A Machine Learning Model for Field Diagnosis of Cassava Mosaic Disease

Jennifer R. Aduwo* and Joyce Nakatumba -Nabende

Department of Computer Science, School of Computing and Informatics Technology, College of Computing and Information Sciences, Makerere University Kampala, Uganda

*Corresponding author: jraduwo [AT] umi.ac.ug

***Abstract*---- Machine learning is a branch of computer science that deals with the study of computer algorithms that improve performance on any task with experience, they are being used for crop disease detection due to their ability to learn a particular disease symptoms and correctly predict a new instance of the same disease with high accuracy. In this study, we investigated cassava mosaic disease, its symptoms that is yellowing of the cassava leaves and distorted shape features and how the disease could be diagnosed automatically to harness the appropriate interventions and treatment. We reviewed and implemented different machine learning techniques for Cassava Mosaic Disease field based diagnosis like: decision tree, artificial neural network, support vector machine, K nearest neighbor and naïve bayes. The experimental results using the combination of color and shape features for CMD diagnosis proved to be very effective and efficient with accuracy rates of over 90% with artificial neural network and K nearest neighbor machine learning techniques. We therefore recommend their implementation for field based diagnosis of CMD.**

Keywords-- Machine learning, Classification, cassava mosaic disease diagnosis

I. INTRODUCTION

Studies indicate that virus diseases are major factors affecting agricultural production in Europe and elsewhere in the world, where losses amount to millions of dollars and more millions are spent annually to control their spread [1]. Cassava mosaic disease (CMD) has been singled out as the biggest economic constraint to the production of cassava in sub-Saharan Africa [2], and a tragic vector-borne disease of any African food crop routinely reducing cassava production continent-wide by 15-25%, but losses as high as 90% could be obtained with susceptible genotypes [3]. CMD has caused an estimated yield loss of over 1.55 billion US dollars a year in Africa [4].

A number of approaches have been implemented in Uganda, Tanzania and Nigeria to manage CMD and other cassava diseases and pests like: i) CMD trained experts visiting cassava fields and manually recording their observations of CMD incidence and severity on paper forms, ii) trained farmer groups sending monthly reports on the status of cassava diseases in their fields through mobile phone-generated text messages (SMS) [3], iii) using Enzyme Linked Immunosorbent Assay (ELISA) and Polymerase Chain Reaction (PCR) for CMD diagnosis[15].

It is observed that trained CMD experts visiting cassava fields results into collection of detailed and reliable data on CMD incidence and severity, however the cost is relatively high leading to limited frequency of data collection to once per year [5]. The use of SMS for monitoring CMD incidence and severity is much less intensive and facilitates cheaper monthly collection of CMD data [5]. Molecular techniques such as ELISA and PCR for CMD diagnosis are expensive and time consuming and their biggest limitation currently is their inability to be automated for rapid plant disease detection. Consequently, there is need for cheap, less time consuming, less labor-intensive, minimal procedures method for quick diagnosis of the CMD based on field images. This would provide information to enable appropriate interventions before detailed laboratory tests [6], the information would be used in planning, implementing, monitoring, communicating and forecasting the CMD prevalence and spread over time [7, 8], however, such information is currently difficult to obtain due to problems with the availability of suitable trained CMD experts, the logistics of transport to the cassava fields, and the time taken to coordinate paper reports. In recent years, the use of machine learning has been extended for crop disease detection. The study explored the extension of the application of standard machine learning techniques for CMD diagnosis using field images based on color and shape features. The goal was to select

the best technique with the highest level of accuracy for prediction of CMD. Machine learning is a branch of computer science that deals with the study of computer algorithms that improve performance on any task with experience [9], for example, a machine learning algorithm trained to detect crop disease would improve its performance based on the number of training examples at its disposal.

II. RELATED WORKS

In this section, we review some researches which have been conducted on machine learning for crop disease detection.
Basvaraj.S.Anami et al., [10] observed that relying on pure naked-eye observation to detect and classify diseases can be expensive. Consequently, they developed a system that implemented color and texture features to recognize and classify different agriculture/horticulture produce into normal and affected entities using neural network classifier. The combination of features that is colour and texture proved to be very effective. The experimental results indicated that the proposed approach significantly supports accuracy in automatic detection of normal and affected produce.

Sannakki S.S et al., [11] proposed an automated system for disease detection and grading in pomegranate plant. The methodology involved image acquisition, enhancement and segmentation using appropriate algorithms. Feature extraction was carried out and selected features were used as input to the support vector machine classifier, which appropriately identified and graded the disease for proper disease treatment advisory. The authors concluded that the system could provide support to farmers during their daily struggle against disease outbreaks.

Anand et al., [12] explored the use of an artificial neural network (ANN) to detect disease in pomegranate plants. The images of the pomegranate leaves were captured, filtered and segmented using Gabor filters. Then, texture and color features were extracted from the result of segmentation and artificial neural network (ANN) was then trained by choosing the feature values that could distinguish the healthy and diseased samples appropriately. Experimental results showed that classification performance by ANN was good with an accuracy of 91%.

Padmavathi, K. [13] used image processing techniques to identify leaf disease of plants and to determine the stage in which the diseases are**.** He used three alternate classification approaches to include; statistical classifier using the Mahalanobis minimum distance method, neural network based classifier using the back propagation algorithm and neural network based classifier using radial basis function. The analyses proved that such methods could be used to identity diseases in plants hence recommending their applications in areas such as precision farming.

III. THE PROPOSED CMD DIAGNOSIS MODEL

The proposed model in this study is depicted in Figure 1 for diagnosis of CMD using cassava field images during early stages of growth that is between one to six months.

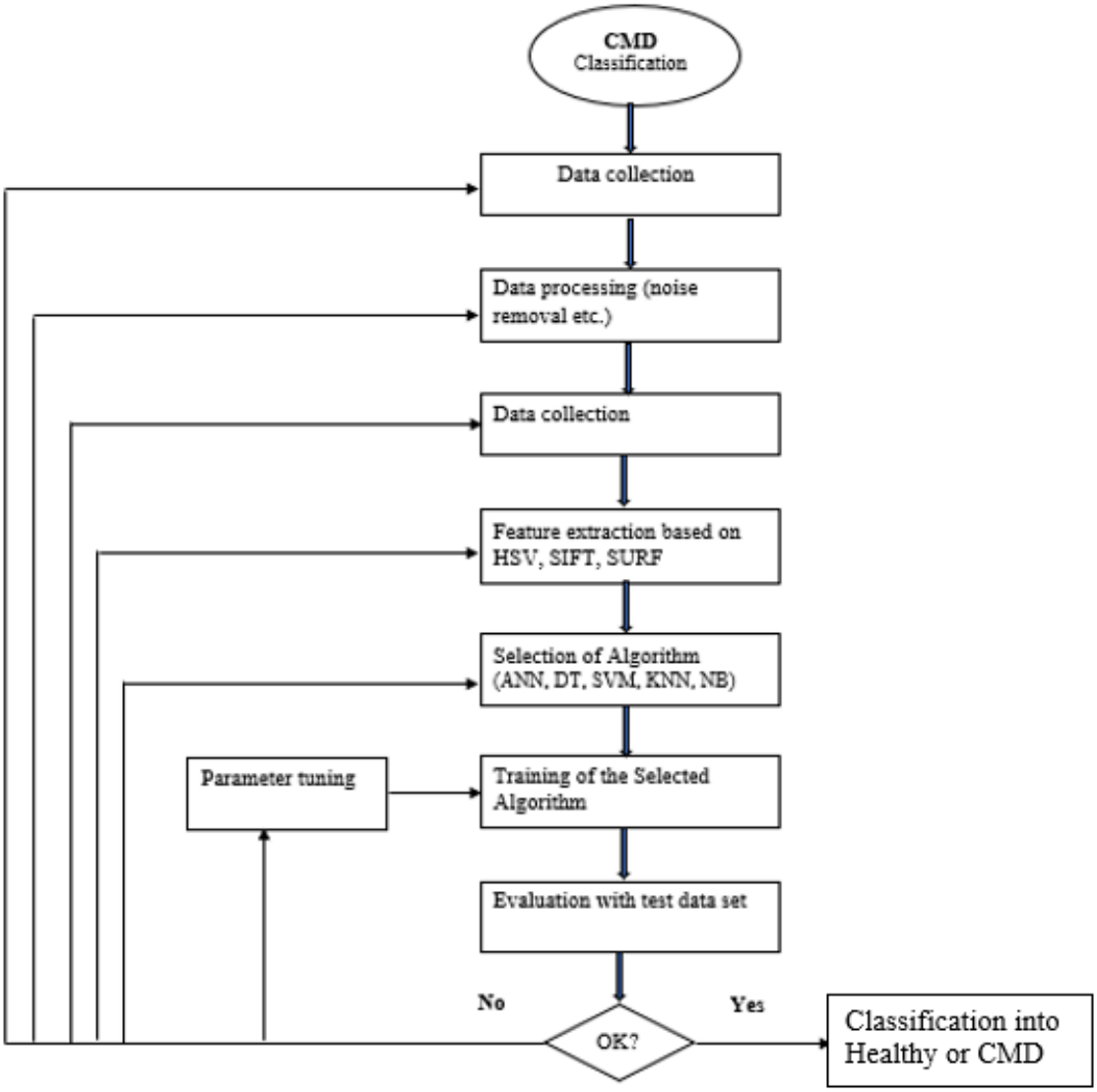

Figure 1: Supervised machine learning process used in the study (adapted from Korsiantis, 2007[14])

The methodology implemented to develop the CMD diagnosis model is explained in this section.

Data collection stage: A total of 248 healthy and 212 diseased cassava leaves samples shown in Figure 2 were collected from National Crops Research Resources Institute (NaCRRI), Uganda using a digital camera under mixed lighting conditions, the images were resized to 640 pixels horizontally and 480 pixels vertically, these were then converted from RGB to gray scale color intensity to enable quick image processing using inbuilt Matlab software version R2015.

Figure 2: Healthy leaf top row and those infected with CMD bottom row taken in situ

Data preprocessing/Back subtraction: Part of data preprocessing is background subtraction (BS) is a technique used in image processing to generate a foreground mask of the object of interest, in other words BS removes anything that can be considered as background given the characteristics of the observed image scene. K-means image segmentation technique was used to remove the image background, this also included clustering all input gray scale images based on pixel intensity using inbuilt Matlab software version R2015

Feature Extraction

Colour (HSV) and shape (SIFT, SURF) features for CMD detection and classification were extracted in Matlab software version R2015. The colour distribution was calculated for each labelled image using 50 histogram bins, and then normalized. The SURF and SIFT shape descriptors were averaged to obtain a representative prototype for each image either healthy or CMD infected. The (HSV) and shape (SIFT, SURF) feature values were stored in the database for the classification experiments.

Classification Experiments

Experiments were conducted using the extracted features that is HSV, SIFT and SURF. The purpose was to use the extracted features and compare the classification performance of the five machine learning techniques that is Decision Tree (DT), Artificial Neural Network (ANN-BPNN, ANN-LMNN), Support Vector Machine (SVM), Naive Bayes (NB) and K-Nearest Neighbor (KNN). These were implemented in MATLAB software version R2015. The performance on the data was obtained using 10-fold cross validation, which is helpful to prevent over fitting. Performance was determined as an average of any 9/10 sample as training set and the rest as testing set repeated 10 times. The standard machine learning techniques were implemented with the selected parameters because they gave the lowest misclassification error on the test data:

- Support Vector Machine (SVM) with polynomial kernel and parameters: C=0; degree =1, γ=0, regularization= 10^{-14} were implemented. The choice of the polynomial kernel is because it is well suited for classification problems where all the training samples are normalized as is the case of the study data set.
- Decision tree (DT) based on Chi-square for forward pruning and fisher binary splitting criterion,
- Naïve Bayes (NB)
- K-Nearest Neighbor (KNN) with k = 3 and,
- Artificial Neural Network (Multi-layer perceptron neural network) with back propagation (BP, and Levenberg-Marquardt (LM), with 5 hidden neurons, number of training epochs=100, regularization= 10^{-14}

IV. EVALUATION OF THE EFFICIENCY AND ACCURACY OF THE MODEL

*We used the test data sets to compare the accuracy of the machine learning algorithms discussed in section 3.0. The performance of the best machine learning techniques for CMD diagnosis were evaluated through the errors generated from the test datasets and measured by: i) accuracy ii) sensitivity and iii) Area Under the Receiver Operating Characteristics Curve (AUC).

Accuracy: is the proportion of the correctly predicted entities to the total predictions made

TP + TN/ (TP + FN + TN + FN)

True Positive (TP): The number of true positives (actual classified positive instances classified as positives) for example the number of unhealthy cassava leaf images identified as CMD infected.

True Negative (TN): The number of true negatives (actual negative instances classified as negatives), for example the number of healthy cassava leaf images identified as healthy.

False Positive (FP): The number of false positives (actual negative instances classified as positives) for example the number of healthy cassava leaf images identified as CMD infected.

False Negative (FN): The number of false negatives (actual positive instances classified as negatives), for example the number of healthy cassava leaf images identified CMD infected.

True Positive Rate (TPR): or Sensitivity/recall: as the proportion of actual positives which are predicted as positive, TP/ (TP + FN)

True Negative Rate (TNR): or Specificity as the proportion of actual negatives that are predicted as negative, TN/ (TN+FP).

Precision or Positive Predictive Value (PPV): is the proportion of actual positives which are predicted positive, TP/ (TP + FP) or TN/(TN + FN). Precision measures the proportion of the entities that are actually called positive, and TPR measures the proportion of true positives.

Mis-classification or error rate: the proportion of the wrongly predicted to the total predictions made, FP + FN/(TP + FN + TN + FN)

Area Under ROC Curve (AUC): the quality of the ROC curve is summed into the AUC. Higher AUC values are between 0.5 for poorest and 1.0 for a perfect model.

V. RESULTS

As seen in Table 1, four machine learning techniques: KNN, SVM, NAIVEB and ANN (BPXN) have the highest rate of accuracy and sensitivity of over 90% among other techniques with only HSV (color) feature set to diagnose the CMD. However depending on colour to diagnose CMD cannot give accurate results because the cassava disease experts depend on colour and shape with their naked eyes to discriminate between healthy and CMD infected leaves.

Table 2 shows results of our model where we implemented a combination of colour and shape features to recognize and classify a cassava leaf into healthy and CMD infected using the classifiers shown in Table 2. Only KNN and ANN presented accuracy rate of 90% and above with HSV-SIFT data set. The experimental results indicated that the proposed approach significantly supports accuracy in automatic diagnosis of a cassava leaf into healthy and CMD infected.

In Table 3, we show the AUC results, which depicts relative tradeoffs between true positives and false positives. The accuracy of the machine learning techniques to classify is measured by the AUC value. AUC values closer to 1 indicate that the classifier reliably distinguishes the two classes for instance healthy and CMD infected.

Whereas values at 0.50 indicates that the classifier is no better than taking chance. Only KNN and ANN presented AUC values of more than 0.95 and above with HSV-SIFT data set. The experimental results showed that the proposed approach significantly supports accuracy in automatic diagnosis of a cassava leaf into healthy and CMD infected using the HSV-SIFT data set.

Table 1: The Comparison of implemented machine learning methods for CMD diagnosis with discrete features

Classifier	**Accuracy %**	**Sensitivity %**	**Accuracy\ %**	**Sensitivity %**	**Accuracy %**	**Sensitivity %**
	HSV		**SIFT**		**SURF**	
KNN	92.2	92.6	81.3	83.0	70.4	70.0
SVM	92.8	93.7	53.9	53.9	53.9	53.9
NAIVEB	91.3	90.2	54.1	54.1	71.7	72.6
DTREE	88.7	86.8	77.8	78.7	70.9	68.9
ANN(BPXN)	93.0	94.2	84.6	84.8	62.8	63.6
ANN (LMN)	88.0	90.4	82.2	81.3	61.5	60.6

Table 2: The Comparison of implemented machine learning methods for CMD diagnosis with combined features

Classifier	**Accuracy %**	**Sensitivity %**	**Accuracy\ %**	**Sensitivity %**	**Accuracy %**	**Sensitivity %**
	HSV -SIFT		**HSV -SURF**		**HSV-SIFT-SURF**	
KNN	93.0	92.2	87.6	72.8	86.3	87.2
SVM	53.9	53.9	53.9	53.9	53.9	53.9
NAIVEB	53.9	53.9	85.9	53.9	53.9	53.9
DTREE	84.8	87.0	85.2	76.9	83.9	84.4
ANN(BPXN)	93.5	93.7	93.5	85.9	92.2	92.9
ANN (LMN)	91.5	90.0	89.8	80.5	90.7	85.0

Table 3: The AUC Evaluation Results of combined features for CMD diagnosis

Classifier	**HSV-SIFT**	**HSV-SURF**	**HSV-SIFT-SURF**
KNN	0.96	0.92	0.94
SVM	0.94	0.93	0.94
NAIVEB	0.50	0.75	0.50
DTREE	0.93	0.93	0.93
ANN(BPXN)	0.97	0.98	0.97
ANN (LMN)	0.86	0.93	0.94

VI. CONCLUSION

In this study, a machine learning Model for field diagnosis of Cassava Mosaic Disease is proposed. A total of 248 healthy and 212 diseased cassava leaves were collected from National Crops Research Resources Institute (NaCRRI), Uganda using a digital camera and a smart phone under mixed lighting conditions. The input images were first pre-processed, then their features were extracted on two parameters namely: color and shape (SIFT and SURF) and then, trained and classification of the same was done using four standard machine learning techniques. The overall system accuracy was measured to be 90% and above. The study recommends only KNN and ANN with HSV-SIFT data set for CMD diagnosis using field images. The model provides one step towards promoting the implementation of machine learning techniques that is KNN and ANN for CMD diagnosis for appropriate interventions. In future, the model could be improved with increased dataset size to improve the overall system performance to diagnose CMD more accurately

REFERENCES

[1] G.W.Otim Nape, T.Alicai and J.M.Thresh, "Changes in the incidence and severity of cassava mosaic virus disease, varietal diversity and cassava production in Uganda", Annals of Applied Biology, pp.313-327, 2005.

[2] P.Ntawuruhunga, G. Okao-Okuja, G. Bembe, A. Bambi, M. J.C. Armand Mvila and J.P Legg, "Incidence and severity of cassava mosaic disease in the republic of Congo", African Crop Science Journal, vol.15, pp.1–9, 2007.

[3] International Institute of Tropical Agriculture. Integrated cassava project: preemptive management of the virulent cassava mosaic disease through an integrated cassava development approach for enhanced rural sector economy in the south and southeast zones of Nigeria. International Institute of Tropical Agriculture, 2003.

[4] J.S.Pita, V.N.Fondong, A. Sangare. G.W.Otim-Nape, S. Ogwal and C.M. Fauquet, "Recombination, pseudore combination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda". Journal of General Virology, pp 655–665, 2001.

[5]S.Bigirimana,W.Tata Hangy, H.Obiero, G. Mkamilo, I.Ndyetabula and S. Jeremiah and T.Alicai, "Cassava Disease Surveillance Surveys 2009, Mapping Report -Great Lakes Cassava Initiative, International Institute of Tropical Agriculture (IITA)". Technical report, 2010.

[6] FAO. "Cassava diseases in Africa, A major threat to food security". Technical report, Cassava diseases in central, eastern and southern Africa (CaCESA), 2015.

[7] J.R.Aduwo and G. Acellam. "Assessing the performance of two artificial neural networks in the classification of cassava mosaic disease.". http://cit.mak.ac.ug/iccir downloads/ICCIR 10, 2010.

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

[9] T.M. Mitchell, "Machine Learning, McGraw-Hill Series in Computer Science". Technical report, WCB/McGraw-Hill, Boston, MA, 1997.

[10]S.Basvaraj,Anami,Pujari, Yakkundimath and Rajesh,Affected Agriculture/horticulture Produce Based on Combined Color and Texture Feature Extraction. 2012.

[11] S.S.Sannakki, V.S. Rajpurohit, V.B. Nargund, R.Arun Kumar and P.S. Yallur, "A hybrid intelligent system for automated pomegranate disease detection and grading". International Journal of Machine Intelligence, vol 3: pp.36–44, 2011.

[12]A.H.Kulkarni and R.K.Ashwin Patil, "Applying image processing technique to detect plant diseases". International Journal of Modern Engineering Research, vol 2, pp.3661-3664, 2012.

[13] K.Padmavathi, "Investigation and monitoring for leaves disease detection and evaluation using image processing". International Research Journal of Engineering Science, Technology and Innovation, vol 1(3): pp 66–70, 2012.

[14] S.Korsiantis, "Machine Learning: A review of Classification Techniques". Information journal, vol 31, pp 249 -268, 2007.

[15] K.G.Mabasa, . "Epidemiology of Cassava Mosaic Disease and Molecular Characterization of Cassava Mosaic Viruses and their associated whitefly (bemisia tabaci) vector in South Africa". Master's thesis, University of the Witwatersrand, Johannesburg, 2007.