The Basis for Development of a Foundational Biomarker Reflectance Signature Database System for Plant Cell Identification, Disease Detection, and Classification Purposes

Jeanette Hariharan
Bioengineering Dept.
Florida Gulf Coast University
Ft. Myers, FL, USA
jhariharan@fgcu.edu

Yiannis Ampatzidis
Agricultural & Bilogical Eng.
Southwest Florida Research and
Edcuation Center
Immokalee, FL, USA
i.ampatzidis@ufl.edu

Jaafar Abdulridha
Agricultural & Bilogical Eng.
Southwest Florida Research and
Edcuation Center
Immokalee, FL, USA
ftash@ufl.edu

Abstract— **The objective of this paper is a novel interpretation of the spectral and imaging data analysis process which takes into account the measurement of the variance caused by disease infestationof a cell. Using multivariate analysis,theKarhounen-LoeveExpansion(KLE) of hyperspectral reflectance data, taken from healthy and diseased states of several plant species, is used to identify a basis setof functions which represent the distribution of reflected signal energy.By spectral decomposition, the eigenvalues are related to the KLEbasis set. The eigenvalues can be used to identify the KLE eigenvectorswhich comprise the highest variation in the data. These componentscan be interpreted as the weighted variables which carry with themmost of the information on the reflectance spectrum of the cell. Fromindications presented by this multivariate KLE analysis, a frequencyreconstruction is adapted to convert the eigenvector information to awave function. This reconstruction via KLE and frequency transformation forms the signature identification process fordeveloping a database of healthy cell reflectance pattern features andvariations produced by disease or other factors. These frequencyspectra can be used as average signature reflectance patterns for cellidentification, classification and biomarkers for diseases. The defining of these spectral identification biomarkers or signatures, is purposefulsince it could lead to less invasive techniques for classification and disease diagnostics. The techniques used to determine thesereflectance spectra require a unique and rarely used transformationmethod. These processes need further testing and verification throughmultiple refinements of this procedure.**

Keywords—Multivariate analysis, hyperspectral,spectra, plant disease, KLE, biomarker

I. Introduction

Many diseases associated with plants are being detected using advanced and sophisticated hyperspectral data analysis approaches [1-12]. Several common analysis techniques include normalization and derivative spectra enhancement using various methods, such as finite differencing [3], complex step derivative [8-9], and derivative spectral shape equation [5]. Wavelet Transform has been used and compared to derivative spectra enhancement, and shown to be very successful in spectral regions of interest, and is becoming more commonly used as an alternative to spectral derivative methods [12]. Interpolating polynomials are used to smooth the data and better represent enhanced spectra [10-11]. Multivariate analysis can be used to gain a better understanding of spectral variance between diseased and healthy reflectance properties [2,5,6].

In this study we are examining two healthy plant species, two diseases associated with these plants and two abiotic stressors that mimic disease symptom properties in reflectance spectrum. The Sugar Belle tangerine and the canker diseased states of this plant at various stages of the disease infestation is the first plant reflectance that are studied. The second is the avocado plant and the Laurel wilt (Lw) disease infection of this plant as well as iron (Fe) and nitrogen (N) deficiencies in the avocado that present similar symptoms as the Lw. We chose to study these two groups of plants first, since these two plants are the highest yield crops in Florida. Also, the hyperspectral data for these two crops was readily obtained. Healthy and Lw infected avocado data was collected through hyperspectral camera data collection techniques. Lw is an almost always lethal disease in the avocado, caused by infestation of Asian fungus Raffaelea lauricola delivered to the plant via injection by the redbay ambrosia beetle through its feeding apparatus tissue, mycangia [13]. Other avocado leaves that were abiologically stressed and suffered mineral deficiencies (iron (Fe), and nitrogen(N)). Both deficient groups were included in the test sample data.

Citrus bacterial canker (CBC), caused by *Xanthomonas citri subsp. citri (Xcc; syn. X. axonopodis pv. citri),* is a serious disease of citrus worldwide [15]. The Florida Sugarbelle citrus crop has suffered serious consequences over the last decade and methods to detect the disease early is critical. Symptoms include necrotic, raised lesions with yellow halos leaves, and twigs [17]. The bacterium is dispersed by wind and rain and prefers humid-wet climates [15]. On severely infected trees, the pathogen can

978-1-7281-3783-4/20/$31.00 ©2020 IEEE

cause severe premature leaf and fruit drop, twig dieback, blemished fruit, and tree decline, resulting in significant economic impacts [17]. Visually, a plant may look healthy, but in fact, the bacterial growth stages take a few months to show symptoms. Late symptoms of this disease may appear within only a few months from the infection.

Hyperspectral data from infected and healthy plants were collected and verified for registration of data. Radiance to reflection conversion was done. Normalized data was used for all calculations. Preprocessing methods that were used are described in section II. Smoothing of the data by either a median filter or other higher order filter such as the Savitzky-Golay filter are often performed on the data for spectral feature enhancement and denoising of the data [1-3]. In our case, an interpolating polynomial was used to better fit the data and allow for analytical derivatives to be taken.

Using multivariate analysis, we are able to prove that there exist specific variance patterns in a diseased verses healthy cell, especially over a mean statistical population (ref. Figs. 3,6). Taking the eigenvector components and establishing a basis vector for classification of core healthy species and for the disease factor species, fundamental waveforms, or basis signatures were calculated to help identify disease and establish baseline biomarkers for two different diseases in two different species of plants.

II. Study Site and methods

A. Plant and Sample Select

To induce Lw, ten plants were randomly selected and inoculated with Raffaelea lauricola. Four holes were drilled into different sides of each plant's stem with a 7/64″ drill bit 5 cm above the graft union, and 25 μL of inoculum prepared with a hem cytometer at a concentration of 30,000 CFUs/mL was pipetted into each hole and wrapped with parafilm to seal the wound. Four leaves per plant were sampled from each of 10H and the inoculated plants

Tangerine Sugar Belle leaves infected with canker disease and healthy leaves and fruits were collected from an experimental orchard at the University of Florida's Southwest Florida Research and Education Center (SWFREC), Immokalee, Florida, USA, for laboratory assessment on October 2018. Four trees were selected, and 10 leaves were collected from each tree in different disease severity stages including (i) asymptomatic stage (leaves without visible symptom); (ii) early stage (symptoms appear as slightly raised, small, blister-like chlorotic lesions); and (iii) late stage (lesions turn tan and then brown, and the edges appear water-soaked and develop a yellow halo). The UAV data was collected in the same field (October 2018, between 10 a.m. to 2 p.m.

B. Spectral Data Collection

1) Spectral data collection of avocado

Five scans per leaf (five locations per leaf) were taken for each healthy, Lw, and deficient leaf samples. These samples were collected between 350 and 2500 nm utilizing a spectrometer (SVC HR-1024, Spectra Vista Cooperation, NY) with 1.3- nm average spectral resolution and 4° fields of view in laboratory conditions. For data analysis, only the spectral range of 400–970 nm was used. Two halogen light sources were used to create optimal conditions for performing the scans and reducing errors. The SVC device was situated so that the lens was 50 cm above the sample pointing down at it. Spectral signatures were calibrated with a barium sulphate standard reflectance panel (Spectral Reflectance Target, CSTM-SRT-99-100, Spectra Vista Cooperation, NY) before and immediately after every 8 samples measurements. Black fabric was utilized as background

2) Outdoor Hyperspectral Data Collection

Hyperspectral data was collected by using a UAV (DJI Matrice 600, Pro Hexacopter) and the same hyperspectral camera, Resonon Pika L 2.4.The UAV-based imaging system includes (i) a Resonon Pika L 2.4 hyperspectral camera (Spectronon Pro, Resonon, Bozeman, MT); (ii) visible-near infrared (V-NIR) objective lenses for the Pika L camera with a focal length of 23 mm, field of view (FOV) of 13.1 degrees, and instantaneous field of view (IFOV) of 0.52 mrad; and (iii) a global positioning system (GPS) and the inertial measurement unit IMU (DJI) flight control system for multi-rotor aircraft, to record sensor position and orientation. Data was collected at 30 m above the ground with a speed of 1.5s/h. The positions of the infected trees were known (leaves were collected and identified by PCR). The maps and images were analyzed by the Spectronon software after hyperspectral data were acquired. Calibration corrections were performed using Resonon hyperspectral data analysis software (Spectronon Pro, Resonon, Bozeman, MT). Georectification and radiometric correction plugins, from the Spectronon Pro software, were used to correct the GPS/IMU and the radiometric data, respectively. The regions of interest were selected manually (and randomly) for each tree, and 20 spectral scans were performed to ensure that the entire canopy was covered spectrally. The regions of interest were then exported as a text file. Pixel-based reflectance data was mixed for each class.

C. Higher Order Spectra Enhancement

1) Normalization Data Analysis

The first step in the data enhancement process was performing a Standard Normal transformation of the reflectance data, x. The Standard Normal transformation was used to provide preservation of data integrity, and restructure the data into a reasonable population domain. The Standard Normal transformation used is given by its probability density function of (1):

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\zeta} \quad (1)$$

Where

$$\zeta = \frac{(X-\mu)^2}{2\sigma} \quad (2)$$

Represents the normal distribution parameter; μ represents the mean of the population and σ represents the standard deviation.

2) Higher-Order Spectral Analysis

Divided differences will reveal with higher resolution where inflection points, local minima, and maxima occur in healthy data samples, and how these points vary with the various categories of disease/deficient data sets that are tested. These variants prove to be unique to the signature attributes and are distinguished in the multivariate analysis and characteristic polynomial fit process later developed in this work. Considering how these higher-order data functions correlate at regions of interest in the spectrum allow for the categorization of plants into their respective states (i.e., healthy, deficient, diseased). Methods for second order forward differencing were applied previously [17] for enhancing spectra for the avocado data. In order to reduce the effects of numerical differentiation noise sensitivity, an interpolating polynomial is estimated prior to numerical differentiation. In the case for the citrus data, a five point centered difference formula using Stirling's formula of divided differences was used to approximate the third order polynomial associated with each state of the hyperspectral data. This was done to increase the degree of error to $O(\xi^4)$, where ξ is a reflectance value between steps, $x_0 - h < \xi < x_0 + h$ The third order polynomial is then approximated by

$$P_3(x) = f[x_0] + \frac{sh}{2}(f[x_{-1}, x_0] + f[x_0, x_1]) + s^2h^2 f[x_{-1}, x_0, x_1] + \frac{s(s^2-1)h^3}{2}(f[x_{-1}, x_0, x_1, x_2] + f[x_{-2}, x_{-1}, x_0, x_1]) \quad (3)$$

With the general central- difference formula being given by:

$$f[x_{n-k}, \dots\ x_{n,\dots}\ x_{n+k}] = \frac{1}{2k!h^k}\delta^k f(x_n) \quad (4)$$

for step size h,

$$h = x_{i+1} - x_i \quad (5)$$

for each $i = 0,1,\dots,n-1$ and $x = x_0 + sh,$ so that the difference $x - x_i = (s - i)h.$ The average step size over all the computations is 2.1854 nm.

3) Analytical Derivative Analysis

The interpolating polynomial provided the necessary smoothing of the data such that it was straight forward to apply a derivative formula to the interpolated data. A Newton's second order divided difference approximation was used to provide accuracy for not only the interior points, but also for the endpoints. The interior points were found by differentiating the three point formula and setting $x_0 = x - h, x_1 = x, and\ x_2 = x + h$. To achieve $O(h^2)$ approximations for the second derivatives at the endpoints, a four point formula was used. These methods are derived in the literature [19] and are given for our calculations in Table 1.

NEWTON'S SECONDD ORDER DIVIDED DIFFERENCES' FORMULAS FOR P_3''

TABLE I. NEWTON'S THIRD ORDER DIVIDED DIFFERENCES' FORMULAS FOR P_3''

Section of Polynomial	**NEWTON'S SECOND ORDER DIVIDED DIFFERENCES' FORMULAS**	
	P_3''	*Error*
Interior points	$\frac{1}{h^2}(P_3(x+h) - 2P_3(x) + P_3(x-h))$	$O(h^2)$
Endpoint (left)	$\frac{1}{h^2}(2P_3(x) - 5P_3(x+h) + 4P_3(x+2h) - P_3(x+3h))$	$O(h^2)$
Endpoint (right)	$\frac{1}{h^2}(2P_3(x) - 5P_3(x-h) + 4P_3(x-2h) - P_3(x-3h))$	$O(h^2)$

4) Multi-Variate Analysis

The formal process of defining spectrums for healthy plants, the detection of diseased species by stochastic analysis based on these spectrums, and classification established through variation of defined signature of healthy verses diseased plant specimens is the premise for this paper. By using multivariate analysis with K-means clustering and applying an optimal orthogonal basis vector for classification, we are able to distinguish between infected and healthy citrus plants. The multivariate approach uses several spectral bands, with the X-variate matrix based on twenty wavebands. The cross-covariance matrix is derived through variance (σ_i^2) and covariance (σ_{ij}) of the X-variate matrix. The eigenvectors are used to distinguish the modal components associated with greatest variance between infected and healthy plant species. We obtain the X-variate matrix, for leaf reflectance at varying wavelengths:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} \quad (6)$$

We can represent this matrix in vector form as X = [x_1 , x_2 ,… x_n]. The data for the reflection coefficients are given by the matrix elements and represent data in the wavebands from 488-569 nm. The cross-covariance matrix is then obtained from cov$(x_i, x_j) = E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)^T] = \sigma_{x_i x_j}^2$

$$C = \begin{bmatrix} cov(x_1, x_1 & \cdots & cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ cov(x_n, x_1) & \cdots & cov(x_n, x_n) \end{bmatrix} = \begin{bmatrix} \sigma_{x_1}^2 & \cdots & \sigma_{x_1}\sigma_{x_n} \\ \vdots & \ddots & \vdots \\ \sigma_{x_n}\sigma_{x_1} & \cdots & \sigma_{x_n}^2 \end{bmatrix} \quad (7)$$

Where σ is the covariance of the x independent variates.

The KLE is an optimal transformation along all orthogonal component vectors. The KLE process determines a formal decorrelation of signal energy into a redistribution that weighs more heavily the components of highest energy contribution for a particular system. Therefore, the KLE realizes the lowest order model for L that adequately describes the main functional contributors to the equilibrium of a system, in our case, a plant reflectance spectrum which relates to its cellular dynamics.

A KLE estimate of the covariance matrix can be approximated and applied as described by:

$$Y = \varepsilon_{i1}x_1 + \varepsilon_{i2}x_2 + \ldots \varepsilon_{in}x_n \qquad (8)$$

Then by Spectral Decomposition Theorem,

$$C\boldsymbol{\varepsilon_i} = \lambda_i \boldsymbol{\varepsilon_i} \qquad (9)$$

The λ_i are the eigenvalues which correspond to the eigenvectors of the cross-covariance matrix.
The variance for the Y_i component of the linear transformation model is equal to the variance of the corresponding eigenvector equation, which is equal to the eigenvalue:

$$var(Y_i) = var(e_{i1}x_1 + e_{i2}x_2 + \cdots e_{in}x_n) = \lambda_i \quad (10)$$

An approximation matrix, Z, can be formulated in accordance with the explained percentage of the variance given by:

$$L = \frac{\sum_{i=1}^{\ell}\lambda_i}{\sum_{j=1}^{n}\lambda_j} \qquad (10)$$

Where $\ell = number\ of\ primary\ modal\ components < n$
The denominator of L is also known as the trace and equals the sum of the eigenvalues. The goal is to maximize L with the least amount of modal components of the numerator.
ℓ represents the optimal number of eigenvalues necessary to accurately approximate the signal energy spectrum.
Our approximation matrix then becomes:

$$\boldsymbol{\hat{Z}} = \sum_{i=1}^{\ell}\boldsymbol{\varepsilon_i Y} = \sum_{i=1}^{\ell}\boldsymbol{\varepsilon_i \varepsilon_i^T X} = I_l\boldsymbol{X} \qquad (11)$$

The $\boldsymbol{\hat{Z}}$ matrix is a reconstructed KLE approximation of the original multivariate data and contains only the pertinent feature components of the data minus the extraneous noise and other factors that account for less than (100-L)% of the data. The eigenvectors associated with this matrix are fundamentally important to the signature extraction process, which is developed in section II-D.

D. Signature Extraction Methodology

An approximation for the plant cell reflectance can be mathematically modeled by applying basic signal processing techniques from hyperspectral data, or other sensing devices. Signatures of average reflectance for specific plants can be found as well as variations caused by disease and other environmental factors. The novelty that we address in this paper is that of data significance and data reduction, to provide a feature extraction analysis that most accurately represents the frequency spectrum of a plant, and variations of that signature, caused by disease, and/or other possible stressors. Large sequences that represent a signal usually contain aspects of the signal that are irrelevant, noise-injected and erroneous. Methods to eliminate the singularities and irrelevant discriminant nodes are undertaken in a lean operation after Fourier decompositon is applied. The time domain signal is extracted and reduced series frequency and phase representation (signature) is verified as a feature identification method to be used for classification and diagnostic purposes.

1) Frequency Analysis

The variate data is given per wavelength in the spatial domain. A translation to frequency domain was carried out to better understand the frequency component aspect of the data. From this translation, the frequency decomposition was carried out. For specific Regions of Interest (ROIs), the Fourier coefficients were obtained via inverse Fourier transform per derivation given below in equations (12) -(16). Several ROIs were selected where the second order derivative critical points were of significance. The focus is to evaluate these regions, develop the Fourier spectrum representation (i.e. signature) within these regions, and use these mapping as proof of concept for the total spectral signature representation of the specific cell reflectance spectra.
Fig. 1 shows an example of a normalized frequency graph of the Sugarbelle tangerine.
A Fourier Transform of the data is carried out by sampling N =4096 data points in the region of normalized frequency 0-2π radians/sample. A truncated series expansion can then be written using the method of Fourier transform in this case:

$$x(t) = \sum_{n=0}^{N} a_n e^{2\pi j\ nt} \qquad (12)$$

$$x(t_m) = \sum_{n=0}^{N} a_n e^{2\pi j f_n t_m} \qquad (13)$$

For

$$m = 0, \frac{1}{f_s}, \frac{2}{f_s}, \ldots \frac{(N-1)}{f_s}$$

Where f_0 = Lowest frequency , $f_s = sampling\ rate$,
$f_n = nf_0$, for n=1,2 … N
t_m = discrete-time steps, a_n = Fourier coefficients

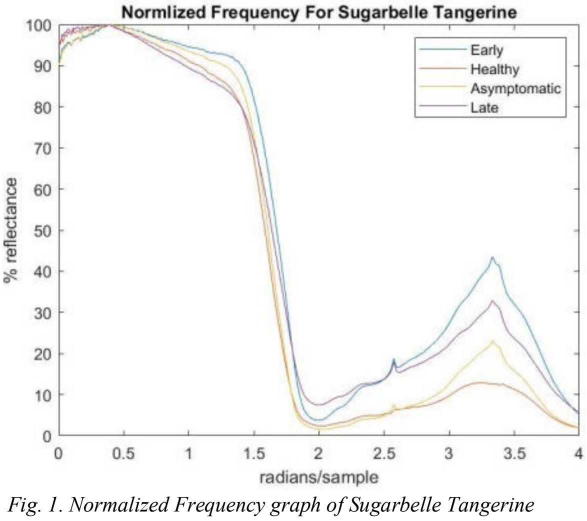

Fig. 1. Normalized Frequency graph of Sugarbelle Tangerine

Then define $\mathrm{x_n}$, for the sampled frequency data to be:

$$(x_{\mathrm{n}}) = \sum_{n=-N}^{N} a_n A_{nm} \qquad (14)$$

$$\hat{\mathrm{x}} = \hat{\mathrm{a}}\boldsymbol{A} \qquad (15)$$

Where **A** = sampled frequency data amplitudes
$\hat{\mathrm{x}}$ = sampled frequency component vector of $\mathrm{x_n}$values
$\hat{\mathrm{a}}$ = Fourier coefficients vector for (12) and (13)
Solving for the coefficient vector:

$$\hat{a} = \boldsymbol{A}^{-1}\hat{x} \qquad (16)$$

The method can be applied for various expansion sets. Using wavelet packet transform, a similar nonlinear scalar wave function can be applied to form a complete orthonormal basis system of functions. We conjecture that other orthogonal basis sets can also be formulated and verified, with similar efficiency. The next step in this design process is the projection of these higher dimensional functions onto a lower dimensional space by truncating the series with optimal preservation of signal signature information.

2) KLE Eigenvalues Association

Having found a way to estimate the frequency spectrum of the plant reflectance signature, it is desirable to accurately extract the feature frequencies that best describe the signature. We consider how the eigenvalues from the KLE effects the dimensional reduction in the Fourier domain. The greatest eigenvalue (by spectral decomposition) is significant in the sense that it relates to the component of greatest energy level of the signal sequence, whether that sequence be developed by a Taylor series or transcendental function expansion. Relating this to our estimate of the cross-covariance matrix (6):

$$x(t) = a_{\lambda_0} e^{2\pi j f_{\lambda_0} t+\varphi} + a_{\lambda_1} e^{2\pi j \lambda_1 t+\varphi} + a_{\lambda_2} e^{2\pi j f_{\lambda_2} t+\varphi} + \dots a_{\lambda_l} e^{2\pi j f_{\lambda_\ell} t+\varphi} \qquad (17)$$

Since we are only interested in the real, positive frequencies of the spectrum, (17) can be rewritten as:

$$x(t) = \mathbb{R}\{\textstyle\sum_{i=0}^{\ell} a_{\lambda_i} e^{2\pi j f_{\lambda_i} t+\varphi}\} \qquad (18)$$

$$x(t) = \textstyle\sum_{i=0}^{\ell} a_{\lambda_i} \cos(2\pi f_{\lambda_i} t + \varphi) \qquad (19)$$

Where

$f_{\lambda_1}, f_{\lambda_2} \dots f_{\lambda_\ell}$ *are the energy frequencies corresponding with the maximum to* ℓ *absolute value eignenvalues of the cross covariance matrix*
$a_{\lambda_0}, a_{\lambda_1}, \dots a_{\lambda_\ell}$
represent the Fourier coefficients at the related energy frequencies
And φ = phase angle

To find the corresponding truncated frequency series, using inverse transformation of (19)

$$X(f) = \mathbb{R}\left\{\tfrac{1}{l}\textstyle\sum_{i=0}^{\ell} a_{\lambda_i} e^{-2\pi j f_{\lambda_i} t+\varphi}\right\} = \tfrac{1}{\ell}\textstyle\sum_{i=0}^{\ell} a_{\lambda_i} \cos(-2\pi f_{\lambda_i} t + \varphi) \qquad (20)$$

$$X(f) = \tfrac{1}{\ell}\textstyle\sum_{i=0}^{\ell} a_{\lambda_i} \cos(2\pi f_{\lambda_i} t + \varphi) \qquad (21)$$

These energy frequencies are associated with the eigenvalues of highest to lowest magnitude ($f_{\lambda_1}, f_{\lambda_2} \dots f_{\lambda_\ell}$) of the cross-covariance matrix as found by the linear regression approximation found in (10). Projecting these onto the Fourier domain space, the frequency components associated with these higher ranked eigenvalue variables (frequencies, in our case), are mapped onto the frequency domain. The sequence is then truncated such that the frequencies associated with the highest signal energy are used to approximate the truncated series.

III. Results

A. Multivariate Analysis of Citrus Canker

In the case of the lab data for the population of citrus leaves, a KLE process was performed and the first two modal components were found to carry ~96% of the variance of the linear regression system model. A K-means clustering algorithm was used to distinguish healthy from infected citrus canker. An optimal orthonormal basis vector was also used in this classification model. Out of over 300 leaves tested, less than one percent were categorized incorrectly. This is shown in Fig. 3 below.

This combination of preprocessing, multi-variate analysis for data reduction, K-Means clustering with applications of shortest Euclidean distance formula and application of the orthonormal basis vector for verification and correction, have proved to be accurate for classification over 99.98% of the population.

To derive the optimized spectrum, the data for each averaged spectra (healthy, canker) was transformed to the frequency domain and sampled at 651.9 rad/sample in normalized frequency space. The time domain spectra are shown in Fig. 4.

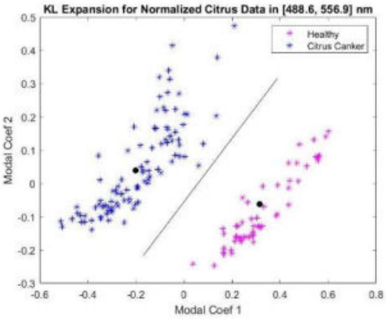

Fig. 3. KLE with K-Means clustering and Orthonormal Basis·Vector

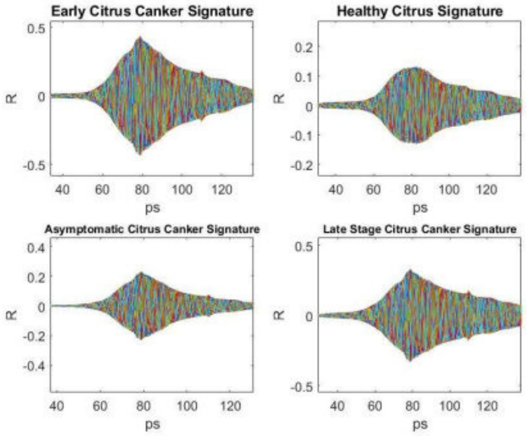

Fig. 4. Time Domain Signature for healthy and citrus canker (various stages)

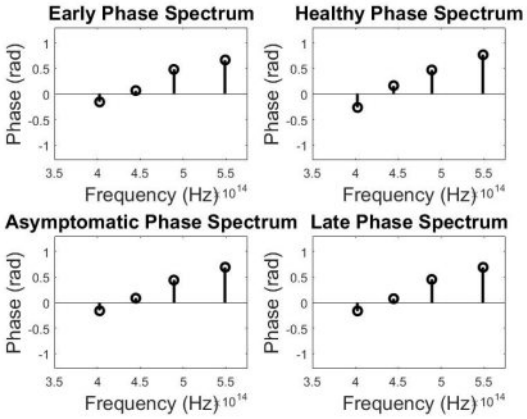

Fig. 5. Reduced Phase Domain Signature for healthy and citrus canker (various stages) after KLE reduction technique

It can be noted that the morphological effects of the disease change the spectrum in the 40-120 ps region in particular. The envelope becomes peaky as soon as disease is present in the cell. Applying the KLE, at the ROI 402-548 THz, and producing the phase spectrum for ℓ=4, the graphs of the phase magnitude are shown in Fig. 5. The equations for the signatures is found by using the first four highest eigenvalue frequencies (normalized) which are given at 3.342, 1.423, 2.558 and 1.967 related to the covariance matrix approximation. They are given here in (22) -(24), representing healthy, early-stage, and asymptomatic respectively.

$$X(f_H) = 10^{-2}\{3.32\cos(6.684\pi t + .767) + .82\cos(4.47\pi t - .268) + 11.72\cos(8.04\pi t + .469) + 1.72\cos(6.18\pi t + .1615)\} \qquad (22)$$

$$X(f_{\mathrm{E}}) = 10^{-2}\{3.37\cos(6.684\pi t + .6657) + .83\cos(4.47\pi t - .1583) + 10.71\cos(8.04\pi t + .4797) + 1.99\cos(6.18\pi t + .0654)\} \qquad (23)$$

$$X(f_A) = 10^{-2}\{3.43\cos(6.684\pi t + .692) + .83\cos(4.47\pi t - .1633) + 11.79\cos(8.04\pi t + .4382) + 1.71\cos(6.18\pi t + .0835)\} \qquad (24)$$

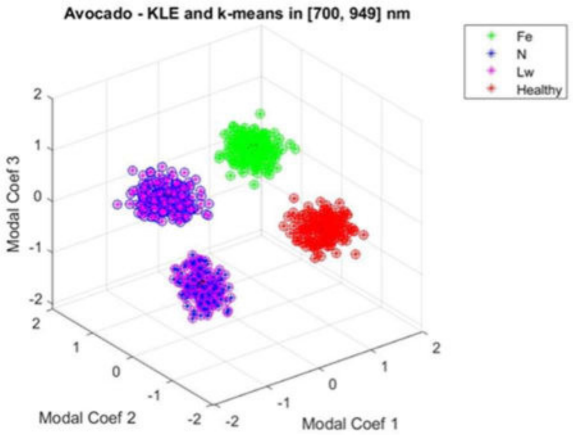

Fig. 6. KLE with K-Means clustering for third order Avocado data for Fe-deficient, N-deficient, Healthy and Laurel Wilt

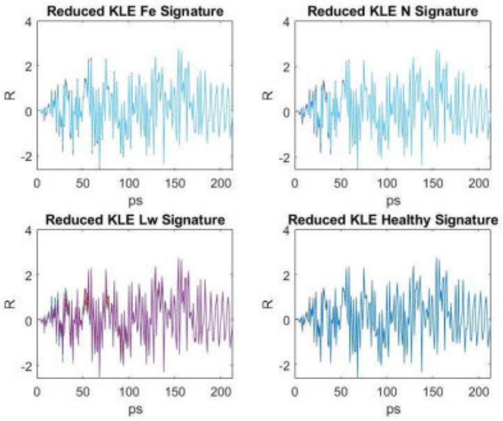

Fig. 7. Reduced Time Domain Signature for healthy and Lw, N, Fe deficiencies (KLE reduction technique)

B. *Multivariate Analysis of Avocado Laurel wilt, Fe-deficiencies, N-deficiencies*

The results for the avocado data set in distinguishing Laurel Wilt, Fe deficient, N deficient from healthy are shown in Fig. 6 and 7. The results were 100% accurate in distinguishing various stages of the Lw disease and deficient avocados from healthy avocados. Using multivariate analysis with the Euclidean distance formula, with respect to the centroids of each group, we were able to distinguish variance in the data presented in the KLE third dimensional space. Three dimensions were used to represent the variance described by the first three principal components, or 47.9926% of the data. Adding up to seven modal components only increased the representative variance to 62.41% (i.e. total variance ~76.9%). Using the first three eigenvalues we were able to establish 100% classification accuracy using multivariate analysis with applications of the K nearest neighbor (KNN) distance formula.

Others [6,11] used neural networks to optimize the linear regression coefficients or eigenvectors of the system which can be associated with the cross-covariance matrix. By spectral decomposition, the eigenvalues are calculated and used to associate these to the frequencies associated with the corresponding frequencies of the Fourier transform. This reduced, optimized spectrum is then used to distinguish categories of healthy from infected species. For the avocado, it is shown that using the first three modal components of the population variance matrix, we are able to distinguish healthy from two various deficiencies and the Laurel Wilt disease.

IV. Conclusions

We have shown in this paper a novel approach to finding optimized frequency spectrums (i.e. signatures) of plants, and how these spectrums vary when various diseases and nutrient deficiencies are present. The method uses multivariate analysis, in particular, KLE to define the highest absolute value of the eigenvectors responsible for the fundamental reflection pattern of the cell and how these patterns are interrupted and changed by disease and malnutrition effects. The application of KLE and spectral decomposition to define the principle eigenvalues of the cross-covariance matrix, played a major role in developing a series truncation process in the frequency domain. Realizing the value of this concept, a relationship between the effective eigenvalues and the primary frequency component transformation process, has allowed us to develop spectral identification features or biomarkers that can be used as healthy plant and disease signatures for classification and diagnostic purposes.

A spectral dictionary or database for classification purposes of diseases in plants is also the basic premise set forth in this work. Generating a database of healthy and disease spectra or signatures could be used for diagnostics, based on the underlying principles of optical reflection theory. These signature databases can be useful in disease determination since it would utilize less invasive methods and only rests on the principle of understanding frequency signature components of both disease and healthy specimens. Matched filtering, correlation, convolution and neural network principles could then be readily applied for classification, based on biomarker signature.

The other noteworthy conjecture from this research is the premise that once a disease factor enters a cell, it fundamentally changes the natural vibration of the cell, thereby varying its optical reflectance. By examining the changes in the cell signature, a healthy vs. diseased state of a cell can be differentiated. Thus, we can build a case for *spectral identifiers* or *biomarkers of a disease* based on its optimized Fourier spectrum.

V. REFERENCES

[1] D.D.Ye, L.J.Sun, W.Y. Tan, W.K. Che, and M.C. Yang, "Detecting and classifying minor bruised potato based on hyperspectral imaging," Chemom. Intell. Lab. Syst. 2018, 177, 129–139.

[2] N. Susic, U. Zibrat, S. Sirca, P. Strajnar, J. Razinger, M. Knapic, A. Voncina, G. Urek, and B. G. Stare, "Discrimination between abiotic and biotic drought stress in tomatoes using hyperspectral imaging," Sens. Actuators B Chem. 2018, 273, pp.842–852.

[3] J. Hariharan, J. Fuller, Y. Ampatzidis, J. Abdulridha, and A. Lerwill, "Finite Difference Analysis and Bivariate Correlation of Hyperspectral Data for Detecting Laurel Wilt Disease and Nutritional Deficiency in Avocado," Remote Sensing Journal, 24July2019.

[4] K. L. Lai, J. L. Crassidis,"Extensions-of-the-first-and-second-complex-derivative", 2008 Journal of Computational and Applied Mathematics, Vol. 19, Is 1, 15, Sept 2008, pp. 276-293.

[5] E. Paine, E. Slonecker, N. Simon, B. Rosen, R. Resmini, D. Allen, "Optical Characterization of two cyanobacteria genera, Aphanizomenon and Microcystis, with hyperspectralmicroscopy," Journal of Applied Remote Sensing, Jul-Sep 2018 Vol. 12(3).

[6] R. Saracoglu, "Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction," Engineering Applications of Artificial Intelligence, Vol. 25, Issue 7, Oct 2012, pp. 1523-1528.

[7] S. Treguier, C. Levasseur-Garcia, "Chapter 8 – Disease Identification: A Review of Vibrational Spectroscopy Applications," Comprehensive Analytical Chemistry, Vol 80, 2018, pp. 195-225.

[8] J. Martins, P. Sturdza, J. Alonso, "The Complex-Step Derivative Approximation," ACM Transactions on Mathematical Software, Vol. 29, No. 3, September 2003, Pages 245–262.

[9] F. Nikolovski, I. Stojkovska, I, Complex-step derivative approximation in noisy environment, Journal of Computational and Applied Math," Vol 327, 1Jan2018, pp. 64-78.

[10] A. Savitzky, M.J.E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," Analytical Chemistry, 36(8), 1627-39.

[11] A. Marco, J.J. Martinez, "Accurate computation of the Moore-Penrose inverse of strictly totally positive matrices," Journal of Computational and Applied Mathematics, Vol. 350, April 2019, pp. 299-308.

[12] E. Dinc, Z. Yazan, "Wavelet Transform-Based UV Spectroscopy for Pharmaceutical Analysis," Frontiers in Chemistry*,* 26 Oct 2018.

[13] R.C. Ploetz, J. Hulk, M. Wingfield, and Z.W. de Beer, "Destructive tree diseases associated
with ambrosia and bark beetles: Black swan events in tree pathology," Plant Disease, 95, 2013, pp. 856–872.

[14] R.C. Ploetz, J.M. Pérez-Martinez, J.A. Smith, M. Hughes, T.J. Dreaden, S.A. Inch, and Y. Fu,
"Responses of avocado to laurel wilt, caused by Raffaelea lauricola," Plant Pathology, 61, 2012, pp. 801–808.

[15] C.H. Bock, P.E. Parker, T.R. Gottwald, "Effect of stimulated wind-driven rain on duration and distance of
dispersal of Xanthomonas axonompodis pv. citri from canker-infected citrus trees," Plant Dis. 2005, 89, pp. 71–80.

[16] J.S. Hartung, J.F. Daniel, O.P. Pruvost, "Detection of anthomonas-campestris pv. citri by the polymerase chain-reaction method," Appl. Environ. Microbiol, 1993, 59, pp. 1143–1148.

[17] S. Duan, H.G. Jia, Z.Q. Pang, D. Teper, F. White, J. Jones, C.Y. Zhou, N. Wang, "Functional characterization of the citrus canker susceptibility gene CsLOB1," Mol. Plant Pathol. 2018, 19, pp. 1908–1916.

[18] A. Subasi, A. Alkan, E. Koklukaya, M.K. Kiymik, "Wavelet neural network classification of EEG signals by using AR model with MLE preprocessing," Neural Networks, vol. 18, Issue 7, Sept. 2005, pp.985-997.

[19] M. Maron, "Numerical Analysis a Practical Approach", MacMillan Publishing Co., Inc. NY, NY 1982.

[20] D. Hackman, "Karhunen-Loeve expansions of Levy Processes", Communications in Statistics, March 2016