (2019). Analysis of changes in oil palm canopy architecture from basal stem rot using terrestrial laser scanner. *Plant Disease*, *103*(12), 3218–3225. https://doi.org/10.1094/PDIS-10-18-1721-RE

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31.

Bejo, S., Abdol-Lajis, G., Abd-Aziz, S., *et al*. (2018). Detecting Basal Stem Rot (BSR) disease at oil palm tree using thermal imaging technique. *Proceedings of the 14th International Conference on Precision Agriculture*, International Society of Precision Agriculture: Monticello, Illinois, U.S.A.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., *et al*. (2008). Support vector machines and kernels for computational biology. *PLoS Computational Biology*, *4*(10).

Boilot, P., Hines, E.L., Gardner, J.W., *et al.* 2002. Classification of bacteria responsible for ENT and eye infections using the Cyranose system. IEEE Sensors Journal2 (3), 247–253.

Bonate, P. L. (2011). Variance models, weighting, and transformations. In *Pharmacokinetic-Pharmacodynamic Modeling and Simulation* (pp. 131–156). Springer, Boston, MA.

Cattell, R. B., & Jaspers, J. (1967). A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*.

Chong, K. P., Dayou, J., & Alexander, A. (2017). Pathogenic nature of Ganoderma boninense and basal stem rot disease. In *Detection and Control of Ganoderma boninense in Oil Palm Crop* (pp. 5–12). Springer, Cham.

Clay, J. (2011). Freeze the footprint of food. *Nature*, *475*(7356), 287.

David, C. C. & Jacobs, D. J. (2014). Principal component analysis: A method for determining the essential dynamics of proteins. In *Protein Dynamics* (pp. 193–226). Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-658-0_11.

Ebong, P. E., Owu, D. U., & Isong, E. U. (1999). Influence of palm oil (Elaesis guineensis) on health. *Plant Foods for Human Nutrition*, *53*(3), 209–222.

Fausett, L. (1994). *Fundamentals of neural networks: Architectures, algorithms, and applications*. Prentice-Hall, Inc.

Fife, B. (2017). *The shocking truth about palm oil: How it affects your health and the environment*. Piccadilly Books, Ltd.

Flood, J., Bridge, P. D., & Holderness, M. (2000). *Ganoderma diseases of perennial crops*. CABI.

Golhani, K., Balasundram, S. K., Vadamalai, G., *et al*. (2018). A review of neural networks in plant disease detection using hyperspectral data. *Information Processing in Agriculture*, *5*(3), 354–371. https://doi.org/10.1016/j.inpa.2018.05.002

Gu, J., Wang, Z., Kuen, J., *et al*. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Guyon, I., Weston, J., Barnhill, S., *et al*. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422.

Hashim, I. C., Shariff, A. R. M., Bejo, S. K., *et al*. (2018a). Classification for non-infected and infected ganoderma boninense of oil palm trees using ALOS PALSAR-2 backscattering coefficient. *IOP Conference Series: Earth and Environmental Science*, *169*(1), 012066.

Hashim, C. I., Rashid, M. S. A., Bejo, S. K., *et al*. (2018b) Severity of Ganoderma boninense disease classification using SAR data. In: *39th Asian Conference on Remote Sensing (ACRS 2018)*, 15-19 Oct. 2018, Renaissance Kuala Lumpur Hotel, Malaysia. (pp. 2492–2499).

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Holland, S. M. (2008). *Principal components analysis (PCA)*. Department of Geology, University of Georgia, Athens, GA, 30602–32501.

Husin, N. H., Khairunniza-Bejo, S., Abdullah, A. F., *et al*. (2020). Classification of basal stem rot disease in oil palm plantations using terrestrial laser scanning data and machine learning. *Agronomy*, *10*(11), 1624.

Husin, N. A., Khairunniza-Bejo, S., Abdullah, A. F., *et al*. (2021). Multi-temporal analysis of terrestrial laser scanning data to detect basal stem rot in oil palm trees. *Precision Agriculture*, 1–26.

Kaestner, C. A. A. (2013). Support vector machines and kernel functions for text processing. *Revista de Informática Teórica e Aplicada*, *20*(3), 130–154.

Khaled, A. Y., Aziz, S. A., Bejo, S. K., *et al*. (2018). Spectral features selection and classification of oil palm leaves infected by Basal stem rot (BSR) disease using dielectric spectroscopy. *Computers and Electronics in Agriculture*, *144*, 297–309. https://doi.org/10.1016/j.compag.2017.11.012.

Lelong, C. C., Roger, J. M., Brégand, S., *et al.* (2010). Evaluation of oil-palm fungal disease infestation with canopy hyperspectral reflectance data. *Sensors, 10*(1): 734–747. https://doi.org/10.3390/s100100734

Li, J., Qi, G., Zhao, D., *et al*. (2013). *Semantic web and web science*. Springer: New York.

Liaghat, S., Ehsani, R., Mansor, S., *et al*. (2014). Early detection of basal stem rot disease (Ganoderma) in oil palms based on hyperspectral reflectance data using pattern recognition algorithms. *International Journal of Remote Sensing*, *35*(10), 3427–3439. https://doi.org/10.1080/01431161.2014.903353

Liaghat, S., Mansor, S. B., Shafri, H. Z. M., *et al*. (2011). Early detection of oil palm fungal disease infestation using a mid-infrared spectroscopy technique. In *International Conference on Precision Agriculture,* Indianapolis, Indiana, USA.

Liakos, K., Busato, P., Moshou, D., *et al*. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8): 2674. https://doi.org/10.3390/s18082674

Mahmood, D. Y., & Hussein, M. A. (2013). Intrusion detection system based on K-star classifier and feature set reduction. *International Organization of Scientific Research Journal of Computer Engineering*, *15*(5), 107–112.

Markom, M. A., Shakaff, A. M., Adom, A. H., *et al*. (2009). Intelligent electronic nose system for basal stem rot disease detection. *Computers and Electronics in Agriculture*, *66*(2), 140–146.

McMorrow, J. (2001). Linear regression modelling for the estimation of oil palm age from Landsat TM. *International Journal of Remote Sensing*, *22*(12), 2243–2264.

Mielke, T. (2018). *Oil World weekly*. ISTA Mielke Gmbh, Hamburg. Retrieved from https://www.oilworld.biz/t/publications/weekly.

Malaysian Palm Oil Council (MPOC). (2017). *Palm oil*. Retrieved from http://www.mpoc.org.my/Palm_Oil.aspx

Naher, L., Yusuf, U. K., Ismail, A., *et al*. (2013). Ecological status of 'Ganoderma' and basal stem rot disease of oil palms ('Elaeis guineensis' Jacq.). *Australian Journal of Crop Science*, *7*(11): 1723.

O'Grady, M., Vahdat-Nejad, H., Wolf, K. H., *et al*. (2013). *Evolving ambient intelligence*: AmI 2013 Workshops, Dublin, Ireland, December 3–5, 2013. Revised Selected Papers (Vol. 413). Springer.

Okamoto, K. (2001). *Global environment remote sensing*. IOS Press.

Panigrahi, N. (2014). *Computing in geographic information systems*. CRC Press.

Paterson, R. R. M. (2007). Ganoderma disease of oil palm —A white rot perspective necessary for integrated control. *Crop Protection*, *26*(9), 1369–1376.

Pinto, J. E., Ramírez, J. M., & Arguello, H. (2019). Classification of oil palm diseases via spectral unmixing and convolutional neural networks. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI* (Vol. 11149, pp. 94-100). SPIE.

Pornsuriya, C., Sunpapao, A., Srihanant, N., *et al*. (2013). A survey of diseases and disorders in oil palms of Southern Thailand. *Plant Pathology Journal*, *12*(4), 169–175.

Rees, R. W., Flood, J., Hasan, Y., *et al*. (2007). Effects of inoculum potential, shading and soil temperature on root infection of oil palm seedlings by the basal stem rot pathogen Ganoderma boninense. *Plant Pathology*, *56*(5): 862–870.

Samsudin, S. H., Shafri, H. Z., Hamedianfar, A., *et al*. (2015). Spectral feature selection and classification of roofing materials using field spectroscopy data. *Journal of Applied Remote Sensing*, *9*(1), 095079.

Santoso, H., Tani, H., & 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. https://doi.org/10.1080/01431161.2017.1331474

Santoso, H., Tani, H., Wang, X., *et al*. (2019). Classifying the severity of basal stem rot disease in oil palm plantations using WorldView-3 imagery and machine learning algorithms. *International Journal of Remote Sensing*, *40*(19): 7624–7646. https://doi.org/10.1080/01431161.2018.1541368

Shafri, H. Z., Hamdan, N., & Saripan, M. I. (2011). Semi-automatic detection and counting of oil palm trees from high spatial resolution airborne imagery. *International Journal of Remote Sensing*, *32*(8): 2095–2115.

Shanmuganathan, S., & Samarasinghe, S. (Eds.). (2016). *Artificial neural network modelling* (Vol. 628). Springer.

Turner, P.D. (1965) The incidence of Ganoderma disease of oil palm in Malaya and its relation to previous crop. *Annals of Applied Biology*, *55*, 417–423.

USDA. (2019). Oilseeds: World Market and Trade, March 2019. Foreign Agricultural Service, USDA: Washington, DC.