Chapter 4

Machine Learning for diagnosis of disease in plants using spectral data

Abstract

Automating crop disease diagnosis is becoming an increasingly important task, in particular for areas where there is a scarcity of experts. Several attempts made have centered on the analysis of leaf images particularly for diseases that manifest on the aerial part of the plant. It has always been a challenge to get the right dataset and extract the relevant features from the images that can 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 Chapter we take a look at the use of spectral data collected from leaves of the plant. We analyse data from visibly affected parts of the leaf and parts of the leaves that appear to be healthy, visibly. We analyse the obtained data by prototype based classification methods and standard classification models in a three-class classification problem. Results point towards significant improvement in performance using spectral data and the possibility of early detection of disease before the crops become symptomatic, which for practical reasons is highly significant.
4.1 Introduction

The state of the art process of identifying diseases in plants in the field is commonly 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 affected by the disease or not.

This research builds on previous work defined in Chapter 2 and considers work by other groups that has focused on automating the detection of cassava diseases e.g. (Ramcharan et al. 2017). 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 e.g Figure 4.1 illustrates the sever effects caused by CBSD disease detected in late stages.

![Figure 4.1: Crop effect as a result of late diagnosis (a - clean cassava tuber; b,c,d - severe effects caused by CBSD disease)](image)

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 leverage spectroscopy in this study to exploit 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 different leaves, diseases and health as well as from different parts of the leaves; and differently affected parts of the leaf and analyse the associated spectral information.
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. In this current work 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. Our results indicate the possibility of this technique actually working for early detection of disease using spectrometry.

First, we present some work that has already been done in the area of spectrometry for inferring disease in other crops.

Imaging spectroscopy has received broad interest in various sectors of agricultural research, including soil science (Wetterlind et al. 2013, Raphael 2011) and crop disease monitoring. A good review of some of the imaging spectroscopy technologies used in this context can be found in (Sindhuja et al. 2010). Generally, they belong to two categories: spectroscopic and imaging-based, and volatile profiling-based plant disease detection methods.

A multi-spectral imaging system for the diagnosis of plant diseases and insect pests is presented in (Feng, Liao, Liang, Zhao and Dai 2009). The same technique has been applied in diagnosing cucumber diseases (Feng, Liao, Zhao, Luo and Li 2009). The use of fluorescence spectroscopy has also attracted many researchers (Pedros et al. 2008, Schweiger et al. 1996, Lang et al. 1991). This technology involves using a beam of light, usually ultraviolet light, that excites the electrons in molecules of certain compounds and causes them to emit light; typically, but not necessarily, visible light. This technique has been used to detect mechanical and disease stresses in citrus plants (Belasque et al. 2008). The same technology has also been applied in (Wetterich et al. 2013) where computer vision and machine learning techniques were developed to detect Huanglongbing Citrus Disease.

A field imaging spectroscopy system to retrieve chlorophyll content from soybean leaves is discussed in (Bo et al. 2014). The methods used in this application were multiple linear regression, partial least squares regression and support vector machine regression. Methods for early detection of rice blast using near-infrared hyper-spectral image are presented (Yang et al. 2012). The use of Near Infrared Spectroscopy to analyse Cold Rice Blast has been discussed (Tan et al. 2012).

Overall, spectroscopy as a tool for measuring the state of biological sample is becoming prevalent and in this work we show first attempts at harnessing it for address viral disease diagnosis in cassava. However, by nature spectroscopy data comes in large dimensions and to address the disease classification challenges in cassava, we investigate feature dimensionality and classification techniques. For example, (Villmann and Merényi 2008) present machine learning approaches and pattern recognition for spectral data. The use of Relevance Learning to classify diverse
samples such as Scotch Whisky, wine and other datasets has also been considered in literature (Melchert et al. 2016a, Backhaus et al. 2012).

4.2 Materials & Methods

Here we describe the data collection process, pre-processing and analysis of the acquired data.

4.2.1 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 4.2 illustrates the data collection pipeline for automated 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.

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 (CID Bio-Science Inc 2010). 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 and across several cassava varieties (Nakabonge et al. 2018), five were considered. Within a variety, three plants were considered and three leaves for each plant were sampled. For each leaf, two readings were taken on each leaf lobe: one on the good part (not visibly showing symptoms) and another one 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. Note that
this was taken care of during validation of the models, so that we never trained and tested on data from the same plant. In total, 760 data points were collected for evenly distributed disease classes. Figure 4.3 illustrates what the terminology of good and bad leaf parts.

![Figure 4.2: Cassava disease automated diagnostic pipeline as described in 4.2](image)

### 4.2.2 Image data processing

Following methodologies from previous work (Mwebaze and Biehl 2016) 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 are considered. SIFT features descriptors that comes in 128 dimensions are also extracted. Both color and SIFT features are computed using OpenCV libraries.
4. Machine Learning for diagnosis of disease in plants using spectral data

Figure 4.3: Data collection in the field & depiction of good and bad part of leaf

(Bradski 2000).

4.2.3 Spectral data pre-processing

Spectral data is very sensitive not only in acquisition but also at analysis stage. In this section we discuss the different data scaling and transformation techniques we found useful during analysis as we classified two cassava diseases (CBSD and CMD) as well as healthy crops. We begin by defining basic concepts.

Data scaling

Two terms; data normalization and standardization are often used interchangeably but mean different things. Normalization refers to re-scaling values into ranges of, for instance, zero to one while standardization means re-scaling values to display, e.g., zero mean and unit variance in a given data set. Under standardization, we investigated on methods including; mean, median, standard deviation, data shifting and Z-score. Considering data sets of the following form:

\[ \{x^\mu, y^\mu\}_{\mu=1}^P \]  

where \( x^\mu \in \mathbb{R}^N \) are feature vectors, the labels \( y^\mu \in 1, 2, ...C \) specify their class membership and \( P \) is the number of observations (spectra). The mean \( \vartheta \) of the spectra is computed as:

\[ \vartheta = \frac{\sum_{\mu=1}^P x^\mu_i}{P} \]  

Dimensionality reduction

In the experiments, we performed data transformation in order to identify useful low-dimensional representations of feature vectors. Spectral data come with thousands of dimensions that are also highly correlated with each other. Therefore, understanding the underlying correlation patterns is useful in representing the data set compactly with fewer variables as well as separating out relevant signals from unwanted noise and also reducing the computational complexity. In this chapter...
4.2. Materials & Methods

The standard deviation $\delta$ is given by the following equation.

$$\delta = \frac{\sum_{\mu=1}^{P} (x_i^\mu - \vartheta)^2}{P - 1}$$

(4.3)

We also experimented with signal shift given by:

$$\tilde{x}_i^\mu = x_i^\mu + (k - \vartheta(x_i^\mu))$$

(4.4)

where $k$ is a positive constant and $\vartheta$ the mean of data samples.

Z-score is one of the widely used data standardisation methods. Z-score ($z_i$) is computed by subtracting the sample mean ($\vartheta$) from the target data point ($x_i^\mu$) and dividing by the target standard deviation ($\delta$):

$$z_i = \frac{x_i^\mu - \vartheta_i}{\delta_i}$$

(4.5)

Under normalization, we experimented with $L1$ normalization ($x^{\mu(L1)}$) and $L2$ normalization ($x^{\mu(L2)}$).

$$x^{\mu(L1)} = \frac{x_i^\mu}{\sum_{\mu=1}^{P} (x_i^\mu)}$$

(4.6)

$L2$ normalization divides a data point ($x_i^\mu$) by the square root of the sum of the squared intensities of the spectrum:

$$x^{\mu(L2)} = \frac{x_i^\mu}{\sqrt{\sum_{\mu=1}^{P} (x_i^\mu)^2}}$$

(4.7)

where $i \in \{1, 2, \ldots, N\}$ and $\mu \in \{1, 2, \ldots, P\}$. More on these normalization techniques for spectral data can be found in (Randolph 2006).

**Dimensionality reduction**

In the experiments, we performed data transformation in order to identify useful low-dimensional representations of feature vectors. Spectral data come with thousands of dimensions that are also highly correlated with each other. Therefore, understanding the underlying correlation patterns is useful in representing the data set compactly with fewer variables as well as separating out relevant signals from unwanted noise and also reducing the computational complexity. In this chapter
we focus on dimension reduction as an important aspect for practical deployment purposes. We considered two approaches for the dimensionality reduction.

Our first approach is functional approximation by Chebyshev polynomials. The functional nature of the data can be exploited systematically by using appropriate representations. In particular, Chebyshev polynomials of the first kind have been employed for a set of basis functions and showed good classification performance in several applications, see (Melchert et al. 2016a, Melchert et al. 2016b).

The assumption that available spectral data $x^\mu$ result from equidistant sampling of the real physical spectrum $f^\mu(\lambda)$ of a leaf sample motivates the following relation:

$$x^\mu_i = f^\mu(\lambda_i), \ i = 1, 2, \ldots, N$$ (4.8)

where $\lambda_i$ represents sampling points of the spectrum defined by the employed sensor.

Given a suitable set of basis function $g_k(\lambda)$ it is possible to represent $f^\mu(\lambda)$ as a weighted sum:

$$f^\mu(\lambda) = \sum_{k=0}^{\infty} c_k^\mu g_k(\lambda).$$ (4.9)

Restricting the maximum number of basis functions to a finite number $n$, Eq.(4.9) yields an approximation $\hat{f}^\mu(\lambda)$ of the original spectrum.

The computation of the coefficients $c_k^\mu, k = 1, 2, \ldots, n$ can be formulated as an optimization problem, achieving best approximation quality (e.g. mean square error) for a given number of basis functions. For Chebyshev polynomials of the first kind, the computation of coefficients can be done in an effective manner employing a single transformation matrix:

$$c^\mu = C x^\mu \text{ where } x \in \mathbb{R}^N,$$ (4.10)

$$C \in \mathbb{R}^{n \times N} \text{ and } c \in \mathbb{R}^n.$$ (4.11)

In practice, setting $n \ll N$ yields an efficient dimensionality reduction which does not require prior knowledge of the data. Furthermore, this includes an implicit denoising of the spectral information by discarding higher order polynomials (Melchert et al. 2016a).
4.2. Materials & Methods

Our second approach to dimensionality reduction was the use of principal component analysis (PCA). PCA is a widely used standard technique for correlation analysis and dimensional reduction e.g. (Vasan and Surendiran 2016). The technique is a linear projection operator that transforms random variables into a new random variable that has maximum variation.

Consider a data matrix \(X (M \times N)\), where \(M\) is the number of points, and \(N\) is data dimensions.

The transformation is defined by a set of \(N\)-dimensional vectors of weights or coefficients \(w(\kappa) = (w_1, \ldots, w_N)\) that map each row vector \(x(i)\) of \(X\) to a new vector of principal component scores \(t(i) = (t_1, \ldots, t_l)\), given by

\[ t_k(i) = x(i) \cdot w(k) \] for \(i = 1, \ldots, n\) \(k = 1, \ldots, l\)

in such a way that the individual variables \((t_1, \ldots, t_l)\) of \(t\) considered over the data set successively inherit the maximum possible variance from \(X\), with the coefficient vectors \(w(k)\) being orthogonal unit vectors.

For dimensionality reduction, PCA transforms \(X\) into \(Y (M \times N')\). However, not all the principal components need to be kept. Keeping only the first principal components yields \(N' \leq N\).

These techniques are also used in the next chapters, therefore reference to this section will be required for definitions.

In our experiments we used the following pre-processing pipeline: For spectral data shown in Figure 4.4, we considered the region between the wavelengths 400 nm – 900 nm after truncating the smallest and largest wavelengths where the measurement appears to be very noisy. Next step was smoothing the spectra. We compared two filtering techniques: median (Arias-Castro and Donoho 2009, Hamza et al. 1999) and average (Smith 1999). For both, we used a window size of 15 nm. Our experiments showed that average filtering yielded better classification results for this window size. As a consequence, average filtering was applied on all the data. After this step, we performed data scaling and we applied data shift because of its good accuracy and feature presentation compared to other methods. We then apply dimensionality reduction techniques that could also be interpreted as optional pre-processing step but will be discussed separately in Chapter 5.

4.2.4 Training a diagnosis classifier

Several options abound for which type of model to train for this kind of data. Prior work has used convolutional neural networks (CNNs) (Ramcharan et al. 2017, Sladojevic et al. 2016) and prototype based methods with great success (Mwebaze et al. 2011, Mwebaze and Biehl 2016). 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 (Pedregosa et al. 2011). For our experiments we use the following: (i) K-Nearest Neighbour (KNN) (Altman 1992) because it is very similar in flavor to prototype based methods, (ii) Linear Support Vector Machine (SVM) (Gunn 1998) because it has shown good performance previously (Mwebaze and Biehl 2016), and (iii) Decision trees, because these have also previously (Owomugisha and Mwebaze 2016) shown good performance particularly the Extremely Randomized Trees (Extra trees) algorithm (Geurts et al. 2006).

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 these methods are very intuitive because of the training process. A prototype is trained in the space of the data. For deployment purposes, prototype-based methods are also 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 in (Kohonen 1986) and since then various modifications have been suggested in the literature all aiming at better convergence or favorable generalization behaviour (Sato and Yamada 1995, Schneider et al. 2007). In LVQ, a particular classification task is defined by a set of $M$ prototype vectors.
$w^j \in \mathbb{R}^N$ which carry labels $c(w^j) \in \{1, 2, \ldots, C\}$ such that $W = \{w^j, c(w^j)\}^M_{j=1}$. 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 $x \in \mathbb{R}^N$ to the closest prototype with respect to some meaningful distance measure.

Most frequently, standard Euclidean distance $d(w, x)$ is employed. The corresponding NPC assigns $x$ to the class $c(w^L)$ of the closest prototype with $d^A(x, w^L) \leq d^A(x, w^j)$ for all $j$.

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

The GMLVQ algorithm proposed in (Schneider et al. 2007) employs a full matrix $\Lambda \in \mathbb{R}^{N \times N}$ of relevances that represents the importance of single features and their combinations in the classification task. Here, the distance measure $d^\Lambda(x, w)$ is defined as:

$$d^\Lambda(x, w) = (x - w)^\top \Lambda (x - w),$$  \hspace{1cm} (4.12)

where the parameterization $\Lambda = \Omega^\top \Omega$ guarantees that $d^\Lambda(x, w) \geq 0$ for arbitrary matrices $\Omega \in \mathbb{R}^{N \times N}$. 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 in (Sato and Yamada 1995):

$$E(W) = \sum_{\mu=1}^P \Phi \left( \frac{d^\Lambda_j(x_i) - d^K_j(x_i)}{d^\Lambda_j(x_i) + d^K_j(x_i)} \right)$$  \hspace{1cm} (4.13)

where $d^\Lambda_j$ denotes the distance to the closest correct prototype with $c(w^\mu) = y^\mu$ and $d^K_j$ is the distance to the closest incorrect prototype ($c(w^\mu) \neq y^\mu$). The modulation function $\Phi$ 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 (Hammer et al. 2005, Schneider et al. 2007) also provides us with a way of reducing the dimensionality of the spectral data in this case as the trained obtained relevance matrix is typically of low rank (Biehl et al. 2016). In the next chapter, we extend this method to identify relevant wavelengths that are most important for classifying the different diseases. The major goal is to extend this to the construction of a simpler, cheaper spectrometry tool that offers analysis in a limited wavelength band.
Validation

For all the models we train, we carry out a 10-fold cross validation and average the performance over the folds. Previously (Chapter 3), we explored with KNN, linear SVC and Extra trees. Here, we employ the following parameters: $K=15$ for the KNN, $C = 1$ 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 (Biehl 2016).

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) and 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 cross-validation as implemented in the SciKit-learn toolbox (Pedregosa et al. 2011).

4.3 Results

4.3.1 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 were 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 were not visibly infected (good part). We run the battery of algorithms on the two datasets and present the results in Table 4.1. 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 4.1: Spectral data dependence on leaf quality (Healthy vs. CBSD, CMD). Overall cross-validation accuracy scores (%) for the different algorithms.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Leaf part</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bad</td>
<td>Good</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>0.919</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>Linear SVC</td>
<td>0.941</td>
<td>0.957</td>
<td></td>
</tr>
<tr>
<td>Extra trees</td>
<td>0.917</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td>GMLVQ</td>
<td>0.937</td>
<td>0.973</td>
<td></td>
</tr>
</tbody>
</table>

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

4.3.2 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 4.4 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 overall cross-validation 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. Another observation from this work is that, spectral data works for “good” and “bad” parts while image data can only work on bad parts.

4.3.3 PCA spectral features

Using PCA, we are able to reduce the spectral data dimension from 2554 to 30 principal components. In Figure 4.5, we illustrated 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 4.4 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
Table 4.2: Confusion matrix (%) for Bad part of leaf with GMLVQ

<table>
<thead>
<tr>
<th></th>
<th>Healthy</th>
<th>CBSD</th>
<th>CMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>98.70</td>
<td>1.30</td>
<td>0</td>
</tr>
<tr>
<td>CBSD</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>CMD</td>
<td>0</td>
<td>17.04</td>
<td>82.96</td>
</tr>
</tbody>
</table>

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

4.3.2 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 4.4 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 overall cross-validation 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. Another observation from this work is that, spectral data works for “good” and “bad” parts while image data can only work on bad parts.

4.3.3 PCA spectral features

Using PCA, we are able to reduce the spectral data dimension from 2554 to 30 principal components. In Figure 4.5, we illustrated 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 4.4 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

Table 4.3: Confusion matrix (%) for Good part of leaf with GMLVQ

<table>
<thead>
<tr>
<th></th>
<th>Healthy</th>
<th>CBSD</th>
<th>CMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CBSD</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>CMD</td>
<td>0</td>
<td>8.2</td>
<td>91.8</td>
</tr>
</tbody>
</table>
Table 4.4: Overall cross-validation accuracy scores (%) with different data features (Healthy vs. CBSD, CMD)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Color</th>
<th>SIFT</th>
<th>Spectral Original</th>
<th>Spectral PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.705</td>
<td>0.849</td>
<td>0.923</td>
<td>0.932</td>
</tr>
<tr>
<td>Linear SVC</td>
<td>0.738</td>
<td>0.895</td>
<td>0.957</td>
<td>0.959</td>
</tr>
<tr>
<td>Extra trees</td>
<td>0.803</td>
<td>0.889</td>
<td>0.927</td>
<td>0.944</td>
</tr>
<tr>
<td>GMLVQ</td>
<td>0.742</td>
<td>0.901</td>
<td>0.973</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Figure 4.5: Overall accuracy (%) with increasing number of principal components with GMLVQ algorithm.

there is a reduction in noise in the data when we do a PCA transformation, however the corresponding disadvantage is that we are 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 compensated for by the reduced time to do the prediction.

4.4 Discussion

This chapter 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 4.1 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 4.2 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 disease visibly infected part of the plant.

Table 4.4 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. As mentioned image data is also prone to occlusion making it less accurate in prediction. The GMLVQ also provides superior performance compared to other algorithms, see Table 4.3 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 4.4 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.

### 4.5 Conclusion

This chapter has 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 demonstrated the consistency of spectral data collection from different parts of the leaf. Of particular 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.
Matrix relevance learning for multi-class classification with spectral data

Abstract

We discuss the use of matrix relevance learning, a popular extension to prototype learning algorithms, applied to a three-class classification task of diagnosing cassava diseases using spectral data. Previously this diagnosis has been done using plant image data taken with a smartphone. However, for this method, disease symptoms need to be visible and, once symptoms have manifested, the root of the plant is already affected and can no longer be used as food.

This research is premised on the hypothesis that diseased crops without visible symptoms can be detected using spectral information, allowing for early action measures. We analyze visible and near-infrared spectra captured from leaves infected with two common cassava diseases (cassava brown streak disease and cassava mosaic virus disease) found in Sub-Saharan Africa as well as from healthy plants. The spectral data come with thousands of dimensions, therefore different wavelengths are analyzed in order to identify the most relevant spectral bands. To cope with the nominally high number of input dimensions of data, functional decomposition of the spectra is considered. The outlined classification task is addressed using Generalized Matrix Relevance Learning Vector Quantization and compared with the standard classification techniques performed in the space of expansion coefficients.