Identifying Agathosma Leaves using Hyperspectral Imagery and Classification Techniques

B. T. Abe and J. A. Jordaan

***Abstract*— We investigated the ability of hyperspectral data to identify Agathosma Betulina and Agathosma Crenulata plants. The plants have been used as traditional medicines to heal diseases such as urinary tract infections, stomach complaints, for washing and cleaning wounds, kidney diseases, and symptomatic relief of rheumatism. The species are normally identified on the basis of their shapes. A. Betulina has round-leaves while A. Crenulata has oval-leaves. This recognition based on morphology is no longer adequate because of extensive cultivation. New hybrids of the leaves now exist which are not easily separable. In this study, hyperspectral image and classification techniques are used to separate the plants. The Agathosma species imagery was used to generate datasets for the classification procedure. The images were processed using Local Polynomial Approximation (LPA) and Principal Component Analysis. Random Forest (RF) and Support Vector Machine (SVM) classifiers were used for the data separation. The results obtained reveal that the classifiers perform better on the LPA processed data as compared to PCA.**

***Index Terms*— Agathosma Betulina, Agathosma Crenulata, classifiers, Hyperspectral Image**

I. Introduction

PLANTS are very essential aspect of life that provide food, medicine, energy, oxygen, wood, healing, among others. In South Africa, plants are used as traditional or alternative medicine and are used by traditional healers for health care (Van Wyk et al., 1997; Thring and Weitz, 2006). The country has a number of natural plants-based products used for traditional remedies as a source of commercial products. Among the natural plants is the Agathosma (A) species, which has two popular types namely: A. Betulina (round-leaf buchu) also known as "bergboegoe" and A. Crenulata (oval-leaf buchu) known as anysboegoe (Street and Prinsloo, 2013). These two types are probably some of the best-known plants used for medicinal purposes (Moolla and Viljoen, 2008).

Manuscript received July 2016. This work was supported in part by the Department of Higher Education and Training, Tshwane University of Technology and the National Research Foundation (NRF), South Africa.

B. T. Abe is with the Tshwane University of Technology, Department of Electrical Engineering, eMalahleni campus, 1035, South Africa. (Corresponding author phone: number +27761304108; e-mail: abe_tolulope@yahoo.com).

J. A. Jordaan is with the Tshwane University of Technology, Department of Electrical Engineering, eMalahleni campus, (e-mail: jordaan. jaco@gmail.com).

Apart from the fact that these plants are used for medicine, they are also use for producing oil. A. Betulina produces high quality oil and is widely available in the market while A. Crenulata product is of lower quality. The species are normally separated based on leaf morphology. Because of extensive cultivation of the spices for oil production, new hybrids of the leaves now exist which are not easily separable by morphological techniques. So, in extreme cases where the leaves are obviously different, the two types of leaves could be distinguished visually with the eyes, but in many cases the leaves are difficult to classify. It is therefore imperative to identify the species through a better technique for the separability. Figure 1 (a) and (b) shows the pictures of Agathosma plants/leaves demonstrating the leave shapes.

Figure 1: Agathosma plants (a) Betulina (b) Crenulata

In this research, we aim at exploring classifiers as tools for separating the species by using the data of the leaves captured with a hyperspectral imagery system.

Hyperspectral images always contain dozens to hundreds of spectral features that are usually used for quantitative and qualitative analysis of numerous targets and materials across the electromagnetic wavelength spectrum (Abe *et al*. 2012; Landman *et al*. 2015; van der Meer *et al*., 2012). Classification based on a leaf image is the main method for leaf plant recognition and classification (Kulkarni *et al*., 2013). This is achieved using sample leaves imagery transferred into a computer for the computer to extract useful information with the application of image processing techniques and consequently identifying the leaf using machine-learning techniques.

Principal component analysis and Local Polynomial approximation techniques are used for the data processing. The remaining part of the paper is structured as follows; section two discusses Local Polynomial approximation (LPA) method, how LPA and PCA were applied to

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

hyperspectral images for data analysis are presented in section three. Section four presents the classification methods. The results obtained are discussed in section five, while section six concludes the work.

II. LOCAL POLYNOMIAL APPROXIMATION

Differentiation and smoothing of data by making use of piecewise polynomials is widely used. Assuming that we have a data set of evenly spaced data points $\{x_k\}$, this data could be smoothed or differentiated with a filter and a polynomial. These filters were used by Savitzky and Golay (Gorry, 1990), The idea behind them is the regression of the raw data within a moving interval of data points, also called a window, and the length of this interval is called the window length. According to (Gorry, 1990) and (Bialkowski, 1989) the data within the window is fitted with a local polynomial function by making use of a least squares technique. From this one obtains the filter coefficients, which define the filter's impulse response.

Three types of data windows could be used. These are left, central or right-sided window. For the *left-sided* window the data sample of interest (the sample we want to smooth) is the right-most sample in the window. The rest of the samples are therefore to the left of the sample to be smoothed. Similarly for the *central* window the data sample of interest is the one in the middle of the window and for the *right-sided* window the sample of interest is the left-most sample in the window. Only the central window is used in this paper.

The derivation of the local polynomial approximation (LPA) model is shown next. Assume the data sample that should be smoothed have an index $k = 0$, which only refers to the samples in the data window. In continuous time a power series of orthogonal polynomial basis functions is given by $f(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots + c_p t^p$. Sampling the data ponts with period T, the time t could be represented by $t = kT$. The coefficients of the polynomial are c_i, and p is the polynomial order. The sampled equation is $f(k) = c_0 + c_1 kT + c_2 (kT)^2 + \ldots + c_p (kT)^p$.

For the central window, solving the coefficients c_i, using a least-squares criterion, produces the following objective function $J = \sum_{k=-w_n}^{w_n} (x_k - f(k))^2$, where $2w_n + 1$ is the window length, x_k is the $k-$th measured data point within the window and $f(k)$ is the value for the data point based on the polynomial model. To obtain the best fit, the objective function is minimized by setting the gradient equal to zero $\frac{\partial J}{\partial c_i} = \mathbf{0}$, and solving the set of equations in the unknown coefficients c_i.

For the Savitzky-Golay method (Gorry, 1990), we are only interested in the data point where $k = 0$. For $k = 0$, the *s*-th derivative only requires an expression for c_s. Considering a single polynomial term $f_i(k) = c_i (kT)^i$ we can write the s-th order derivative as follows:

$$f_i^{(s)}(k) = (i-s+1)(i-s+2)\ldots(i-1)i c_i (kT)^{i-s} = \frac{i!}{(i-s)!} c_i (kT)^{i-s}.$$

Including all the LPA terms we can write the s-th order derivative of the function as $f^{(s)}(k) = \sum_{i=s}^{p} f_i^{(s)}(kT)$.

For more explanations and derivations on LPA, see (Jordaan, 2006).

III. DATA ANALYSIS

A. LPA Applied to Hyperspectral Data

The leaf samples are scanned to give the hyperspectral data, where each leaf is made up of a series of pixels and each pixel has a set of intensities over the different spectral bands. To apply LPA to the hyperspectral data, each pixel with its full spectrum of intensity values is treated as a separate data set. This means that for each pixel, a local polynomial is fitted in a sliding window moving over the spectrum of the pixel. This is done in order to extract, as data features, the derivatives of the data. These features could then used in classification of the different leaves.

For this study, the following settings for LPA were used: the window length is 11 ($w_n = 5$), and only the first order derivative is used. Since the focus of the paper is not on window length, the effect of window length on the classification results is not discussed nor investigated. The number of spectral values per pixel is 256. Therefore, there will also be 256 derivative values per pixel.

B. Application of PCA to Hyperspectral data

Principal component analysis model is used to extract the spectral patterns of the species and this provides visual plots. Near infrared (NIR) spectra patterns of nine leaf samples per species are chosen to reveal the features to identify the species. Figure 2 (a) – (d) show the leaf samples NIR spectra of the Betulina (BTL) and Crenulata (CLT).

Because the features of the leaves cannot be detected clearly from the spectral patterns, Principal Component Analysis (PCA) was conducted on the data to give the visual plots for clearer observation. The PCA was applied with mean centered on the leaves. The aim is to reduce dimensionality of the hyperspectral dataset by decomposing unified variables into a new set of uncorrelated coordinates for arrangement in a way that the first few components have the variation of the data (Sandasi *et al*., 2014).

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

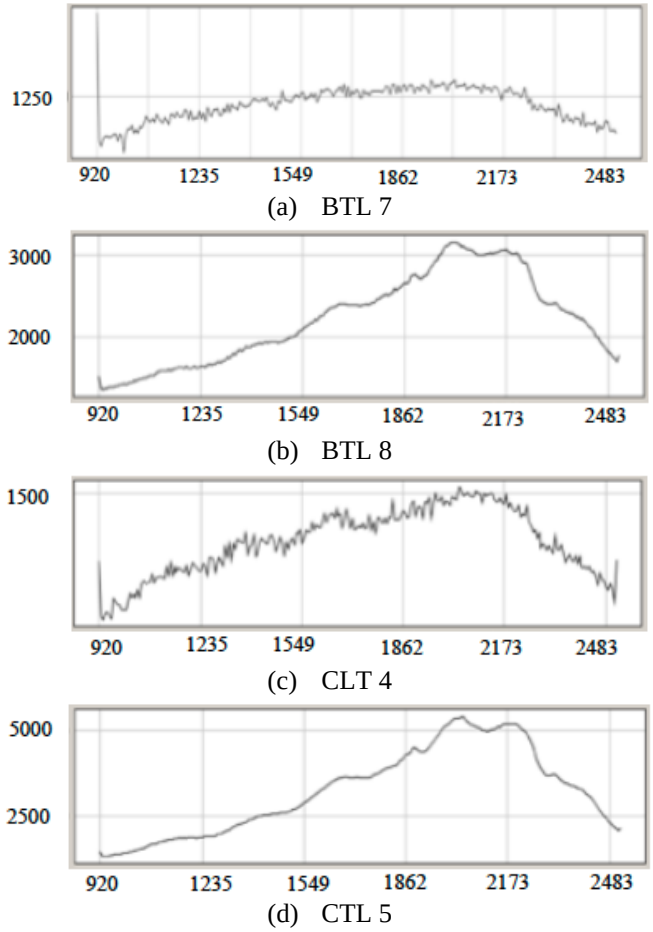


Figure 2: Near Infrared imaging spectra of A. Betulina (a–b) and A. Crenulata (c - d) species

IV. Classification Methods

The *A. Betulina* and *A. Crenulata* datasets generated from LPA and PCA has been used for the experiment. Hyperspectral image of nine reference leaves per species were used to obtain the datasets. The images were captured using the macro lens (10mm) with a 1cm field of view. Each pixel represented a NIR spectrum ranging from 1000 – 25000nm. Random Forests (Breiman, 2001) and Support vector machine (Vapnik, 1999) classifiers were used to conduct the classification process. Their predictions are examined to see how the classifiers are able to separable the leaves for recognition. The WEKA (Witten and Frank, 1999) data mining tool was used for classification. To train and test the classifiers, 25,603 instances and 9 features per species were identified with 70% of the dataset used for training and the remaining for testing. Table 1 presents how the instances are distributed.

V. Results and Discussion

This section present results and discussion of our experiment for the leaves identification as conducted by the classifiers. Table 2 presents the experimental results as predicted by the classifiers. Based on the classifiers' performance, the results reveal that LPA technique for data preparation identified the leaves better than the PCA.

TABLE I
Data Distribution

Betulina (BTL)	Number of attributes	Crenulata (CLT)	Number of attributes
BTL 1	1869	CLT 1	1451
BTL 2	1761	CLT 2	1726
BTL 3	1687	CLT 3	1515
BTL 4	2056	CLT 4	962
BTL 5	584	CLT 5	893
BTL 6	1419	CLT 6	1291
BTL 7	1716	CLT 7	1611
BTL 8	1600	CLT 8	1490
BTL 9	540	CLT 9	1432
Total	13232	Total	12371
TOTAL in all = 25,603			

TABLE 2
Accuracy Results of the Classifiers

Classifier	LPA Accuracy	PCA Accuracy
Random Forest	88.0094 %	77.6071 %
Support Vector Machine	81.6691 %	71.423 %

In general, the results reveal that the classifiers performance on the dataset processed using LPA technique is better with 88% and 82% (approx.) accuracy for RF and SVM respectively as compared to PCA processed data. In addition, it was observed during the experimental results that the classifiers spent less time on training and testing the model on training split using LPA dataset as compared to PCA dataset.

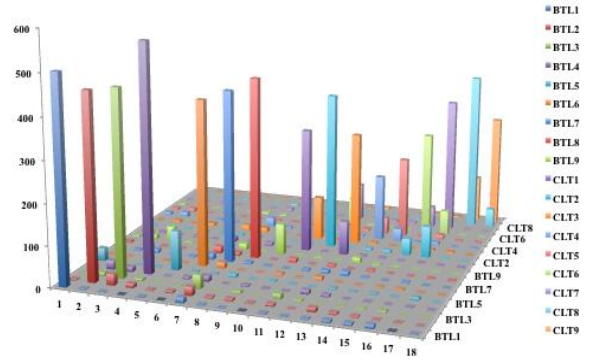
For better clearance of the result, Figure 3 (a) –(d) presents the graphical representations of the classifiers performances. From the graphs, it can be deduced that with the PCA charts, the pixel are more scattered and they are not very coherent as compared to LPA charts.

VI. Conclusion

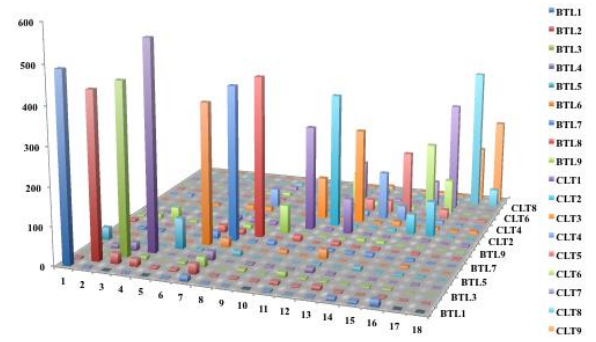
In this study we used hyperspectral image data and machine learning to identify *A. betulina* and *A. crenulata* leaves. The LPA and PCA were used for the data processing. Random Forest and Support Vector Machines classifiers were used for classification. The accuracy results generated by the classifiers with the data processed by LPA technique are better than the PCA processed data. This implies that the experiment conducted using hyperspectral imagery is a feasible alternating technique for classifying the leaves. Our experiment also shows that the technique is cost effective.

Acknowledgment

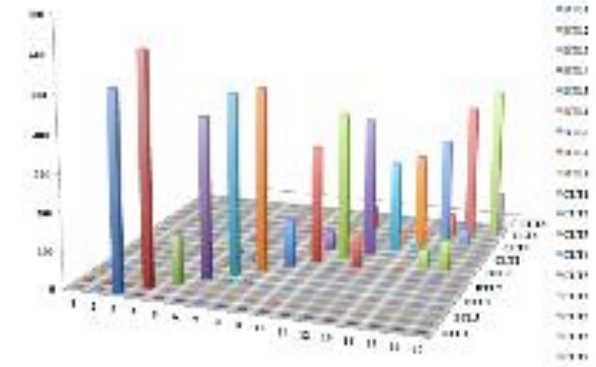
This work is based on the research supported by the Department of Higher Education and Training, Tshwane University of Technology and the National Research Foundation of South Africa (Grant specific unique reference number (UID) 85745). The Grant holders acknowledge that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the author(s), and that the NRF accepts no liability whatsoever in this regard.



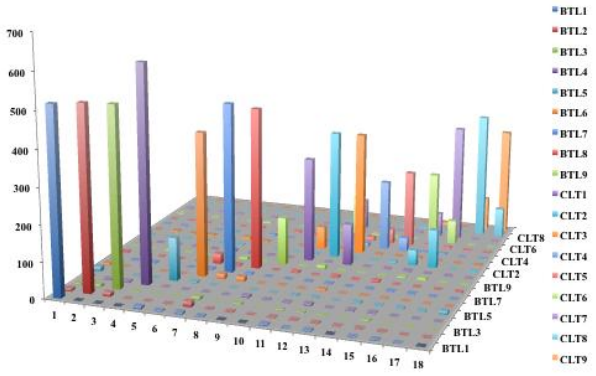
(a) Random Forest accuracy results using PCA



(b) Support Vector Machine accuracy results using PCA



(c) Random Forest accuracy results using LPA



(d) Support Vector Machine accuracy results using LPA

Figure 2 (a) – (d): Classification accuracy results

REFERENCES

[1] B. T. Abe, O. O. Olugbara and T. Marwala, "Hyperspectral Image Classification using Random Forest and Neural Network," *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2012, WCECS 2012*, 24-26 October, San Francisco, USA, 2012, pp. 522-527.

[2] S. Bialkowski, "Generalized Digital Smoothing Filters Made Easy by Matrix Calculations," *Analytical Chemistry*, vol. 61, no. 11, June 1989. pp. 1308–1310.

[3] L. Breiman, "Random forests," *Machine Learning*, 45, 1, 2001, pp. 5–32

[4] P. Gorry, "General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method," *Analytical Chemistry*, vol. 62, no. 6, March 1990, pp. 570–573

[5] J.A. Jordaan, "Fast and Accurate Spectral Estimation Algorithms for Power System Applications," Doctoral Thesis, Tshwane University of Technology, South Africa, 2006.

[6] Kulkarni *et al*., "A Leaf Recognition Technique for Plant Classification Using RBPNN and Zernike," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, Issue 1, pp. 984 – 988

[7] T. Landmann, *et al*., "Application of hyperspectral remote sensing for flower mapping in African savannas," *Remote Sensing of Environment*, vol. **166**, 2015, pp. 50-60.

[8] A. Moolla and A.M. Viljoen, " 'Buchu' – Agathosma betulina and Agathosma crenulata (Rutaceae): A review", *Elsevier Journal of Ethnopharmacology*, vol. 119, 2008, pp. 413–419

[9] Sandasi *et al*., "Hyperspectral imaging and chemometric modeling of echinacea - a novel approach in the quality control of herbal medicines," *Molecules*, 19(9), 2014, pp13104-21.

[10] R.A. Street, G. Prinsloo, "Commercially important medicinal plants of South Africa: a review," *J. Chem.*, vol. 2013, pp. 1–16.

[11] T.S.A. Thring, F.M. Weitz, "Medicinal plant use in the Bredasdorp/Elim region of the Southern Overberg in the Western Cape Province of South Africa," *Journal of Ethnopharmacology*, vol. 103, 2006, pp. 261–275

[12] F.D. Van derMeer, H.M.A. van derWerff, F.J.A. van Ruitenbeek, C.A. Hecker, W.H. Bakker, M.F. Noomen, *et al*., "Multi- and hyperspectral geologic remote sensing: A review," *International Journal of Applied Earth Observation and Geoinformation*, 14, 2012, pp. 112–128.

[13] B.E. Van Wyk, B. Van Oudtshoorn, N. Gericke, "Medicinal Plants of South Africa," Briza Publications, Pretoria, 1997.

[14] V. Vapnik, "The nature of statistical learning theory," *second edition. New York: Springer-Verlag*, 1999.

[15] I. H. Witten and H. Frank, "Data mining: practical machine learning tools and techniques with java implementations," San Francisco: Morgan Kaufmann, 1999.

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)