Machine Learning for diagnosis of disease in plants using spectral data

Godliver Owomugisha[1], Friedrich Melchert[2], Ernest Mwebaze[3], John A Quinn[4] and Michael Biehl[5]

[1,2,5]*University of Groningen*
Johann Bernoulli Institute for Mathematics and Computer Science,
P.O. Box 407, 9700 AK Groningen, The Netherlands

[2]*Fraunhofer Institute for Factory Operation and Automation IFF,*
Sandtorstrasse 22, 39106 Magdeburg, Germany

[1,3,4]*Makerere University, School of Computing & Informatics Technology*
P.O. Box 7062 Kampala, Uganda

[1]*Busitema University, Faculty of Engineering*
P. O. Box 236, Tororo, Uganda

Email: [g.owomugisha, m.biehl]@rug.nl, friedrich.melchert@iff.fraunhofer.de, [emwebaze, jquinn]@cit.ac.ug

Abstract*—Automating crop disease diagnosis is an important task, particularly for regions with few experts. Most current methods detect disease by analyzing leaf images, particularly for diseases that manifest on the aerial part of the plant. To train a good classifier one requires a huge image dataset and the appropriate methods to extract relevant features from the images that represent the disease unambiguously. Image data also tends to be prone to effects of occlusion that make consistent analysis of the data hard. In this paper we take a look at the use of spectral data collected from leaves of a plant. We analyse spectral data from visibly diseased parts of a leaf as well as parts that are visibly healthy. We employ prototype based classification methods and standard classification models in a three-class classification problem configuration. Results presented show significant improvement in performance when spectral data is used and the possibility of early detection of disease before the crops become visibly symptomatic, which for practical reasons is very important.

Keywords: Spectral data, disease diagnosis, crop images, prototype-based classification, neural networks.

1. Introduction

The state of the art method of identifying diseases in plants in the field is by use of visual symptoms which an agricultural expert is able to relate to particular diseases in the plant. For places where experts are not available or where farmer knowledge is insufficient, other methods for carrying out field-based diagnoses are a critical need. Computational work in this area has been towards automating this process through building machine learning models that can take an image of a leaf and predict whether the plant is infected with a particular disease or not.

In this work, we focus on an important crop for Sub-Saharan Africa and other regions, *Cassava (Manihot esculenta)*. Cassava is the second most important food crop in Sub-Saharan Africa especially amongst smallholder farmers because it can easily be grown in poor soils and requires few inputs. It is also a very important food security crop for the same reasons. Although cassava is known to survive under harsh conditions, its productivity has greatly been affected by pests and viral diseases in recent years causing losses in the millions of dollars [1]. Our work in this paper particulary looks at two critical viral diseases in Cassava; Cassava mosaic disease (CMD) and Cassava brown streak disease (CBSD).

This research builds on previous work in the area [2], [3], [4] and considers work by other groups that has focused on automating the detection of cassava diseases e.g. [5]. Most of the earlier work considers the use of leaf images as the key data input into the model and in order to be effective, diseases symptoms need to be visible or in advanced stages. From a practical point of view, however, once symptoms have manifested, little can be done to save the situation since the disease has spread to almost all the neighboring plants.

Spectroscopy is a field aimed at studying how different materials interact with light, particularly which wavelengths will be absorbed or reflected by a material once the material is exposed to rays of light. We leaverage spectroscopy in this study to attempt to understand how plants manifesting different diseases interact with light. Our hypothesis is that disease causes several metabolic changes in the biology of the leaf that can be teased out through spectroscopy. To this end we collect spectral data from diseased and healthy cassava leaves.

A key outlook from this work is the possibility of detecting disease earlier or before a diseased plant is symptomatic. This has implications in the timeliness and effectiveness of interventions that can be applied to the crops. We test this hypothesis by looking at leaves on visibly diseased plants that still look healthy, so we know they are infected but are not yet symptomatic, as well as looking at visibly diseased parts of the plant as depicted in Figure 2. Our results indicate the possibility of this technique actually working for early

detection of disease using spectrometry.

In the following sections we discuss some work that has already been done in this area, looking at image based techniques for diagnosis as well as the use of spectrometry for inferring disease in other crops. We also discuss our data collection, experiments and results from applying several algorithms to this data. We end the paper with a brief discussion of the results and conclusion.

2. Related work

Several attempts have been made to diagnose disease using leaf image data. Image data presents a natural means in this context because the disease manifests visibly on the leaf. Spectrometry goes further by potentially capturing underlying mechanisms in the leaf that are associated with the disease. We review some of the work in these two broad methodologies.

2.1. Disease detection using image data

One of the first pieces of work in this regard was published by Aduwo et. al. [2], who present the use of computer vision to diagnose cassava diseases. They used leaf images of cassava plants taken in a lab setting with uniform lighting and background. A two-class algorithm was developed that detects whether an image of a leaf is from a diseased or health plant. Three sets of features were extracted from leaf images including features related to the hue and intensity of the image (HSV), and features that capture interest keypoints on the image, including Scale Invariant Feature Tranformation (SIFT) features and Speeded-Up Robust Features (SURF).

The classification methods used were improved upon in subsequent work [3], [6], [7] where improved implementations of prototype-based classification schemes were employed. Several other work extended the two-class problem to a multi-class problem with multiple diseases and different severity levels of disease [4]. Several other features were used in these extensions of the initial work by Aduwo et. al.

More recent image based approaches for cassava disease detection are based on deep learning, e.g. Ramcharan et. al. [5], where several images of the different diseases in cassava were used to build a deep neural network that was able to detect disease with relatively good performance. Further attempts at using deep neural networks have also been shown to work in other studies [8]. A key advantage of the use of deep neural networks is that the features corresponding to the disease need not be hand crafted, the model is able to learn the relevant features given sufficient training data. The drawback with these methods is that significant amounts of training data is required to build these networks.

Several other image based approaches to crop disease detection have been suggested in the literature, see e.g. [9], [10], [11].

Obviously, any image based technique, whether it is combined with machine learning or not, relies on the presence of visual symptoms. However, once symptoms have manifested, not much can be done to control the situation. For some of the diseases in cassava for example, the root of the plant is already affected and cannot be used for food consumption. Frequently the disease has also spread across neighboring plants. The need for early detection of disease before the plant is symptomatic is profound. One direction we investigate in this paper is the use of spectrometry. Obtaining a spectral signature of a leaf, we surmise, will be more informative of the state of disease of the plant than image data particularly if we want to determine disease before the plant is visibly symptomatic.

2.2. Spectrometry for disease diagnosis

Imaging spectroscopy has received broad interest in various sectors of agricultural research, including soil science [12], [13] and crop disease monitoring. A good review of some of the imaging spectroscopy technologies used can be found here [14].

The range of work done in this regard is diverse. Feng et. al. [15] present a multispectral imaging system for the diagnosis of plant diseases and insect pests. They apply the same suite of methods in diagnosing cucumber diseases as well [16]. Spectroscopy has also been used to detect mechanical and disease stresses in citrus plants [17], [18].

Further examples include Bo et. al. [19] who present a field imaging spectroscopy system that was used to predict the chlorophyll content from soybean leaves using linear regression, partial least squares regression and support vector machine regression. In [20], methods for early detection of rice blast using near-infrared hyper-spectral images are also presented.

Overall, spectroscopy as a tool for measuring the state of a material is becoming prevalent and in this work we show first attempts at leveraging it to detect viral diseases in cassava. Spectral data being generally high dimensional, we also present feature engineering methods we employed to take care of the dimensionality.

3. Experiments

Here we describe the data collection process, preprocessing and analysis of the acquired data.

3.1. Data collection

To carry out the experiments, two types of data were collected, each dataset broken up into different categories to represent the disease classes. Figure 1 illustrates the data collection pipeline for automating cassava disease diagnosis. The first type of data consisted of 760 images of cassava leaves in the field, taken using a smartphone camera with a resolution of 72dpi. The leaf images were evenly split in three categories; (i) those representing Cassava brown streak

disease (CBSD), (ii) those representing Cassava mosaic virus disease (CMD) and (iii) those representing healthy control plants (HC). Experiments on this data focused on image-based techniques of disease diagnosis.

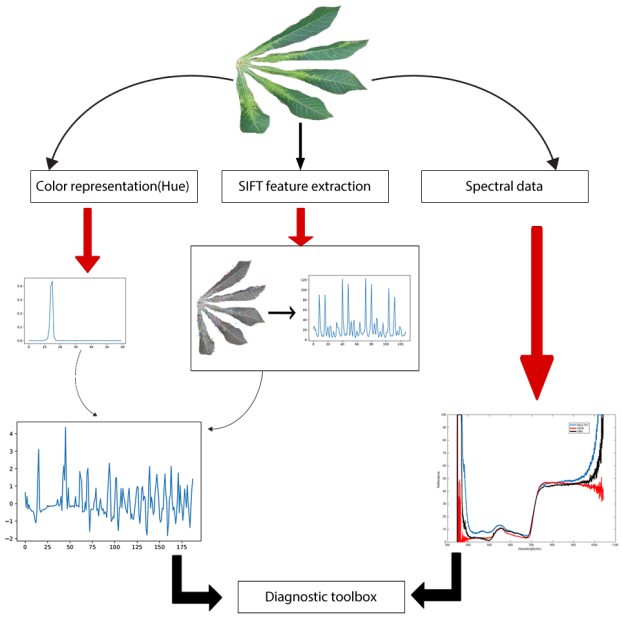

Figure 1. Cassava disease automated diagnostic pipeline as described in Section 3.1

The second type of data acquired was spectrometry data corresponding to the leaves from which the image data was collected. This data was acquired with the use of a CI-710 miniature leaf spectrometer [21]. The device is USB powered and portable so it can be used to collect field measurements. Specialised software that comes with the device allows us to collect the spectra from the leaves. From experiments carried out in the field, we realised that several parameters influence the intensity and shape of the spectra obtained, illumination being of particular importance. For this reason, we collected data directly in the fields under natural light. We also focused on reflectance mode since previous measures and experiments did not show significant difference between reflectance and transmission spectra obtained for these leaves.

We collected data for plants aged 6 to 9 months. At this age, diseased plants manifest symptoms. We collected data across five cassava varieties. For each variety, three plants were considered and of each plant, three leaves were sampled. The cassava leaf has multiple lobes, thus for each leaf, two readings were taken on each leaf lobe: one on the *good* part (not visibly showing symptoms) and the other on the *bad* part (part showing visible disease symptoms). Because the spectrometer takes readings on a small area of the plant about 7.6 mm in diameter, readings for every leaf lobe were recorded in order to achieve a representative and reliable sampling representing a single leaf. Note that this was taken care of during validation of the models, so that

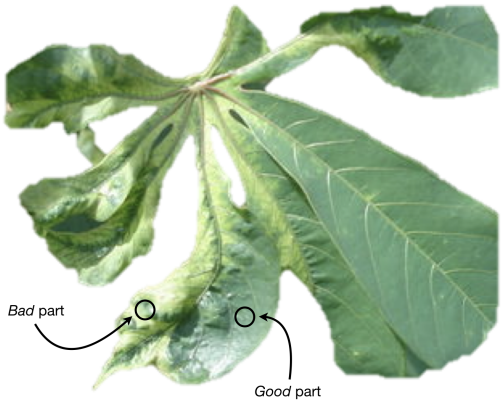

Figure 2. Depiction of good and bad part of leaf

we never trained and tested on data from the same plant. In total, 760 data points were collected for evenly distributed disease classes. Figure 2 illustrates the good and bad leaf parts of interest.

3.2. Feature extraction

3.2.1. Image data feature extraction. Following methodologies from previous work [3] on cassava disease diagnosis using leaf images, we extracted color (HSV) and SIFT features because they have been shown to accurately capture the manifestation of the different diseases in the leaves of cassava plants. For color, a Hue, Saturation, Value (HSV) color transformation of the image is computed. Of the three components, Hue has been found to be more significant and histograms of 60 bins of this component were considered. SIFT feature descriptors of 128 dimensions were also extracted. Both color and SIFT features were computed using the standard OpenCV toolbox [22].

3.2.2. Spectral data pre-processing. A single spectrogram representing one reading on a leaf presented as a 2,554 dimensional vector with noisy components at each end of the spectrogram. The first pre-processing step is to truncate the spectrogram to within the limits of operation as set for the equipment [21] which is an interval of wavelengths from 400nms to 900nms. For all spectra hence, the region of the spectrum between the wavelengths 400 nms – 900nms was considered for the next processing steps. A further pre-processing step done was smoothening the spectra. For this, we compared two filtering techniques: median filtering [23] and average filtering [24]. For both, we used a window size of 15 nm. Preliminary experiments indicated that the use of average filtering yielded better classification results. As a consequence, average filtering was applied to the spectral data.

In the experiments, we compare performance when using the high dimensional spectral data and when using a reduced dimension dataset. Dimensionality reduction is important for practical deployment purposes. Here, we apply Principal Component Analysis (PCA). PCA is a standard technique

for correlation analysis and dimensionality reduction which has been widely used [25]. PCA can be used to project high-dimensional data linearly to a low-dimensional space in which most of the statistical variation is preserved [26], [27].

3.3. Training a diagnosis classifier

Several options abound for which type of model to train for this kind of data. Previous work has used convolutional neural networks (CNNs) and prototype based methods with great success. We are restricted in the use of CNNs here because of the limited size of our dataset. Our choice was thus prototype based Learning Vector Quantization (LVQ). We compare this method with some standard machine learning algorithms from the SciKit-learn toolbox [28]. For our experiments we use the following: (i) K-Nearest Neighbour (KNN) because it is very similar in flavor to prototype based methods, (ii) Linear Support Vector Machine (SVM) because it has shown good performance previously [3], and (iii) Decision trees, because these have also previously shown good performance particularly the Extremely Randomized Trees (Extra trees) algorithm [4].

3.3.1. Prototype-based classification methods. As a set of methods that have given good performance in previous classification tasks related to cassava images, we give a small review of the motivation behind prototype-based classification methods. Suffice to say that because these methods essentially train a prototype which is in the space of the data, they are very intuitive and for deployment purposes very simple to integrate into a diagnosis pipeline like that on a smartphone.

The simplest prototype-based classification algorithm, Learning Vector Quantization (LVQ) was introduced by Kohonen [29] in 1986 and since then various modifications have been suggested in the literature all aiming at better convergence or favorable generalization [30], [31]. In LVQ, a particular classification task is defined by a set of M prototype vectors $\mathrm{w}^j \in \mathrm{R}^N$ which carry labels $c(\mathrm{w}^j) \in \{1, 2, ...C\}$ such that $W = \{\mathrm{w}^j, c(\mathrm{w}^j)\}_{j=1}^{M}$. The system can be set up with one or more prototype vectors per class. For this experiment, we considered one prototype vector for each class.

A nearest prototype classifier (NPC) assigns a given feature vector $\mathrm{x} \in \mathrm{R}^N$ to the *closest* prototype with respect to some meaningful distance measure. Most frequently, standard Euclidean distance $d(\mathrm{w}, \mathrm{x})$ is employed. The corresponding NPC assigns x to the class $c(\mathrm{w}^{\mathrm{L}})$ of the closest prototype with $d^{\Lambda}(\mathrm{x}, \mathrm{w}^{\mathrm{L}}) \leq d^{\Lambda}(\mathrm{x}, \mathrm{w}^{\mathrm{j}})$ for all j.

An important conceptual extension of the basic LVQ concept is so-called relevance learning: There, an adaptive distance d^{Λ} is used where Λ denotes a set of adjustable parameters which are adapted, together with the prototypes, in a data-driven training process.

The GMLVQ algorithm proposed in [31] employs a full matrix $\Lambda \in \mathrm{R}^{N \times N}$ of relevances that represents the importance of single features and their combinations in the classification task. Here, the distance measure $d^{\Lambda}(\mathrm{x}, \mathrm{w})$ is defined as:

$$d^{\Lambda}(\mathrm{x}, \mathrm{w}) = (\mathrm{x} - \mathrm{w})^{\top} \Lambda (\mathrm{x} - \mathrm{w}), \tag{1}$$

where the parameterization $\Lambda = \Omega^{\top}\Omega$ guarantees that $d^{\Lambda}(\mathrm{x}, \mathrm{w}) \geq 0$ for arbitrary matrices $\Omega \in \mathrm{R}^{N \times N}$. In order to avoid numerical degeneracies, a normalization constraint of the form

$$\sum_{i=1}^{N} \Lambda_{ii} = \sum_{i,j=1}^{N} \Omega_{ij}^2 = 1$$

is imposed. In GMLVQ, the training process is guided by the optimization of a cost function of the form suggested by [30]:

$$E(W) = \sum_{\mu=1}^{p} \Phi\left(\frac{d_J^{\Lambda}(\mathrm{x}_i) - d_K^{\Lambda}(\mathrm{x}_i)}{d_J^{\Lambda}(\mathrm{x}_i) + d_K^{\Lambda}(\mathrm{x}_i)}\right) \tag{2}$$

where d_J^{Λ} denotes the distance to the closest correct prototype with $\mathrm{c}(\mathrm{w}^{\mathrm{J}}) = \mathrm{y}^{\mu}$ and d_K^{Λ} is the distance to the closest incorrect prototype $\left(\mathrm{c}(\mathrm{w}^{\mathrm{J}}) \neq \mathrm{y}^{\mu}\right)$. The modulation function Φ is frequently chosen to be a sigmoidal function. Here we resort to the identity $\Phi(x) = x$ in order to avoid the introduction and tuning of additional parameters.

This model based on learning a relevance matrix [31], [32] also provides us with a way of reducing the dimensionality of the spectral data in this case. Part of our future work will be to extend this method to identify relevant wavelengths that are most important for classifying the different diseases. One can then extend this to the construction of a simpler, cheaper spectrometry tool that offers analysis in a limited wavelength band.

3.3.2. Validation. For all the models we train, we carry out a 10-fold cross validation and average the performance over the folds. We employ parameter K=15 for the KNN algorithm, C = 1 for the linear SVC and 200 estimators for the Extra trees algorithm. For the GMLVQ algorithm we employ standard parameters used in the GMLVQ tool box which is available online [33].

A particular precaution had to be made for the spectral data. Since the data collection process involved picking more than one sample from a particular plant, it was important to choose a validation strategy that matches this condition in order to avoid training and testing on data from the same plant. We kept track of the class label (HC, CBSD and CMD) as well as the unique plant labels (also called groups). During training, the cross validation splits were based on plant groups and the validation scheme was Shuffle-Group(s)-Out crossvalidation as implemented in the SciKit-learn toolbox [34].

4. Results

4.1. *Good* vs. *bad* part of leaves in spectral data

A key aspect of this work was to figure out whether the location of where spectra was taken from a leaf matters, particularly if there is a significant difference between taking

spectral data from visibly infected parts of the leaf (bad part) or from parts of the leaf that are not visibly infected (good part). We run the battery of algorithms on the two datasets and present the results in Table 1. The results (accuracy scores) point towards a marginal difference between the two parts for some of the algorithms, SVC, KNN and Extra trees, but show significant difference for GMLVQ. All results presented here are for a multi-class problem and the two datasets are composed of three classes (Healthy, CBSD and CMD disease).

TABLE 1. SPECTRAL DATA DEPENDENCE ON LEAF QUALITY (HEALTHY VS. CBSD, CMD)

Classifier	Leaf part	
	Bad	***Good***
KNN	0.919	0.923
Linear SVC	0.941	0.957
Extra trees	0.917	0.927
GMLVQ	0.937	0.973

TABLE 2. CONFUSION MATRIX FOR *Bad* PART OF LEAF WITH GMLVQ

	Healthy	**CBSD**	**CMD**
Healthy	**98.70**	1.30	0
CBSD	0	**100**	0
CMD	0	17.04	**82.96**

Given this initial analysis, for the rest of the experiments we use the spectral data taken from the *good* part of the leaf.

4.2. Image-based features vs spectral data

Our hypothesis is that spectral data can offer better representation of the inherent disease in the plant than image data. Our first experiment was to test this hypothesis. Table 4 gives a depiction of the results. As is evident we see superior performance of each of the algorithms on spectral data than on the color and SIFT features extracted from the images. The metric used is the accuracy score. From the results, it appears spectral data is a more useful representation of the cassava plants than image data. A drawback one immediately sees is that the dimensionality of the spectral data (2554 features) presents a challenge.

4.3. PCA spectral features

Using the technique for dimensionality reduction described earlier, PCA, we are able to reduce the spectral data dimension from 2554 to 30 principal components. In Figure 3, we illustrate the performance for n-principal components thus justifying the choice for using 30 principal components. We use these as features in the training of the battery of classification algorithms. Table 4 shows the results from these experiments. Generally we see a marginal improvement in all algorithms when this reduced feature set is used. A clear advantage from this is that there is a reduction in noise in the data when we do a PCA transformation, however the corresponding disadvantage is that we are not not able to identify the relevant wavelengths critical for classification of the different diseases. Also for a live system this introduces a computational penalty in transforming the data, however which could be offset by the reduced time to do the prediction.

TABLE 3. CONFUSION MATRIX FOR *Good* PART OF LEAF WITH GMLVQ

	Healthy	**CBSD**	**CMD**
Healthy	**100**	0	0
CBSD	0	**100**	0
CMD	0	8.2	**91.8**

TABLE 4. PERFORMANCE WITH DIFFERENT DATA FEATURES (HEALTHY VS. CBSD, CMD)

Classifier	Color	SIFT	Spectral	
			Original	***PCA***
KNN	0.705	0.849	0.923	0.932
Linear SVC	0.738	0.895	0.957	0.959
Extra trees	0.803	0.889	0.927	0.944
GMLVQ	0.742	0.901	0.973	1.000

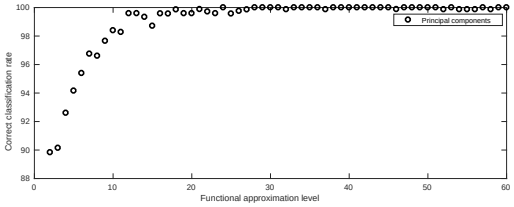

Figure 3. Performance with increasing number of principal components

5. Discussion

This paper has introduced a method based on spectrometry which constitutes a novel approach in the context of field based diagnosis of cassava disease. To the best of our knowledge, we are not aware of any other work combining spectrometry and classification for cassava diseases as presented here. Our experimental results are promising. The first result in Table 1 comparing the bad and good parts of a leaf was a bit surprising. In these results, we noticed a marginal difference in the performance albeit for one algorithm where there was a significant increase in performance for the experiment using the spectral data from the good part. Analysis of the confusion matrix, Table 2 gives a glimpse at why this may be so; for experiments with the bad part we observe the classifier confusing CMD and CBSD diseases which could result from the metabolic mechanisms that represent disease being obscured by the visibly infected part of the plant.

Table 4 provides evidence of the superiority of spectral data compared to image data for classification of viral disease in cassava. One explanation is that spectral data captures the inherent metabolic changes related to the disease infecting the plant, and probably different diseases

manifest differently in different plants. As mentioned image data is also prone to occlusion making it less accurate in prediction. The GMLVQ algorithm however provides superior performance compared to other algoirthms as shown in Table 3 probably because the nature of the data allows for formulation of very representative prototypes.

A practical problem with the use of the spectral data is the large dimension of the data. A possible solution is to use PCA for the extraction of the most relevant information. Table 4 presents results of running the same battery of algorithms on the PCA representation of the spectral data (30 features). We observe very high accuracies for the reduced set. For practical purposes, a model based on a reduced set of features is best. But we lose interpretability of the features, in this particular case, the wavelength band that would be critical for detection of a particular disease. However, GMLVQ also provides us with another advantage to reconstruct the original features using the coefficients thus 30 principal components are a good representation for our problem.

6. Conclusion

In this paper we have shown the efficacy of using spectral data to do field diagnosis of disease compared with image data, the de-facto automated diagnosis methodology. Experiments show a significant gain in prediction accuracy for disease with spectral data. This work has also demostrated the consistency of spectral data collection from different parts of the leaf. Particularly of interest is the collection of spectral data from the *good* part of the leaf which has implications for doing detection of disease in the plants before they are symptomatic. This will form the crust of our future work.

Acknowledgments

The authors would like to thank their partners at the Uganda National Crop Resources Research Institute (NaCRRI) for granting us permission to access cassava fields to collect data for this study. The authors particularly extend their thanks to Dr. Ephraim Nuwamanya of NaCRRI for the support he showed us in the data collection process. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. The work is supported by a grant from the Bill and Melinda Gates Foundation (OPP1112548) to whom we are ever so grateful for the support.

References

[1] E. Nuwamanya, Y. Baguma, E. Atwijukire, S. Acheng, and T. Alicai, "Competitive commercial agriculture in sub saharan africa," *International Journal of Plant Physiology and Biochemistry*, vol. 7(2), pp. 12–22, 2015.

[2] J. R. Aduwo, E. Mwebaze, and J. A. Quinn, "Automated vision-based diagnosis of cassava mosaic disease," *Industrial Conference on Data Mining - Workshops*, pp. 114–122, 2010.

[3] E. Mwebaze and M. Biehl, "Prototype-based classification for image analysis and its application to crop disease diagnosis," *Advances in Self-Organizing Maps and Learning Vector Quantization - Proceedings of the 11th International Workshop WSOM 2016*, pp. 329–339, January 2016.

[4] G. Owomugisha and E. Mwebaze, "Machine learning for plant disease incidence and severity measurements from leaf images," *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 158–163, 2016.

[5] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," *Frontiers in Plant Science*, vol. 8, p. 1852, 2017.

[6] E. Mwebaze, P. Schneider, F.-M. Schleif, J. Aduwo, J. Quinn, S. Haase, T. Villmann, and M. Biehl, "Divergence-based classification in learning vector quantization," *Neurocomputing*, vol. 74, no. 9, pp. 1429 – 1435, 2011.

[7] E. Mwebaze, M. Biehl, G. Bearda, and D. Zuehlke, "Combining dissimilarity measures for prototypebased classification," *European Symposium on Artificial Neural Networks (ESANN)*, vol. 23, pp. 31–36, 2015.

[8] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, p. 11, 2016.

[9] K. K. R. Gokulakrishnan, "Detecting the plant diseases and issues by image processing technique and broadcasting," *International Journal of Science and Research*, vol. 3, 2014.

[10] S. D. Khirade and A. B. Patil, "Plant disease detection using image processing," *Proceedings of the 2015 International Conference on Computing Communication Control and Automation*, pp. 768–771, 2015.

[11] M. Nixon and A. S. Aguado, "Feature extraction & image processing for computer vision, third edition," *Academic Press*, 2012.

[12] W. Johanna, B. Stenberg, and R. A. V. Rossel, "Soil analysis using visible and near infrared spectroscopy," *Plant Mineral Nutrients: Methods and Protocols*, 2013.

[13] L. Raphael, "Application of ftir spectroscopy to agricultural soils analysis," *Chapter from the Book Fourier Transforms*, 2011.

[14] S. Sindhuja, M. Ashish, E. Reza, and D. Cristina, "A review of advanced techniques for detecting plant diseases," *Computers and Electronics in Agriculture*, vol. 72, pp. 1 – 13, 2010.

[15] J. Feng, N.-F. Liao, M.-Y. Liang, B. Zhao, and Z.-F. Dai, "Multispectral imaging system for the plant diseases and insect pests diagnosis," *Guang Pu Xue Yu Guang Pu Fen Xi*, 2009.

[16] J. Feng, N.-F. Liao, B. Zhao, Y.-D. Luo, and B.-J. Li, "Cucumber diseases diagnosis using multispectral imaging technique," *Guang pu xue yu guang pu fen xi = Guang pu*, 2009.

[17] J. J. Belasque, M. C. G. Gasparoto, and L. G. Marcassa, "Detection of mechanical and disease stresses in citrus plants by fluorescence spectroscopy," *Applied Optics*, vol. 47, no. 11, pp. 1922–1926, 2008.

[18] C. B. Wetterich, R. Kumar, S. Sankaran, J. J. Belasque, R. Ehsani, and L. G. Marcassa, "A comparative study on application of computer vision and fluorescence imaging spectroscopy for detection of huanglongbing citrus disease in the usa and brazil," *Journal of Spectroscopy*, 2013.

[19] L. Bo, Y. Yue-Min, L. Ru, S. Wen-Jing, and W. Ke-Lin, "Plant leaf chlorophyll content retrieval based on a field imaging spectroscopy system," *Sensors*, vol. 14, no. 10, 2014.

[20] Y. Yang, R. Chai, and Y. He, "Early detection of rice blast (pyricularia) at seedling stage in nipponbare rice variety using near-infrared hyper-spectral image," *African Journal of Biotechnology*, vol. 11, pp. 6809–6817, 2012.

[21] C. B.-S. Inc, "Ci-710 miniature leaf spectrometer," 2010. [Online]. Available: http://www.cid-inc.com

[22] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[23] E. Arias-Castro and D. L. Donoho, "Does median filtering truly preserve edges better than linear filtering?" *The Annals of Statistics*, 2009.

[24] S. W. Smith, "Moving average filters," *The Scientist and Engineer's Guide to Digital Signal Processing*, pp. 277–284, 1999.

[25] K. K. Vasan and B. Surendiran, "Dimensionality reduction using principal component analysis for network intrusion detection," *Perspectives in Science*, vol. 8, pp. 510 – 512, 2016.

[26] I. Jolliffe, "Principal component analysis," *Springer Verlag*, 2002.

[27] K. Korjus, "Machine learning-principal component analysis," *University of Tartu-Institute of Computer Science courses - Spring Technical Report*, 2016.

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

[29] T. Kohonen, "Learning vector quantization for pattern recognition," *Technical Report TKKF-A601, Helsinki Univeristy of Technology, Espoo, Finland.*, 1986.

[30] A. Sato and K. Yamada, "Generalized learning vector quantization," *Hasselmo (Eds.), NIPS*, pp. 423–429, 1995.

[31] P. Schneider, M. Biehl, and B. Hammer, "Relevance matrices in lvq," *Proc. European Symposium on Artificial Neural Networks*, pp. 37–42, 2007.

[32] B. Hammer, S. Marc, and V. Thomas, "On the generalization ability of grlvq networks," *Neural Process. Lett.*, pp. 109–120, 2005.

[33] M. Biehl, "A no-nonsense gmlvq toolbox," *University of Groningen, The Netherlands*, 2016. [Online]. Available: http://matlabserver.cs.rug.nl/gmlvqweb/web/

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