Early detection of plant diseases using spectral data

Owomugisha, Godliver; Nuwamanya, Ephraim; Quinn, John A.; Biehl, Michael; Mwebaze, Ernest

Published in:
APPIS 2020

DOI:
10.1145/3378184.3378222

EARLY DETECTION OF PLANT DISEASES USING SPECTRAL DATA

A PREPRINT SUBMITTED TO APPIS2020, JANUARY 08,2020

Godliver Owomugisha[1,2,3], **Ephraim Nuwamanya**[4], **John A. Quinn**[2], **Michael Biehl**[1], **Ernest Mwebaze**[2]
1. University of Groningen, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,
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
4. National Crops Resources Research Institute (NaCRRI), P.O Box 7084, Kampala, Uganda
[g.owomugisha, m.biehl]@rug.nl, [emwebaze, jquinn]@cit.ac.ug, nuwamanyaephraim@gmail.com

ABSTRACT

Early detection of crop disease is an essential step in food security. Usually, the detection becomes possible in a stage where disease symptoms are already visible on the aerial part of the plant. However, once the disease has manifested in different parts of the plant, little can be done to salvage the situation. Here, we suggest that the use of visible and near infrared spectral information facilitates disease detection in cassava crops before symptoms can be seen by the human eye. To test this hypothesis, we grow cassava plants in a screen house where they are inoculated with disease viruses. We monitor the plants over time collecting both spectra and plant tissue for wet chemistry analysis. Our results demonstrate that suitably trained classifiers are indeed able to detect cassava diseases. Specifically, we consider Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) applied to original spectra and, alternatively, in combination with dimension reduction by Principal Component Analysis (PCA). We show that successful detection is possible shortly after the infection can be confirmed by wet lab chemistry, several weeks before symptoms manifest on the plants.

***K*eywords** Crop disease detection, Spectral data, Prototype learning, LVQ, GMLVQ

1 Introduction

Early detection of disease in crops is of paramount importance for food security, particularly in Sub-Saharan Africa. In this paper we focus on a key food security crop: cassava (Manihot escuenta). This crop is grown predominantly by smallholder farmers because it can easily be grown even when environmental conditions are tough and also because it requires few inputs. However, productivity has been limited by several other factors, susceptibility to pests and diseases being the most severe one.

In this paper, we focus on developing methods for diagnosing cassava diseases before they are visibly symptomatic on the plant. We build on the previous studies e.g. [1], [2] and [3] that showed the superiority of spectral data over raw image data when used for detection of diseases in cassava. The uniqueness in the current work is the ability to identify cassava brown streak disease (CBSD) at a much earlier stage in the lifecyle of an inoculated cassava plant, before disease symptoms become visible to the human eye. Different sets of plants were grown in a screen house and a portion of them inoculated and monitored regularly over time. At each monitoring epoch, wet chemistry was applied to determine the state of health of the plants. This work is guided by our hypothesis that crop diseases cause several metabolic changes in the metabolism of the leaf which can be detected at an early stage using a spectrometer. Unlike in the previous experimental work [3], where we considered visibly diseased mature plants aged 6 - 9 months and grown in open fields, this work is based on a controlled experiment in a screen house environment. The controlled setup rules out the influence of other diseases, pests or severe weather conditions while at the same time providing us with time series data that allows us to determine how soon we can diagnose disease in a non symptomatic plant.

Presently, CBSD is one of the most severe disease in cassava and it is transmitted from garden to garden by vectors called white flies [4, 5]. The disease can also be transmitted through infected plant cuttings. The disease symptoms consist of a characteristic yellow or necrotic vein banding which may enlarge and coalesce to form comparatively large yellow patches of the leaf. Tuberous root symptoms consist of dark-brown necrotic areas within the tuber and reduction in root size. However, leaf and/or stem symptoms can occur without the development of tuber symptoms [6].

Most current methods of diagnosis rely on visual inspection of the plants by agricultural experts. However, the process is tedious and there is frequently a significant degree of disagreement between the experts' assessment. Thus, the introduction of efficient, reproducible computational diagnoses has been aimed at in recent years, including our study [7] where diagnosis was done using plant image data taken with a smartphone. The limiting factor of such systems is that disease symptoms have to be visible in the photographic image. Once symptoms have manifested, the root of the plant is already affected and can no longer be used as food particularly for CBSD disease.

The following sections describe the controlled experimental setup we used to study the detection of CBSD. Specifically, section 2 presents research done in relation to molecular and biological characterization of CBSD disease, we discuss further studies done in the area of disease diagnosis using near infra-red spectrometry. Section 3 describes the materials and methods used. Section 4 presents our results in two parts: wet chemistry lab and machine learning analysis. Finally, we discuss the results and conclude in section 5.

2 Related work

There has been a lot of work on automating the detection of disease in crops from images. Most of these rely on the visible symptoms of the disease being present and manifest on the leaves or stem of the plant. A smaller number of studies has investigated detection of disease before its symptomatic on the plant. The gold standard is the use of wet lab chemistry to determine the presence or absence of disease in the plant material. This tends to be a destructive process.

These studies tend to be carried out in screen houses where conditions are well controlled. Healthy crops in the screen house are generally inoculated with disease and measurements are taken over time. For cassava previous studies have mainly relied on non-vector transmission of disease to the plants [8, 9, 10]. In this method virulent isolates of disease are grafted on to healthy plants. Vector based methods in cassava include using diseased whiteflies to infect the cassava. While this tends to be the natural means of transmission, it is hard to replicated in a screen house. We employed non-vector methods in our study.

Once the crops in a screen house are infected, then comes the task of determining the state of health of the plant. A common non-destructive method of doing this is spectroscopy. Several studies have been carried out using spectroscopy to agriculture including some previous work [3]. Schweiger et. al. [11] broadly categorize these techniques as follows: fluorescence spectroscopy, multispectral or hyperspectral imaging, infrared spectroscopy, visible/multiband spectroscopy among others.

A lot of work in this area has been done using the Miniature Leaf Spectrometer CI-710 (CID- Bio-Science) [12] as the collection instrument. It is portable and non-destructive of the material being measured. Most of this work tends to be be focused on determining the chlorophyll levels of the plant [13, 14]. In this study we used this miniature leaf spectrometer for data collection.

3 Materials and methods

3.1 Experimental design and data collection

The experiment was conducted in a controlled screen house environment. This setup rules out the influence of other diseases, pests or severe weather conditions. Cassava stems of variety Narocass 1 and NASE 14 were acquired from different fields and first tested to confirm that they were from healthy plants. All planting materials were thoroughly cleaned, which included the sterilization of the soil to ensure that no gaps led to disease transmission.

Initially, twenty seven (27) healthy cassava stems were planted. At week four (4) of growth, plants were split into two (2) separated groups. The first one (10 plants) was reserved as a healthy control class (HC) and no disease inoculation was applied to the group. The second group of plants (17 plants) was infected with the cassava brown streak disease(CBSD) virus. CBSD virus was transmitted to these plants using a non-vector technique also known as grafting inoculation rated amongst the most efficient ways of inoculation [8, 9].

Guided by agricultural and bio-chemical experts from the Uganda National Crop Resources Research Institute (NaCRRI), the process of data collection targeted two (2) sets of data which were essential in this study. The first set of data was collected using the leaf spectrometer [12], another set of data was provided by the bio-chemical experts using wet

chemistry lab results based on real-time RT-PCR of CBSD virus. In addition, a visual assessment of the plant status was provided by the agricultural experts. Data was collected for a period of fifteen (15) consecutive week inclusive of the first week before plant disease inoculation. For each week, three lower leaves on each plant were identified and tagged. Spectral data and tissue samples were collected on all the three leaves. In total, the number of samples collected per week per class were 30 and 51 data points for HC and CBSD, respectively. Tissue samples were used for the lab tests. This process was repeated for the entire 15 weeks of the experiment. Figure 1 shows the data collection with the leaf spectrometer.

Figure 1: Data collection with a CI-710 miniature leaf spectrometer [12].

3.2 Confirmation of CBSD transmission

Usually, when a plant is infected with a certain virus, its DNA begins to alter and a specific type of protein molecules are produced and introduced to the plant by the pathogen during infection. The confirmation of a successful transmission of the target viruses can be determined in two ways (molecular-based disease detection and PCR-based disease detection techniques). These methods have been investigated on previously including the study in [11]. In this section, we discuss the protocols we acquired to confirm the presence of CBSD in our study.

3.2.1 Ribonucleic acid (RNA) extraction

RNA is nucleic acid molecule similar to DNA but containing ribose rather than deoxyribose. RNA was extracted using the protocol explained in [15] with a few modifications. Cassava leaves were ground to powder with liquid nitrogen in a sterile mortar using a sterile pestle. 2 mL of grinding buffer was added (2% CTAB, 2% polyvinylpolypyrrolidone, 100 mm Tris pH 8.0, 20 mm EDTA, 1.4 M NaCl and 20 mm DDT added fresh). The suspension (800 L) was transferred to a 2 mL micrfuge tube and then incubated at 65 °C for 30 min.

After incubation, equal volume (800 ul) of chloroform: Iso amyl alcohol (24:1) was added and mixed by inverting the tube. The phases were separated by microfuge at 13,000 rpm for 10 min. The upper aqueous layer was transferred to a new sterile 1.5 mL eppendorf tube and chloroform extraction repeated. Ethanol precipitation of the nucleic acids was performed with 0.5 volumes of 5M NaCl and 2 volumes of ice cold ethanol at -200 ° C for 30 min. The nucleic acid was collected by microfuge. The nucleic acid was collected by microfuge at 10,000 rpm for 10 min and resuspended in 0.5 - 1.0 mL of 2 M LiCl. The nucleic acid was left in the LiCl overnight at 40°C. The RNA was pelleted in a microfuge at 13,000 rpm for 30 min at 40 °C. The LiCl was removed, the pellet washed with 70% ethanol, dried and resuspended in 60 L of RNase-free water.

The RNA concentration was measured using a nanodrop spectrophotomer 2000 and 5 ul of RNA of each sample was run on 1% agarose and viewed under UV to check whether it was degraded or not using the SYNGEN gel documentation system.

3.2.2 Real-Time Polymerase Chain Reaction (RT-PCR)

The reactions were prepared in a 96 well plate and analyzed with RT-PCR to detect the two viruses CBSV and UCBSV. As a control, a COX assay was also carried out. COX is a widely used housekeeping gene for normalizing cycle threshold (Ct) values. The COX assay was performed to see if there was cDNA in the samples. Three master mixes were made (CBSV, UCBSV and COX) with the final concentration of 10 ul 2x Sso advanced Universal SYBR green super mix, 1 ul of 10 pmol/ul forward primer, 1 ul of 10 pmol/ul reverse primer, 6 ul of nuclease free water and 2 ul of cDNA per reaction. The Real-Time amplification program was set; initial denaturation 95 °C for 30 min followed

by 40 cycles of Denaturation at 95 °C for 10 sec and annealing at 56 °C for 30 sec. cDNA from CBSV- and UCBSV-infected plants were used as positive controls. A negative control with all the reagents and sterile distilled water instead of cDNA was used.

3.3 Data pre-processing and feature extraction

The process of data collection generated two (2) sets of data which were essential for analysis; spectral and RT-PCR data. RT-PCR data was analyzed by the bio-chemists and together with visual symptom scoring, we consider this data as the ground truth information in our experiment. In this section we discuss the pre-processing and feature extraction techniques we applied for spectral data used in the machine learning models. We follow the same pre-processing approach defined in [3]. A typical spectrogram is representing the two classes; healthy and CBSD see Figure 2. The intensities corresponding to the smallest and largest wavelengths are affected by significant noise. By truncating the spectrogram, we selected a wavelength range of 400 - 900 nm for subsequent analysis. This truncation provided a range of 500 nm, corresponding to 2500 equally spaced feature dimensions, which was still quite high. The spectrogram had many perturbations from small noise added to each wavelength. Consequently, the next pre-processing step aimed at smoothing the data over a small window of wavelengths. Average filtering was applied on all the data and a window size of 15 nm was used.

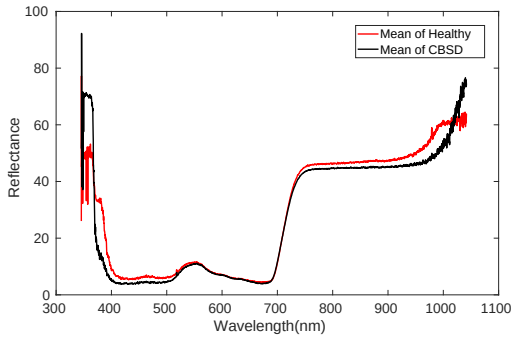

Figure 2: Spectral data in original form. Mean spectra of healthy samples and diseased samples are shown, respectively.

3.4 Dimensionality reduction

Spectral data of the type considered here are nominally high-dimensional. As a consequence, the naive application of machine learning techniques will result in classifiers with a very large number of adjustable parameters, which causes problems ranging from computationally expensive training to a potentially increased risk of over-fitting.

In this study, we employed standard PCA for the purpose of dimension reduction. PCA is a well known and widely used technique for correlation analysis and dimensional reduction, e.g. [16]. The technique identifies a linear transformation of vectors such that the orthogonal projections are ordered according to the variation in the data.

Original vectors $\hat{\mathrm{x}} \in \mathbb{R}^{\hat{N}}$ are mapped to the coefficient space via

$$\mathrm{x} = \Psi\, \hat{\mathrm{x}} \in \mathbb{R}^{N} \quad \text{where the matrix } \Psi \in \mathbb{R}^{\hat{N} \times N} \tag{1}$$

is obtained from a given set of P $\hat{N}$-dim. vectors and comprises $N \leq P$ orthogonal principal components.

In our analysis of $\hat{N}$-dimensional spectra, the vectors $\mathbf{x} \in \mathbb{R}^{N}$ with $N < \hat{N}$ serve as lower-dimensional representatives of the data. Classifiers are trained in the N-dim. coefficient space as described in the following sections. In previous, similar studies we have demonstrated that the consideration of $N = 30$ coefficients is sufficient and yields near optimal results [17]. For comparison, we consider also systems which operate in the original $\hat{N}$-dim. feature space.

3.5 Prototype-based model for disease classification

In this section, we describe the specific prototype-based model in use [18]. Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) [18, 19, 20] has displayed superior performance in related classification problems considered earlier. In the present study, we also compare the performance of GMLVQ with that of standard techniques like K-Nearest Neighbour (KNN) classification [21], Linear Support Vector Machine (SVM) [22] and Extremely Randomized Trees (Extra trees)) [23].

We consider dataset of the general form:

$$\{\mathrm{x}^{\mu}, y^{\mu}\}_{\mu=1}^{\mathrm{P}} \tag{2}$$

where $\mathrm{x}^{\mu} \in \mathbb{R}^N$ represents a feature vector and the label $y^{\mu} \in 1, 2, ...C$ specifies its class membership. These data are generally standardized by performing a z-score operation.

Learning Vector Quantization (LVQ) is a family of prototype-based supervised classification algorithms first introduced in 1986 [24]. The prototypes of an LVQ system are defined as a set $W = \{\mathrm{w}^j, c(\mathrm{w}^j)\}_{j=1}^{M}$ of M vectors $\mathrm{w}^j \in \mathbb{R}^N$ which carry labels $c(\mathrm{w}^j) \in \{1, 2, ...C\}$. The system can be set up with one or several prototype vectors per class. The vectors w^j are defined in the feature space of observed data and ideally serve as typical representatives of their classes. Together with a given distance measure $d(\mathrm{w}, \mathrm{x})$, they parametrize the classification scheme: To predict the class of an arbitrary data point $\mathrm{x} \in \mathbb{R}^N$, its distance from all prototypes in the system is computed and x is assigned to the class $c(\mathrm{w}^{\mathrm{L}})$ of the nearest prototype with $d(\mathrm{x}, \mathrm{w}^{\mathrm{L}}) \leq d(\mathrm{x}, \mathrm{w}^{\mathrm{j}})$ for all j.

An important extension of the basic concept is relevance learning [18], in which an adaptive distance $d^{\Lambda}(\mathrm{x}, \mathrm{w}^{\mathrm{j}})$ is used, where Λ denotes a set of adjustable parameters which are optimized together with the prototypes in a data-driven training process. Specifically, the GMLVQ algorithm proposed in [19, 19] employs a matrix $\Lambda \in \mathbb{R}^{N \times N}$ of coefficients which defines the distance measure

$$d^{\Lambda}(\mathrm{x}, \mathrm{w}) = (\mathrm{x} - \mathrm{w})^{\top} \Lambda \, (\mathrm{x} - \mathrm{w}) = \sum_{i,j=1}^{N} (x_i - w_i) \, \Lambda_{ij} (x_j - w_j). \tag{3}$$

A parameterization of the form $\Lambda = \Omega^{\top} \Omega$ guarantees that $d^{\Lambda}(\mathrm{x}, \mathrm{w}) \geq 0$ with unrestricted matrices $\Omega \in \mathbb{R}^{N \times N}$. In order to avoid numerical degeneracies, a normalization constraint of the following form is imposed:

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

In GMLVQ, the training process is guided by the optimization of a cost function of the form suggested in [25]:

$$E(W, \Omega) = \sum_{\mu=1}^{p} \Phi \left(\frac{d_J^{\Lambda}(\mathrm{x}^{\mu}) - d_K^{\Lambda}(\mathrm{x}^{\mu})}{d_J^{\Lambda}(\mathrm{x}^{\mu}) + d_K^{\Lambda}(\mathrm{x}^{\mu})} \right). \tag{4}$$

In the sum over all available training examples, d_j^{Λ} denotes the distance from the closest correct prototype with $\mathrm{c}(\mathrm{w}^{\mathrm{J}}) = \mathrm{y}^{\mu}$ and d_K^{Λ} is the distance from the closest incorrect prototype with $\mathrm{c}(\mathrm{w}^{\mathrm{J}}) \neq \mathrm{y}^{\mu}$, respectively. The modulation function Φ is frequently chosen to be a sigmoidal function [25]. Here, we resort to the identy function $\Phi(z) = z$ as a simple choice.

Compared to many other classifiers, the prototype-based approach in combination with relevance learning offers several advantages, among them its intuitive interpetability and the ability to infer feature relevances from the training data [18]. The elements Λ_{ij} of the relevance matrix in Eq. (3) quantify the contribution of pairs of features to the distance measure. In particular, diagonal elements $\Lambda_{ii} = \sum_j \Omega_{ij}^2$ summarize the importance of an individual feature in the classification task, see [18] for a detailed discussion and illustrative examples.

If the vectors $\mathrm{x} \in \mathbb{R}^N$ result from a linear transformation of the form (1), it is still possible to obtain the relevance matrix in terms of the original data, e.g. high-dim. spectra $\hat{\mathrm{x}} \in \mathbb{R}^{\hat{N}}$, after training in the N-dim. space. Note that with $\mathrm{x} = \Psi \, \hat{\mathrm{x}}$ and prototypes $\mathrm{w} = \Psi \, \hat{\mathrm{w}}$ we have

$$d^{\Lambda}(\mathrm{x}, \mathrm{w}) = (\mathrm{x} - \mathrm{w})^{\top} \Lambda \, (\mathrm{x} - \mathrm{w}) = (\hat{\mathrm{x}} - \hat{\mathrm{w}})^{\top} \Psi^{\top} \, \Lambda \, \Psi \, (\hat{\mathrm{x}} - \hat{\mathrm{w}}).$$

Hence, the matrix $\hat{\Lambda} = \Psi^{\top} \Lambda \Psi$ can be constructed after training in coefficient space, which represents the relevances in the original, high-dim. feature space. In particular, its diagonal elements Λ_{ii} can be interpreted as the relevance of original features, e.g. particular wavelengths in the spectra. Similarly, low-dim. prototypes can be transformed back to their high-dim. counterparts $\hat{\mathrm{w}} \in \mathbb{R}^{\hat{N}}$, see [3] for details.

As illustrated in Figure 3, the interpretation of the relevance matrix can facilitate the selection of favorable features which ultimately helps to improve performance and to reduce the computational costs of the actual classification.

3.6 Training and validation

The training and validation strategy corresponded to Leave-One-Out cross-validation, where all data of a particular plant were disregarded in the corresponding training process.

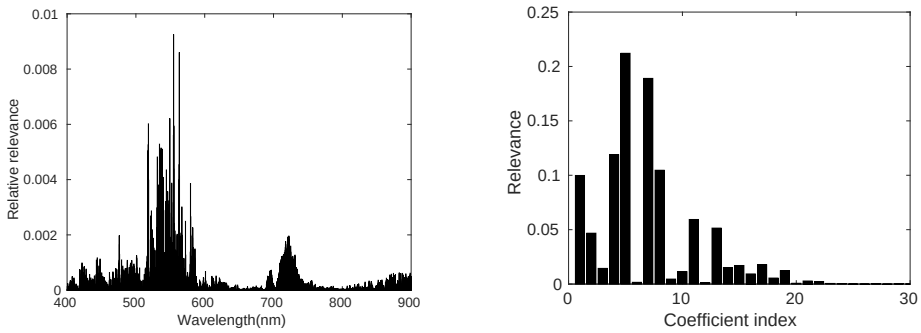

Figure 3: Feature relevance as quantified by diagonal elements of Λ, cf. Eq. (2) (left), feature representation in the coefficient space with PCA (right). In [17], sections 3.5 and 3.6, we explain the feature selection process where spectral bands 500 - 600 nm were found to be more relevant.

Data for two classes: healthy (HC) and CBSD was grouped by unique plant labels in order to avoid training and testing on data from the same plant. During training, the partitioning was based on plant groups and the validation scheme was based on Shuffle-Group(s)-Out cross-validation. Shuffle-split is an alternative to k-fold cross-validation that allows a finer control on the number of iterations and the proportion of samples on each side of the train / test split. Combined as Shuffle-Group(s)-Out cross-validation ensures that the same group is not represented in both testing and training sets. In our case, the groups were the plant ID since we obtained data from the same plant for consecutive weeks. The technique makes it possible to detect this kind of over-fitting situations. Both, PCA and z-score transformations were computed from training data only and then applied to test / validation data. We employed the standard Scikit-learn [26] implementation of this cross-validation scheme for the algorithms that were implemented using Scikit-learn. In a similar way, this validation strategy was implemented for LVQ in MATLAB(R2016a) for the open source GMLVQ toolbox [27] that we employed for the GMLVQ algorithm. GMLVQ experiments were done with one prototype per class and batch gradient distance with adaptive step size control. If not specified otherwise, we used default parameters as suggested in the documentation of the toolbox [27].

4 Results

The main objective of this study was to explore the potential use of spectral data for the early detection of disease before symptoms become visible. Our analysis considers data from healthy plants as well as plants with a virus transmitted to them. We monitor changes, with emphasis on the time before visual symptoms occur. Our analysis combines i) the use of machine learning techniques to detect viral load using spectral data and ii) results provided by the bio-chemical tests carried out on the leaf samples to evaluate our model. Table 1 shows the results of the experiment presenting different algorithms. For all further investigations, we have used GMLVQ due to its superior performance in this classification task.

Table 1: Accuracy in original feature space vs. dimensional reduction obtained on average over validation. (a) using original data full spectra between 400 - 900 nm and (b) using original data between 500-600 nm. In the coefficient space we use 30 dimensions for all algorithms.

Classifier	Original space (a)	PCA (a)	500 - 600 nm (b)	PCA (b)
KNN	0.695	0.707	0.711	0.735
Extra Trees	0.748	0.731	0.766	0.708
LinearSVM	0.812	0.638	0.780	0.641
GMLVQ	0.848	0.929	0.831	0.995

Figure 4 (top) shows results from the chemistry tests carried out on the leaf samples. The graph shows the onset of detection of disease in the plants in week 11 using wet chemistry in the lab. The lower panels of Fig. 4 display predictions of the GMLVQ algorithm over the same time span in terms of a continuous score $0 \leq S \leq 1$. We compare the use of original spectra (middle panel) and the combination of PCA with GMLVQ in 30-dimensional coefficient space (bottom).

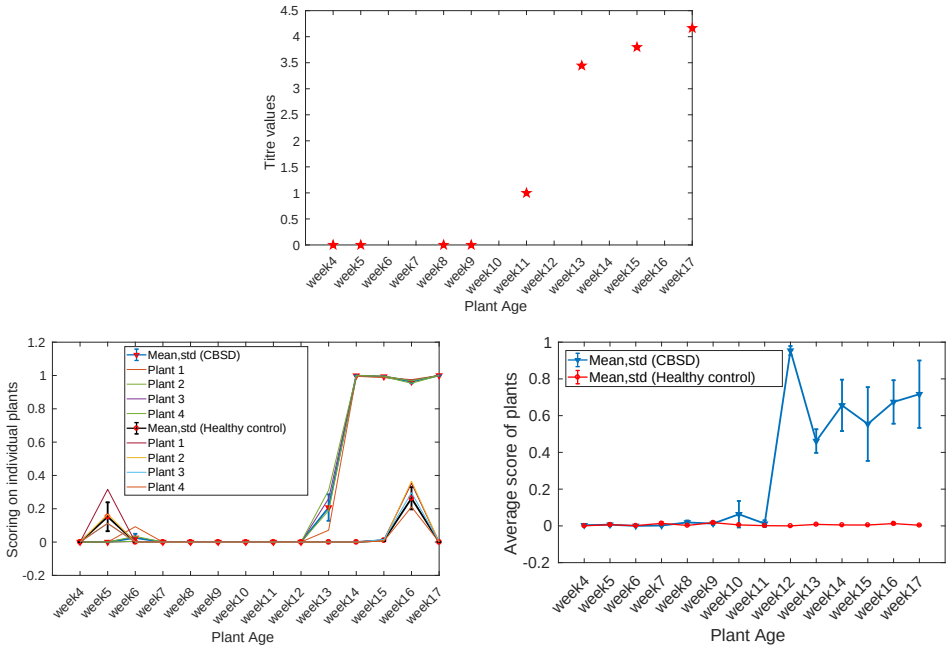

Figure 4: The top graph illustrates the ground truth in terms of virus load based on RT-PCR analysis. The lower panels display GMLVQ scores S, Eq. (5), for individual plants (middle panel) and on average over classes (bottom). The middle panel corresponds to the original space with wavelengths 500-600 nm. The bottom graph shows results of combining GMLVQ with PCA with 30 coefficients.

The GMLVQ score can be interpreted as a proxy for the likelihood of disease, which is computed as:

$$S(\mathrm{x}) = \frac{1}{2}\left[1 + \frac{d(\mathrm{x}, \mathrm{w}^1) - d(\mathrm{x}, \mathrm{w}^2)}{d(\mathrm{x}, \mathrm{w}^1) + d(\mathrm{x}, \mathrm{w}^2)}\right]. \tag{5}$$

A value $S = 0$ indicates that the feature vector x is very likely from class 1, i.e. healthy, while a large value close to $S = 1$ means the plant was classified as diseased, class 2, with high certainty.

For the GMLVQ system based on original spectra we observe that a clear signal of the disease is present in week 14, well before plants become symptomatic (visually) in week 20.

In the plot (bottom panel), we show that the use of PCA enhances the performance further and facilitates the detection of the disease as early as week 12.

5 Discussion and Outlook

We have presented an early disease diagnosis approach to detect cassava CBSD disease before symptoms can be seen by the human eye. Our experimental results show that use spectral data and GMLVQ classification tool show that the presence of the disease can be detected from leaf spectra several weeks before the appearance of visual symptoms.

The experiments were carried out using the most relevant spectral bands, 500 – 600 nm, which had been identified in our previous studies. Classification based on a reduced number of features, as obtained by PCA, not only simplified the classifier, but also made it possible to detect the virus even earlier than by use of original spectra.

Future work should focus on transferring the method from the controlled screen house environment to the field. The ultimate goal is a practical, robust method which requires only low-cost hardware and little computational power. The reliable detection of CBSD and other viral diseases in cassava, before symptoms become visible, would be highly desirable. An early detection of the disease would facilitate treatment and protection measures several weeks before the current practice and could contribute significantly to higher yield and greater food security.

References

[1] Ernest Mwebaze and Michael Biehl. Prototype-based classification for image analysis and its application to crop disease diagnosis. *Advances in Self-Organizing Maps and Learning Vector Quantization*, 428:329–339, 2016.

[2] E. Nuwamanya, P. R. Rubaihayo, S. Mukasa, S. Kyamanywa, J. Hawumba, and Y. Baguma. Influence of spectral properties on cassava leaf development and metabolism. *African Journal of Biotechnology*, 13(7):834–843, 2014.

[3] G. Owomugisha, F. Melchert, E. Mwebaze, J.A. Quinn, and M. Biehl. Machine learning for diagnosis of disease in plants using spectral data. *Proc. Intl. Conf. Artificial Intelligence (ICAI'15)*, pages 9–15, 2018.

[4] Moffat K. Njoroge, D.L. Mutisya, D.W. Miano, and D.C. Kilalo. Whitefly species efficiency in transmitting cassava mosaic and brown streak virus diseases. *Cogent Biology*, 3(1):1311499, 2017.

[5] J.M. Thresh, D. Fargette, and William Otim-Nape. The viruses and virus diseases of cassava in africa. *African Crop Science Journal*, 2(4), 01 1994.

[6] R.J Hillocks, Raya. M., and J. M 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*, 42(4):285–289, 1996.

[7] 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)*, pages 158–163, 2016.

[8] G.M. Rwegasira and C. Mer. Efficiency of non-vector methods of cassava brown streak virus transmission to susceptible cassava plants. *African Journal of Food, Agriculture, Nutrition and Development*, 15(4):10335–10351, 2015.

[9] H. Wagaba, G. Beyene, C. Trembley, T. Alicai, C.M. Fauquet, and N.J. Taylor. Efficient transmission of cassava brown streak disease viral pathogens by chip bud grafting. *BMC Research Notes*, 6(1):516, 2013.

[10] M.N. Maruthi, S. Bouvaine, H.A. Tufan, I.U. Mohammed, and R.J. Hillocks. Transcriptional response of virus-infected cassava and identification of putative sources of resistance for cassava brown streak disease. *PLOS ONE*, 9:1–9, 05 2014.

[11] Sankaran Sindhuja, Mishra Ashish, Ehsani Reza, and Davis Cristina. A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture*, 72(1):1 – 13, 2010.

[12] CID-Bio-Science. Ci-710 miniature leaf spectrometer, 2010.

[13] Jakub Oliwa, Andrzej Kornas, and Andrzej Skoczowski. Morphogenesis of sporotrophophyll leaves in platycerium bifurcatum depends on the red/far-red ratio in the light spectrum. *Acta Physiologiae Plantarum*, 38:247, 2016.

[14] Luiz Oliveira, Marcio Leles Romarco de Oliveira, Francisco Sérgio Gomes, and Reynaldo Campos Santana. Estimating foliar nitrogen in eucalyptus using vegetation indexes. *Scientia Agricola*, 74:142–147, 03 2017.

[15] WA Monger, S Seal, AM Isaac, and GD Foster. Molecular characterization of the cassava brown streak virus coat protein. *Plant Pathology*, 50 (4):527 – 534, 8 2001.

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

[17] G. Owomugisha, F. Melchert, E. Mwebaze, J.A. Quinn, and M. Biehl. Matrix relevance learning for multi-class classification with spectral data. *Submitted*, 10 2019.

[18] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7:92–111, 2016.

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

[20] P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in Learning Vector Quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[21] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46:175–185, 1992.

[22] Steve R. Gunn. Support vector machines for classification and regression. *University of Southampton, Technical Report*, 1998.

[23] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

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

[25] Atsushi Sato and Keiji Yamada. Generalized learning vector quantization. *Hasselmo (Eds.), NIPS*, pages 423–429, 1995.

[26] 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*, 12:2825–2830, 2011.

[27] Michael Biehl. A no-nonsense gmlvq toolbox. *University of Groningen, The Netherlands.*, 2017.