Machine Learning for Plant Disease Incidence and Severity Measurements from Leaf Images

Godliver Owomugisha[1,3], and Ernest Mwebaze[2]
[1]University of Groningen
Johann Bernoulli Institute for Mathematics and Computer Science
P.O. Box 407, 9700 AK Groningen, The Netherlands
[2]Makerere University, School of Computing & Informatics Technology
P.O. Box 7062 Kampala, Uganda
[3]Busitema University, Faculty of Engineering
P. O. Box 236, Tororo, Uganda
[1]ogodliver@gmail.com, [2]emwebaze@cit.ac.ug

***Abstract*—In many fields, superior gains have been obtained by leveraging the computational power of machine learning techniques to solve expert tasks. In this paper we present an application of machine learning to agriculture, solving a particular problem of diagnosis of crop disease based on plant images taken with a smartphone. Two pieces of information are important here; the disease incidence and disease severity. We present a classification system that trains a 5 class classification system to determine the state of disease of a plant. The 5 classes represent a health class and 4 disease classes. We further extend the classification system to classify different severity levels for any of the 4 diseases. Severity levels are assigned classes 1 - 5, 1 being a healthy plant, 5 being a severely diseased plant. We present ways of extracting different features from leaf images and show how different extraction methods result in different performance of the classifier. We finally present the smartphone-based system that uses the classification model learnt to do real-time prediction of the state of health of a farmers garden. This works by the farmer uploading an image of a plant in his garden and obtaining a disease score from a remote server.**

I. Introduction

Automation of expert tasks in various sectors is on the increase in part due to advances in machine learning. In this paper we tackle the challenge of automating diagnosis of cassava viral diseases in plants from images of the leaves of the plant taken *in situ*. Two outputs are of interest to the agricultural researcher and farmers who will use such a system; (1) a system that can determine the type of disease (incidence) affecting the crops and (2) a system that can determine the severity of that particular disease.

For this system, we look at the four major diseases affecting the cassava plant *(Manihot esculenta Cranz)* in Africa; Cassava brown streak disease (CBSD), Cassava mosaic disease (CMD), Cassava Bacterial Blight (CBB) and Cassava green mite (CGM). This presents as a multi-class classification system. Presently severity of disease is scored from 1 to 5, 1 representing a healthy plant and 5 a severely diseased plant. For each category of disease, we thus have other sub-classes that represent how severe the disease is. This paper extends previous work in this field and introduces the determination of the severity of disease from leaf images of diseased cassava plants using machine learning techniques.

A. Problem Context

Cassava is the second most important food crop in sub-Saharan Africa after maize [1], [2]. The crop continues to gain importance in Africa as a staple food eaten by more than 500 million people a day in Africa [3] because of its resilience under harsh environments, and its tolerance to extreme ecological stress conditions and poor soils. As such, the crop has exponentially gained the authority to curb food insecurity and rural poverty. This has made Cassava an ideal crop for small-holder farmers.

The crop is presently cultivated in around 40 African countries where it has historically played an important famine-prevention role. In Eastern and Southern Africa where drought is a recurrent problem [4] cassava is also the preferred staple food. However, crop yield is severly threatened by various pests and diseases particularly CMD, CBSD, CGM and CBB. Of the four, CMD and CBSD are the most devastating diseases to the cassava yield in Eastern and Central Africa [5], [6] and the greatest threats to the food security and livelihoods of over 200 million people.

B. Current methods of diagnosis

The current methods used for diagnosis involve experts traveling to disparate parts of the country and visually scoring the plants by looking at the disease symptoms manifested on the leaves. This method tends to be erratic and very subjective; it is not uncommon for experts to disagree on a score for a particular plant. With our work, we can enable experts to have a more reliable way of scoring disease as well as enabling farmers in remote places do diagnosis of their crops without need of an expert.

C. Related work

Application of machine learning in agriculture for purposes of diagnosis is still a young field. Some related research has been done already in other crops as well as in cassava including [7], [8], [9], [10]. A common thread in this work is the use of small samples in the training of the algorithms. Also for most they present a binary classification problem attempting to distinguish healthy from diseased plants. For some of

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.126

the previous studies, images were also taken in controlled environments where the light and image background could be controlled.

With the advent of deep learning and convolutional neural networks, the last couple of years has seen the research extend to using these deep networks to make inferences on disease in plants from images [11], [12]. This process automates the in some sense the feature extraction process that needs to be done. Results indicate improving levels of accuracy though with a penalty due to the expense in terms of processing time required for training these networks. Many other digital image processing techniques have been used in the literature. For brevity we will not cite all here but good reviews of the techniques can be found here [13].

This research therefore builds on some of this previous research to determine the state of health of cassava plants from a large set of images (over 7K), captured *in situ* using a smartphone. The large dataset also enables us to score the severity of disease based on the leaf image. We explore the use of some already existing techniques that have been applied to solve the problem and others that have not been used in this area. We use different feature extraction techniques to extract from the images, color, interest points and shape information and apply a battery of standard machine learning algorithms to the combined featureset. We apply these techniques to a large dataset of expert labeled leaf images of different cassava plant diseases and severities.

The different sections explain how we go about with this analysis. In section 2 we describe the data and the data collection protocols. In section 3 we discuss the different feature extraction mechanisms employed. Section 4 and 5 we delve into the classification of disease and severities and section 6 we discuss the deployment of the system for use with a smartphone.

The economic importance of diagnosing disease in cassava particularly for Africa cannot be overstated. The normal life span of a cassava plant is 9 - 12 months. Early detection of disease in the garden can lead the farmer to apply early interventions to save time and/or money.

II. The Leaf Image Data

The data we used consists of 7,386 images of leaves of cassava plants. The images are in 5 categories; the healthy class of images (1476 examples) and the four classes of diseased images representing the 4 diseases; CMD (3012 images), CBSD (1751 images), CBB (425 images), and CGM (722 images). Figure 1 depicts typical leaf images of the 4 disease classes. For the 4 disease classes, each data subset is broken down further into 4 subsets representing disease severities 2 - 5 (severity level 1 is the healthy class).

The data was collected during a national pest and disease survey by the National Crops Resources Research Institute (NaCRRI) using smartphones. NaCRRI is the government body of Uganda responsible for agricultural research in the country. All the images collected were manually labelled by experts from NaCRRI who scored each of the images for disease incidence and severity.

A. Disease leaf symptoms

Each of the diseases cause some unique symptomatic features to appear on the leaves as shown in Figure 1. We explain what these symptoms are and how we extract representative features in the next section. The four major diseases affecting cassava and their symptoms include:

1) Cassava mosaic disease (CMD): This disease is the most widespread cassava disease in East Africa and sub-Saharan Africa and this greatly affects production of cassava. CMD produces a variety of foliar symptoms that include mosaic, mottling, misshapen and twisted leaflets, and an overall reduction in size of leaves and plants [14]. Leaves affected by this disease have patches of normal green color mixed with different proportions of yellow and white depending on the severity. These chlorotic patches indicate reduced amounts of chlorophyll in the leaves, which affects photosynthesis and thus limits crop yield.

2) Cassava brown streak disease (CBSD): CBSD is presently the most severe of the cassava diseases. It is vectored by white flies and can also be transmitted through infected cuttings. The disease is very common in East Africa and in other cassava growing countries in sub-Saharan Africa. The CBSD leaf symptoms consist of a characteristic yellow or necrotic vein banding which may enlarge and coalesce to form comparatively large yellow patches. Tuberous root symptoms consist of dark-brown necrotic areas within the tuber and reduction in root size and according to [15], leaf and/or stem symptoms can occur without the development of tuber symptoms.

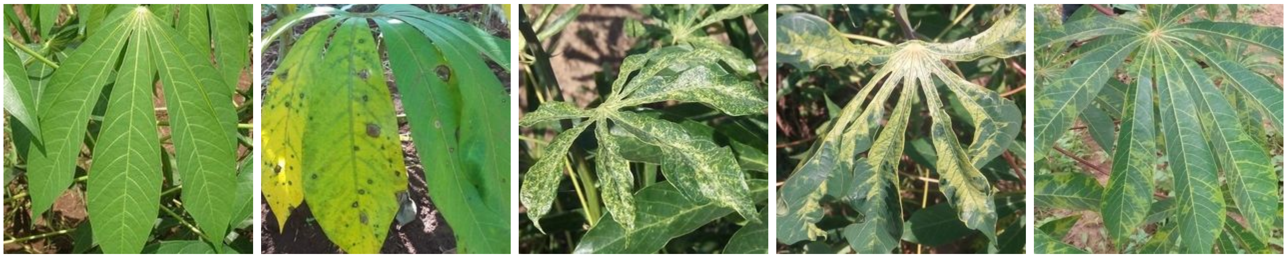
3) Cassava bacterial blight (CBB): CBB is a major bacterial disease. This disease is favored by wet conditions, however large variations in the predominance and severity of symptoms can vary depending on location, season and aggressiveness of the bacterial strains. CBB leaf symptoms include; black leaf spots and blights, angular leaf spots, premature drying and shedding of leaves due to wilting of young leaves and severe attack.

4) Cassava green mite (CGM): This disease causes white spotting of leaves, which increase from the initial small spots to cover the entire leaf thus loss of chlorophyll. Leaves damaged by CGM may also show mottled symptoms which can be confused with symptoms of cassava mosaic disease (CMD). Severely damaged leaves shrink, dry out and fall off, which can cause a characteristic candle-stick appearance.

III. Feature extraction

In order to be able to determine the state of disease based on a leaf image, we need to extract representative disease features from the image. The viral diseases in cassava manifest mainly with color and shape deformations on the leaf. In previous work [7], [8] we have extracted features that represent color and shape particularly Hue histograms, Histograms of Oriented Gradient (HOG) [16], Scale Invariant Feature Transforms (SIFT) [17] and Speeded Up Robust Features (SURF) [18] features on comparatively smaller datasets.

We have had good results in the past with Color and SIFT features. For this work we require a system that can be implemented on a server or mobile phone that can support this



(a) Healthy

(b) CBB

(c) CGM

(d) CMD

(e) CBSD

Fig. 1: Sample images associated with the five disease classes of the classification problem.

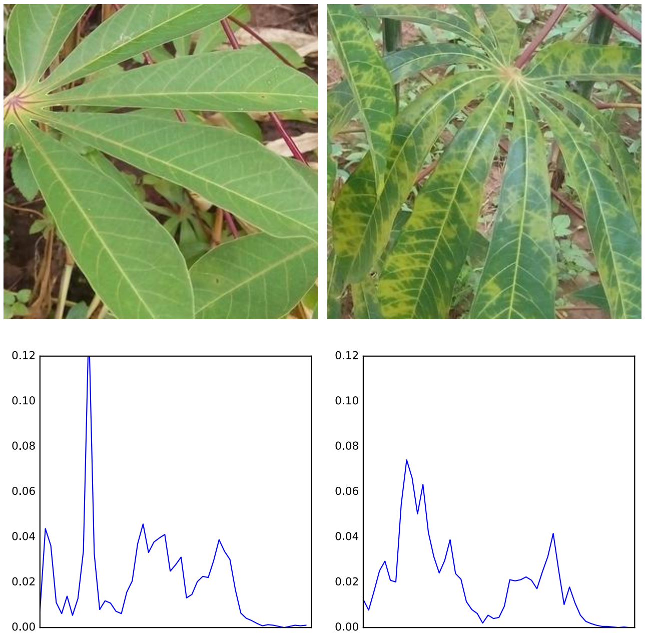


Fig. 2: Examples of histograms (bottom) extracted from the corresponding healthy and diseased images (top).

remote diagnosis by small holder farmers in Africa. For this reason we required open source feature extraction tools. We thus settled for Color and Oriented FAST and Rotated BRIEF (ORB) [19] features. SIFT and SURF are patented and thus not free for commercial use.

A. *Color feature extraction*

For the four types of diseases, color is an important feature because the diseases tend to eat away at the chlorophyll of the leaf giving it a yellowish hue. To extract these features we do an HSV color transformation of the image and calculate the normalized hue histogram of the image using 50 bins. Figure 2 depicts two sample images; a healthy image and a diseased image, and their corresponding histograms extracted.

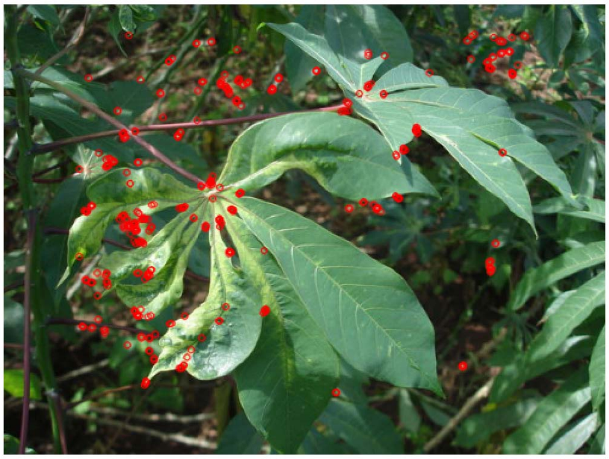


Fig. 3: Image with ORB interest keypoints identified

B. *ORB feature extraction*

ORB features offer a good alternative to the non-free SIFT and SURF features both in computation cost and matching performance [19]. ORB is a combination of a popular keypoint detector algorithm, Features from Accelerated Segment Test (FAST) and a well known feature description algorithm, Binary Robust Independent Elementary Features (BRIEF). The ORB algorithm tends to be superior to the two however because it solves some of the problems of FAST e.g. computation of orientations, as well as some of the drawbacks of BRIEF e.g. poor performance on rotation. Combining the two also results in a more powerful algorithm.

Figure 3 shows a depiction of ORB keypoints detected on an image representing one of the viral cassava diseases.

The ORB detection algorithm identifies interest keypoints on the image. As seen in the Figure 3 the keypoints are scattered throughout the image with most centered round the deformed part of the leaf. Each point is a 32 vector that describes that particular keypoint at that particular location uniquely. In order to get a uniform representative feature vector of the image, we apply the bag-of-visual words technique that clusters the different keypoints around 120 clusters representing the image. This forms a dictionary that is trained uniquely for each disease class.

To represent a new image using ORB features, keypoint descriptors are extracted from the image and then mapped to the cluster centers in the dictionary.

C. The extracted data

From the feature extraction process we derived two datasets, a 7386 × 50 dataset representing the color hue histograms and a 7386 × 120 dataset representing the generated ORB feature vectors. The 7386 records represent 5 classes; the healthy class (1476 examples), the CBB disease class (425 examples), the CGM disease class (722 examples), the CMD class (3012 examples) and the CBSD disease class (1751 examples).

IV. Classification of Disease Incidence

Our task here is to take features derived from the leaf images representing the different diseases and train a suitable classifier that can offer good performance. We used the scikit-learn[1] machine learning toolbox to train suitable classifiers. Three classifiers were trained and used;

1) LinearSVC: A linear Support Vector Classifier was trained on the data. To obtain appropriate algorithm parameters, a grid search over a limited parameter space of C was done for both ORB and color features, $C \in [1, 10, 100, 1000]$. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A suitable parameter C of 100 was obtained for both featuresets. For all the other parameters we used the defaults from sklearn. Results in Table I represent the 10-fold cross validated performance of the algorithm on this 5-class problem.

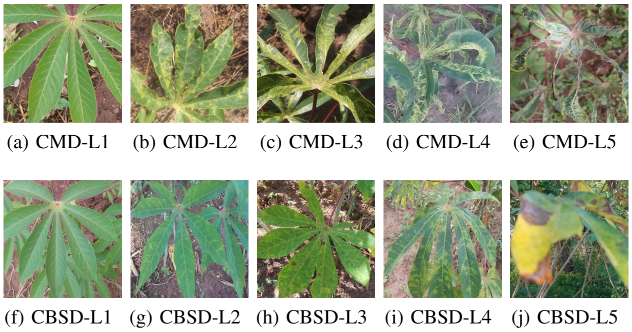
2) KNN: A K-Nearest Neighbour algorithm was fitted to the data as well. The appropriate value of K was obtained by doing a grid search over a limited space of possible K values for both ORB and color features, $K \in [1 \dots 12]$. The appropriate K was found to be 1 for ORB and 10 for color features. All other parameters were taken from the default sklearn parameters. Table I shows the corresponding results.

3) Extra Trees: Extremely Randomized Trees have been shown in the literature to perform well because they average over very many weak learners on various subsamples of the data. We find the appropriate number of trees in the forest to use using grid search of 5 parameters for ORB features $n\_estimators \in [10, 20, 30, 40, 50]$, and 7 parameters for color features $n\_estimators \in [50, 100, 200, 300, 400, 500, 600]$. The optimal number of trees we find is 30 for ORB and 400 for color. We use default parameters for the rest. Table I shows the corresponding results.

Table I shows the performance of the algorithms on the whole dataset. Results presented are of the 10-fold cross-validated accuracy score of the different algorithms applied to the data with a 95% confidence interval. We note a very high performance for the ORB generated features for both algorithms.

	LinearSVC	ExtraTrees	k-NN
Color	80	48.94	44.68
ORB	99.98	99.88	100

TABLE I: Overall 10-fold cross-validated accuracy scores (%) for different algorithms applied to the different leaf image representations.



(a) CMD-L1 (b) CMD-L2 (c) CMD-L3 (d) CMD-L4 (e) CMD-L5

(f) CBSD-L1 (g) CBSD-L2 (h) CBSD-L3 (i) CBSD-L4 (j) CBSD-L5

Fig. 4: Sample images associated with the five severity levels for CMD (top) and CBSD (bottom).

V. Classification of disease severity

Knowing the presence or absence of disease (incidence) is important for the farmer, however knowing the severity of disease is critical if appropriate and timely interventions are to be taken to prevent crop yield loss. In the previous section images representing different severities were merged together. Here we split up each of the classes into 4 subclasses; the healthy class, severity level 2, severity level 3 and severity level 4; severity-4 possessing the most severe symptoms of the 4. We did not include severity level 5 for this analysis because of the low quantities of images representing this severity class for all diseases.

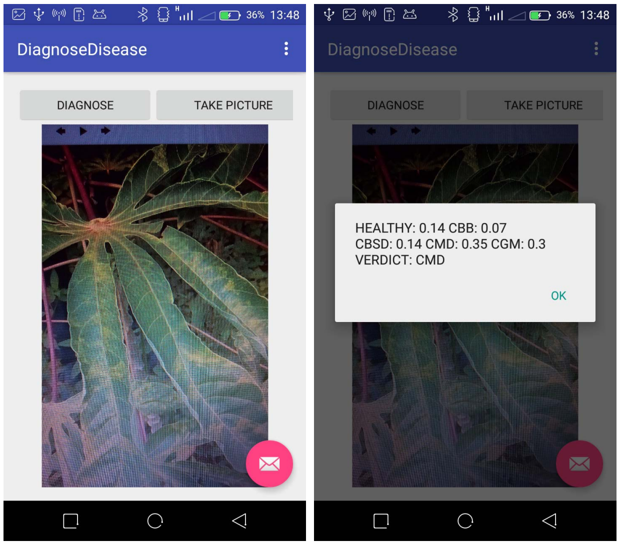
Figure 4 depicts images that represent the different severities for the two most common diseases; CMD and CBSD. Severity of disease is assigned from severity levels 1 - 5 with 1 representing a healthy leaf and 5 a severely diseased leaf. The cross validated performance of a Linear SVC classifier applied to each of the disease categories is particularly quite impressive for the ORB features compared to other features extracted. We obtain accuracy scores of close to 99 %.

We also investigate the performance when all disease categories are combined. Again we observe strong evidence of high discriminatory power of our algorithm for these particular ORB feature representations in the region of 99 % cross-validated score for accuracy.

VI. System Deployment

The goal is to translate this work to a usable application that a small holder farmer or researcher can use in the field to diagnose disease in his garden, both the incidence and severity of disease. To this end we implemented a smartphone-based diagnostic system which a farmer with a smartphone can use to get the state of health of his garden in real-time.

[1] http://www.scikit-learn.org



HEALTHY: 0.14 CBB: 0.07
CBSD: 0.14 CMD: 0.35 CGM: 0.3
VERDICT: CMD

Fig. 5: Screenshots of the smartphone application for remote diagnosis of crop health.

The way the system works is that the farmer using his smartphone can take an image of the diseased crop in his garden, have that uploaded to a server which automatically classifiers the disease and level of severity and relays this information to the farmer in real-time. By using the application at different locations in his field, the farmer is able to get a sense of the state of health of his garden and plan appropriate interventions. Figure 5 depicts screenshots of the smartphone application in use.

The application uses a client server architecture with an Android app as the client and a *falcon* rest backend python framework acting as the server. The server runs the disease diagnosis algorithm and provides results after analysing a cassava leaf image. We use *Retrofit*, a type-safe REST client for Android, as the networking library to make the HTTP calls.

VII. Discussion

In this paper we have presented a smartphone-based diagnostic system for cassava crop health that leaverages machine learning to solve the problem of identifying disease in the field from analysis of plant leaf images. We have shown how we extract the relevant features that represent disease from the leaf images and train machine learning algorithms to be able to differentiate diseases based on these features.

Different feature extraction techniques we selected and tested. Particularly we extract color hue and ORB shape/interest keypoint features from the leaf images. We found ORB to be a fast and reliable replacement for SIFT and SURF which are patented and non-free for this application.

Results indicate vastly varying performance for the Color and the ORB featuresets. It is likely color which performed well in previous studies fails here because all diseases tend to present with a yellowish color. Previously color performend well for the problem of differenting between a diseased and healthy leaf. For differentiating between two or more diseases, it appears not to do well. ORB on the other hand offers a much superior performance when the feature vectors are extracted using the bag-of-visual words approach.

We also present results obtained from applying different algorithms in a multi-class classification system for diagnosing the severity of disease based on the leaf images. Again we notice considerable performance with the ORB features for all algorithms. However the range of severities used in this work is not complete due to insufficient data in severity level 5. However for practical purposes this may not be an issue since most times by the time a plant gets to severity level 5 it is clearly visibly sick and can only be uprooted as an intervention.

Results for the ORB features are unusually high so we further investigated this result. As is evident, cross-validation and use of different classifiers gives similar results, so it appears we are not overfitting the data. We looked through the images and noticed there were some repetitions of images resulting from data collectors taking more than one picture of the same image to improve clarity. The performance shown is after removing duplicate data from the derived featuresets. On average we notice about 40 duplicates in the whole derived dataset of 7386 samples, so this again doesnot account for the unusally high performance. The classes are generally highly imbalanced but even with down sampling performance does not change much. We thus conclude that the feature extraction with ORB and bag-of-visual words offered the superior advantage in this case.

We conclude by embedding this work into a smartphone based diagnostic system for farmers in remote places. Particular dependencies of the system are the farmer must have a smartphone and a working data connection. Some of the future work will be in implementing a low power first pass offline version of the application on the smartphone that can give a preliminary diagnosis that can be ratified once the device gets online.

Acknowledgment

The authors gratefully acknowledge the team at National Crops Resources Research Institute (NaCCRI) for their unconditional assistance especially in the data collection and annotation stages. The authors would also like to acknowledge the Bill and Melinda Gates Foundation who provided the requisite funding for this project.

References

[1] S. Katrine, M. Hailu, and D. Spurling, “Raising the productivity of women farmers in sub-saharan africa,” *World Bank discussion papers*, vol. 230, 1994.

[2] C. Poulton, G. Tyler, P. Hazell, A. Dorward, J. Kydd, and M. Stockbridge, “Competitive commercial agriculture in sub?saharan africa,” *Centre for Environmental Policy, Imperial College London, Wye, Ashford, Kent, TN25 5AH,UK*, 2006.

[3] L. McCandless, “Nextgen cassava, why is cassava important?” *Cornell University*, 2012. [Online]. Available: http://www.nextgencassava.org/index.html

[4] FAO and IFAD, “A review of cassava in africa with country case studies on nigeria, ghana, the united republic of tanzania, uganda and benin,” *Proceedings of the Validation Forum on the Global Cassava Development Strategy*, vol. 2, 2005.

[5] 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.

[6] G. M. Rwegasira and C. M. E. Rey, "Response of selected cassava varieties to the incidence and severity of cassava brown streak disease in tanzania," *Journal of Agricultural Science*, vol. 4, no. 7, 2012.

[7] 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.

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

[9] E. Mwebaze, P. Schneider, F. Schleif, J. R. Aduwo, J. A. Quinn, S. Haase, T. Villmann, and M. Biehl, "Divergence-based classification in learning vector quantization," *Neurocomputing*, vol. 74, pp. 1429–1435, 2011.

[10] J. Tuhaise, J. A. Quinn, and E. Mwebaze, "Pixel classification methods for automatic symptom measurement of cassava brown streak disease," *Proceding of the 1st International Conference on the Use of Mobile ICT in Africa*, 2014.

[11] 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*, vol. 2016, p. 11, 2016.

[12] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *CoRR*, vol. abs/1604.03169, 2016.

[13] A. Barbedo and J. Garcia, "Digital image processing techniques for detecting, quantifying and classifying plant diseases," *SpringerPlus*, vol. 2, no. 1, pp. 1–12, 2013.

[14] I. Abdullahi, G. Atiri, and A. Dixon, "Effects of cassava genotype, climate and the bemisia tabaci vector population on the development of african cassava mosaic geminivirus (acmv)," *Acta Agronomica Hungarica*, pp. 285–289, 2003.

[15] R. Hillocks, M. Raya, and J. Thresh, "The association between root necrosis and above-ground symptoms of brown streak virus infection of cassava in southern tanzania," *International Journal of Pest Management*, pp. 285–289, 1996.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893 vol. 1.

[17] D. G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the International Conference on Computer Vision*, pp. 1150–1157, 1999.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding*, vol. 110(3), pp. 346–359, 2008.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2564–2571.