# Distinguishing Healthy and Infected Vegetable Crops using Hyperspectral Leaf Reflectance

**Anjana Ghule[1], Prof. Dr. R. R. Deshmukh[2] and Chitra Gaikwad[3]**

[1] Assistant Professor, IT Department, GEC Aurangabad. Email: anjanaghule@gmail.com
[2] Professor, CS and IT Department, Dr. B.A.M. University Aurangabad. Email: rrdeshmukh.csit@bamu.ac.in
[3] Assistant Professor, IT Department, GEC Aurangabad. Email: gaikwadchitra@gmail.com

**Hyperspectral Remote Sensing provides data over a large number of contiguous wavebands. Primary objective of this research was to reduce the dimensionality and distinguish healthy and infected vegetable plants by selecting significant spectral region and subset of wavelengths. For the experimentation, spectral measurements of healthy and diseased leaves of Brinjal, Cluster Beans and Long Beans were divided into five regions of electromagnetic spectrum, i) Visible: 350nm-670nm ii) Red Edge: 671nm-780nm iii) Near Infrared: 781nm-1000nm iv) Shortwave I: 1301nm-1500nm and v) Shortwave II: 1701 nm-1900 nm. PCA and LDA methods were used as pre-filters for dimensionality reduction on each region, before applying Random Forest classifier. The results obtained revealed that classification performance of Visible and Red Edge regions was better than NIR, SW1 and SWII regions. The PCA method provided better accuracy and kappa values as compared to LDA.**

**Keywords:** Hyperspectral Remote Sensing, Spectral subset selection, Vegetable crops, Dimensionality reduction, Classification

## 1. Introduction

India, the second largest populated country and majority of Indians prefer vegetarian food. Broadly we can say that the health of Indians depends on healthy crops. Various types of vegetables are grown in different regions of India. Considering the public health and economy of the country, it is essential to monitor the health of vegetables and crops.

Hyperspectral Remote Sensing has extensive applications in the vegetation and agriculture over conventional methods [1]. Crop species discrimination, monitoring of crop growth, crop identification, crop type classification, vegetation health monitoring is also very crucial in the field of agriculture[2]. Plant health or canopy health has direct impact on the ecosystem. Major part of the plant above the soil is covered with leaves. Plant health can be monitored by studying leaves and we can consider leaves as representatives of the canopy. For the study of crop health, it is important to characterize and quantify vegetation parameters. Nutrient deficiencies, water deficiencies, floods and disease are major factors of crop stress [3].

Crops have different biophysical and biochemical characteristics, and this has a direct impact on the spectral profile. Measurement and study of attributes and their inter-relationship can provide significant information about plant productivity, plant health, stress and availability of nutrients. The visible domain is in the range of 400nm -700nm., major photosynthetic pigments (chl a) chlorophyll a and (chl b) chlorophyll b shows absorption in this region. In 700-1300nm NIR region, absorption is very low as cellulose and leaf pigments are almost transparent. 1300-2500nm SWIR region is largely influenced by water contents. Other parameters such as protein, cellulose, lignin and starch also influence SWIR region [3][4].

Red Edge Position is extensively used for the assessment of crop stress. Sudden inflection is observed between 680 - 780 nm in the reflectance spectra of vegetation, this is called as Red Edge Region. Red Edge Position is the wavelength at which maximum magnitude of First Derivative curve (FD) is observed in the red-edge region [5].When single peak in the First Derivative curve is obtained Maximum First Derivative (MFD), Maximum First Derivative Spread –mean (MFDS-m) methods are used to find Red Edge Position for the study of crop health [6].

Hyperspectral reflectance provides measurements over large number of narrow and contiguous bands, in various regions of electromagnetic spectrum. However, high correlation and redundant neighboring wavebands makes further analysis challenging [7]. Therefore, to improve the classification speed and accuracy, selection of significant

wavelengths is used as a pre-requisite for eliminating curse of dimensionality in the applications of hyperspectral remote sensing. [8][9].

To discriminate between healthy and unhealthy crops, reflectance at specific wavelengths in Visible, Red Edge, NIR and SWIR regions are found useful. In ground based hyperspectral remote sensing, researchers have also used various non-REP methods to assess the crops health. Principal Component Analysis and Linear Discriminant analysis were used to find significant wavebands and hyperspectral vegetation indices to discriminate between the potato species and to assess disease, water and nutrient stress of potato crop [10]. To Discriminate Orchard Species, ANOVA and PCA and Random Forest classifier were applied one by one for reducing the number of wavelengths and obtained optimal discriminating wavelengths without losing significant wavelengths, Discriminant Analysis was performed to check spectral separability [11]. For the discrimination between healthy and infected oil palms, both spectral reflectance and first derivatives transform of the samples was recorded. Optimal spectral bands were selected using ANOVA and, classification was done using a maximum likelihood classifier [12].

Random Forest classifier is an ensemble machine learning method and has been widely used for Hyperspectral data. This tree based classifier has an ability of noise reduction and removing irrelevant features. Here we have also used two pre-filtering approaches PCA and LDA for reducing dimensions before the classifier is build. Performance of PCA and LDA was compared by classification accuracy measures.

## 2. Material and Methods

### 2.1 Study area and sample collection

Aurangabad city (Lat 19.846011 and Long 75.282556) is situated in Maharashtra, India. In-field hyperspectral spectral data was collected 27th August 2016 from a farm of 100 Acres, where different vegetables are grown. Healthy and infected plants of Brinjal, Cluster Beans and Long Beans were identified by visual judgment. Leaves without stalks were hand-picked and transferred in sealed polythene bags. Total 25 leaf samples of each were collected. Leaf spectra were measured within 3 hrs after collection.

### 2.2 Spectral Data Acquisition

#### 2.2.1 Leaf Spectra Measurements

The spectral reflectance of leaf sample was acquired with an ASD FieldSpecPro spectroradiometer, in the Multispectral Lab of Computer Science and Information Technology Department of Dr. Babasaheb Ambedkar Marathwada University, Aurangbad. This instrument acquires reflectance in 350-2500 nm spectral range and sampling step is 1 nm.

Spectral measurements were recorded in the dark room using a fiber optic with a 8° field of view. The leaf blade was kept 15cm below the sensor. The target was illuminated by 50 W halogen lamp. White reference scans were done for optimization of the signal and calibration of accuracy. The spectral data was collected in the range of 350 to 2500 nm, containing reflectance at 2151 wavelengths. Reflectance spectra of the samples was recorded with 10 iterations each using RS3 software [13].

Fig:1 shows spectral reflectance curves of healthy and infected leaves of Brinjal, Long Beans and Cluster Beans in the range of 350nm-2500nm.

#### 2.2.2 Spectral Data Preprocessing

The raw .asd files were further processed by ViewSpec Pro version 6.2. Splice Correction was done for removing noise and interpolating missing data. Reflectance curves were exported as text data files. These data files were transferred to Excel software and files were converted to .csv format. In spectral data processing, if we represent the dataset by matrix X, then each row of the matrix Xj is contains the intensity value corresponding each wavelength for ith sample. Each column Xi, is an intensity value of a wavelength for all samples.

### 2.3. Dimensionality Reduction and Classification

Hyperspetral data has a curse of dimensionality. A reflectance spectrum comprised 2151 wavelengths was available for experimentation, but not all were useful. Selection of significant wavelengths is a crucial task. This study aspires to obtain subset of significant wavelengths for discrimination between healthy and unhealthy vegetable leaves. For achieving this, dimensionality reduction process was carried out and most significant features were selected in the different regions of the electromagnetic spectrum. LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis) were applied to find discriminative wavelegths. Performance of these methods was evaluated using Random Forest supervised classifier on the basis of Accuracy and Kappa values.

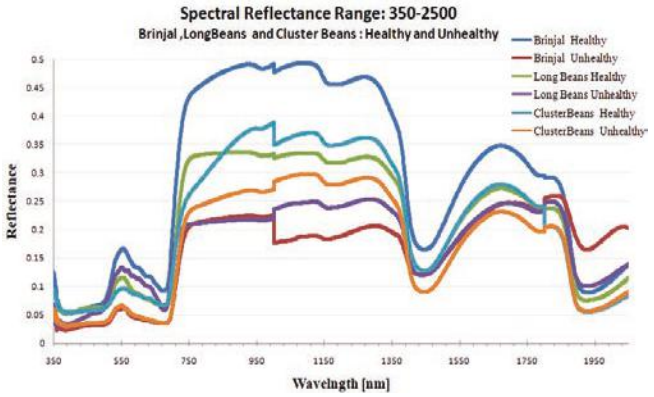


***Fig.1: Spectral Reflectance of healthy and unhealthy leaves of Brinjal, Long Beans and Cluster Beans in the range of 350 nm -2500nm. (Wavelength vs Reflectance)***

#### 2.3.1. Linear Discriminant Analysis (LDA)

LDA is a supervised dimensionality reduction method which is based on Fisher Criterion. This method aims to find a linear transformation by projecting high dimensional space onto a low-dimensional space, in which, it maximizes the Between-Class scatter and minimizes the Within-Class scatter. Two measures are defined for all samples in both the classes i) Sb: between class scatter matrix and ii) Sw : within-class scatter matrix.

$$S_b = \sum_{k=1}^{c} n_k (\mu_k - \mu)(\mu_k - \mu)^T \quad (1)$$

$$S_w = \sum_{k=1}^{c} \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \quad (2)$$

where $C_k$ is the index, $\mu_k$ is mean and $n_k$ is number of samples, in the $k^{th}$ class respectively.

$$\mu = \sum_{k=1}^{c} n_k \mu_k \quad (3)$$

is the overall mean of original data space[14][15].

#### 2.3.2. Principal Component Analysis (PCA)

PCA transforms original data set X, containing *p* column vectors into another dataset Y is having *d* column vectors, where d< <p. First Principal Component has the highest variance with original data set. The second PC, is orthogonal to the first PC and also has highest variance to the first PC. Likewise each PC is orthogonal to the previous one, this property eliminates the problem of colinearity. This variance is captured in first few PCs' which represents the original data set. Higher numbered PCs are ignored, as they contain very less information [16].

If $Y_1$ is the first single largest variance among all linear combination, it is called as first PC and calculates as

$$Y_1 = e_1^T X \quad (4)$$

$i^{th}$ PC is calculated as,

$$Y_i = e_i^T X \quad (5)$$

where $e_i$ is called $i^{th}$ loading vector and $Y_i$ is $i^{th}$ PC.

#### 2.3.3. Random Forest Classifier (RF)

Random forest creates set of multiple decision trees. It selects samples randomly and generates separate tree for each sample. Prediction result of each tree is used to provide vote for selecting best classification. Prediction results having majority of votes is selected. RF does not suffer from overfitting because of randomly selected samples and number of trees participating in voting by prediction result. Given a training set $X = x_1, ..., x_i$ with response variable $Y = y_1, ..., y_i$, discrimination function is defines as

$$H(x) = argmax_Y \sum_{i=1}^{k} (I(hi(X, \theta_k) = Y) \quad (6)$$

Where I.() is the indicator function, hi is $i^{th}$ single decision tree, Y is the class label and $argmax_Y$ is the Y value by maximizing

$$\sum_{i=1}^{k} (I(hi(X, \theta_k) = Y) \quad (7)$$

## 3. Experimentation and Results

As discussed in section 2.2 data set is pre-processed and following steps were carried out for further experimentation 1) Mean of multiple iterations of each sample was obtained 2) Reflectance spectra was divided into five distinct regions of electromagnetic spectrum and subset of wavelengths from different regions were selected viz. i) Visible : 350nm-670nm ii) Red Edge : 671nm-780nm iii) Near Infrared : 781nm-1000nm iv) Shortwave I: 1301nm-1500nm and v) Shortwave II: 1701 nm-1900 nm.

Further processing was done with Scikit-Learn library of python. In the first step, training and testing data was split. In the second step, constants and quasi-constant features were removed. In the third step, duplicate variables were removed. In the fourth step highly correlated features were detected and removed.

To reduce processing time and improve accuracy LDA and PCA were applied for dimensionality reduction, before applying Random Forest classifier. First five components of LDA and PCA are used for classification.

In case of Brinjal, when reflectance spectra of healthy leaves was visually compared with diseased spectra, decrease in the reflectance magnitude in the visible, Red Edge , NIR regions and increase in the reflectance of SWII regions was observed, where as not much difference was observed in the first water absorption SWI region. As shown in Table no 1, highest accuracy of .95 was obtained in the visible region in both cases. Here minimum classification accuracy was obtained in the NIR region.

**Table 1: Comparison of classification accuracy of Brinjal**

| Regions | Filter | Accuracy | Kappa |
|---|---|---|---|
| Visible | LDA | 0.95 | 0.89 |
| | PCA | 0.95 | 0.9 |
| Red Edge | LDA | 0.89 | 0.75 |
| | PCA | 0.9 | 0.76 |
| N IR | LDA | 0.61 | 0.11 |
| | PCA | 0.65 | 0.14 |
| SW I | LDA | 0.89 | 0.75 |
| | PCA | 0.88 | 0.75 |
| SW II | LDA | 0.89 | 0.75 |
| | PCA | 0.94 | 0.75 |

In Spectral reflectance of Cluster Beans, it is seen that reflectance magnitude of visible, RE and NIR region was decreased, whereas increase in reflectance of water absorption SW1 and SW2 region. But in some leaves difference in the

NIR region of some diseased leaves samples was observed. Table 2 displays, classification results. Accuracy of .97 was obtained in the visible region using both PCA and LDA, but in RE and NIR regions classification accuracy of PCA was greater than LDA components, which was .96 and .95 respectively. Average accuracy of classification of Cluster Beans, considering five regions was .90 for LDA and .92 for PCA.

**Table 2: Comparison of classification accuracy of Cluster Beans**

| Regions | Filter | Accuracy | Kappa |
|---|---|---|---|
| Visible | LDA | 0.97 | 0.92 |
| | PCA | 0.97 | 0.93 |
| Red Edge | LDA | 0.91 | 0.8 |
| | PCA | 0.96 | 0.92 |
| N IR | LDA | 0.94 | 0.88 |
| | PCA | 0.95 | 0.9 |
| SW I | LDA | 0.87 | 0.72 |
| | PCA | 0.87 | 0.72 |
| SW II | LDA | 0.84 | 0.64 |
| | PCA | 0.86 | 0.7 |

On visual judgment of healthy and diseased Long Beans leaf spectras, it is observed that there increase in reflectance magnitude of Visible, decrease in RE and NIR regions of diseased leaves, where was no difference in the SW1 and SW2 regions. As mentioned results in Table 3, maximum classification accuracy of 1 was obtained in RE region for PC components, average accuracy was .88 and .90 respectively for LDA and PCA.

Table 4 shows the mean classification accuracy obtained by RF classifier after PCA and LDA filter methods.

**Table 3: Comparison of classification accuracy of Long Beans**

| Regions | Filter | Accuracy | Kappa |
|---|---|---|---|
| Visible | LDA | 0.94 | 0.87 |
| | PCA | 0.94 | 0.87 |
| Red Edge | LDA | 0.95 | 0.89 |
| | PCA | 1 | 1 |
| N IR | LDA | 0.77 | 0.4 |
| | PCA | 0.8 | 0.49 |
| SW I | LDA | 0.89 | 0.75 |
| | PCA | 0.9 | 0.78 |
| SW II | LDA | 0.89 | 0.75 |
| | PCA | 0.89 | 0.75 |

**Table 4: Comparison of Classification performance (mean values of accuracy and Kappa)**

| Regions | Filter | Accuracy | Kappa |
|---|---|---|---|
| Brinjal | LDA | 0.846 | 0.65 |
| | PCA | 0.864 | 0.66 |
| Cluster Beans | LDA | 0.9 | 0.79 |
| | PCA | 0.92 | 0.83 |
| Long Beans | LDA | 0.89 | 0.73 |
| | PCA | 0.9 | 0.78 |

As shown in the Fig.2 , Mean accuracy and mean Kappa values of PCA filtering are better compared to LDA

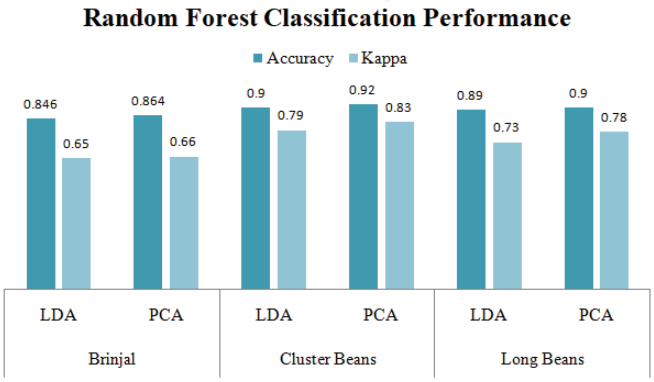


***Fig. 2: Random Forest Classification Performance***

## 4. Discussion and Conclusion

Hyperspectral Remote sensing provides large number of features in the form of wavelengths. Number of wavelengths was greater than sample size. Selection of significant wavelengths for a particular application is a critical job. This experiment was done for narrowing down to a small region and selecting only few features useful for further classification of healthy and unhealthy vegetable leaves. Two filter methods LDA and PCA were applied and Random Forest classification was performed on five different regions.

Experimental results have shown that, for distinguishing healthy and unhealthy vegetable leaves, we can achieve good results by focusing on Visible and Red Edge regions of the spectral signature. Performance of PCA and LDA was similar in some cases, but overall accuracy of Random Forest classification was better with PCA.

## 5. Acknowledgement

Department of Science and Technology - Funds for Infrastructure under Science and Technology (DST-FIST) has supported the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India , under section No. SR/ FST/ ET1340. This research work has been done under the same support. Authors wish to extend sincere gratitude towards the Department and University authorities for providing required resources.

## References:


1] Prasad S. Thenkabail, M. K. (2014). Hyperspectral Hyperion Images and Spectral Libraries of Agricultural Crops,Hyperspectral Remote Sensing of Vegetation and Agricultural Crops. *Photogrammetric Engineering & Remote Sensing , 80* (8), 694-723.12.

2] Prasad S. Thenkabail, M.I.(2013). Selection of hyperspectral narrowbands (HNBs) and composition of hyperspectral twoband vegetation indices (HVIs) for biophysical characterization and discrimination of crop types using field reflectance and Hyperion/ EO-1 data. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing,6* (2) , 427-439.

3] R. N. Sahoo, S. S. (2015). Hyperspectral remote sensing of agriculture. *CURRENT SCIENCE,Special Section Hyperspectral Remote Sensing , 108* (5), 848-859.

4] Christoph Römer, M. W. (2012). Early Drought Stress Detection in Cereals: Simplex Volume Maximisation for Hyperspectral Image Analysis. *Functional Plant Biology , 39* (11), 878–890.

5] Dawson, T. &. (1998). A new technique for interpolating the reflectance red edge position. *International Journal of Remote Sensing , 19* (11), 2133-2139.

6] Anjana Ghule, P. D. (2019). MFDS-m Red Edge Position Detection Algorithm for Discrimination between Healthy and Unhealthy Vegetable Plants. *International Conference on Recent Trends in Image Processing & Pattern Recognition*, (pp. 372-379).

7] Mozhgan Abbasi, ,. J. (2019). Optimal SpectralWavelengths for Discriminating Orchard Species Using Multivariate Statistical Techniques. *Remote Sensing , 12* (63), 1-16.

8] Prospere K., M. K. (2014). Plant Species Discrimination in a Tropical Wetland Using In Situ Hyperspectral Data. *Remote Sensing , 6* (9), 8494-8523.

9] Mirzaei M., M. S. (2019). Scenario-based discrimination of common grapevine varieties using in-field hyperspectral data in the western of Iran. *International Journal of Applied Earth Observation and Geoinformation , 80*, 26–37.

10] Shibendu Shankar Ray, J. P. (2010). Use of Hyperspectral Remote Sensing Data for Crop Stress Detection: Ground-Based Studies. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science , 38* (8), 562-570.

11] Junshi Xia, P. D. (2014). Hyperspectral Remote Sensing Image Classification Based on Rotation Forest. *IEEE Geoscience And Remote Sensing Letters , 11* (1), 239-243.

12] Helmi Z. M. Shafri, M. I. (2011). Spectral Discrimination of Healthy and Ganoderma-Infected Oil Palms from Hyperspectral Data. *International Journal of Remote Sensing , 32* (22), 7111–7129.

*13] http://www.asdi.com/applications/remote-sensing/spectral-remote-sensing.* (n.d.). Retrieved Janurary 21, 2016, from http://www.asdi.com/applications/remote-sensing/spectral- remote-sensing.

14] Aleix M. Martinez, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 23* (2), 228-233.

15] PU, R. (2009). Broadleaf species recognition with in situ hyperspectral data. *International Journal of Remote Sensing , 30* (11), 2759 — 2779.

16] Tom Howley a, M. G.-L. (2005). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 363–370). Cambridge: Proceedings of AI. ■


## About the Authors

**Mrs. Anjana N Ghule** (CSI LM. - 95255) M.E. Computer Science and Engineering from Dr. B. A. M University Aurangabad (MS) India. Currently working as an Assistant Professor, Department of Information Technology, Government College of Engineering, Aurangabad, (MS), India. Life member CSI, ISTE and IETE. EC member, IETE Aurangabad Centre. Published research papers in International conferences and reputed journals. Her area of research is Remote Sensing. She can be reached at anghule@geca.ac.in.

**Prof. Dr. R. R. Deshmuk** (CSI LM - 00100518 ) M.E., M.Sc. (CSE) Ph.D. FIETE, PEIN Fellow, Working as Professor, Department of CSIT, Dr. B.A.M. University, Aurangabad, (MS), India. Coordinator of DST-FIST program, University Coordinator, MHRD GIAN program, Chairman, IETE Aurangabad Centre 2014–2018, Organized Zonal ISF-2016, Sectional President, ICT Section ISCA-2019, Coordinator, महाराष्ट्र राज्य मराठी विश्वकोश निर्मिती मंडळ, महाराष्ट्र शासन, Life member ISCA, CSI, ISTE, IEEE, IAEng, CSTA, IDES, ACEE. Management Council, Senate, Academic Council Member at University, Edited Twelve books, Published more than 225 research papers in reputed Journals, Editor-In-Chief, CSI Journal of Computing. Visited Russia, USA, China, Spain, Philippines, Uzbekistan, Thailand for academic work. He can be reached at rrdeshmukh.csit@bamu.ac.in

**Mrs. Chitra M. Gaikwad** [CSI-1191698] -Assistant Professor, Department of Information Technology, Government College of Engineering, Aurangabad, Maharashtra, India. She has 8 paper publications in International Journals, 3 in International Conference and 1 in National Conference. Her research interests include Image Processing and Remote Sensing. She can be reached at cmgaikwad@geca.ac.in